Efficient Learning for Bounded-gate Quantum Circuits with Incoherent Measurements *

Yuxuan Du, Hsin-Yuan Huang, Min-Hsiu Hsieh, Dacheng Tao

Introduction: Advancing efficient methodologies to characterize the behavior of quantum computers is an endless pursuit in quantum science, with pivotal outcomes contributing to designing improved quantum devices and identifying computational merits. In this context, quantum tomography [1, 2, 3] and classical simulators [4, 5, 6] have been two standard approaches. Despite their advancements, (shadow) tomography-based methods are quantum resources intensive, necessitating extensively interactive access to quantum computers, and classical simulators are confined to handling specific classes of quantum states. Accordingly, there is a pressing need for innovative approaches to effectively uncover the dynamics of modern quantum computers with hundreds or thousands of qubits [7, 8]. Recently, machine learning (ML) has emerged as a leading candidate toward this goal [9]. Distinct from prior purely classical or quantum methods, ML-based approaches synergy the power of classical learners and quantum computers. Empirical studies have showcased the superiority of ML compared to traditional methods in many substantial tasks, such as real-time feedback control of quantum systems [10, 11], correlations and entanglement quantification [12, 13], and enhancement of quantum algorithms [14, 15]. However, in contrast to the empirical successes, the theoretical foundation of these ML-based methods holds far-reaching consequences, where rigorous performance guarantees or scaling results remain unknown. A pivotal and elusive question in the field is:

Can we design a provably efficient learning model for predicting the dynamics of large-qubit circuits?

The goal of this work is to answer the above question under a practical scenario. That is, we consider a wide class of quantum circuits featuring bounded-gates with incoherent measurements. The focus on this setting stems from their extensive relevance in early-stage quantum computing, including practical utility identification [7], variational quantum algorithms [16], and quantum system certification [17]. As a result, any progress on this front would enhance our confidence to utilize ML to comprehend large-qubit devices.

Problem setup: Early-stage quantum computers share several common features, including a large qubit count, a bounded gate number, and incoherent and local measurements. In consideration of these traits, the behavior of early-stage quantum computers can be encapsulated within a unified model. Particularly, the state space associated to an N-qubit quantum circuit with a bounded-gate number G yields $\mathcal{Q} = \left\{\rho(\boldsymbol{x}) = U(\boldsymbol{x})\rho_0 U(\boldsymbol{x})^\dagger \middle| \boldsymbol{x} \in [-\pi,\pi]^d\right\}$, where ρ_0 denotes an N-qubit state and $U(\boldsymbol{x})$ refers to the *bounded-gate quantum circuit* depending on the classical input $\boldsymbol{x}$ with d dimensions. Due to the universality of Clifford gates with RZ gates [18], the bounded-gate quantum circuit can always be expressed as $U(\boldsymbol{x}) = \prod_{l=1}^{d}(\mathrm{RZ}(\boldsymbol{x}_l)u_e)$, where u_e is composed of Clifford (abbreviated as CI) gates and the identity gate $\mathbb{I}_2$ with $\mathrm{CI} = \{H, S, \mathrm{CNOT}\}$. When $\rho(\boldsymbol{x}) \sim \mathcal{Q}$ is measured by O, an observable sampled from a prior distribution $\mathbb{D}_O$, its *incoherent dynamics* are described the mean-value space

$$\mathcal{F} = \left\{ f(\boldsymbol{x}, O) = \mathrm{Tr}(\rho(\boldsymbol{x})O) \middle| \boldsymbol{x} \in [-\pi,\pi]^d, O \sim \mathbb{D}_O \right\}. \tag{1}$$

Without loss of generality, we suppose O is constituted by multiple local observables with a bounded norm, i.e., $O = \sum_{i=1}^{q} O_i$ and $\sum_l \|O_i\|_\infty \leq B$, and the maximum locality of $\{O_i\}$ is K. This formalism encompasses diverse tasks in quantum computing, e.g., variational quantum algorithms, numerous applications of classical shadow, and quantum system certification.

According to the unified model formulated above, here we explore the learnability of $\mathcal{F}$ in Eq. (1) by separately assessing the required *sample complexity* and *runtime complexity* of training a classical learner $h(\boldsymbol{x}, O)$ that attains a low *average* prediction error with $f(\boldsymbol{x}, O)$. Mathematically, denoting $\mathbb{D}_{\mathcal{X}}$ as the data distribution, the average performance of the learner is expected to satisfy

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}_{\mathcal{X}}} |h(\boldsymbol{x}, O) - f(\boldsymbol{x}, O)|^2 \leq \epsilon, \ \forall O \sim \mathbb{D}_O. \tag{2}$$

*The technical version of the submission is also attached.

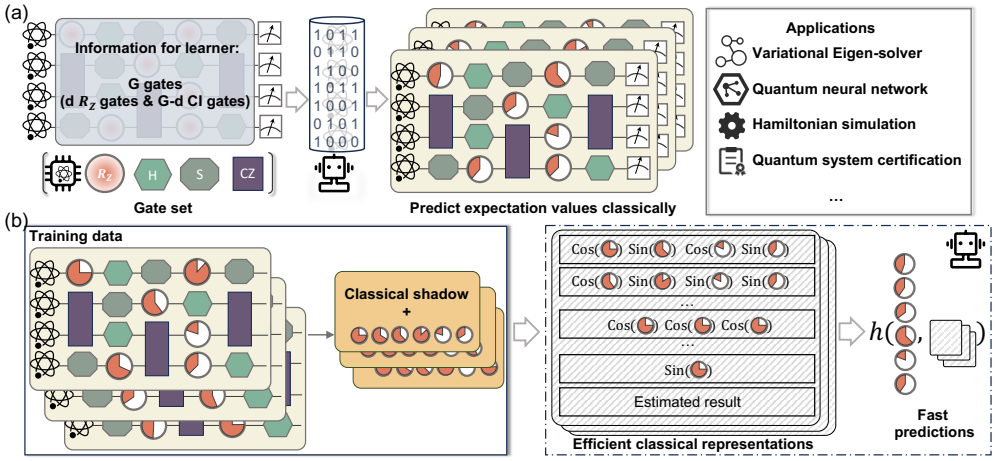

Figure 1: **Learning protocols for quantum circuits with the bounded number of gates**. (a) Given a circuit composed of finite Clifford gates and RZ gates, a classical learner feeds n classical inputs, i.e., n tuples of the varied angles of RZ gates, to the quantum device and collects the relevant measured results as data labels. The collected n labeled data are used to train a prediction model h such that the linear properties of the generated state over new input $\boldsymbol{x}$, i.e., $\mathrm{Tr}(\rho(\boldsymbol{x})O)$ with O being an observable sampling from a prior distribution, can be accurately estimated. (b) First, the learner collects the training dataset, i.e., n labeled data via classical shadow. Then, the learner applies shadow estimation and the trigonometric monomial expansion to the collected dataset to obtain classical representations, where any new input of the explored quantum circuits can be efficiently predicted offline.

Remark. Our analysis adheres to the conventions of most ML algorithms to train $h(\boldsymbol{x}, O)$. Namely, as shown in Fig. 1(a), to accommodate the constraints of modern quantum devices, the training data are exclusively gathered by simple measurements [19]. In addition, the classical learner is kept unaware of the circuit layout details, except for the gate count G and the dimension of classical controls d, and can solely access estimated information about quantum states through *finite* and *incoherent measurements*. Last, considering the scarce and costly nature of modern quantum computers, the preferred mode for the prediction model $h(\boldsymbol{x}, O)$ is *offline*, enabling predictions for new inputs $\boldsymbol{x}$ solely on the classical side.

Overview of main results: Our main results consist of two parts: (♣) Analyzing the learnability of bounded-gate quantum circuits $\mathcal{F}$ in terms of the sample and runtime complexities; and (♠) Devising a provably efficient protocol to learn $\mathcal{F}$ when $\boldsymbol{x}$ is sampled from the uniform distribution. Concretely, for Outcome (♣), our analysis uncovers that (i) with high probability, the sample complexity scaling linearly in d is *necessary and sufficient* to achieve a small prediction error on average, while the corresponding computational complexity may scale *exponentially* in d. The exponential separation of the sample and computational complexity pinpoints the importance of crafting provably efficient algorithms to learn $\mathcal{F}$ in Eq. (1). To address this issue, in Outcome (♠), we harness the concept of classical shadow and truncated trigonometric expansion to devise a kernel-based learning model capable of trading off prediction error and computational complexity, transitioning from exponential to polynomial scaling in many practical settings when $\mathbb{D}_{\mathcal{X}}$ in Eq. (2) is restricted to be the uniform distribution, i.e., $\boldsymbol{x} \sim [-\pi, \pi]^d$. These results advance two crucial realms in quantum computation: the exploration of quantum algorithms with practical utilities and learning-based quantum system certification. In the rest of this extended abstract, we elaborate on outcomes (♣) and (♠) and their implications.

■ *Outcome (♣)* is obtained by the following theorem.

Theorem 1 (informal) *Following notations in Eq. (1), let* $\mathcal{T} = \{\boldsymbol{x}^{(i)}, \tilde{f}_T(\boldsymbol{x}^{(i)})\}_{i=1}^n$ *be a dataset containing* n *training examples with* $\boldsymbol{x}^{(i)} \in [\pi, \pi]^d$ *and* $\tilde{f}_T(\boldsymbol{x}^{(i)})$ *being the estimation of* $f(\boldsymbol{x}^{(i)}, O)$ *using* T *incoherent measurements with* $T \leq (N \log 2)/\epsilon$. *Then, the training data size* $\widetilde{\Omega}\left(\frac{(1-\epsilon)d}{\epsilon T}\right) \leq n \leq \widetilde{\mathcal{O}}\left(\frac{B^2 d + B^2 NG}{\epsilon}\right)$ *is sufficient and necessary to achieve Eq. (2) with high probability. However, there exists a class of* G*-bounded-gate quantum circuits that no algorithm can achieve Eq. (2) in a polynomial time.*

The implications of Theorem 1 are summarized below.

• The exponential separation in terms of the sample and computational complexities underscores the non-trivial nature of learning the incoherent dynamics of bounded-gate quantum circuits. That is, while a

linear number of training examples with d is sufficient and necessary to guarantee a satisfactory prediction accuracy, identifying these training examples may be computationally hard.

• Our finding also aligns with a recent study exploring the learnability of quantum states with bounded-gate complexity, which, like our case, is sample-efficient but computationally demanding [20].

■ *Outcome* (♠). Our proposal, as depicted in Fig. 1(b), contains two steps: (i) Collect training data from the exploited quantum device; (ii) Construct the learning model and use it to predict new inputs. In Step (i), the learner feeds different $\boldsymbol{x}^{(i)} \in [-\pi, \pi]^d$ to the circuit and collects classical information of $\rho(\boldsymbol{x}^{(i)})$ under Pauli-based classical shadow with T snapshots [21], denoted by $\tilde{\rho}_T(\boldsymbol{x}^{(i)})$. In this way, the learner constructs the training dataset $\mathcal{T}_{\mathsf{s}} = \{\boldsymbol{x}^{(i)} \to \tilde{\rho}_T(\boldsymbol{x}^{(i)})\}_{i=1}^n$ with n training examples. Then, in Step (ii), the learner utilizes $\mathcal{T}_{\mathsf{s}}$ to build a kernel-based ML model h_{s}, i.e., given a new input $\boldsymbol{x}$, its prediction yields

$$h_{\mathsf{s}}(\boldsymbol{x}, O) = \frac{1}{n} \sum_{i=1}^{n} \kappa_\Lambda \left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) g\left(\boldsymbol{x}^{(i)}, O\right), \tag{3}$$

where $g(\boldsymbol{x}^{(i)}, O) = \mathrm{Tr}(\tilde{\rho}_T(\boldsymbol{x}^{(i)})O)$ refers to the shadow estimation of $\mathrm{Tr}(\rho(\boldsymbol{x}^{(i)})O)$, $\kappa_\Lambda(\boldsymbol{x}, \boldsymbol{x}^{(i)})$ is the *truncated trigonometric monomial kernel* with $\kappa_\Lambda \left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) = \sum_{\boldsymbol{\omega}, \|\boldsymbol{\omega}\|_0 \le \Lambda} 2^{\|\boldsymbol{\omega}\|_0} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \Phi_{\boldsymbol{\omega}}\left(\boldsymbol{x}^{(i)}\right) \in \mathbb{R}$, and $\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})$ with $\boldsymbol{\omega} \in \{0, 1, -1\}^d$ is the trigonometric monomial basis with values $\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) = \prod_{i=1}^d \begin{cases} 1 & \text{if } \boldsymbol{\omega}_i = 0 \\ \cos(\boldsymbol{x}_i) & \text{if } \boldsymbol{\omega}_i = 1 \\ \sin(\boldsymbol{x}_i) & \text{if } \boldsymbol{\omega}_i = -1 \end{cases}$. The subsequent theorem provides a provable guarantee for the proposed method.

Theorem 2 (Informal) *Following notations in Eqs. (1)-(3), denote* $\mathfrak{C}(\Lambda) = \{\boldsymbol{\omega} | \boldsymbol{\omega} \in \{0, \pm 1\}^d, \text{ s.t. } \|\boldsymbol{\omega}\|_0 \le \Lambda\}$. *Suppose* $\mathbb{E}_{\boldsymbol{x} \sim [-\pi, \pi]^d} \|\nabla_{\boldsymbol{x}} \mathrm{Tr}(\rho(\boldsymbol{x})O)\|_2^2 \le C$. *Then, when the frequency is truncated to* $\Lambda = 4C/\epsilon$ *and the number of training examples is* $n = |\mathfrak{C}(\Lambda)| 2B^2 9^K \epsilon^{-1} \log(2|\mathfrak{C}(\Lambda)|/\delta)$, *even for* $T = 1$ *snapshot per training example, with probability at least* $1 - \delta$, $\mathbb{E}_{\boldsymbol{x} \sim [-\pi, \pi]^d} |h_{\mathsf{s}}(\boldsymbol{x}, O) - f(\rho(\boldsymbol{x}), O)|^2 \le \epsilon$.

The implications of Theorem 2 are as follows.

• Both sample and computational complexities of h_{s} are dominated by the cardinality of $\mathfrak{C}(\Lambda)$, or equivalently the quantity C as $\Lambda = 4C/\epsilon$. Notably, the quantity C, in many practical scenarios, can be effectively bounded. One such scenario involves the advancement of variational quantum algorithms (VQAs) [16, 22, 23], as recent studies have shown that the gradients norm of numerous VQAs yields $C \le 1/\mathrm{poly}(N)$ [24] or $C \le 1/\exp(N)$, a.k.a, barren plateaus [25, 26].

• The exponential dependence of the locality K arises from the use of Pauli-based classical shadow in h_{s}. A potential solution to address this issue is adopting advanced variants of classical shadow to enhance the sample complexity bounds [27, 28].

• Our proposal and the relevant theoretical findings can be effectively extended to a broader context, i.e., the bounded-gate circuits composed of CI gates alongside parameterized multi-qubit gates generated by arbitrary Pauli strings.

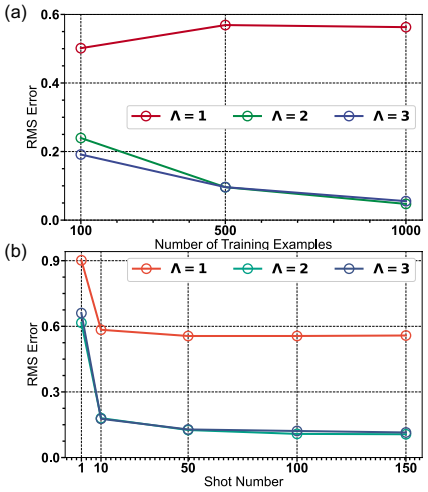

Figure 2: **Predicting properties of rotational 60-qubit GHZ states.** (a) The root mean squared (RMS) prediction error of the trained ML model with varied truncation Λ and number of training examples n. (b) The RMS prediction error of the trained ML model with varied truncation Λ and shot number T.

■ *Numerical results*. We apply the proposed ML model to predict the properties of N-qubit rotational GHZ states with $|\mathrm{GHZ}(\boldsymbol{x})\rangle = U(\boldsymbol{x})(|0\rangle^{\otimes N} + |1\rangle^{\otimes N})/\sqrt{2}$, where $U(\boldsymbol{x}) = \mathrm{RY}_1(\boldsymbol{x}_1) \otimes \mathrm{RY}_{30}(\boldsymbol{x}_2) \otimes \mathrm{RY}_{60}(\boldsymbol{x}_3)$, the subscript j refers to apply RY gate to the j-th qubit, and $N = 60$. At the training stage, we constitute the dataset $\mathcal{T}_s$ containing $n = 30$ examples. Here we exploit how the prediction error depends on the truncation value Λ, the number of training examples n, and the shot number T. Fig. 2(a) showcases that when $T = 1000$ and the training data are sufficient (i.e., $n \ge 500$), the root mean squared (RMS) prediction error on 10 unseen test examples is almost the same for the full expansion (i.e., $\Lambda = 3$) and the proper truncation (i.e., $\Lambda = 2$). Besides, Fig. 2(b) indicates that when $n = 500$, the prediction error on the same test examples reaches a low value for both $\Lambda = 3$ and $\Lambda = 2$ once the shot number T exceeds a threshold value ($T \ge 50$). These results are consistent with our theoretical analysis.

References

[1] Ulf Leonhardt. Quantum-state tomography and discrete wigner function. *Physical review letters*, 74(21):4101, 1995.

[2] Joseph B Altepeter, David Branning, Evan Jeffrey, TC Wei, Paul G Kwiat, Robert T Thew, Jeremy L O'Brien, Michael A Nielsen, and Andrew G White. Ancilla-assisted quantum process tomography. *Physical Review Letters*, 90(19):193601, 2003.

[3] Scott Aaronson. Shadow tomography of quantum states. In *Proceedings of the 50th annual ACM SIGACT symposium on theory of computing*, pages 325–338, 2018.

[4] Igor L Markov and Yaoyun Shi. Simulating quantum computation by contracting tensor networks. *SIAM Journal on Computing*, 38(3):963–981, 2008.

[5] Sergey Bravyi and David Gosset. Improved classical simulation of quantum circuits dominated by clifford gates. *Physical review letters*, 116(25):250501, 2016.

[6] J Ignacio Cirac, David Perez-Garcia, Norbert Schuch, and Frank Verstraete. Matrix product states and projected entangled pair states: Concepts, symmetries, theorems. *Reviews of Modern Physics*, 93(4):045003, 2021.

[7] Youngseok Kim, Andrew Eddins, Sajant Anand, Ken Xuan Wei, Ewout Van Den Berg, Sami Rosenblatt, Hasan Nayfeh, Yantao Wu, Michael Zaletel, Kristan Temme, et al. Evidence for the utility of quantum computing before fault tolerance. *Nature*, 618(7965):500–505, 2023.

[8] Dolev Bluvstein, Simon J Evered, Alexandra A Geim, Sophie H Li, Hengyun Zhou, Tom Manovitz, Sepehr Ebadi, Madelyn Cain, Marcin Kalinowski, Dominik Hangleiter, et al. Logical quantum processor based on reconfigurable atom arrays. *Nature*, pages 1–3, 2023.

[9] Valentin Gebhart, Raffaele Santagati, Antonio Andrea Gentile, Erik M Gauger, David Craig, Natalia Ares, Leonardo Banchi, Florian Marquardt, Luca Pezzè, and Cristian Bonato. Learning quantum systems. *Nature Reviews Physics*, 5(3):141–156, 2023.

[10] Giacomo Torlai and Roger G Melko. Neural decoder for topological codes. *Physical review letters*, 119(3):030501, 2017.

[11] Kevin Reuer, Jonas Landgraf, Thomas Fösel, James O'Sullivan, Liberto Beltrán, Abdulkadir Akin, Graham J Norris, Ants Remm, Michael Kerschbaum, Jean-Claude Besse, et al. Realizing a deep reinforcement learning agent for real-time quantum feedback. *Nature Communications*, 14(1):7138, 2023.

[12] Askery Canabarro, Samuraí Brito, and Rafael Chaves. Machine learning nonlocal correlations. *Physical review letters*, 122(20):200401, 2019.

[13] Dominik Koutnỳ, Laia Ginés, Magdalena Moczała-Dusanowska, Sven Höfling, Christian Schneider, Ana Predojević, and Miroslav Ježek. Deep learning of quantum entanglement from incomplete measurements. *Science Advances*, 9(29):eadd7131, 2023.

[14] Elizabeth R Bennewitz, Florian Hopfmueller, Bohdan Kulchytskyy, Juan Carrasquilla, and Pooya Ronagh. Neural error mitigation of near-term quantum simulations. *Nature Machine Intelligence*, 4(7):618–624, 2022.

[15] Shi-Xin Zhang, Zhou-Quan Wan, Chee-Kong Lee, Chang-Yu Hsieh, Shengyu Zhang, and Hong Yao. Variational quantum-neural hybrid eigensolver. *Physical Review Letters*, 128(12):120502, 2022.

[16] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021.

[17] Jens Eisert, Dominik Hangleiter, Nathan Walk, Ingo Roth, Damian Markham, Rhea Parekh, Ulysse Chabaud, and Elham Kashefi. Quantum certification and benchmarking. *Nature Reviews Physics*, 2(7):382–390, 2020.

[18] Neil J Ross and Peter Selinger. Optimal ancilla-free clifford+ t approximation of z-rotations. *arXiv preprint arXiv:1403.2975*, 2014.

[19] Yuxuan Du, Yibo Yang, Tongliang Liu, Zhouchen Lin, Bernard Ghanem, and Dacheng Tao. Shadownet for data-centric quantum system learning. *arXiv preprint arXiv:2308.11290*, 2023.

[20] Haimeng Zhao, Laura Lewis, Ishaan Kannan, Yihui Quek, Hsin-Yuan Huang, and Matthias C Caro. Learning quantum states and unitaries of bounded gate complexity. *arXiv preprint arXiv:2310.19882*, 2023.

[21] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.

[22] Jinkai Tian, Xiaoyu Sun, Yuxuan Du, Shanshan Zhao, Qing Liu, Kaining Zhang, Wei Yi, Wanrong Huang, Chaoyue Wang, Xingyao Wu, et al. Recent advances for quantum neural networks in generative learning. *arXiv preprint arXiv:2206.03066*, 2022.

[23] 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(1):015004, 2022.

[24] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J Coles, and Marco Cerezo. Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 3(6):542–551, 2023.

[25] 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):1–6, 2018.

[26] Marco 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):1–12, 2021.

[27] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Efficient estimation of pauli observables by derandomization. *Physical review letters*, 127(3):030503, 2021.

[28] H Chau Nguyen, Jan Lennart Bönsel, Jonathan Steinberg, and Otfried Gühne. Optimizing shadow tomography with generalized measurements. *Physical Review Letters*, 129(22):220502, 2022.

Efficient learning for bounded-gate quantum circuits with incoherent measurements

Yuxuan Du,[1] Hsin-Yuan Huang,[2,3] Min-Hsiu Hsieh,[4] and Dacheng Tao[1]

[1]*School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore*
[2]*Institute for Quantum Information and Matter, Caltech, Pasadena, California, USA*
[3]*Google Quantum AI*
[4]*Foxconn Research, Taipei, Taiwan*

The vast and complicated large-qubit state space forbids us to comprehensively capture the dynamics of modern quantum computers via classical simulations or quantum tomography. However, recent progress in quantum learning theory invokes a crucial question: given a quantum circuit containing d tunable RZ gates and $G-d$ Clifford gates, can a learner efficiently predict its linear properties with new classical controls, after learning from data obtained by incoherently measuring states generated by the same circuit but with different classical controls? In this work, we prove that the sample complexity scaling linearly in d is necessary and sufficient to achieve a small prediction error, while the corresponding computational complexity may scale exponentially in d. Building upon these derived complexity bounds, we further harness the concept of classical shadow and truncated trigonometric expansion to devise a kernel-based learning model capable of trading off prediction error and computational complexity, transitioning from exponential to polynomial scaling in many practical settings. Our results advance two crucial realms in quantum computation: the exploration of quantum algorithms with practical utilities and learning-based quantum system certification. We conduct numerical simulations to validate our proposals across diverse scenarios, encompassing quantum information processing protocols, Hamiltonian simulation, and variational quantum algorithms up to 60 qubits.

I. INTRODUCTION

Advancing efficient methodologies to characterize the behavior of quantum computers is an endless pursuit in quantum science, with pivotal outcomes contributing to designing improved quantum devices and identifying computational merits. In this context, quantum tomography and classical simulators have been two standard approaches. Quantum tomography, spanning state [1], process [2, 3], and shadow tomography [4, 5], provides concrete ways to benchmark current quantum computers. Classical simulators, transitioning from state-vector simulation to tensor network methods [6–9] and Clifford-based simulators [10–12], not only facilitate the design of quantum algorithms without direct access to quantum resources [13] but also push the boundaries to unlock practical utilities [14]. Despite their advancements, (shadow) tomography-based methods are quantum resources intensive, necessitating extensively interactive access to quantum computers, and classical simulators are confined to handling specific classes of quantum states. Accordingly, there is a pressing need for innovative approaches to effectively uncover the dynamics of modern quantum computers with hundreds or thousands of qubits [15, 16].

Machine learning (ML) is fueling a new paradigm for comprehending quantum computers [17]. This hybrid approach, distinct from prior purely classical or quantum methods, synergies the power of classical learners and quantum computers. Concisely, it begins by collecting samples from target quantum devices to train a classical learner, and then uses the trained learner to predict unseen data from the same data distribution without direct access to quantum computers or by utilizing few quantum resources. Empirical studies have showcased the superiority of ML compared to traditional methods in many substantial tasks, such as real-time feedback control of quantum systems [18–21], correlations and entanglement quantification [22–24], and enhancement of quantum algorithms [25–28]. However, in contrast to the empirical successes, the theoretical foundation of these ML-based methods holds far-reaching consequences, where rigorous performance guarantees or scaling results remain unknown. A pivotal and elusive question in the field is:

can we design a provably efficient learning algorithm for predicting the dynamics of large-qubit quantum circuits?

Any progress on this front would enhance our confidence to utilize ML to comprehend large-qubit devices.

Here we answer the above question affirmatively for a wide class of quantum circuits featuring bounded-gates with incoherent measurements. The focus on these quantum circuits stems from their extensive relevance in early-stage quantum computing, including applications such as practical utility identification [15], variational quantum algorithms [29], and quantum system certification [30]. Concretely, we first explore the learnability of a class of N-qubit quantum circuits, which consists of an arbitrary initial state, a unitary composed of d tunable RZ gates and $G-d$ Clifford gates, followed by an incoherent measure operator O. Our analysis uncovers that (i) with high probability, $\widetilde{\Omega}(\frac{(1-\epsilon)d}{\epsilon T}) \leq n \leq \widetilde{\mathcal{O}}(\frac{B^2 d + B^2 NG}{\epsilon^2})$ training examples are sufficient and necessary to achieve an ϵ-prediction

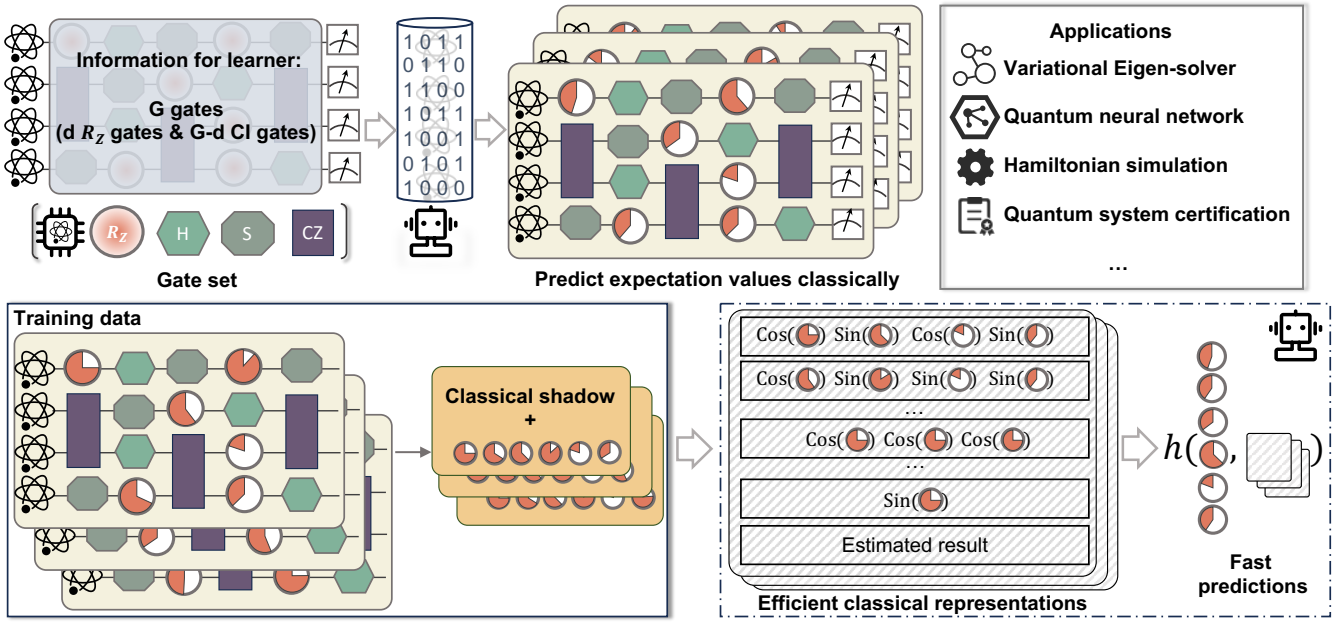

Figure 1: **Learning protocols for quantum circuits with the bounded number of non-Clifford gates**. (a) The visualization of learning bounded-gate quantum circuits with incoherent measurements. Given a circuit composed of finite Clifford gates and RZ gates, a classical learner feeds n classical inputs, i.e., n tuples of the varied angles of RZ gates, to the quantum device and collects the relevant measured results as data labels. The collected n labeled data are used to train a prediction model h such that the linear properties of the generated state over new input $\boldsymbol{x}$, i.e., $\mathrm{Tr}(\rho(\boldsymbol{x})O)$ with O being an observable sampling from a prior distribution, can be accurately estimated. Following conventions, the interaction between the learner and the system is restrictive in which the learner can only access the quantum computer via incoherent and finite measurements, and the circuit layout is unnecessarily available to the learner except for the number of total gates G and the number of tunable RZ gates d. (b) The proposed kernel-based ML model consists of two steps. First, the learner collects the training dataset, i.e., n labeled data via classical shadow. Then, the learner applies shadow estimation and the trigonometric monomial expansion to the collected training dataset to obtain efficient classical representations, where any new input of the explored quantum circuits can be efficiently predicted offline.

error on average, where B refers to the bounded norm of O and T is the number of incoherent measurements; (ii) there exists a class of G-bounded-gate quantum circuits that no polynomial runtime algorithms can predict their outputs within an ϵ-prediction error. These results not only enrich the landscape of quantum learning theory by shedding light on the learnability of quantum circuits, but also invoke the necessity of developing a learning algorithm capable of addressing the exponential gap between sample efficiency and computational complexity.

To address the above computational challenge, we further utilize the concepts of classical shadow [31] and trigonometric expansion to design a kernel-based ML model that effectively balances prediction accuracy with computational demands. We prove that when the classical inputs are sampled from the uniform distribution, with high probability, the runtime and memory cost of the proposed ML model is no larger than $\widetilde{\mathcal{O}}(TNB^2c/\epsilon)$ for a large constant c to achieve an ϵ prediction error in many practical scenarios. Our proposed algorithm paves a new way of predicting the dynamics of large-qubit quantum circuits in a provably efficient manner.

II. LEARNABILITY OF BOUNDED-GATE QUANTUM CIRCUITS

Early-stage quantum computers share several common features, including a large qubit count, a bounded gate number, and incoherent and local measurements. In consideration of these traits, the behavior of early-stage quantum computers can be encapsulated within a unified model. Particularly, the state space associated to an N-qubit quantum circuit with a bounded-gate number G yields

$$\mathcal{Q} = \left\{ \rho(\boldsymbol{x}) = U(\boldsymbol{x})\rho_0 U(\boldsymbol{x})^\dagger \middle| \boldsymbol{x} \in [-\pi, \pi]^d \right\}, \tag{1}$$

where ρ_0 denotes an N-qubit state and $U(\boldsymbol{x})$ refers to the bounded-gate quantum circuit depending on the classical input $\boldsymbol{x}$ with d dimensions. Due to the universality of Clifford gates with RZ gates [32], the bounded-gate quantum

circuit can always be expressed as

$$U(\boldsymbol{x}) = \prod_{l=1}^{d} (\mathrm{RZ}(\boldsymbol{x}_l) u_e), \tag{2}$$

where u_e is composed of Clifford (abbreviated as CI) gates and the identity gate $\mathbb{I}_2$ with $\mathrm{CI} = \{H, S, \mathrm{CNOT}\}$. When the state $\rho(\boldsymbol{x}) \sim \mathcal{Q}$ is measured by an observable O sampled from a prior distribution $\mathbb{D}_O$, its *incoherent dynamics* are described the mean-value space

$$\mathcal{F} = \left\{ f(\boldsymbol{x}, O) = \mathrm{Tr}(\rho(\boldsymbol{x}) O) \middle| \boldsymbol{x} \in [-\pi, \pi]^d, O \sim \mathbb{D}_O \right\}. \tag{3}$$

Without loss of generality, we suppose O is constituted by multiple local observables with a bounded norm, i.e., $O = \sum_{i=1}^{q} O_i$ and $\sum_l \|O_i\|_\infty \leq B$, and the maximum locality of $\{O_i\}$ is K. This formalism encompasses diverse tasks in quantum computing, e.g., variational quantum algorithms [29], numerous applications of classical shadow [33], and quantum system certification [30] (see Supplementary Material (SM) A for the elaboration).

According to the unified model formulated above, here we explore the learnability of $\mathcal{F}$ in Eq. (3) by separately assessing the required *sample complexity* and *runtime complexity* of training a classical learner $h(\boldsymbol{x}, O)$ that attains a low *average* prediction error with $f(\boldsymbol{x}, O)$. Mathematically, the average performance of the learner is expected to satisfy

$$\mathbb{E}_{\boldsymbol{x} \sim [-\pi, \pi]^d} |h(\boldsymbol{x}, O) - f(\boldsymbol{x}, O)|^2 \leq \epsilon, \ \forall O \sim \mathbb{D}_O, \tag{4}$$

where the classical control $\boldsymbol{x}$ is uniformly sampled from $[-\pi, \pi]^d$. Note that our analysis adheres to the conventions of most ML algorithms to train $h(\boldsymbol{x}, O)$. Namely, as shown in Fig. 1(a), to accommodate the constraints of modern quantum devices, the training data are exclusively gathered by simple measurements [27]. In addition, the classical learner is kept unaware of the circuit layout details, except for the gate count G and the dimension of classical controls d, and can solely access estimated information about quantum states through *finite* and *incoherent measurements*. Last, considering the scarce and costly nature of modern quantum computers, the preferred mode for the prediction model $h(\boldsymbol{x}, O)$ is *offline*, enabling predictions for new inputs $\boldsymbol{x}$ solely on the classical side.

The following theorem demonstrates the learnability of $\mathcal{F}$ in Eq. (3), where the formal statement and the proof are presented in SM B-E.

Theorem 1 (informal)**.** *Following notations in Eq. (3), let $\mathcal{T} = \{\boldsymbol{x}^{(i)}, \tilde{f}_T(\boldsymbol{x}^{(i)})\}_{i=1}^n$ be a dataset containing n training examples with $\boldsymbol{x}^{(i)} \in [\pi, \pi]^d$ and $\tilde{f}_T(\boldsymbol{x}^{(i)})$ being the estimation of $f(\boldsymbol{x}^{(i)}, O)$ using T incoherent measurements with $T \leq (N \log 2)/\epsilon$. Then, the training data size*

$$\widetilde{\Omega}\left(\frac{(1-\epsilon)d}{\epsilon T}\right) \leq n \leq \widetilde{\mathcal{O}}\left(\frac{B^2 d + B^2 N G}{\epsilon}\right) \tag{5}$$

is sufficient and necessary to achieve Eq. (4) with high probability. However, there exists a class of G-bounded-gate quantum circuits that no algorithm can achieve Eq. (4) in a polynomial time.

The exponential separation in terms of the sample and computational complexities underscores the non-trivial nature of learning the incoherent dynamics of bounded-gate quantum circuits. While the matched upper and lower bounds in Eq. (5) indicate that a linear number of training examples with d is sufficient and necessary to guarantee a satisfactory prediction accuracy, the derived exponential runtime cost hints that identifying these training examples may be computationally hard. These results enrich quantum learning theory [34, 35], especially for the learnability of quantum circuits. Our finding also aligns with a recent study exploring the learnability of quantum states with bounded-gate complexity, which, like our case, is sample-efficient but computationally demanding [36]. We note that the upper bound does not depend non-trivially on T, so we omit it. Besides, our future work will explore novel techniques to match the factors N, B, and G in the lower and upper bounds, whereas such deviations do not affect our key results.

III. A PROVABLY EFFICIENT PROTOCOL TO LEARN BOUNDED-GATE QUANTUM CIRCUITS

The exponential separation of the sample and computational complexity pinpoints the importance of crafting provably efficient algorithms to learn $\mathcal{F}$ in Eq. (3). To address this issue, here we devise a kernel-based ML protocol adept

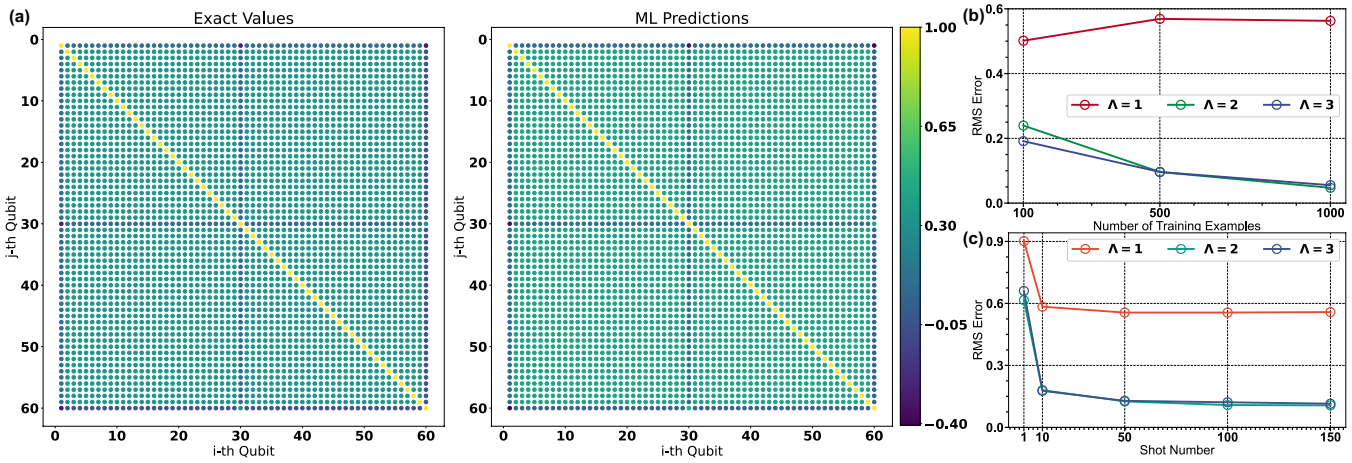

Figure 2: **Numerical results of predicting properties of rotational 60-qubit GHZ states**. (a) TWO-POINT CORRELATION. Exact values and ML predictions of the expectation value of the correlation function $C_{ij} = (X_iX_j + Y_iY_j + Z_iZ_j)/3$ for all qubit pairs of 60-qubit GHZ states. The node's color indicates the exact value and predicted value of the two subplots, respectively. (b-c) PREDICTION ERROR. Subplot (b) depicts the root mean squared (RMS) prediction error of the trained ML model with varied truncation Λ and number of training examples n. Subplot (c) shows the RMS prediction error of the trained ML model with varied truncation Λ and the shot number T. The setting $\Lambda = 3$ refers to the full expansion.

at balancing the average prediction error ϵ and the computational complexity, making a transition from exponential to polynomial scaling with d when $\mathbb{D}_{\mathcal{X}}$ is restricted to be the uniform distribution, i.e., $\boldsymbol{x} \sim [-\pi, \pi]^d$.

Our proposal, as depicted in Fig. 1(b), contains two steps: (i) Collect training data from the exploited quantum device; (ii) Construct the learning model and use it to predict new inputs. In Step (i), the learner feeds different $\boldsymbol{x}^{(i)} \in [-\pi, \pi]^d$ to the circuit and collects classical information of $\rho(\boldsymbol{x}^{(i)})$ under Pauli-based classical shadow with T snapshots, denoted by $\tilde{\rho}_T(\boldsymbol{x}^{(i)})$. In this way, the learner constructs the training dataset $\mathcal{T}_{\mathsf{s}} = \{\boldsymbol{x}^{(i)} \to \tilde{\rho}_T(\boldsymbol{x}^{(i)})\}_{i=1}^n$ with n training examples. Then, in Step (ii), the learner utilizes $\mathcal{T}_{\mathsf{s}}$ to build a kernel-based ML model h_{s}, i.e., given a new input $\boldsymbol{x}$, its prediction yields

$$h_{\mathsf{s}}(\boldsymbol{x}, O) = \frac{1}{n} \sum_{i=1}^{n} \kappa_\Lambda \left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) g\left(\boldsymbol{x}^{(i)}, O\right), \tag{6}$$

where $g(\boldsymbol{x}^{(i)}, O) = \mathrm{Tr}(\tilde{\rho}_T(\boldsymbol{x}^{(i)})O)$ refers to the shadow estimation of $\mathrm{Tr}(\rho(\boldsymbol{x}^{(i)})O)$, $\kappa_\Lambda(\boldsymbol{x}, \boldsymbol{x}^{(i)})$ is the *truncated trigonometric monomial kernel* with

$$\kappa_\Lambda \left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) = \sum_{\boldsymbol{\omega}, \|\boldsymbol{\omega}\|_0 \le \Lambda} 2^{\|\boldsymbol{\omega}\|_0} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \Phi_{\boldsymbol{\omega}}\left(\boldsymbol{x}^{(i)}\right) \in \mathbb{R}, \tag{7}$$

and $\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})$ with $\boldsymbol{\omega} \in \{0, 1, -1\}^d$ is the trigonometric monomial basis with values

$$\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) = \prod_{i=1}^{d} \begin{cases} 1 & \text{if } \boldsymbol{\omega}_i = 0 \\ \cos(\boldsymbol{x}_i) & \text{if } \boldsymbol{\omega}_i = 1 \\ \sin(\boldsymbol{x}_i) & \text{if } \boldsymbol{\omega}_i = -1 \end{cases}. \tag{8}$$

A distinctive aspect of our proposal is its capability to predict the incoherent dynamics $\mathrm{Tr}(\rho(\boldsymbol{x})O)$ across various O purely on the classical side. This is achieved by storing the shadow information $\{\tilde{\rho}_T(\boldsymbol{x}^{(i)})\}$ into the classical memory, where the shadow estimation $\{g(\boldsymbol{x}^{(i)}, O)\}$ for different $\{O\}$ can be efficiently manipulated through classical post-processing.

With the full expansion, the cardinality of the frequency set $\{\boldsymbol{\omega}\}$ in Eq. (7) is 3^d when $\Lambda = d$, impeding the computational efficiency of our proposal. To remedy this, judicious frequency truncation is necessary to strike a balance between prediction error and computational complexity. The subsequent theorem provides a provable guarantee for this tradeoff relation, with the formal description and proof deferred to SM F.

Theorem 2 (Informal)**.** *Following notations in Eqs. (3)-(6), denote* $\mathfrak{C}(\Lambda) = \{\boldsymbol{\omega} | \boldsymbol{\omega} \in \{0, \pm 1\}^d, \; s.t. \; \|\boldsymbol{\omega}\|_0 \le \Lambda\}$. *Suppose* $\mathbb{E}_{\boldsymbol{x} \sim [-\pi,\pi]^d} \|\nabla_{\boldsymbol{x}} \mathrm{Tr}(\rho(\boldsymbol{x})O)\|_2^2 \le C$. *Then, when the frequency is truncated to* $\Lambda = 4C/\epsilon$ *and the number*

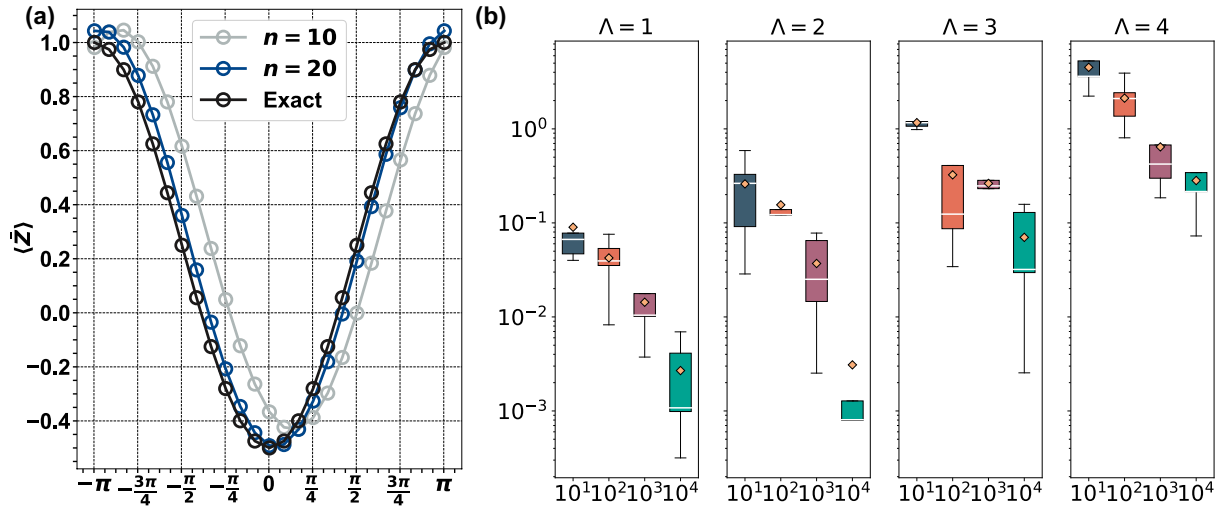

Figure 3: **Numerical results of predicting properties of quantum states evolved by 60-qubit global Hamiltonians**. (a) PREDICTION ON THE EVOLVED STATE WITH $d = 1$. The notation $n = a$ refers that the number of training examples used to form the classical representation is a. The y-axis denotes the magnetization $\langle \bar{Z} \rangle = \frac{1}{60} \sum_i \langle Z_i \rangle$. (b) PREDICTION ERROR ON THE EVOLVED STATE WITH $d = 30$ WITH DIFFERENT SIZES OF THE TRAINING DATASET. The meaning of notations in the box plot is as follows. The label '$\Lambda = a$' stands for that the truncation value is set as a. The x-axis refers to the varied number of training examples n, ranging from $n = 10$ to $n = 10^4$. The y-axis refers to the achieved RMS prediction error. All results are obtained using the same random seeds.

of training examples is $n = |\mathfrak{C}(\Lambda)| 2B^2 9^K \epsilon^{-1} \log(2|\mathfrak{C}(\Lambda)|/\delta)$*, even for* $T = 1$ *snapshot per training example, with probability at least* $1 - \delta$*,*

$$\mathbb{E}_{\boldsymbol{x} \sim [-\pi,\pi]^d} \left| h_{\mathsf{s}}(\boldsymbol{x}, O) - f(\rho(\boldsymbol{x}), O) \right|^2 \leq \epsilon. \tag{9}$$

The obtained results reveal that both sample and computational complexities of h_{s} are predominantly influenced by the cardinality of $\mathfrak{C}(\Lambda)$, or equivalently the quantity C as $\Lambda = 4C/\epsilon$. That is a polynomial scaling of $|\mathfrak{C}(\Lambda)|$ with N and d can ensure both a polynomial runtime cost to obtain $\kappa_\Lambda(\boldsymbol{x}, \boldsymbol{x}^{(i)}) g(\boldsymbol{x}^{(i)}, O)$ and a polynomial sample complexity n, leading to an *overall polynomial computational complexity* of our proposal (see SM G for details). In contrast, for an unbounded C such that $|\mathfrak{C}(\Lambda)|$ exponentially scales with d, the computational overhead of h_{s} becomes prohibitive for a large d, aligning with the observations from Theorem 1.

We next underscore that in many practical scenarios, the quantity C can be effectively bounded, allowing the proposed ML model to serve as a valuable complement to quantum tomography and classical simulations in comprehending quantum devices. One such scenario involves characterizing near-Clifford quantum circuits consisting of many CI gates and few non-Clifford gates, which find applications in quantum error correction [37, 38] and efficient implementation of approximate t-designs [39]. In this context, adopting the full expansion with $\Lambda = d$ is also computationally efficient, as $|\mathfrak{C}(d)| \sim O(N)$. Meanwhile, when focused on a specific class of quantum circuits described as CI + RZ with a fixed layout, our model surpasses classical simulation methods [40, 41] in runtime complexity by eliminating the dependence on the number of Clifford gates $G - d$.

Another scenario involves the advancement of variational quantum algorithms (VQAs) [29, 42, 43], a leading candidate of leveraging near-term quantum computers for practical utility in machine learning [44], quantum chemistry, and combinatorial optimization [45]. Recent studies have shown that in numerous VQAs, the gradients norm of trainable parameters yields $C \leq 1/\mathrm{poly}(N)$ [46–49] or $C \leq 1/\exp(N)$, a.k.a, barren plateaus [50–53]. These insights, coupled with the results in Theorem 2, suggest that our model can be used to pretrain VQA algorithms on the classical side to obtain effective initialization parameters before quantum computer execution, preserving access to quantum resources [54, 55]. Theoretically, our model broadens the class of VQAs amenable to classical simulation, pushing the boundaries of achieving practical utility with VQAs [56, 57].

Last, the complexity bound in Theorem 2 hints that the locality of the observable K is another factor dominating the performance of h_{s}. This exponential dependence arises from the use of Pauli-based classical shadow, and two potential solutions can be employed to alleviate this influence. The first solution involves adopting advanced variants of classical shadow to enhance the sample complexity bounds [58–61]. The second solution entails utilizing classical simulators to directly compute the quantity $\{\mathrm{Tr}(\rho(\boldsymbol{x}^{(i)})O)\}$ instead of shadow estimation $\{g(\boldsymbol{x}^{(i)}, O)\}$ in Eq. (6), with the sample complexity summarized in the following corollary.

Corollary 1 (Informal)**.** *Following notations in Theorem 2, when* $\{\mathrm{Tr}(\rho(\boldsymbol{x}^{(i)})O)\}_i$ *are computed by classical simulators and* $n \sim \widetilde{\mathcal{O}}(3^d B^2 d/\epsilon)$*, with high probability, the average prediction error is upper bounded by* ϵ*.*

Although using the classical simulators can improve the dependency of the locality of observable and remove the necessity of quantum resources, the price to pay is increasing the computational overhead and only restricting to a small constant d. Refer to SM H for the proofs, implementation details, and more discussions.

Remark. For clarity, we concentrate on illustrating how the proposed ML model applied to the bounded-gate circuits comprising CI gates alongside RZ gates. In SM I, we illustrate that our proposal and the relevant theoretical findings can be effectively extended to a broader context, i.e., the bounded-gate circuits composed of CI gates alongside parameterized multi-qubit gates generated by arbitrary Pauli strings.

IV. NUMERICAL RESULTS

We conduct numerical simulations on 60-qubit quantum circuits to assess the efficacy of the proposed ML model. The omitted details, as well as the demonstration of classically optimizing VQAs with smaller qubit sizes, are provided in SM J.

We first use the proposed ML model to predict the properties of rotational GHZ states. Mathematically, we define an N-qubit rotational GHZ states with $N = 60$ as $|\mathrm{GHZ}(\boldsymbol{x})\rangle = U(\boldsymbol{x})(|0\rangle^{\otimes N} + |1\rangle^{\otimes N})/\sqrt{2}$, where $U(\boldsymbol{x}) = \mathrm{RY}_1(\boldsymbol{x}_1) \otimes \mathrm{RY}_{30}(\boldsymbol{x}_2) \otimes \mathrm{RY}_{60}(\boldsymbol{x}_3)$ and the subscript j refers to apply RY gate to the j-th qubit. At the training stage, we constitute the dataset $\mathcal{T}_s$ containing $n = 30$ examples, where each example is obtained by uniformly sampling $\boldsymbol{x}$ from $[-\pi, \pi]^3$ and applying classical shadow to $|\mathrm{GHZ}(\boldsymbol{x})\rangle$ with the shot number $T = 1000$.

The first subtask is predicting a two-point correlation function, i.e., the expectation value of $C_{ij} = (X_iX_j + Y_iY_j + Z_iZ_j)/3$ for each pair of qubits (i, j), at new values of $\boldsymbol{x}$. To do so, the proposed ML model leverages $\mathcal{T}_s$ to form the classical representations with $\Lambda = 3$ and exploits these representations to proceed with prediction at $\boldsymbol{x}$. Fig. 2(a) depicts the predicted and actual values of the correlation function for a particular value of $\boldsymbol{x}$, showing reasonable agreement. The second subtask is exploiting how the prediction error depends on the truncation value Λ, the number of training examples n, and the shot number T. Fig. 2(b) showcases that when $T = 1000$ and the training data are sufficient (i.e., $n \geq 500$), the root mean squared (RMS) prediction error on 10 unseen test examples is almost the same for the full expansion (i.e., $\Lambda = 3$) and the proper truncation (i.e., $\Lambda = 2$). Besides, Fig. 2(c) indicates that when $n = 500$, the prediction error on the same test examples reaches a low value for both $\Lambda = 3$ and $\Lambda = 2$ once the shot number T exceeds a threshold value ($T \geq 50$). These results are consistent with our theoretical analysis.

We next apply the proposed ML model to predict properties of the state evolved by a global Hamiltonian $\mathsf{H} = J_1 \otimes_{i=1}^N Z_i + \sum_{i=1}^N X_i$, where J_1 is a predefined constant and N is the qubit count with $N = 60$. The initial state is fixed to be $|0\rangle^{\otimes N}$ and the circuit $U(\boldsymbol{x})$ is the Trotterized time evolution of H. By properly selecting the evolution time at each Trotter step and J_1, the Trotterized time-evolution circuit takes the form as $U(\boldsymbol{x}) = \prod_{j=1}^d (e^{-\imath \boldsymbol{x}_j \otimes_{i=1}^N Z_i} \otimes_{i=1}^N \mathrm{RX}(\pi/3))$. We consider two settings of the total number of Trotter steps, i.e., $d = 1$ and $d = 30$. For both cases, at the training stage, we constitute the dataset $\mathcal{T}_\mathsf{s}$ following the same rule presented in the last simulation. The only difference is the dataset size and the shot number, which is $n = 20$ and $T = 500$ in the former case, and $n = 50000$ and $T = 50$ in the latter case.

The first subtask is predicting the magnetization with $\langle \bar{Z} \rangle = \frac{1}{60} \sum_i \langle Z_i \rangle$ when $d = 1$. The comparison between the prediction with $\Lambda = 1$ (full expansion) and the exact value is shown in Fig. 3(a). Namely, we select 25 test examples evenly distributed across the interval $[-\pi, \pi]$. By increasing the number of training examples n from 10 to 20, the prediction values of the proposed ML model almost match the exact results.

The second subtask is much more difficult, which is predicting the magnetization with $\langle \bar{Z} \rangle$ when $d = 30$. In this scenario, full expansion becomes computationally infeasible, where the cardinality of frequency set is 3^{30}. In this regard, we evaluate the performance of the proposed ML models by adopting low truncation values with $\Lambda \in \{1, 2, 3, 4\}$. Fig. 3(b) demonstrates the RMS prediction error of our proposal with the varied number of training examples, i.e., $n \in \{10, 10^2, 10^3, 10^4\}$. An immediate observation is that with the increased n, the averaged prediction error continuously decreases for all settings of Λ. In addition, the achieved results hint that the proposed model with a larger truncation value Λ may require more training examples to surpass the one with a smaller truncation value. For instance, the proposed ML model with $\Lambda = 2$ attains a better performance compared to $\Lambda = 1$ when $n = 10^4$. It is expected that the model with $\Lambda = 3, 4$ can attain better performance by further increasing n.

V. DISCUSSION AND OUTLOOK

In this study, we prove that learning bounded-gate quantum circuits with incoherent measurements is sample efficient but computationally hard. Furthermore, we devise a provably efficient ML algorithm to predict the incoherent dynamics of bounded-gate quantum circuits, transitioning from exponential to polynomial scaling. The achieved results provide both theoretical insights and practical applications, demonstrating the efficacy of ML in comprehending and advancing quantum computing.

Several crucial research avenues merit further exploration. First, our study addresses the scenario of noiseless quantum operations. An important and promising area for future investigation is the development of provably efficient learning algorithms capable of predicting the incoherent dynamics of noisy bounded-gate quantum circuits [62–64]. Secondly, it is vital to extend our theoretical results from average-case scenarios to worst-case scenarios, wherein classical control can be sampled from arbitrary distributions rather than solely from the uniform distribution [65, 66]. Such extensions would deepen our understanding of the capabilities and limitations of employing machine learning to comprehend quantum circuits. Moreover, there exists ample opportunity to enhance our proposed learning algorithm by exploring alternative kernel methods, such as the positive good kernels [67] adopted in Ref. [68]. In addition, independent of this work, a crucial research topic is understanding the hardness of classically simulating the incoherent dynamics of bounded-gate quantum circuits with simple input quantum states in the measure of the averaged prediction error. Last, it would be intriguing to explore whether deep learning algorithms [69] can achieve provably improved prediction performance and efficiency for specific classes of quantum circuits.

[1] Ulf Leonhardt. Quantum-state tomography and discrete wigner function. *Physical review letters*, 74(21):4101, 1995.

[2] Joseph B Altepeter, David Branning, Evan Jeffrey, TC Wei, Paul G Kwiat, Robert T Thew, Jeremy L O'Brien, Michael A Nielsen, and Andrew G White. Ancilla-assisted quantum process tomography. *Physical Review Letters*, 90(19):193601, 2003.

[3] Masoud Mohseni, Ali T Rezakhani, and Daniel A Lidar. Quantum-process tomography: Resource analysis of different strategies. *Physical Review A*, 77(3):032322, 2008.

[4] Scott Aaronson. The learnability of quantum states. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2088):3089–3114, 2007.

[5] Scott Aaronson. Shadow tomography of quantum states. In *Proceedings of the 50th annual ACM SIGACT symposium on theory of computing*, pages 325–338, 2018.

[6] Igor L Markov and Yaoyun Shi. Simulating quantum computation by contracting tensor networks. *SIAM Journal on Computing*, 38(3):963–981, 2008.

[7] Benjamin Villalonga, Sergio Boixo, Bron Nelson, Christopher Henze, Eleanor Rieffel, Rupak Biswas, and Salvatore Mandrà. A flexible high-performance simulator for verifying and benchmarking quantum circuits implemented on real hardware. *npj Quantum Information*, 5(1):86, 2019.

[8] J Ignacio Cirac, David Perez-Garcia, Norbert Schuch, and Frank Verstraete. Matrix product states and projected entangled pair states: Concepts, symmetries, theorems. *Reviews of Modern Physics*, 93(4):045003, 2021.

[9] Feng Pan and Pan Zhang. Simulation of quantum circuits using the big-batch tensor network method. *Physical Review Letters*, 128(3):030501, 2022.

[10] Sergey Bravyi and David Gosset. Improved classical simulation of quantum circuits dominated by clifford gates. *Physical review letters*, 116(25):250501, 2016.

[11] Sergey Bravyi, Dan Browne, Padraic Calpin, Earl Campbell, David Gosset, and Mark Howard. Simulation of quantum circuits by low-rank stabilizer decompositions. *Quantum*, 3:181, 2019.

[12] Tomislav Begušić and Garnet Kin Chan. Fast classical simulation of evidence for the utility of quantum computing before fault tolerance. *arXiv preprint arXiv:2306.16372*, 2023.

[13] Jielun Chen, EM Stoudenmire, Steven R White, et al. Quantum fourier transform has small entanglement. *PRX Quantum*, 4(4):040318, 2023.

[14] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.

[15] Youngseok Kim, Andrew Eddins, Sajant Anand, Ken Xuan Wei, Ewout Van Den Berg, Sami Rosenblatt, Hasan Nayfeh, Yantao Wu, Michael Zaletel, Kristan Temme, et al. Evidence for the utility of quantum computing before fault tolerance. *Nature*, 618(7965):500–505, 2023.

[16] Dolev Bluvstein, Simon J Evered, Alexandra A Geim, Sophie H Li, Hengyun Zhou, Tom Manovitz, Sepehr Ebadi, Madelyn Cain, Marcin Kalinowski, Dominik Hangleiter, et al. Logical quantum processor based on reconfigurable atom arrays. *Nature*, pages 1–3, 2023.

[17] Valentin Gebhart, Raffaele Santagati, Antonio Andrea Gentile, Erik M Gauger, David Craig, Natalia Ares, Leonardo Banchi, Florian Marquardt, Luca Pezzè, and Cristian Bonato. Learning quantum systems. *Nature Reviews Physics*,

5(3):141–156, 2023.
[18] Giacomo Torlai and Roger G Melko. Neural decoder for topological codes. *Physical review letters*, 119(3):030501, 2017.
[19] Yexiong Zeng, Zheng-Yang Zhou, Enrico Rinaldi, Clemens Gneiting, and Franco Nori. Approximate autonomous quantum error correction with reinforcement learning. *Physical Review Letters*, 131(5):050601, 2023.
[20] Hossein Dehghani, Ali Lavasani, Mohammad Hafezi, and Michael J Gullans. Neural-network decoders for measurement induced phase transitions. *Nature Communications*, 14(1):2918, 2023.
[21] Kevin Reuer, Jonas Landgraf, Thomas Fösel, James O'Sullivan, Liberto Beltrán, Abdulkadir Akin, Graham J Norris, Ants Remm, Michael Kerschbaum, Jean-Claude Besse, et al. Realizing a deep reinforcement learning agent for real-time quantum feedback. *Nature Communications*, 14(1):7138, 2023.
[22] Askery Canabarro, Samuraí Brito, and Rafael Chaves. Machine learning nonlocal correlations. *Physical review letters*, 122(20):200401, 2019.
[23] Zhenyu Chen, Xiaodie Lin, and Zhaohui Wei. Certifying unknown genuine multipartite entanglement by neural networks. *Quantum Science and Technology*, 2022.
[24] Dominik Koutnỳ, Laia Ginés, Magdalena Moczała-Dusanowska, Sven Höfling, Christian Schneider, Ana Predojević, and Miroslav Ježek. Deep learning of quantum entanglement from incomplete measurements. *Science Advances*, 9(29):eadd7131, 2023.
[25] Elizabeth R Bennewitz, Florian Hopfmueller, Bohdan Kulchytskyy, Juan Carrasquilla, and Pooya Ronagh. Neural error mitigation of near-term quantum simulations. *Nature Machine Intelligence*, 4(7):618–624, 2022.
[26] Shi-Xin Zhang, Zhou-Quan Wan, Chee-Kong Lee, Chang-Yu Hsieh, Shengyu Zhang, and Hong Yao. Variational quantum-neural hybrid eigensolver. *Physical Review Letters*, 128(12):120502, 2022.
[27] Yuxuan Du, Yibo Yang, Tongliang Liu, Zhouchen Lin, Bernard Ghanem, and Dacheng Tao. Shadownet for data-centric quantum system learning. *arXiv preprint arXiv:2308.11290*, 2023.
[28] Victor Wei, WA Coish, Pooya Ronagh, and Christine A Muschik. Neural-shadow quantum state tomography. *arXiv preprint arXiv:2305.01078*, 2023.
[29] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021.
[30] Jens Eisert, Dominik Hangleiter, Nathan Walk, Ingo Roth, Damian Markham, Rhea Parekh, Ulysse Chabaud, and Elham Kashefi. Quantum certification and benchmarking. *Nature Reviews Physics*, 2(7):382–390, 2020.
[31] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.
[32] Neil J Ross and Peter Selinger. Optimal ancilla-free clifford+ t approximation of z-rotations. *arXiv preprint arXiv:1403.2975*, 2014.
[33] Andreas Elben, Steven T Flammia, Hsin-Yuan Huang, Richard Kueng, John Preskill, Benoît Vermersch, and Peter Zoller. The randomized measurement toolbox. *Nature Reviews Physics*, pages 1–16, 2022.
[34] Anurag Anshu and Srinivasan Arunachalam. A survey on the complexity of learning quantum states. *Nature Reviews Physics*, pages 1–11, 2023.
[35] Leonardo Banchi, Jason Luke Pereira, Sharu Theresa Jose, and Osvaldo Simeone. Statistical complexity of quantum learning. *arXiv preprint arXiv:2309.11617*, 2023.
[36] Haimeng Zhao, Laura Lewis, Ishaan Kannan, Yihui Quek, Hsin-Yuan Huang, and Matthias C Caro. Learning quantum states and unitaries of bounded gate complexity. *arXiv preprint arXiv:2310.19882*, 2023.
[37] A Robert Calderbank, Eric M Rains, Peter W Shor, and Neil JA Sloane. Quantum error correction and orthogonal geometry. *Physical Review Letters*, 78(3):405, 1997.
[38] Barbara M Terhal. Quantum error correction for quantum memories. *Reviews of Modern Physics*, 87(2):307, 2015.
[39] Jonas Haferkamp, Felipe Montealegre-Mora, Markus Heinrich, Jens Eisert, David Gross, and Ingo Roth. Quantum homeopathy works: Efficient unitary designs with a system-size independent number of non-clifford gates. *arXiv preprint arXiv:2002.09524*, 2020.
[40] Scott Aaronson and Daniel Gottesman. Improved simulation of stabilizer circuits. *Physical Review A*, 70(5):052328, 2004.
[41] Ching-Yi Lai and Hao-Chung Cheng. Learning quantum circuits of some t gates. *IEEE Transactions on Information Theory*, 68(6):3951–3964, 2022.
[42] Jinkai Tian, Xiaoyu Sun, Yuxuan Du, Shanshan Zhao, Qing Liu, Kaining Zhang, Wei Yi, Wanrong Huang, Chaoyue Wang, Xingyao Wu, et al. Recent advances for quantum neural networks in generative learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
[43] 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(1):015004, 2022.
[44] Weikang Li and Dong-Ling Deng. Recent advances for quantum classifiers. *Science China Physics, Mechanics & Astronomy*, 65(2):1–23, 2022.
[45] Edward Farhi and Aram W Harrow. Quantum supremacy through the quantum approximate optimization algorithm. *arXiv preprint arXiv:1602.07674*, 2016.
[46] Arthur Pesah, Marco 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(4):041011, 2021.
[47] Kaining Zhang, Min-Hsiu Hsieh, Liu Liu, and Dacheng Tao. Toward trainability of deep quantum neural networks. *arXiv preprint arXiv:2112.15002*, 2021.

[48] Xinbiao Wang, Junyu Liu, Tongliang Liu, Yong Luo, Yuxuan Du, and Dacheng Tao. Symmetric pruning in quantum neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

[49] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J Coles, and Marco Cerezo. Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 3(6):542–551, 2023.

[50] 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):1–6, 2018.

[51] Marco 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):1–12, 2021.

[52] Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. Entanglement-induced barren plateaus. *PRX Quantum*, 2(4):040316, 2021.

[53] 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(4):045015, 2022.

[54] James Dborin, Fergus Barratt, Vinul Wimalaweera, Lewis Wright, and Andrew G Green. Matrix product state pre-training for quantum machine learning. *Quantum Science and Technology*, 7(3):035014, 2022.

[55] Manuel S Rudolph, Jacob Miller, Danial Motlagh, Jing Chen, Atithi Acharya, and Alejandro Perdomo-Ortiz. Synergistic pretraining of parametrized quantum circuits via tensor networks. *Nature Communications*, 14(1):8367, 2023.

[56] Franz J Schreiber, Jens Eisert, and Johannes Jakob Meyer. Classical surrogates for quantum learning models. *Physical Review Letters*, 131(10):100803, 2023.

[57] M Cerezo, Martin Larocca, Diego García-Martín, NL Diaz, Paolo Braccia, Enrico Fontana, Manuel S Rudolph, Pablo Bermejo, Aroosa Ijaz, Supanut 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.

[58] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Efficient estimation of pauli observables by derandomization. *Physical review letters*, 127(3):030503, 2021.

[59] H Chau Nguyen, Jan Lennart Bönsel, Jonathan Steinberg, and Otfried Gühne. Optimizing shadow tomography with generalized measurements. *Physical Review Letters*, 129(22):220502, 2022.

[60] Qingyue Zhang, Qing Liu, and You Zhou. Minimal clifford shadow estimation by mutually unbiased bases. *arXiv preprint arXiv:2310.18749*, 2023.

[61] Benoît Vermersch, Aniket Rath, Bharathan Sundar, Cyril Branciard, John Preskill, and Andreas Elben. Enhanced estimation of quantum properties with common randomized measurements. *arXiv preprint arXiv:2304.12292*, 2023.

[62] Yuguo Shao, Fuchuan Wei, Song Cheng, and Zhengwei Liu. Simulating quantum mean values in noisy variational quantum algorithms: A polynomial-scale approach. *arXiv preprint arXiv:2306.05804*, 2023.

[63] Enrico Fontana, Manuel S Rudolph, Ross Duncan, Ivan Rungger, and Cristina Cîrstoiu. Classical simulations of noisy variational quantum circuits. *arXiv preprint arXiv:2306.05400*, 2023.

[64] Hsin-Yuan Huang, Sitan Chen, and John Preskill. Learning to predict arbitrary quantum processes. *PRX Quantum*, 4(4):040337, 2023.

[65] Srinivasan Arunachalam and Ronald de Wolf. Guest column: A survey of quantum learning theory. *ACM Sigact News*, 48(2):41–67, 2017.

[66] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Physical Review Letters*, 126(19):190505, 2021.

[67] Elias M Stein and Rami Shakarchi. *Fourier analysis: an introduction*, volume 1. Princeton University Press, 2011.

[68] Yanming Che, Clemens Gneiting, and Franco Nori. Exponentially improved efficient machine learning for quantum many-body states with provable guarantees. *arXiv preprint arXiv:2304.04353*, 2023.

[69] Yang Qian, Yuxuan Du, Zhenliang He, Min-hsiu Hsieh, and Dacheng Tao. Multimodal deep representation learning for quantum cross-platform verification. *arXiv preprint arXiv:2311.03713*, 2023.

[70] Hsin-Yuan Huang. Learning quantum states from their classical shadows. *Nature Reviews Physics*, 4(2):81–81, 2022.

[71] Maria Schuld and Nathan Killoran. Is quantum advantage the right goal for quantum machine learning? *PRX Quantum*, 3:030101, Jul 2022.

[72] Ryan Sweke, Erik Recio, Sofiene Jerbi, Elies Gil-Fuster, Bryce Fuller, Jens Eisert, and Johannes Jakob Meyer. Potential and limitations of random fourier features for dequantizing quantum machine learning. *arXiv preprint arXiv:2309.11647*, 2023.

[73] Manuel S Rudolph, Enrico Fontana, Zoë Holmes, and Lukasz Cincio. Classical surrogate simulation of quantum systems with lowesa. *arXiv preprint arXiv:2308.09109*, 2023.

[74] Gregory Boyd and Bálint Koczor. Training variational quantum circuits with covar: covariance root finding with classical shadows. *Physical Review X*, 12(4):041022, 2022.

[75] Stefan H Sack, Raimel A Medina, Alexios A Michailidis, Richard Kueng, and Maksym Serbyn. Avoiding barren plateaus using classical shadows. *PRX Quantum*, 3(2):020365, 2022.

[76] Sofiene Jerbi, Casper Gyurik, Simon C Marshall, Riccardo Molteni, and Vedran Dunjko. Shadows of quantum machine learning. *arXiv preprint arXiv:2306.00061*, 2023.

[77] Steven T Flammia and Yi-Kai Liu. Direct fidelity estimation from few pauli measurements. *Physical review letters*, 106(23):230501, 2011.

[78] Michał Horodecki, Paweł Horodecki, and Ryszard Horodecki. Separability of n-particle mixed states: necessary and sufficient conditions in terms of linear maps. *Physics Letters A*, 283(1-2):1–7, 2001.

[79] Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V Albert, and John Preskill. Provably efficient machine learning for quantum many-body problems. *Science*, 377(6613):eabk3333, 2022.

[80] Laura Lewis, Hsin-Yuan Huang, Viet T. Tran, Sebastian Lehner, Richard Kueng, and John Preskill. Improved machine learning algorithm for predicting ground state properties. *Nature Communications*, 15(1):895, 2024.

[81] Emilio Onorati, Cambyse Rouzé, Daniel Stilck França, and James D Watson. Efficient learning of ground & thermal states within phases of matter. *arXiv preprint arXiv:2301.12946*, 2023.

[82] Emilio Onorati, Cambyse Rouzé, Daniel Stilck França, and James D Watson. Provably efficient learning of phases of matter via dissipative evolutions. *arXiv preprint arXiv:2311.07506*, 2023.

[83] Hsin-Yuan Huang, Yunchao Liu, Michael Broughton, Isaac Kim, Anurag Anshu, Zeph Landau, and Jarrod R McClean. Learning shallow quantum circuits. *arXiv preprint arXiv:2401.10095*, 2024.

[84] Alba Cervera-Lierta, Jakob S Kottmann, and Alán Aspuru-Guzik. Meta-variational quantum eigensolver: Learning energy profiles of parameterized hamiltonians for quantum simulation. *PRX Quantum*, 2(2):020329, 2021.

[85] He-Liang Huang, Yuxuan Du, Ming Gong, Youwei Zhao, Yulin Wu, Chaoyue Wang, Shaowei Li, Futian Liang, Jin Lin, Yu Xu, et al. Experimental quantum generative adversarial networks for image generation. *Physical Review Applied*, 16(2):024051, 2021.

[86] Shi-Xin Zhang, Chang-Yu Hsieh, Shengyu Zhang, and Hong Yao. Neural predictor based quantum architecture search. *Machine Learning: Science and Technology*, 2(4):045027, 2021.

[87] Nishant Jain, Brian Coyle, Elham Kashefi, and Niraj Kumar. Graph neural network initialisation of quantum approximate optimisation. *Quantum*, 6:861, 2022.

[88] Cong Lei, Yuxuan Du, Peng Mi, Jun Yu, and Tongliang Liu. Neural auto-designer for enhanced quantum kernels. *arXiv preprint arXiv:2401.11098*, 2024.

[89] Enrico Fontana, Ivan Rungger, Ross Duncan, and Cristina Cîrstoiu. Efficient recovery of variational quantum algorithms landscapes using classical signal processing. *arXiv preprint arXiv:2208.05958*, 2022.

[90] Nikita A. Nemkov, Evgeniy O. Kiktenko, and Aleksey K. Fedorov. Fourier expansion in variational quantum algorithms. *Phys. Rev. A*, 108:032406, Sep 2023.

[91] Jonas Landman, Slimane Thabet, Constantin Dalyac, Hela Mhiri, and Elham Kashefi. Classically Approximating Variational Quantum Machine Learning with Random Fourier Features, 2022. arXiv:2210.13200v1.

[92] Berta Casas and Alba Cervera-Lierta. Multi-dimensional fourier series with quantum circuits. *arXiv preprint arXiv:2302.03389*, 2023.

[93] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nature Physics*, 14(5):447–450, 2018.

[94] Shahnawaz Ahmed, Carlos Sánchez Muñoz, Franco Nori, and Anton Frisk Kockum. Quantum state tomography with conditional generative adversarial networks. *Physical Review Letters*, 127(14):140502, 2021.

[95] Peter Cha, Paul Ginsparg, Felix Wu, Juan Carrasquilla, Peter L McMahon, and Eun-Ah Kim. Attention-based quantum tomography. *Machine Learning: Science and Technology*, 3(1):01LT01, 2021.

[96] Giuseppe Carleo, Yusuke Nomura, and Masatoshi Imada. Constructing exact representations of quantum many-body systems with deep neural networks. *Nature communications*, 9(1):1–11, 2018.

[97] Juan Carrasquilla, Giacomo Torlai, Roger G Melko, and Leandro Aolita. Reconstructing quantum states with generative models. *Nature Machine Intelligence*, 1(3):155–161, 2019.

[98] Alistair WR Smith, Johnnie Gray, and MS Kim. Efficient quantum state sample tomography with basis-dependent neural networks. *PRX Quantum*, 2(2):020348, 2021.

[99] Tobias Schmale, Moritz Reh, and Martin Gärttner. Efficient quantum state tomography with convolutional neural networks. *npj Quantum Information*, 8(1):115, 2022.

[100] Xiaoqian Zhang, Maolin Luo, Zhaodi Wen, Qin Feng, Shengshi Pang, Weiqi Luo, and Xiaoqi Zhou. Direct fidelity estimation of quantum states using machine learning. *Physical Review Letters*, 127(13):130503, 2021.

[101] Yan Zhu, Ya-Dong Wu, Ge Bai, Dong-Sheng Wang, Yuexuan Wang, and Giulio Chiribella. Flexible learning of quantum states with generative query neural networks. *Nature Communications*, 13(1):6222, 2022.

[102] Amit Nir, Eran Sela, Roy Beck, and Yohai Bar-Sinai. Machine-learning iterative calculation of entropy for physical systems. *Proceedings of the National Academy of Sciences*, 117(48):30234–30240, 2020.

[103] Tailong Xiao, Jingzheng Huang, Hongjing Li, Jianping Fan, and Guihua Zeng. Intelligent certification for quantum simulators via machine learning. *npj Quantum Information*, 8(1):138, 2022.

[104] Ya-Dong Wu, Yan Zhu, Ge Bai, Yuexuan Wang, and Giulio Chiribella. Quantum similarity testing with convolutional neural networks. *Physical Review Letters*, 130(21):210601, 2023.

[105] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[106] Xinbiao Wang, Yuxuan Du, Zhuozhuo Tu, Yong Luo, Xiao Yuan, and Dacheng Tao. Transition role of entangled data in quantum machine learning. *arXiv preprint arXiv:2306.03481*, 2023.

[107] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[108] 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(1):4919, 2022.

[109] Yuxuan Du, Zhuozhuo Tu, Xiao Yuan, and Dacheng Tao. Efficient measure for the expressivity of variational quantum algorithms. *Physical Review Letters*, 128(8):080506, 2022.

[110] Yuxuan Du, Zhuozhuo Tu, Bujiao Wu, Xiao Yuan, and Dacheng Tao. Power of quantum generative learning. *arXiv preprint arXiv:2205.04730*, 2022.

[111] Riccardo Molteni, Casper Gyurik, and Vedran Dunjko. Exponential quantum advantages in learning quantum observables from classical data. *arXiv preprint arXiv:2405.02027*, 2024.

[112] John Watrous. Quantum computational complexity. *arXiv preprint arXiv:0804.3401*, 2008.

[113] Giacomo Torlai and Matthew Fishman. PastaQ: A package for simulation, tomography and analysis of quantum computers, 2020.

[114] 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(7671):242–246, 2017.

[115] 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(7747):209–212, 2019.

[116] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3):032331, 2019.

[117] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[118] Yuxuan Du, Yibo Yang, Dacheng Tao, and Min-Hsiu Hsieh. Problem-dependent power of quantum neural networks on multiclass classification. *Physical Review Letters*, 131(14):140601, 2023.

[119] Elies Gil-Fuster, Jens Eisert, and Carlos Bravo-Prieto. Understanding quantum machine learning also requires rethinking generalization. *Nature Communications*, 15(1):1–12, 2024.

[120] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018.

[121] 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.

[122] Iris Cong, Soonwon Choi, and Mikhail D Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, 2019.

[123] Johannes Herrmann, Sergi Masot Llima, Ants Remm, Petr Zapletal, Nathan A McMahon, Colin Scarato, François Swiadek, Christian Kraglund Andersen, Christoph Hellings, Sebastian Krinner, et al. Realizing quantum convolutional neural networks on a superconducting quantum processor to recognize quantum phases. *Nature communications*, 13(1):4144, 2022.

Supplementary Material of "Efficient learning for bounded-gate quantum circuits with incoherent measurements"

Contents

SM A: Preliminary and literature review

This section contains the necessary backgrounds of the whole work. Specifically, in Supplementary Material (SM) A 1, we present the basic information of classical shadow. Then, in SM A 2, we recap the trigonometric expansion of quantum circuits implemented by RZ and Clifford gates. Afterward, in SM A 3, we elaborate on the generality of the explored problem in our work for early-stage quantum computation. Last, we provide a literature review in SM A 4.

1. Classical shadow and its application in estimating incoherent dynamics of quantum states

Classical shadow represents a computationally and memory-efficient approach for storing quantum states on classical computers, primarily used for estimating the expectation values of local observables [31]. The fundamental principle of classical shadow lies in the 'measure first and ask questions later' strategy. In this subsection, we outline the utilization of classical shadow to estimate linear functions under Pauli-based measurements. Interested readers can refer to tutorials and surveys [33, 70] for more comprehensive details.

Formalism of classical shadow. The general scheme of classical shadow for an unknown N-qubit state ρ is repeating the following procedure T times. At each time, the state ρ is first operated with a unitary U randomly sampled from the predefined unitary ensemble $\mathcal{U}$ and then each qubit is measured under the Z basis to obtain an N-bit string denoted by $\boldsymbol{b} \in \{0,1\}^N$. There exists a linear map $\mathcal{M}(\cdot)$ satisfying

$$\mathcal{M}(\rho) = \mathbb{E}_{U\sim\mathcal{U}}\mathbb{E}_{\boldsymbol{b}\sim\mathbb{P}(\boldsymbol{b})} U^\dagger |\boldsymbol{b}\rangle \langle \boldsymbol{b}| U = \sum_{\boldsymbol{b}} \int \mathsf{d}U U^\dagger |\boldsymbol{b}\rangle \langle \boldsymbol{b}| U \langle \boldsymbol{b}| U\rho U^\dagger |\boldsymbol{b}\rangle , \tag{A1}$$

where $\mathbb{P}(\boldsymbol{b}) = \langle \boldsymbol{b}| U\rho U^\dagger |\boldsymbol{b}\rangle$. Such a linear map implies that the unknown state ρ can be formulated as

$$\rho = \sum_{\boldsymbol{b}} \int \mathsf{d}U \mathcal{M}^{-1}\Big(U^\dagger |\boldsymbol{b}\rangle \langle \boldsymbol{b}| U\Big) \langle \boldsymbol{b}| U\rho U^\dagger |\boldsymbol{b}\rangle . \tag{A2}$$

In other words, the state ρ can be estimated by sampling the snapshot with M times following the distribution $\mathbb{P}(\boldsymbol{b})$. Define the t-th snapshot as $U_t^\dagger |\boldsymbol{b}_t\rangle \langle \boldsymbol{b}_t| U_t$ with $t \in [T]$ and $U_t \sim \mathcal{U}$. The estimated state of ρ is

$$\hat{\rho}_T = \frac{1}{T}\sum_{t=1}^{T} \hat{\rho}_t, \text{ with } \hat{\rho}_t = \mathcal{M}^{-1}(U_t^\dagger |\boldsymbol{b}_t\rangle \langle \boldsymbol{b}_t| U_t). \tag{A3}$$

As pointed out in Ref. [31], the random Pauli basis is not only experimentally friendly but also enables a succinct form of the classical shadow. When Pauli-based measurements are adopted, it is equivalent to setting the unitary ensemble $\mathcal{U}$ as single-qubit Clifford gates, i.e., $U_t = U_{1,t} \otimes \cdots U_{j,t} \cdots \otimes U_{N,t} \sim \mathcal{U} = \mathrm{CI}(2)^{\otimes N}$ with uniform weights. In this case, the inverse snapshot takes the form as

$$\hat{\rho}_t = \mathcal{M}^{-1}\left(\bigotimes_{j=1}^{N} U_{j,t}^\dagger |\boldsymbol{b}_{j,t}\rangle\langle \boldsymbol{b}_{j,t}| U_{j,t}\right) = \bigotimes_{j=1}^{N} \mathcal{D}_{1/3}^{-1}\left(U_{j,t}^\dagger |\boldsymbol{b}_{j,t}\rangle\langle \boldsymbol{b}_{j,t}| U_{j,t}\right) = \bigotimes_{j=1}^{N}\left(3U_{j,t}^\dagger |\boldsymbol{b}_{j,t}\rangle\langle \boldsymbol{b}_{j,t}| U_{j,t} - \mathbb{I}_2\right), \tag{A4}$$

where $\mathcal{D}_{1/3}^{-1}(Y) = 3Y - \mathrm{Tr}(Y)\mathbb{I}$.

Estimate incoherent dynamics (linear properties of quantum states). The tensor product form of the classical shadow in Eq. (A4) allows an efficient procedure to predict linear properties of the state ρ. A typical instance is estimating the expectation value $\mathrm{Tr}(\rho O)$ with O being a local observable. Mathematically, suppose the local observable to be a Pauli string, i.e., $O = P_1 \otimes ...P_i... \otimes P_N$ with $P_i \in \{X, Y, Z, \mathbb{I}\}$ for $\forall i \in [N]$, the estimation of the classical shadow is

$$\mathrm{Tr}\left(\hat{\rho}_T(P_1 \otimes ...P_i... \otimes P_N)\right) = \frac{1}{T}\sum_{t=1}^{T}\prod_{j=1}^{N} \mathrm{Tr}\left(\left(3U_{j,t}^\dagger |\boldsymbol{b}_{j,t}\rangle\langle \boldsymbol{b}_{j,t}| U_{j,t} - \mathbb{I}_2\right) P_j\right), \tag{A5}$$

which is memory and computation efficient. Namely, one only needs $\mathcal{O}(NT)$ bits to store $\hat{\rho}_T$ and $\mathcal{O}(NT)$ computational time to load $\hat{\rho}_T$ to the classical memory. Next, when the locality of the observable $O = \sum_i O_i$ is $K \sim \mathcal{O}(1)$, the shadow estimation of the expectation value, i.e., $\mathrm{Tr}(\hat{\rho}O_i)$, can be performed in $\mathcal{O}(T)$ time after $\hat{\rho}_T$ is loaded into the classical memory [31].

2. Pauli transfer matrix and Trigonometric expansion of RZ + CI quantum circuits

Pauli Transfer Matrix. Here we use Pauli-Liouville representation to reformulate the quantum state and the observable. Specifically, an N-qubit state can be represented as a 4^N-dimensional vector, whose i-th entry is

$$|\cdot\rangle\rangle = \mathrm{Tr}(\cdot P_i) \text{ with } P_i \in \{\mathbb{I}, X, Y, Z\}^N. \tag{A6}$$

For example, the state $|0\rangle^{\otimes N}$ satisfies $(|0\rangle\langle 0|)^{\otimes N} = ((\mathbb{I}+Z)/2)^{\otimes N}$, which indicates that its representation under the Pauli basis yields

$$|0\rangle^{\otimes N} \equiv |\mathbf{0}\rangle\rangle = ((\mathbb{I}+Z)/2)^{\otimes N} = \left([1,0,0,1]^\top\right)^{\otimes N}. \tag{A7}$$

Similarly, the normalized Pauli operator O under the Pauli basis yields

$$|O\rangle\rangle = \left[\mathrm{Tr}\left(OP_1\right), \cdots, \mathrm{Tr}\left(OP_{4^N}\right)\right]^\top. \tag{A8}$$

The unitary operator can also be represented by Pauli basis, i.e., given a parameterized circuit $U(\boldsymbol{\theta})$, its Pauli Transfer Matrix (PTM) $\mathfrak{U}(\boldsymbol{\theta})$ yields

$$[\mathfrak{U}(\boldsymbol{\theta})]_{ij} = \langle\langle P_i|\,\mathfrak{U}(\boldsymbol{\theta})|P_j\rangle\rangle = \mathrm{Tr}\left(P_i U(\boldsymbol{\theta}) P_j U(\boldsymbol{\theta})^\dagger\right). \tag{A9}$$

For example, the PTM representation of $\mathrm{RZ}(\boldsymbol{x}_j)$ gate is

$$\mathfrak{R}_{\mathrm{Z}}(\boldsymbol{x}_j) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\boldsymbol{x}_j) & \sin(\boldsymbol{x}_j) & 0 \\ 0 & -\sin(\boldsymbol{x}_j) & \cos(\boldsymbol{x}_j) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = D_0 + \cos(\boldsymbol{x}_j) D_1 + \sin(\boldsymbol{x}_j) D_2, \tag{A10}$$

where $D_0 = \begin{bmatrix} 1&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&1 \end{bmatrix}$, $D_1 = \begin{bmatrix} 0&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&0 \end{bmatrix}$, and $D_2 = \begin{bmatrix} 0&0&0&0 \\ 0&0&-1&0 \\ 0&1&0&0 \\ 0&0&0&0 \end{bmatrix}$.

Trigonometric expansion of $\mathrm{RZ}+\mathbf{CI}$ quantum circuits. A critical research line in quantum computing involves determining if variational quantum algorithms can provide meaningful advantages over state-of-the-art classical methods through dequantization via Fourier expansion [71, 72]. Despite differing objectives, the techniques developed, particularly the Low-Weight Efficient Simulation Algorithm (LOWESA) [63, 73], serve as inspiration for our approach to predicting the incoherent dynamics of bounded-gate quantum circuits. In the following context, we delve into the mechanics of LOWESA and postpone the discussion on the connection and difference between our work and this research line to SM A 4.

We now recap the mechanism of LOWESA. When an N-qubit quantum circuit $U(\boldsymbol{x})$ is composed of d RZ gates and $G-d$ CI gates, the state representation under Pauli-basis expansion yields

$$\rho(\boldsymbol{x}) = U(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N} U(\boldsymbol{x})^\dagger = \sum_{\boldsymbol{\omega}} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \langle\langle 0|\,\mathfrak{U}_{\boldsymbol{\omega}}^\dagger \equiv \sum_{\boldsymbol{\omega}} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \rho_{\boldsymbol{\omega}}. \tag{A11}$$

The notation $\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})$ with $\boldsymbol{\omega} \in \{0, \pm 1\}^d$ refers to the *trigonometric monomial basis* with values

$$\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) = \prod_{i=1}^{d} \begin{cases} 1 & \text{if } \boldsymbol{\omega}_i = 0 \\ \cos(\boldsymbol{x}_i) & \text{if } \boldsymbol{\omega}_i = 1 \\ \sin(\boldsymbol{x}_i) & \text{if } \boldsymbol{\omega}_i = -1 \end{cases}. \tag{A12}$$

Moreover, $\mathfrak{U}_{\boldsymbol{\omega}}$ in Eq. (A11) refers to the *purely-Clifford circuit* for the path indexed by $\boldsymbol{\omega}$ in the sense that in the path $\boldsymbol{\omega}$, each RZ gate at the i-th position with $i \in [d]$ is replaced by the operator $D_{\boldsymbol{\omega}_i}$ in Eq. (A10). The quantum mean value under an observable O is

$$f(\boldsymbol{x}, O) \equiv \mathrm{Tr}(\rho(\boldsymbol{x}) O) = \sum_{\boldsymbol{\omega}} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \langle\langle 0|\,\mathfrak{U}_{\boldsymbol{\omega}}^\dagger\,|O\rangle\rangle. \tag{A13}$$

It is noteworthy that LOWESA is a purely classical approach to estimating expectation values of simple variational quantum circuits. The term 'simple' is reflected by the fact that even for low dimensional classical controls (i.e., a small d), its computational complexity is prohibited by other two key factors: the initial state and the number of CI gates $G-d$. This is because when the initial state $(|0\rangle\langle 0|)^{\otimes N}$ is substitute a complicated one (e.g., a state with exponentially many terms under Pauli basis), and $G-d$ becomes large, the computational overhead of calculating $\langle\langle 0|\,\mathfrak{U}_{\boldsymbol{\omega}}^\dagger\,|O\rangle\rangle$ is unaffordable by classical simulators.

3. Generality of exploring the learnability of $\mathcal{F}$ for early-stage quantum computation

Here we discuss the generality of the explored mean-value space in the main text, i.e.,

$$\mathcal{F} = \left\{ f(\boldsymbol{x}, O) = \mathrm{Tr}\left(U(\boldsymbol{x})\rho_0 U(\boldsymbol{x})^\dagger O\right) \middle| \boldsymbol{x} \in [-\pi, \pi]^d, O \sim \mathbb{D}_O \right\}, \tag{A14}$$

where $U(\boldsymbol{x}) = \prod_{l=1}^{d}(\mathrm{RZ}(\boldsymbol{x}_l)u_e)$ is composed of d $\mathrm{RZ}(\boldsymbol{x}_l)$ gates and $G-d$ CI gates denoted by U_e with $\mathrm{CI} = \{H, S, \mathrm{CNOT}\}$. This formalism encompasses many applications for early-stage quantum computing, including variational quantum algorithms, classical-shadow-based algorithms, and quantum system certification. A common feature of these applications is that the exploited quantum circuit is generally hardware-efficient, due to the limited coherence time and qubit connectivity of early-stage quantum devices. As a result, for a specified quantum device, its executable quantum circuits can be described by $\mathrm{RZ}+\mathrm{CI}$ gates with a fixed layout but with different angles, i.e., $U(\boldsymbol{x})$ in Eq. (A14). Given an unknown state ρ_0 evolved by $U(\boldsymbol{x})$ and incoherently measured by an observable O, it forms the mean-value space $\mathcal{F}$ explored in this work. In the following, we detail how $\mathcal{F}$ relates to variational quantum algorithms, classical-shadow-based applications, and quantum system certification, respectively.

Variational quantum algorithms. We briefly review the mechanism of variational quantum algorithms (VQAs). Interested readers can refer to the surveys [29, 43] for detailed information. VQAs generally consist of an N-qubit circuit and a classical optimizer. In the training stage, VQAs follow an iterative manner to proceed with optimization, where the optimizer continuously leverages the output of the quantum circuit to update trainable parameters of the adopted ansatz, i.e., $\boldsymbol{x}$ of $U(\boldsymbol{x})$, to minimize a predefined objective function $\mathcal{L}(\cdot)$. Mathematically, at the t-th iteration, the updating rule for the trainable parameters $\boldsymbol{x}$ is

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \eta \frac{\partial \mathcal{L}\left(f(\boldsymbol{x}^{(t)}, O), c_1\right)}{\partial \boldsymbol{x}}, \tag{A15}$$

where η is the learning rate, $c_1 \in \mathbb{R}$ is the target label, and $f(\boldsymbol{x}^{(t)}, O)$ amounts to the output of the quantum circuit defined in Eq. (A14). The optimization is terminated when the training loss is lower than a threshold or the total number of iterations T exceeds a predefined value. Two main protocols of VQAs are quantum neural networks (QNNs) and variational quantum Eigen-solver (VQEs). The former is utilized to solve machine learning tasks such as classification and regression, and the latter is exploited to estimate the ground state energy of a given Hamiltonian. Notably, when QNNs are applied, the classical control $\boldsymbol{x}$ should be divided into two parts, where the first part is used to encode training examples (without updating) and the second part refers to the trainable parameters. Refer to SM J for details.

Classical-shadow-based applications. A major application of classical shadow, as introduced in SM A 1, is estimating linear properties of quantum states, i.e., $\mathrm{Tr}(\rho O)$. When the measured states ρ are generated by early-stage quantum devices, whose circuit can be described by $\mathrm{RZ}+\mathrm{CI}$ gates with a fixed layout but with different angles, the formed function space coincides with $\mathcal{F}$ in Eq. (A14). It is noteworthy that classical-shadow-based applications are highly entangled with variational quantum algorithms [74, 75], quantum system certification (explained below), and two novel quantum machine learning protocols—flipped model and shadow model [76]. For the flipped model, the training examples $\boldsymbol{z}$ are encoded into the measurement observables rather than the quantum state, i.e., the prediction of the flipped model yields $\mathrm{Tr}(U(\boldsymbol{x})\rho_0 U(\boldsymbol{x})^\dagger O(\boldsymbol{z}))$, where $\boldsymbol{x}$ refers to trainable parameters. As the mean-value space $\mathcal{F}$ in Eq. (A14) supports a class of measurement operators O sampled from a distribution $\mathbb{D}_O$, the achieved results in this work can be harnessed to empower the flipped model. Similarly, the key concept of the shadow model is using classical shadow to obtain the classical description of the state $U(\boldsymbol{x})\rho_0 U(\boldsymbol{x})^\dagger$, followed by classically estimating $\mathrm{Tr}(U(\boldsymbol{x})\rho_0 U(\boldsymbol{x})^\dagger O(\boldsymbol{z}))$, which is also under the framework of $\mathcal{F}$ in Eq. (A14). Taken together, the developed algorithm in this work can greatly reduce the required quantum resources for developing the relevant algorithms.

Quantum system certification. As quantum devices scale up to larger system sizes, the demand for application-specific certification tools becomes apparent. These tools must surpass standard approaches, such as fully simulating a device on a classical computer or performing full tomographic reconstruction, which incurs exponential computational overhead with the qubit count. In response, various certification protocols have been developed to extract different levels of information from the explored quantum chip by using minimal quantum resources. Among them, a representative class of certification protocols is estimating the linear property of quantum states generated by the employed quantum chip, such as direct fidelity estimation [77], entanglement witnessing [78], and two-point correlator [79]. Notably, all of these tasks can be described by the mean-value space $\mathcal{F}$ in Eq. (A14).

4. Literature review

Relevant prior literature to our study can be categorized into three groups: solving quantum many-body problems using machine learning, learnability of quantum circuits, and enhancing early-stage quantum computing (e.g., variational quantum algorithms and quantum system certification) by classical machine learning. In the following, we elucidate how our study aligns with and distinguishes itself from these earlier works.

Machine learning for quantum many-body problems. The seminal work in this context is presented in Huang et al. [79], demonstrating that machine learning algorithms, informed by data collected in physical experiments, can effectively tackle certain quantum many-body problems that are challenging for classical algorithms. In particular, the l_2 *Dirichlet kernel* is proposed to predict the ground state properties of a class of Hamiltonians.

Let the training data be $\{\boldsymbol{x}^{(i)} \to \tilde{\rho}_T(\boldsymbol{x}^{(i)})\}_{i=1}^n$ with n training examples. Here $\rho(\boldsymbol{x})$ refers to the ground state of Hamiltonian $H(\boldsymbol{x})$ with $\boldsymbol{x} \in [-1,1]^d$ and $\tilde{\rho}_T(\boldsymbol{x})$ amounts to the classical shadow with T snapshots. The predicted ground state representation is given by

$$\hat{\sigma}_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \kappa\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right)\tilde{\rho}_T(\boldsymbol{x}^{(i)}), \quad \text{with } \kappa(\boldsymbol{x}, \boldsymbol{x}^{(i)}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d, \|\boldsymbol{k}\|_2 \le \Lambda} \cos\left(\pi \boldsymbol{k}\cdot\left(\boldsymbol{x}-\boldsymbol{x}^{(i)}\right)\right) \in \mathbb{R}, \tag{A16}$$

where $\kappa(\cdot,\cdot)$ is the l_2 *Dirichlet kernel*. The following theorem indicates the efficiency of learning $\rho(\boldsymbol{x})$ for quantum many-body problems.

Theorem A.1 (Adapted from Theorem 4, [79])**.** *Consider a parametrized family of N-qubit states $\{\rho(\boldsymbol{x}) : \boldsymbol{x} \in [-\pi,\pi]^d\}$ and a sum of local observables $O = \sum_{i=1} O_i$ that obey $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}\|\nabla_{\boldsymbol{x}} \mathrm{Tr}(O\rho(\boldsymbol{x}))\|_2^2 \le C$ (smoothness condition) and $\sum_i \|O_i\|_\infty \le B$ (bounded norm). Then, classical shadow data $\{\boldsymbol{x}^{(i)} \to \tilde{\rho}_T(\boldsymbol{x}^{(i)})\}_{i=1}^n$ with $\boldsymbol{x}^{(i)} \sim \mathrm{Unif}[-\pi,\pi]^d$, $n = B^2 d^{\mathcal{O}(C/\epsilon)}$ (training data size), and $\Lambda = \Theta(\sqrt{C/\epsilon})$ suffices to produce a state prediction model we can learn from classical shadow data to produce a model $\hat{\sigma}_n(\boldsymbol{x})$ in Eq. (A16) that achieves $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} |\mathrm{Tr}(O\hat{\sigma}_n(\boldsymbol{x})) - \mathrm{Tr}(O\rho(\boldsymbol{x}))|^2 \le \epsilon$ with high probability. The classical training time for constructing $\hat{\sigma}_n(\boldsymbol{x})$ and the prediction time for computing $\mathrm{Tr}(O\hat{\sigma}_n(\boldsymbol{x}))$ are both upper bounded by $\mathcal{O}((N+L)B^2 d^{\mathcal{O}(C/\epsilon)})$.*

Recent follow-up works have further augmented the capacity of machine learning in addressing quantum many-body problems. Specifically, Ref. [68] explored a specific class of scenarios in quantum many-body problems and devised *positive good kernels*, achieving a polynomial sample complexity for predicting quantum many-body states and their properties. Furthermore, when concentrating on learning the average of observables with a locality assumption, Refs. [80, 81] achieved a quasi-polynomial sample complexity. Additionally, Ref. [82] extended the results of learning phases of quantum matter characterized by exponential decay of correlations, to the task of learning local expectation values for all states within a phase.

Our study diverges from this line of research in terms of its distinct objectives. While previous works concentrate on predicting the properties of quantum many-body states, our focus is on predicting the incoherent dynamics of bounded-gate quantum circuits. This disparity in objectives results in the utilization of different prior information, with the former relying primarily on assumptions about the explored Hamiltonian, such as locality, while the latter leverages the properties of quantum gates, such as RZ + CI. As a result, the proposed trigonometric monomial kernel in our proposal is more succinct compared to the l_2 Dirichlet kernel in terms of the frequency set, i.e., $\boldsymbol{\omega} \in \{0, \pm 1\}^d$ in Eq. (6) versus $\boldsymbol{k} \in \mathbb{Z}^d$ in Eq. (A16).

Learning quantum circuits. Several studies have been undertaken to explore the learnability of quantum circuits, a critical aspect of quantum learning theory [34, 35]. Notably, in Ref. [36], it was demonstrated that when learning a state generated by a quantum circuit with G two-qubit gates to a small trace distance, a sample complexity scaling linearly in G is both necessary and sufficient. However, the computational complexity for learning states and unitaries must scale exponentially in G. While our work shares similar complexity scaling, it diverges in two key aspects: (1) we address a different problem, focusing on the incoherent dynamics of bounded-gate quantum circuits; and (2) we provide concrete algorithms to balance the trade-off between sample and computational complexities.

In Ref. [83], a polynomial-time classical algorithm was devised to learn the description of any unknown N-qubit shallow quantum circuit U. Furthermore, they developed another polynomial-time classical algorithm to learn the description of any unknown N-qubit state prepared by a shallow quantum circuit U on a 2D lattice. However, unlike their study, our proposed algorithm does not necessitate the circuit to be shallow. Additionally, the key components of our algorithm include classical shadow, kernel method, and trigonometric expansion, whereas their algorithm relies on a quantum circuit representation based on local inversions and a technique to combine these inversions.

Enhancing early-stage quantum computing by machine learning. For clarity, here we separately discuss how previous studies harness machine learning to enhance variational quantum algorithms and quantum system certification.

Variational quantum algorithms. The incorporation of machine learning to improve VQAs follows two main approaches. The first approach involves integrating deep neural networks and variational quantum circuits to form a hybrid learning model capable of addressing diverse computational tasks [84–88]. These learning protocols often lack theoretical guarantees and fall outside the scope of our study.

The second approach entails designing *classical surrogates* capable of inferring the output of VQAs. The outcomes of this approach not only benefit from assessing whether VQAs can offer meaningful advantages over state-of-the-art classical methods but also contribute to conserving quantum resources for the development of novel VQAs. A notable paradigm in this context, complementary to tensor network methods and Clifford-based simulators, involves the dequantization of VQAs via Fourier expansion [72]. In particular, Refs. [63, 73, 89, 90] proposed Fourier-based algorithms to simulate variational quantum algorithms when applied to simulate the expectation value of an observable for an initial state evolved under a unitary quantum circuit, e.g., the tasks covered by variational quantum Eigensolvers and quantum approximate optimization algorithms. Besides, Refs. [56, 91, 92] harness Fourier features to dequantize VQAs, i.e., quantum neural networks, when applied to solve classical machine learning tasks.

Our approach distinguishes itself from previous Fourier-based methods by offering broader applicability beyond the classical simulation of VQAs. A specific illustration is that prior Fourier-based approaches often impose constraints on the initial state, such as requiring it to be a simple product state $(|0\rangle\langle 0|)^{\otimes N}$ interpreted in SM A 2. In contrast, our proposal eliminates this requirement. This advancement is rooted in the hybrid nature of our approach, which utilizes shadow information from quantum computers to build the training dataset. Technically, our method integrates three distinct techniques—classical shadow, kernel method, and trigonometric expansion—paving the way for the development of novel strategies to enhance variational quantum algorithms with provable guarantees.

Quantum system certification. We next explain how our work relates to learning-based quantum system certification. Prior literature related to this topic can be classified into two categories. The first category is using deep learning to improve quantum tomography [93], which has two distinct research lines. The first line involves explicit state reconstruction, where the output of neural networks represents the density matrix of the target quantum state [27, 94, 95]. This line exploits the generalization ability of neural networks, where the optimized neural networks can predict the density matrix of unseen states when sampled from the same distribution of the training data. The second line focuses on implicit state reconstruction, wherein neural networks are optimized to emulate the behavior of a given quantum state [96–99]. Note that the second line merely exploits the fitting power of neural networks, where the optimized neural networks can only reconstruct a single state and do not possess the generalization ability. As such, this research line is beyond the scope of our work. The second category is using deep learning to predict the properties of quantum states, including fidelity estimation [100], energy estimation [101], entropy estimation [102], cross-platform verification [103], and similarity testing [104].

Despite distinct applications, the learning paradigm of these two categories is very similar. In particular, a labeled dataset needs to be first collected to conduct supervised learning. The data features of training examples generally are random measurement results, and the label corresponds to the specific tasks, i.e., the density matrix for quantum state tomography and entanglement entropy for entropy estimation. After training, the optimized neural network can predict the unseen state by feeding into the random measurement results. As addressed in the main text, a critical caveat of learning-based quantum system certification is that it lacks a theoretical guarantee. Our work fills this knowledge gap and provides concrete evidence of using various machine learning techniques to comprehend quantum systems.

Remark. We conclude this section by highlighting the complementary nature of our work to quantum tomography and classical simulators in comprehending the behavior of large-qubit quantum devices. In contrast to tomography-based approaches, the offline capability of our proposal could markedly reduce the quantum resource overhead. Moreover, compared to classical simulators, our proposal offers two notable advantages. Firstly, it outperforms classical simulation methods in terms of runtime complexity by removing the reliance on the number of Clifford gates. Secondly, our approach accommodates arbitrarily complex input states, whereas classical simulators often necessitate simple input states such as the product state.

SM B: Proof of Theorem 1

Theorem (Formal statement of Theorem 1)**.** *Following notations in Eq. (3), denote a dataset* $\{\boldsymbol{x}^{(i)}, \tilde{f}_T(\boldsymbol{x}^{(i)})\}_{i=1}^n$ *containing* n *training examples with* $\boldsymbol{x}^{(i)} \in [\pi, \pi]^d$ *and* $\tilde{f}_T(\boldsymbol{x}^{(i)})$ *being the estimation of* $f(\boldsymbol{x}^{(i)}, O)$ *using* T *incoherent measurements. Then, when* $T \leq \frac{(1-\epsilon^2)}{\epsilon} N \log 2$ *and* $\epsilon \ll 1$*, the training data size*

$$\frac{(1-\epsilon)(C_1 d - \log 2)}{\epsilon T} \leq n \leq \widetilde{\mathcal{O}}\left(\frac{B^2 d + B^2 N G}{\epsilon}\right) \tag{B1}$$

is sufficient and necessary to achieve $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}|h(\boldsymbol{x},O)-\mathrm{Tr}(\rho(\boldsymbol{x}),O)| \leq \epsilon$ *with high probability. However, there exists a class of G-bounded-gate quantum circuits requiring* $\Omega(2^d)$ *runtime cost to achieve* $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}|h(\boldsymbol{x},O)-\mathrm{Tr}(\rho(\boldsymbol{x}),O)| \leq \epsilon$.

The proof of Theorem 1 can be broken down into three parts: the lower bound of sample complexity, the upper bound of sample complexity, and the lower bound of runtime complexity. The corresponding bounds are established in the following three theorems, and the proofs are presented in SM C, SM D, and SM E, respectively.

Theorem B.1 (Lower bound of sample complexity)**.** *Consider that a learner collects a dataset* $\mathcal{T}=\{\boldsymbol{x}^{(i)},\tilde{\rho}_T(\boldsymbol{x}^{(i)})\}_{i=1}^n$ *containing n training examples to predict unseen states* $\rho(\boldsymbol{x})$ *with* $\boldsymbol{x}\in[\pi,\pi]^d$*. Suppose that with high probability, the learned model* $h(\cdot,\cdot)$ *can achieve*

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}|h(\boldsymbol{x},O)-\mathrm{Tr}(\rho(\boldsymbol{x}),O)| \leq \epsilon. \tag{B2}$$

Then, when $T \leq \frac{(1-\epsilon)}{\epsilon}N\log 2$*, the training data size must obey*

$$n \geq \frac{(1-\epsilon)(C_1 d-\log 2)}{\epsilon T}, \tag{B3}$$

where $C_1\in(0,1)$ *is a constant.*

Theorem B.2 (Upper bound of sample complexity)**.** *Consider a learner quires an N-qubit quantum circuits and collects a dataset* $\mathcal{T}=\{\boldsymbol{x}^{(i)},\tilde{\rho}_T(\boldsymbol{x}^{(i)})\}_{i=1}^n$ *containing n training examples to predict unseen states* $\rho(\boldsymbol{x})$ *with* $\boldsymbol{x}\in[\pi,\pi]^d$*. Suppose with high probability, there exists a learned model* $h(\cdot,\cdot)$ *that can achieve*

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}|h(\boldsymbol{x},O)-\mathrm{Tr}(\rho(\boldsymbol{x}),O)|^2 \leq \epsilon. \tag{B4}$$

using training data of size

$$n \leq \widetilde{\mathcal{O}}\left(\frac{B^2 d+B^2 NG}{\epsilon}\right). \tag{B5}$$

Theorem B.3 (Lower bound of runtime complexity)**.** *There exists a class of bounded-gate quantum circuits composed of d* RZ *gates and* $G-d$ *CI gates such that no polynomial-time algorithm exists to achieve*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}_{\mathcal{X}}}|h(\boldsymbol{x},O)-\mathrm{Tr}(\rho(\boldsymbol{x}),O)|^2 \leq \epsilon, \tag{B6}$$

unless $\mathsf{BQP}\subset\mathsf{HeurP/Poly}$.

Proof of Theorem 1. The proof of Theorem 1 can be immediately obtained by combining the results of Theorems B.1, B.2, and B.3. □

SM C: Lower bound of learning bounded-gate quantum circuits with incoherent measurement (Proof of Theorem B.1)

In this section, we comprehend the fundamental limitation of learning quantum circuits with a bounded number of non-Clifford gates when only the classical input control and incoherent measurements are allowed, and the circuit layout is unknown to the learner. As elaborated on in the main text, these restrictions echo most experiments on early-stage quantum computers, i.e., the prediction solely relies on shadow information about the output state without harnessing any prior knowledge about the circuit layout.

Theorem (Restatement of Theorem B.1)**.** *Consider that a learner collects a dataset* $\mathcal{T}=\{\boldsymbol{x}^{(i)},\tilde{\rho}_T(\boldsymbol{x}^{(i)})\}_{i=1}^n$ *containing n training examples to predict unseen states* $\rho(\boldsymbol{x})$ *with* $\boldsymbol{x}\in[\pi,\pi]^d$*. Suppose that with high probability, the learned model* $h(\cdot,\cdot)$ *can achieve*

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}|h(\boldsymbol{x},O)-\mathrm{Tr}(\rho(\boldsymbol{x}),O)| \leq \epsilon. \tag{C1}$$

Then, when $T \leq \frac{(1-\epsilon)}{\epsilon}N\log 2$*, the training data size must obey*

$$n \geq \frac{(1-\epsilon)(C_1 d-\log 2)}{\epsilon T}, \tag{C2}$$

where $C_1\in(0,1)$ *is a constant.*

To reach Theorem B.1, we constitute a class of quantum circuits and analyze the lower bound of the required sample complexity for a learner to achieve the average prediction error ϵ. In particular, we consider a simple family of two-qubit quantum circuits illustrated in Fig. C.4(a), which is composed of d RZ gates and at most $6d$ CI gates. The explicit expression of this class of quantum circuits is

$$\mathcal{F} = \left\{ f_{\boldsymbol{a}}(\boldsymbol{x}) = \mathrm{Tr}\left(W(\boldsymbol{x},\boldsymbol{a})(\rho_0 \otimes |0\rangle\langle 0|)W(\boldsymbol{x},\boldsymbol{a})^\dagger O\right) \middle| \boldsymbol{a} \in \{0,1\}^d, \boldsymbol{x} \sim [-\pi,\pi]^d \right\}, \tag{C3}$$

where $\rho_0 \otimes |0\rangle\langle 0|$ is the initial two-qubit state with

$$\rho_0 = \begin{bmatrix} \left(\left(\sqrt{1-2\sqrt{\epsilon}} + \sqrt{1+2\sqrt{\epsilon}}\right)/2\right)^2 & \sqrt{2\epsilon}/2 \\ \sqrt{2\epsilon}/2 & \left(\left(\sqrt{1-2\sqrt{\epsilon}} - \sqrt{1+2\sqrt{\epsilon}}\right)/2\right)^2 \end{bmatrix}, \tag{C4}$$

$W(\boldsymbol{x},\boldsymbol{a})$ refers to the parameterized unitary depending on the classical control $\boldsymbol{x}$ and the label $\boldsymbol{a}$ with

$$W(\boldsymbol{x},\boldsymbol{a}) = G_d^{\boldsymbol{a}_d}(\mathrm{RZ}(\boldsymbol{x}_d)\otimes\mathbb{I})G_d^{\boldsymbol{a}_d}\cdots G_j^{\boldsymbol{a}_j}(\mathrm{RZ}(\boldsymbol{x}_j)\otimes\mathbb{I})G_j^{\boldsymbol{a}_j}\cdots G_1^{\boldsymbol{a}_1}(\mathrm{RZ}(\boldsymbol{x}_1)\otimes\mathbb{I})G_1^{\boldsymbol{a}_1} \tag{C5}$$

and $G_0^0 = \cdots = G_d^0 = \mathbb{I}_4$ and $G_1^1 = \cdots = G_d^1 = \mathrm{SWAP}$, and $O = B(X \otimes \mathbb{I}_2)$ is the observable.

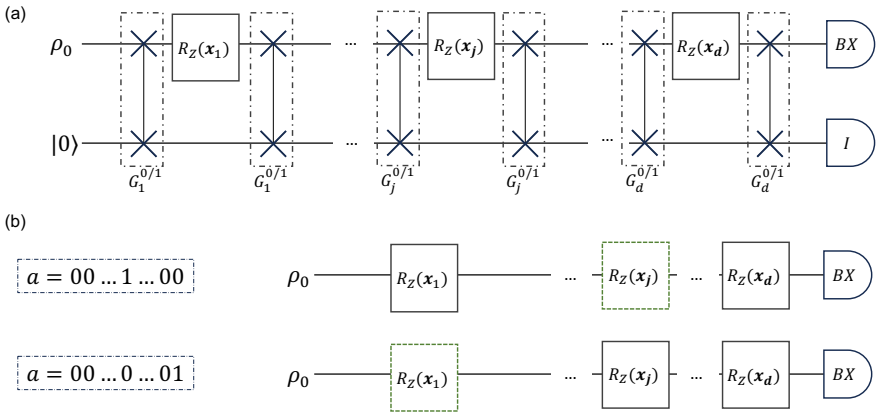

Figure C.4: (a) Illustration of the constructed family of quantum circuits. The dashed box refers that $\forall i \in [d]$, the SWAP gate G_i is optionally applied to the circuit if the bit string satisfies $\boldsymbol{a}_i = 1$. (b) The visualization of the simplified circuit when $\boldsymbol{a}$ = 00...1...0 (upper) and $\boldsymbol{a}$ = 00...0...1 (lower). The dashed green box stands for replacing the RZ gate with the identity gate.

Note that any candidate in $\mathcal{F}$ can be simplified to a single-qubit circuit, as demonstrated in Fig. C.4(b). That is, the equivalent single-qubit circuit only consists of a sequence of $\mathrm{RZ}(\boldsymbol{x}_j)$ gates whose index $\boldsymbol{a}_j$ is zero applied to the initial state ρ_0, followed by the measurement operator BX. Thus, the simplified expression of this family of circuits is

$$\mathcal{F} = \left\{ f_{\boldsymbol{a}}(\boldsymbol{x}) = B\,\mathrm{Tr}\left(V^{\boldsymbol{a}_d}(\boldsymbol{x}_d)\cdots V^{\boldsymbol{a}_1}(\boldsymbol{x}_1)\rho_0 V^{\boldsymbol{a}_1}(\boldsymbol{x}_1)^\dagger\cdots V^{\boldsymbol{a}_d}(\boldsymbol{x}_d)^\dagger X\right) \middle| \boldsymbol{a} \in \{0,1\}^d, \boldsymbol{x} \sim [-\pi,\pi]^d \right\}, \tag{C6}$$

where $V^0(\boldsymbol{x}_j) = \mathrm{RZ}(\boldsymbol{x}_j)$ and $V^1(\boldsymbol{x}_j) = \mathbb{I}_2$ for $\forall j \in [d]$. It is evident that the cardinality of $\mathcal{F}$ is $|\mathcal{F}| = 2^d$.

To proceed further analysis, we now quantify the average distance of different functions in $\mathcal{F}$. To do so, we reformulate $f_{\boldsymbol{a}}(\boldsymbol{x}) \in \mathcal{F}$ based on the trigonometric expansion. According to PTM representations in Eq. (A9), the initial state ρ_0 and Pauli-X can be written as

$$|\rho_0\rangle\rangle = \left[1, \sqrt{2\epsilon}, 0, 0\right]^\top \quad \text{and} \quad |X\rangle\rangle = [0,1,0,0]^\top, \quad \text{respectively.} \tag{C7}$$

Moreover, following the PTM representations of RZ in Eq. (A10), every function in $\mathcal{F}$ satisfies

$$f_{\boldsymbol{a}}(\boldsymbol{x}) = B\left\langle\left\langle \epsilon \middle| \mathfrak{R}_{\mathrm{Z}}{}^\dagger\left(\sum_{j,\boldsymbol{a}_j=0}\boldsymbol{x}_j\right)\middle| X\right\rangle\right\rangle \tag{C8a}$$

$$= \sqrt{2\epsilon}B\cos\left(\sum_{j,\boldsymbol{a}_j=0}\boldsymbol{x}_j\right) \tag{C8b}$$

$$= \sqrt{2\epsilon}B\cos\left((1-\boldsymbol{a}_1)\boldsymbol{x}_1 + (1-\boldsymbol{a}_2)\boldsymbol{x}_2 + \ldots + (1-\boldsymbol{a}_d)\boldsymbol{x}_d\right). \tag{C8c}$$

The derived explicit form allows us to quantify the average discrepancy of any two candidates in $\mathcal{F}$, i.e., given $\forall f_{\boldsymbol{a}}, f_{\boldsymbol{a}'} \in \mathcal{F}$ with $\boldsymbol{a} \neq \boldsymbol{a}'$, we have

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} |f_{\boldsymbol{a}}(\boldsymbol{x}, O) - f_{\boldsymbol{a}'}(\boldsymbol{x}, O)|^2 && \text{(C9a)} \\
= \; & 2\epsilon B^2 \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left| \cos\left((1-\boldsymbol{a}_1)\boldsymbol{x}_1 + ... + (1-\boldsymbol{a}_d)\boldsymbol{x}_d\right) - \cos\left((1-\boldsymbol{a}'_1)\boldsymbol{x}_1 + ... + (1-\boldsymbol{a}'_d)\boldsymbol{x}_d\right) \right|^2 && \text{(C9b)} \\
= \; & 2\epsilon B^2 \left(\frac{1}{2} + \frac{1}{2}\right) - 2\epsilon B^2 \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \cos\left((1-\boldsymbol{a}_1)\boldsymbol{x}_1 + ... + (1-\boldsymbol{a}_d)\boldsymbol{x}_d\right) \cos\left((1-\boldsymbol{a}'_1)\boldsymbol{x}_1 + ... + (1-\boldsymbol{a}'_d)\boldsymbol{x}_d\right) && \text{(C9c)} \\
= \; & 2\epsilon B^2, && \text{(C9d)}
\end{aligned}
$$

where the first equality follows the explicit form of $f_{\boldsymbol{a}}$ and $f_{\boldsymbol{a}'}$ in Eq. (C8a), the second equality employs the two facts: (i) $\forall \boldsymbol{a} \in \{0,1\}^d$, there are $d - \|\boldsymbol{a}\|_0$ effective and independent variables $\{\boldsymbol{x}_j\}$; (ii) these effective variables satisfy

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left|\cos\left((1-\boldsymbol{a}_1)\boldsymbol{x}_1 + ... + (1-\boldsymbol{a}_d)\boldsymbol{x}_d\right)\right|^2 && \text{(C10a)} \\
= \; & \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^{d-\|\boldsymbol{a}\|_0}} \cos^2\left(\boldsymbol{x}_1 + \boldsymbol{x}_2 + ... + \boldsymbol{x}_{d-\|\boldsymbol{a}_0\|}\right) && \text{(C10b)} \\
= \; & \frac{1}{(2\pi)^{d-\|\boldsymbol{a}_0\|}} \int \frac{1 + \cos\left(2\boldsymbol{x}_1 + 2\boldsymbol{x}_2 + ... + 2\boldsymbol{x}_{d-\|\boldsymbol{a}_0\|}\right)}{2} \mathsf{d}\boldsymbol{x} && \text{(C10c)} \\
= \; & \frac{1}{2} + \frac{1}{(2\pi)^{d-\|\boldsymbol{a}_0\|}} \int \frac{\cos(2\boldsymbol{x}_1)\cos\left(2\boldsymbol{x}_2 + ... + 2\boldsymbol{x}_{d-\|\boldsymbol{a}_0\|}\right) - \sin(2\boldsymbol{x}_1)\sin\left(2\boldsymbol{x}_2 + ... + 2\boldsymbol{x}_{d-\|\boldsymbol{a}_0\|}\right)}{2} \mathsf{d}\boldsymbol{x} && \text{(C10d)} \\
= \; & \frac{1}{2} + \frac{1}{(2\pi)^{d-\|\boldsymbol{a}_0\|}} \int_{-\pi}^{\pi} \cos(2\boldsymbol{x}_1)\mathsf{d}\boldsymbol{x}_1 \int \frac{\cos\left(2\boldsymbol{x}_2 + ... + 2\boldsymbol{x}_{d-\|\boldsymbol{a}_0\|}\right)}{2} \mathsf{d}\boldsymbol{x} \\
& - \frac{1}{(2\pi)^{d-\|\boldsymbol{a}_0\|}} \int_{-\pi}^{\pi} \sin(2\boldsymbol{x}_1)\mathsf{d}\boldsymbol{x}_1 \int \frac{\sin\left(2\boldsymbol{x}_2 + ... + 2\boldsymbol{x}_{d-\|\boldsymbol{a}_0\|}\right)}{2} \mathsf{d}\boldsymbol{x} && \text{(C10e)} \\
= \; & \frac{1}{2}, && \text{(C10f)}
\end{aligned}
$$

and the last equality exploits the fact that the second term in Eq. (C9b) is zero as explained below. Specifically, we partition $\boldsymbol{x}_1, ..., \boldsymbol{x}_d$ into two groups, depending on the index list $\boldsymbol{a}$ and $\boldsymbol{a}'$. Denote $R = \sum_{i,\boldsymbol{a}_i=\boldsymbol{a}'_i=0} \boldsymbol{x}_i$ as the summation of entries whose index is zero in both $\boldsymbol{a}$ and $\boldsymbol{a}'$, and $\bar{R} = (1-\boldsymbol{a}_1)\boldsymbol{x}_1 + ... + (1-\boldsymbol{a}_d)\boldsymbol{x}_d - R$ and $\bar{R}' = (1-\boldsymbol{a}'_1)\boldsymbol{x}_1 + ... + (1-\boldsymbol{a}'_d)\boldsymbol{x}_d - R$ as the summation of the rest effective variables with the index list $\boldsymbol{a}$ and $\boldsymbol{a}'$, respectively. Without loss of generality, we suppose $\|\boldsymbol{a}\|_0 > \|\boldsymbol{a}'\|_0$, implying that R contains at least one effective variable. Then, the second term in Eq. (C9b) follows

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \cos\left((1-\boldsymbol{a}_1)\boldsymbol{x}_1 + ... + (1-\boldsymbol{a}_d)\boldsymbol{x}_d\right) \cos\left((1-\boldsymbol{a}'_1)\boldsymbol{x}_1 + ... + (1-\boldsymbol{a}'_d)\boldsymbol{x}_d\right) && \text{(C11a)} \\
= \; & \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \cos(R+\bar{R})\cos(R+\bar{R}') && \text{(C11b)} \\
= \; & \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} [\cos(\bar{R})\cos(R) - \sin(\bar{R})\sin(R)][\cos(R)\cos(\bar{R}') - \sin(R)\sin(\bar{R}')] && \text{(C11c)} \\
= \; & \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} [\cos(\bar{R})\cos^2(R)\cos(\bar{R}') - \cos(\bar{R})\cos(R)\sin(R)\sin(\bar{R}') \\
& - \sin(\bar{R})\sin(R)\cos(R)\cos(\bar{R}') + \sin(\bar{R})\sin^2(R)\sin(\bar{R}')] && \text{(C11d)} \\
= \; & 0, && \text{(C11e)}
\end{aligned}
$$

where the last equality exploits the fact that when R contains at least one effective variable, the symmetric property of integral with respect to trigonometric functions gives $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \cos(R) = \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \sin(R) = 0$.

The result in Eq. (C9d) indicates that taking an expectation over $\boldsymbol{x}$, all concept functions in $\mathcal{F}$ are equally distant by $2\epsilon B^2$. Learning this family of circuits can be reduced to the following *multiple hypothesis testing problem*:

1. Alice randomly and uniformly chooses a target concept $f_{\boldsymbol{a}^*} \in \mathcal{F}$, or equivalently $\boldsymbol{a}^* \in \{0,1\}^d$;

2. The training dataset $\mathcal{T} = \{(\boldsymbol{x}^{(i)}, \hat{\boldsymbol{o}}^{(i)}_{\boldsymbol{a}^*})\}_{i=1}^n$ with the size n is collected based on $f_{\boldsymbol{a}^*}$, i.e., for the i-th training example, the classical input $\boldsymbol{x}^{(i)}$ randomly sampled from $[-\pi,\pi]^d$ is fed into the quantum system described by $f_{\boldsymbol{a}^*}$ and the statistical estimation of $f_{\boldsymbol{a}^*}(\boldsymbol{x}^{(i)}, O)$ is obtained by T shots with $\hat{\boldsymbol{o}}^{(i)}_{\boldsymbol{a}^*} = \frac{1}{T}\sum_t \hat{\boldsymbol{o}}^{(i)}_{\boldsymbol{a}^*,t}$ and $\mathbb{E}(\hat{\boldsymbol{o}}^{(i)}_{\boldsymbol{a}^*}) = f_{\boldsymbol{a}^*}(\boldsymbol{x}^{(i)}, O)$;

3. The learner leverages the training dataset $\mathcal{T}$ to conduct the empirical risk minimization, i.e.,

$$h_{\mathcal{T}} = \arg\min_{h\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell\left(h(\boldsymbol{x}^{(i)}) - \hat{\boldsymbol{o}}_{\boldsymbol{a}^*}^{(i)}\right), \tag{C12}$$

where ℓ refers to the loss function measuring the difference between the prediction $h(\boldsymbol{x}^{(i)})$ and the label $\hat{\boldsymbol{o}}_{\boldsymbol{a}^*}^{(i)}$;

4. The hypothesis testing is conducted to infer the target concept using the learned $h_{\mathcal{T}}$, i.e., the inferred index is

$$\bar{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}, f_{\boldsymbol{a}}\in\mathcal{F}} \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left|h_{\mathcal{T}}(\boldsymbol{x}) - f_{\boldsymbol{a}}(\boldsymbol{x})\right|^2 \tag{C13}$$

and the associated error probability is $\mathbb{P}(\bar{\boldsymbol{a}} \neq \boldsymbol{a}^*)$.

Since the average discrepancy between any two concepts in $\mathcal{F}$ is $2\epsilon B^2$ as indicated in Eq. (C9d), the error probability $\mathbb{P}(\bar{\boldsymbol{a}} \neq \boldsymbol{a}^*)$ becomes zero when $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}|h_{\mathcal{T}}(\boldsymbol{x}) - f_{\boldsymbol{a}^*}(\boldsymbol{x}, O)|^2 < \epsilon B^2$. In this scenario, $h_{\mathcal{T}}$ has a large average prediction error for the data sampled from other $f_{\boldsymbol{a}'} \in \mathcal{F}\setminus\{f_{\boldsymbol{a}^*}\}$ in which the average error is at least ϵ, i.e.,

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}|h_{\mathcal{T}}(\boldsymbol{x})-f_{\boldsymbol{a}'}(\boldsymbol{x},O)|^2 \geq \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}|f_{\boldsymbol{a}'}(\boldsymbol{x},O)-f_{\boldsymbol{a}^*}(\boldsymbol{x},O)|^2 - \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}|h_{\mathcal{T}}(\boldsymbol{x})-f_{\boldsymbol{a}^*}(\boldsymbol{x},O)|^2 > 2\epsilon B^2 - \epsilon B^2 = \epsilon B^2. \tag{C14}$$

The multiple hypothesis testing problem reformulated above enables us to use Fano's inequality to derive the lower bound of sample complexity in learning $f_{\boldsymbol{a}^*} \in \mathcal{F}$. To be concrete, Fano's lemma [105] states that

$$\mathbb{P}[\boldsymbol{a}^* \neq \bar{\boldsymbol{a}}] \geq 1 - \frac{I(\boldsymbol{a}^*;\bar{\boldsymbol{a}}) + \log 2}{\log|\mathcal{F}|}, \tag{C15}$$

where $I(\boldsymbol{a}^*;\bar{\boldsymbol{a}})$ refers to the mutual information between random variables $\boldsymbol{a}^*$ and $\bar{\boldsymbol{a}}$, and $|\mathcal{F}|$ denotes the cardinality of $\mathcal{F}$. In other words, the derivation of the lower bound of the sample complexity n amounts to quantifying the upper bound of $I(\boldsymbol{a}^*;\bar{\boldsymbol{a}})$ and the lower bound of $|\mathcal{F}|$, which motivates the following proof of Theorem B.1.

Remark. For ease of analysis, in the proof of Theorem B.1, we consider the shot number T is *sufficiently large* such that the measured results can be approximated by the normal distribution with the mean $\mu_{\boldsymbol{a}} = f_{\boldsymbol{a}}(\boldsymbol{x})$. Besides, for each $\boldsymbol{a}$, the corresponding variance of measured results is assumed to be equal with the varied $\boldsymbol{x}$, i.e., for all $\boldsymbol{x}$, the variance is $\nu_{\boldsymbol{a}} = \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}\nu_{\boldsymbol{a}}(\boldsymbol{x})$.

Proof of Theorem 2. According to the above elaboration, we next separately quantify $|\mathcal{F}|$ and the mutual information $I(\boldsymbol{a}^*;\bar{\boldsymbol{a}})$, followed by Eq. (C15) to attain the lower bound of sample complexity in learning the incoherent dynamics of bounded-gate quantum circuits.

Cardinality of $\mathcal{F}$. The cardinality of $\mathcal{F}$ can be obtained following its definition in Eq. (C3), i.e.,

$$|\mathcal{F}| = 2^d. \tag{C16}$$

Upper bound of mutual information $I(\boldsymbol{a}^*;\bar{\boldsymbol{a}})$. Recall that the process of learning $f_{\boldsymbol{a}^*}$ implies the Markov chain:

$$\boldsymbol{a}^* \to \rho_{1:n} \to \hat{\boldsymbol{o}}_{1:n} \to \bar{\boldsymbol{a}}, \tag{C17}$$

where for ease of notation, $\rho_{1:n}$ refers to the abbreviation of n resulting states $\rho_{\boldsymbol{a}^*}(\boldsymbol{x}^{(1)}), ..., \rho_{\boldsymbol{a}^*}(\boldsymbol{x}^{(n)})$ before taking the measurements, and $\hat{\boldsymbol{o}}_{1:n}$ refers to the abbreviation of the measured statistic results $\hat{\boldsymbol{o}}_{\boldsymbol{a}^*}^{(1)}, ..., \hat{\boldsymbol{o}}_{\boldsymbol{a}^*}^{(n)}$ of n training examples. Then, according to the data processing inequality, the mutual information $I(\boldsymbol{a}^*;\bar{\boldsymbol{a}})$ is upper bounded by the mutual information between the target index $\boldsymbol{a}^*$ and the measured statistical results $\hat{\boldsymbol{o}}_{1:n}$, i.e.,

$$I(\boldsymbol{a}^*;\bar{\boldsymbol{a}}) \leq I(\boldsymbol{a}^*;\hat{\boldsymbol{o}}_{1:n}). \tag{C18}$$

Moreover, the mutual information on the right-hand side is upper bounded by

$$\begin{aligned} & I(\boldsymbol{a}^*;\hat{\boldsymbol{o}}_{1:n}) && \text{(C19a)}\\ \leq\ & \sum_{i=1}^{n} I(\boldsymbol{a}^*;\hat{\boldsymbol{o}}_i) && \text{(C19b)}\\ \leq\ & \sum_{i=1}^{n} I(\boldsymbol{a}^*;\hat{\boldsymbol{o}}_i, \boldsymbol{x}^{(i)}) && \text{(C19c)} \end{aligned}$$

$$= \sum_{i=1}^{n} I(\boldsymbol{a}^*; \boldsymbol{x}^{(i)}) + I(\boldsymbol{a}^*; \hat{\boldsymbol{o}}_i | \boldsymbol{x}^{(i)}) \tag{C19d}$$

$$= \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{x}^{(i)} \sim [-\pi,\pi]^d} I(\boldsymbol{a}^* | \boldsymbol{x}^{(i)}; \hat{\boldsymbol{o}}_i | \boldsymbol{x}^{(i)}) \tag{C19e}$$

$$= \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{x}^{(i)} \sim [-\pi,\pi]^d} I(\boldsymbol{a}^*; \hat{\boldsymbol{o}}_i | \boldsymbol{x}^{(i)}). \tag{C19f}$$

The first inequality stems from the chain rule and the fact that conditioning reduces entropy, i.e., $I(X; Y_{1:n}) = \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}) - H(Y_i|X, Y_{1:i-1}) = \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}) - H(Y_i|X) \leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|X) = \sum_{i=1}^{n} I(X; Y_i)$, the second inequality exploits the relation $I(X; Y, Z) = H(X) - H(X|Y, Z) \geq H(X) - H(X|Y) = I(X; Y)$, the first equality employs the chain rule of mutual information with $I(X; Y, Z) = I(X; Y) + I(X; Z|Y)$, the second equality adopts the independence between $\boldsymbol{a}^*$ and $\boldsymbol{x}^{(i)}$ with $I(\boldsymbol{a}^*; \boldsymbol{x}^{(i)}) = 0$ and the KL divergence reformulation of mutual information, and the last equality comes from the independence between $\boldsymbol{a}^*$ and $\boldsymbol{x}^{(i)}$.

The above relation hints that the prerequisite to derive the upper bound of $I(\boldsymbol{a}^*, \hat{\boldsymbol{o}}_{1:n})$ is upper bounding $I(\boldsymbol{a}^*; \hat{\boldsymbol{o}}_i | \boldsymbol{x}^{(i)})$. As such, we apply KL divergence formulation to $I(\boldsymbol{a}^*; \hat{\boldsymbol{o}}_i | \boldsymbol{x}^{(i)})$ and obtain

$$I(\boldsymbol{a}^*; \hat{\boldsymbol{o}}_i | \boldsymbol{x}^{(i)}) = \mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{a}^*, \boldsymbol{o}|\boldsymbol{x}^{(i)}} \| P_{\boldsymbol{a}^*} P_{\boldsymbol{o}|\boldsymbol{x}^{(i)}}) \tag{C20a}$$

$$= \sum_{\boldsymbol{a}^*} \int p(\boldsymbol{a}^*, \boldsymbol{o}|\boldsymbol{x}^{(i)}) \log \frac{p(\boldsymbol{a}^*, \boldsymbol{o}|\boldsymbol{x}^{(i)})}{p(\boldsymbol{a}^*) p(\boldsymbol{o}|\boldsymbol{x}^{(i)})} d\boldsymbol{o} \tag{C20b}$$

$$= \sum_{\boldsymbol{a}^*} p(\boldsymbol{a}^*) \int p(\boldsymbol{o}|\boldsymbol{a}^*, \boldsymbol{x}^{(i)}) \log \frac{p(\boldsymbol{o}|\boldsymbol{a}^*, \boldsymbol{x}^{(i)})}{p(\boldsymbol{o}|\boldsymbol{x}^{(i)})} d\boldsymbol{o} \tag{C20c}$$

$$= \frac{1}{|\mathcal{F}|} \sum_{\boldsymbol{a}^*} \mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{o}|\boldsymbol{a}^*, \boldsymbol{x}^{(i)}} \| P_{\boldsymbol{o}|\boldsymbol{x}^{(i)}}) \tag{C20d}$$

$$\leq \frac{1}{|\mathcal{F}|^2} \sum_{\boldsymbol{a}^*, \boldsymbol{a}'} \mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{o}|\boldsymbol{a}^*, \boldsymbol{x}^{(i)}} \| P_{\boldsymbol{o}|\boldsymbol{a}', \boldsymbol{x}^{(i)}}), \tag{C20e}$$

where the third equality uses $p(\boldsymbol{a}^*, \boldsymbol{o}|\boldsymbol{x}^{(i)}) = p(\boldsymbol{a}^*) p(\boldsymbol{o}|\boldsymbol{a}^*, \boldsymbol{x}^{(i)})$, the fourth equality comes from the fact that $\boldsymbol{a}^*$ is uniformly sampled with $p(\boldsymbol{a}^*) = 1/|\mathcal{F}|$, and the inequality employs the property of KL divergence.

Combining the above results, the mutual information is upper bounded by

$$I(\boldsymbol{a}^*; \bar{\boldsymbol{a}}) \leq \frac{1}{|\mathcal{F}|^2} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{x}^{(i)} \sim [-\pi,\pi]^d} \sum_{\boldsymbol{a}^*, \boldsymbol{a}'} \mathrm{D}_{\mathrm{KL}} \left(P_{\boldsymbol{o}|\boldsymbol{a}^*, \boldsymbol{x}^{(i)}} \middle\| P_{\boldsymbol{o}|\boldsymbol{a}', \boldsymbol{x}^{(i)}} \right). \tag{C21}$$

According to our assumption, when the shot number T becomes large, the central limit theorem suggests that $P_{\boldsymbol{o}|\boldsymbol{a}^*, \boldsymbol{x}^{(i)}}$ follows the Gaussian distribution with the mean $\mu_{\boldsymbol{a}^*}(\boldsymbol{x}^{(i)})$ and the variance ν^2. Moreover, assuming that the variance of the varied candidate is the same, we have

$$\mathbb{E}_{\boldsymbol{x}^{(i)} \sim [-\pi,\pi]} \mathrm{D}_{\mathrm{KL}} \left(P_{\boldsymbol{o}|\boldsymbol{a}^*, \boldsymbol{x}^{(i)}} \middle\| P_{\boldsymbol{o}|\boldsymbol{a}', \boldsymbol{x}^{(i)}} \right) = \mathbb{E}_{\boldsymbol{x}^{(i)} \sim [-\pi,\pi]^d} \frac{(\mu_{\boldsymbol{a}^*}(\boldsymbol{x}^{(i)}) - \mu_{\boldsymbol{a}'}(\boldsymbol{x}^{(i)}))^2}{2\nu^2}. \tag{C22}$$

For the nominator in Eq. (C22), the result of Eq. (C9d) gives

$$\mathbb{E}_{\boldsymbol{x}^{(i)} \sim [-\pi,\pi]^d} (\mu_{\boldsymbol{a}^*}(\boldsymbol{x}^{(i)}) - \mu_{\boldsymbol{a}'}(\boldsymbol{x}^{(i)}))^2 = \mathbb{E}_{\boldsymbol{x}^{(i)} \sim [-\pi,\pi]^d} (f_{\boldsymbol{a}^*}(\boldsymbol{x}^{(i)}) - f_{\boldsymbol{a}'}(\boldsymbol{x}^{(i)}))^2 = 2\epsilon B^2. \tag{C23}$$

For the generic variance ν^2, it is derived as follows. When the state $\rho_{\boldsymbol{a}}(\boldsymbol{x})$ is measured by $O = BX$, the probability of measuring the state associated with the eigenvalue $+1$ and the state associated with the eigenvalue -1 is

$$\mathbb{P}(+1) - \mathbb{P}(-1) = \frac{f_{\boldsymbol{a}}(\boldsymbol{x}^{(i)})}{B} \Leftrightarrow \mathbb{P}(+1) = \frac{1 + f_{\boldsymbol{a}}(\boldsymbol{x}^{(i)})/B}{2}, \quad \mathbb{P}(-) = \frac{1 - f_{\boldsymbol{a}}(\boldsymbol{x}^{(i)})/B}{2}. \tag{C24}$$

This property indicates that when the shot number is T, the explicit expression of the variance for the input $\boldsymbol{x}^{(i)}$ is

$$\nu_{\boldsymbol{a}}^2(\boldsymbol{x}^{(i)}) = \frac{1}{T} \left[\left(\mathbb{P}(+) * B^2 + \mathbb{P}(-) * (-B)^2 \right) - f_{\boldsymbol{a}}(\boldsymbol{x}^{(i)})^2 \right] = \frac{B^2}{T} - \frac{f_{\boldsymbol{a}}(\boldsymbol{x}^{(i)})^2}{T}. \tag{C25}$$

Taking expectation over $\boldsymbol{x}$, we have

$$\nu = \mathbb{E}_{\boldsymbol{x}^{(i)}\sim[-\pi,\pi]^d}(\nu^2_{\boldsymbol{a},\boldsymbol{x}^{(i)}}) = \frac{B^2}{T} - \frac{\mathbb{E}_{\boldsymbol{x}^{(i)}\sim[-\pi,\pi]^d}\left(f_{\boldsymbol{a}}(\boldsymbol{x}^{(i)})^2\right)}{T} = \frac{B^2 - B^2\epsilon}{T}, \tag{C26}$$

where the last equality uses the result of Eq. (C10).

In conjunction with Eqs. (C21), (C22), (C23), and (C26), the mutual information is upper bounded by

$$I(\boldsymbol{a}^*; \bar{\boldsymbol{a}}) \leq \frac{n\epsilon T}{1-\epsilon}. \tag{C27}$$

Note that the mutual information cannot continuously enhance with respect to the increased shot number T. When $T \to \infty$, the mutual information is upper bounded by

$$\lim_{T\to\infty} I(\boldsymbol{a}^*; \hat{\boldsymbol{o}}_{1:n}) \leq I(\boldsymbol{a}^*; \rho_{1:n}). \tag{C28}$$

Following the results [106, Lemma 9], the right-hand side term is upper bounded by the number of training examples and the qubit count, i.e.,

$$I(\boldsymbol{a}^*; \rho_{1:n}) \leq nN\log 2. \tag{C29}$$

Taken together, we have

$$I(\boldsymbol{a}^*; \bar{\boldsymbol{a}}) \leq n \cdot \min\left\{\frac{\epsilon T}{1-\epsilon}, N\log 2\right\} \tag{C30}$$

Reusing the Fano's inequality, we obtain

$$I(\boldsymbol{a}^*; \bar{\boldsymbol{a}}) \geq (1 - \mathbb{P}[\boldsymbol{a}^* \neq \bar{\boldsymbol{a}}])\log|\mathcal{F}| - \log 2 \tag{C31a}$$

$$\Rightarrow \quad n \cdot \min\left\{\frac{\epsilon T}{1-\epsilon}, N\log 2\right\} \geq (1 - \mathbb{P}[\boldsymbol{a}^* \neq \bar{\boldsymbol{a}}])d - \log 2 \tag{C31b}$$

$$\Rightarrow \quad n \geq \frac{(1 - \mathbb{P}[\boldsymbol{a}^* \neq \bar{\boldsymbol{a}}])d - \log 2}{\min\left\{\frac{\epsilon T}{1-\epsilon}, N\log 2\right\}}. \tag{C31c}$$

When $T \leq \frac{(1-\epsilon)}{\epsilon} N\log 2$, the lower bound of the sample complexity can be simplified to

$$n \geq (1-\epsilon)\frac{(1 - \mathbb{P}[\boldsymbol{a}^* \neq \bar{\boldsymbol{a}}])d - \log 2}{\epsilon T}. \tag{C32}$$

□

SM D: Upper bound of the sample complexity when incoherently learning the dynamic with bounded-gates (Proof of Theorem B.2)

In this section, we analyze the upper bound of the sample complexity when a learner can predict the output of quantum circuits with bounded gates within a tolerable error. A restatement of the corresponding theorem is as follows.

Theorem (Restatement of Theorem B.2)**.** *Consider a learner quires an N-qubit quantum circuits and collects a dataset $\mathcal{T} = \{\boldsymbol{x}^{(i)}, \tilde{\rho}_T(\boldsymbol{x}^{(i)})\}_{i=1}^n$ containing n training examples to predict unseen states $\rho(\boldsymbol{x})$ with $\boldsymbol{x} \in [\pi,\pi]^d$. Suppose with high probability, there exists a learned model $h(\cdot,\cdot)$ that can achieve*

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} |h(\boldsymbol{x}, O) - \mathrm{Tr}(\rho(\boldsymbol{x}), O)|^2 \leq \epsilon. \tag{D1}$$

using training data of size

$$n \leq \widetilde{\mathcal{O}}\left(\frac{B^2 d + B^2 NG}{\epsilon}\right). \tag{D2}$$

Our proof is rooted in utilizing the packing net and packing number, which are advanced tools broadly used in statistical learning theory [107] and quantum learning theory [66, 108, 109], to quantify the complexity of the class of function represented by bounded-gate quantum circuits. For elucidating, in the following, we first introduce some basic concepts and results that will be employed in our proof, followed by presenting the proof of Theorem B.2. Besides, we emphasize that our current focus is solely on the sample complexity, which implies that the computational cost of preparing the training examples may exhibit exponential scaling with the qubit number as indicated in Theorem B.3.

We now elaborate on how to use packing nets to derive the upper bound of the sample complexity of a learner tasked with predicting the incoherent dynamics of bounded-gate quantum circuits. The formal definition of the packing net and packing number is given below.

Definition 1 (Packing net/number)**.** *Let $(\mathcal{U}, \mathfrak{d})$ be a metric space. The subset $\mathcal{V} \subset \mathcal{U}$ is an ϵ-packing net of $\mathcal{U}$ if for any $A, B \in \mathcal{V}$, $\mathfrak{d}(A, B) \geq \epsilon$. The packing number $\mathcal{M}(\mathcal{U}, \epsilon, \mathfrak{d})$ denotes the largest cardinality of an ϵ-packing net of $\mathcal{U}$.*

Intuitively, an ϵ packing number refers to the maximum number of elements that can be ϵ-separated.

We next leverage this definition to prove Theorem B.2. Following notations introduced in the main text, given an observable O, the function space of quantum circuits with G gates is

$$\mathcal{F} = \left\{ f(\boldsymbol{x}, O) = \mathrm{Tr}(\rho(\boldsymbol{x})O) \middle| \boldsymbol{x} \in [-\pi, \pi]^d, \mathsf{Arc}(\mathrm{RZ}, \mathrm{CI}) \right\}, \tag{D3}$$

where $\boldsymbol{x} \in [-\pi, \pi]^d$ is the classical input control, $\mathsf{Arc}(\mathrm{RZ}, \mathrm{CI})$ refers to the set of circuit layouts formed by d RZ gates and $G - d$ Clifford gates, and $\rho(\boldsymbol{x})$ is the quantum state generated by the bounded gates with $\boldsymbol{x}$, i.e., $\rho(\boldsymbol{x}) = U(\boldsymbol{x})\rho_0 U(\boldsymbol{x})^\dagger$ and the layout of $U(\boldsymbol{x})$ follows an element of $\mathsf{Arc}(\mathrm{RZ}, \mathrm{CI})$. The diversity of circuit layouts stems from the constraint that the learner can solely utilize the classical input and the corresponding measurement results to infer the target concept, where prior information regarding the quantum circuit indicates that it comprises G gates, with d RZ gates and $d - G$ CI gates. From the perspective of the learner, the quantum computer can yield many circuit layouts, leading that the function space $\mathcal{F}$ contains in total $|\mathsf{Arc}(\mathrm{RZ}, \mathrm{CI})|$ circuit layouts.

When the quantum computer is specified with an *unknown but fixed* circuit, the corresponding target concept is denoted by $f^*(\boldsymbol{x}, O) \in \mathcal{F}$. To infer $f^*(\boldsymbol{x}, O)$, in the training data collection procedure, the learner uniformly samples n classical inputs $\{\boldsymbol{x}^{(i)}\}_{i=1}^n$ from $[-\pi, \pi]^d$, sends them into the quantum computer, and obtains the corresponding outcomes $\{\boldsymbol{o}^{(i)}\}_{i=1}^n$. The outcome $\boldsymbol{o}^{(i)}$ for $\forall i \in [n]$ is obtained by performing a *single-shot* measurement of O (i.e., $T = 1$) on $\rho(\boldsymbol{x}^{(i)})$ with $\mathbb{E}[\boldsymbol{o}^{(i)}] = f^*(\boldsymbol{x}^{(i)}, O)$. Let the collected dataset be

$$\mathcal{T} = \left\{ \left(\boldsymbol{x}^{(i)}, \boldsymbol{o}^{(i)} \right) \right\}_{i=1}^n. \tag{D4}$$

What we are interested in here is the required number of training examples n in $\mathcal{T}$ that allows the learner to produce a prediction model $h_{\mathcal{T}}(\cdot)$ whose prediction $h_{\mathcal{T}}(\boldsymbol{x})$ is close to $f^*(\boldsymbol{x})$ on average. More formally, the sample complexity explored here refers to the minimal number of training examples n to ensure that with probability $1 - \delta$, the expected risk satisfies

$$\mathbb{E}_{\boldsymbol{x} \sim [-\pi, \pi]^d} \left| h_{\mathcal{T}}(\boldsymbol{x}) - f^*(\boldsymbol{x}) \right|^2 \leq \mathcal{O}(\epsilon). \tag{D5}$$

Huang et al. [66] presented a method that is efficient in sample complexity (but maybe computationally demanding) for solving this task. In particular, according to Definition 1, denote the 4ϵ-packing net of $\mathcal{F}$ as $\mathcal{F}_{4\epsilon}^p$. The learner conducts the empirical risk minimization on the training dataset $\mathcal{T}$ to estimate the optimal hypothesis, i.e.,

$$\hat{h}_{\mathcal{T}} = \arg\min_{f \in \mathcal{F}_{4\epsilon}^p} \frac{1}{n} \sum_{i=1}^n \left| f(\boldsymbol{x}^{(i)}) - \boldsymbol{o}^{(i)} \right|^2. \tag{D6}$$

Under this formalism, Ref. [66] proved that the required number of training examples n to achieve an ϵ prediction error, as summarized in the following proposition.

Proposition D.1 (Adapted from Proposition 1, [66])**.** *Suppose that the observable O satisfies $\sum_i \|O_i\|_\infty \leq B$. Let $\hat{h}_{\mathcal{T}}$ be an element of the 4ϵ-packing net of $\mathcal{F}$ with the packing number $\mathcal{M}(\mathcal{F}, 4\epsilon, |\cdot|^2)$ that minimizes the empirical training error in Eq. (D6). Then for $\delta \in (0, 1)$, the size of training data*

$$n \geq \frac{38B^2 \log(4\mathcal{M}(\mathcal{F}, 4\epsilon, |\cdot|^2)/\delta)}{\epsilon} \tag{D7}$$

implies

$$\mathbb{E}_{\boldsymbol{x} \sim [-\pi, \pi]^d} \left| \hat{h}_{\mathcal{T}}(\boldsymbol{x}) - f^*(\boldsymbol{x}) \right|^2 \leq 12\epsilon \quad \textit{with probability at least } 1 - \delta. \tag{D8}$$

Remark. Note that although the original proof of Proposition D.1 only concerns $\|O\| \leq 1$, the pertinent proof can be readily extended to the setting of $\sum_i \|O_i\|_\infty \leq B$ concerned in Theorem 1, and therefore we omit it here. Moreover, it only focuses on the case of $T = 1$, but the results still hold for the case of $T > 1$. This is because in the extreme case with $T \to \infty$, the required training data size is reduced to $n \geq 38B^2 \log(2\mathcal{M}(\mathcal{F}, 4\epsilon, \mathfrak{d})/\delta)/\epsilon$ [66, Equation (C66)]. In this regard, the setting of $T \geq 1$ only trivially influences the sample complexity bound (at most logarithmically). For this reason, we omit the relevant analysis in our proof.

Supported by Proposition D.1, the proof of Theorem B.2 amounts to quantifying the upper bound of the packing number $\mathcal{M}(\mathcal{F}, \epsilon, |\cdot|^2)$. Note that different from quantum neural networks [108–110] whose the hypothesis space is continuous and the packing number depends on the value of ϵ, the function class $\mathcal{F}$ explored here is discrete and the number of elements in this class finite. As a result, we have $\mathcal{M}(\mathcal{F}, \epsilon, |\cdot|^2) \leq |\mathsf{Arc}(\mathrm{RZ}, \mathrm{CI})|$ no matter how ϵ is. In light of this fact, we are now ready to show the proof of Theorem B.2.

Proof of Theorem B.2. According to the above explanation, this proof is composed of two parts. The first part is to quantify the upper bound of the packing number $\mathcal{M}(\mathcal{F}, \epsilon, |\cdot|^2)$, or equivalently the cardinality of $\mathsf{Arc}(\mathrm{RZ}, \mathrm{CI})$. And the second part is combining the obtained packing number and Proposition D.1 to acquire the upper bound of the sample complexity to learn $\mathcal{F}$.

We now derive the upper bound of $|\mathsf{Arc}(\mathrm{RZ}, \mathrm{CI})|$. The total number of possible layouts for an N-qubit circuit consisting with d RZ gates and $G - d$ CI gates (i.e., $\mathrm{CI} = \{H, S, \mathrm{CNOT}\}$) is

$$|\mathsf{Arc}(\mathrm{RZ}, \mathrm{CI})| \leq \binom{G}{d} \cdot N^d \cdot 3^{G-d} \cdot \binom{N}{2}^{G-d}, \tag{D9}$$

where the first term $\binom{G}{d}$ computes the number of different arrangements for placing RZ or CI gates at each circuit depth, the second term N^d calculates the total combinations of placing d RZ gates on different qubit wires, the third term 3^{G-d} counts the total combinations of choosing different gates from the CI gate set, and the last term $\binom{N}{2}^{G-d}$ calculates the upper bound for the total combinations of placing the selected CI gates on different qubit wires.

The sample complexity can be efficiently obtained by combining Proposition D.1 and Eq. (D9). That is, the metric entropy of 4ϵ-packing net of $\mathcal{F}$ yields

$$\begin{aligned}
& \log(4\mathcal{M}(\mathcal{F}, 4\epsilon, |\cdot|)/\delta) && \text{(D10a)}\\
\leq\ & \log\left(\binom{G}{d}\right) + d\log(N) + (G-d)\log(3) + (G-d)\log\left(\binom{N}{2}\right) + \log\left(\frac{4}{\delta}\right) && \text{(D10b)}\\
\leq\ & d\log\left(\frac{eG}{d}\right) + d\log(N) + (G-d)\log(3) + 2(G-d)N\log\left(\frac{eN}{2}\right) + \log\left(\frac{4}{\delta}\right), && \text{(D10c)}
\end{aligned}$$

where the second inequality uses $\binom{a}{b} \leq (ea/b)^b$.

The above result, accompanied by Proposition D.1, suggests the upper bound of the sample complexity to achieve the ϵ-prediction error, i.e.,

$$\begin{aligned}
n &= \frac{38B^2\left(d\log\left(\frac{eG}{d}\right) + d\log(N) + (G-d)\log(3) + 2(G-d)N\log\left(\frac{eN}{2}\right) + \log\left(\frac{4}{\delta}\right)\right)}{\epsilon} && \text{(D11a)}\\
&\leq \widetilde{\mathcal{O}}\left(\frac{B^2 d + B^2 NG}{\epsilon}\right). && \text{(D11b)}
\end{aligned}$$

□

SM E: Exponential separation between computational complexity and sample complexity (Proof of Theorem B.3)

We analyze the sample complexity of learning bounded-gate quantum circuits with incoherent measurements and showcase the sample efficiency of this learning task in SM D. However, it is noteworthy that the sample efficiency does not necessarily imply the computational efficiency, as identifying the desired training examples may incur exponential running time. For instance, in the task of pretty-good tomography, the approach outlined in Ref. [4] requires only a linear number of training examples with the qubit count N for a low prediction error, yet the runtime cost scales exponentially with N.

In this section, we prove that learning bounded-gate quantum circuits with incoherent measurements also manifests an exponential separation between sample and computational complexity.

A very recent study [111] explores the computational hardness of learning quantum observables from the measured out data. Here we first introduce some necessary definitions, and then briefly review their key results, followed by elucidating how to generalize such results to our case to complete the proof of Theorem B.3.

The complexity class BQP. BQP (short for bounded-error quantum polynomial time) is the class of promise problems that can be solved by polynomial-time quantum computations that may have some small probability to make an error [112]. Let $A = (A_{\text{yes}}, A_{\text{no}})$ be a promise problem and let $a, b : \mathbb{N} \to [0, 1]$ be functions. Then $A \in \mathsf{BQP}(a, b)$ if and only if there exists a polynomial-time generated family of quantum circuits $\mathcal{Q} = \{\mathcal{Q}_N : N \in \mathbb{N}\}$, where each circuit $\mathcal{Q}_N$ takes N input qubits and produces one output qubit, that satisfies the following properties: (1) if $\boldsymbol{x} \in A_{\text{yes}}$ then $\Pr(\mathcal{Q} \text{ accepts } \boldsymbol{x}) \geq a(|\boldsymbol{x}|)$, and (2) if $\boldsymbol{x} \in A_{\text{no}}$ then $\Pr(\mathcal{Q} \text{ accepts } \boldsymbol{x}) \leq b(|\boldsymbol{x}|)$. The class is defined as $\mathsf{BQP} = \mathsf{BQP}(2/3, 1/3)$.

The complexity class $\mathsf{HeurP}^{\mathbb{D}}/\mathsf{poly}$. The complexity class $\mathsf{HeurP}^{\mathbb{D}}/\mathsf{poly}$ refers to the class of languages L that can be decided in $\mathsf{HeurP}/\mathsf{poly}$ under the distribution $\mathbb{D}$, i.e. $\mathsf{HeurP}^{\mathbb{D}}/\mathsf{poly} = \{\mathsf{L} | (\mathsf{L}, \mathbb{D}) \in \mathsf{HeurP}/\mathsf{poly}\}$.

The formal statement of the result in Ref. [111] is summarized as followa.

Lemma E.1 (Adapted from Lemma 1, [111])**.** *Define the concept class as*

$$\mathcal{F}_{\mathsf{Hard}} = \left\{ f(\boldsymbol{x}^{\perp}) = \mathrm{Tr}\left(U \left|\boldsymbol{x}^{\perp}\right\rangle \left\langle \boldsymbol{x}^{\perp}\right| U^{\dagger} O\right) \middle| \boldsymbol{x}^{\perp} \in \{-1, 1\}^N, O \sim \mathbb{D}_O \right\}, \tag{E1}$$

where $\boldsymbol{x}^{\perp}$ specifies the initial state $|\boldsymbol{x}^{\perp}\rangle$, U denotes the quantum circuit containing at most a polynomial number of quantum gates and O refers to a linear combination of local Pauli strings sampled from a prior distribution $\mathbb{D}_O$.

Under the assumption that there exists a distribution $\mathbb{D}'$ such that $\mathsf{BQP} \not\subset \mathsf{HeurP}^{\mathbb{D}'}/\mathsf{poly}$, then for any BQP-complete language, there exists an associated input distribution $\mathbb{D}$ such that no randomized algorithm that can solely use the collected measure-out data to output a prediction model $h(\boldsymbol{x})$, which learns the concept class $\mathcal{F}_{\mathsf{Hard}}$ in a polynomial time. Note that the learning algorithm is said to successfully learn $\mathcal{F}_{\mathsf{Hard}}$ if for any $f \in \mathcal{F}_{\mathsf{Hard}}$, it can output a prediction model $h(\boldsymbol{x})$ that satisfies $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}[|f(\boldsymbol{x}) - h(\boldsymbol{x})|^2] \leq \epsilon$ with high probability.

It is noteworthy that the terminology 'learning by measure-out data' appeared in the above lemma is identical to the learning paradigm investigated in this work. More precisely, the learner is restrictive to collect information from quantum systems within a polynomial time (i.e., applying a polynomial number of measurement in total). Once the data collection process is completed, the learner utilize the collected data to infer the prediction model $h(\boldsymbol{x})$ on the classical side.

The main idea of the proof for Lemma E.1 is as follows. First, the authors correlate $\mathcal{F}_{\mathsf{Hard}}$ with BQP circuits. In particular, the unitary in $\mathcal{F}_{\mathsf{Hard}}$ is specified to be a family of quantum circuit $\{U_{\mathsf{BQP}^N}\}$, which decides the BQP-complete language L, one circuit per size; the observable is set as $O = Z \otimes I_2 \otimes \cdots \otimes I_2$. In this way, the output of quantum circuits can correctly decide every $\boldsymbol{x}^{\perp} \in \mathsf{L}$, i.e., $f(\boldsymbol{x}^{\perp}) > 0$ if $\boldsymbol{x} \in \mathsf{L}$ and $f(\boldsymbol{x}^{\perp}) < 0$ if $\boldsymbol{x} \notin \mathsf{L}$. Then, the authors show that if there exists a classical algorithm that solely utilizes the measure-out data to learn $\mathcal{F}_{\mathsf{Hard}}$ in a polynomial time, then such algorithm could decide any BQP language on average, which implies $\mathsf{BQP} \subseteq \mathsf{HeurP}^{\mathbb{D}}/\mathsf{poly}$ for any $\mathbb{D}$. However, under widely believed assumptions, for any BQP complete language L there always exists a distribution such that $(\mathsf{L}, \mathbb{D}) \notin \mathsf{HeurP}/\mathsf{poly}$, indicating such classical algorithm cannot exist.

We next generalize the results of Lemma E.1 to show the computational hardness of learning the bounded-gate circuit with incoherent dynamics. To achieve this goal, it is sufficient to show the concept class $\mathcal{F}$ in Eq. (3) contains the class of BQP circuits discussed above. Recall the definition of $\mathcal{F}$. As for the observable $Z \otimes I_2 \otimes \cdots \otimes I_2$, it meets the requirement of the observable defined in $\mathcal{F}$, which is formed by Pauli operators with a bounded norm. As for the state $U_{\mathsf{BQP}} |\boldsymbol{x}^{\perp}\rangle$, it can also be expressed by the bounded-gate circuit $\{\mathrm{RZ} + \mathrm{CI}\}$ with the initial state $|0\rangle^{\otimes N}$. Specifically, the bounded-gate circuit is decomposed into two parts, where the first part is used to prepare the state $|\boldsymbol{x}^{\perp}\rangle$ and the second part is to prepare the BQP circuit. For the first part, N RZ gates and $\mathcal{O}(N)$ CI gates associated with a proper distribution are sufficient to prepare any input state $|\boldsymbol{x}^{\perp}\rangle$ with $\boldsymbol{x}^{\perp} \in \{0, 1\}^N$. For the second part, since U_{BQP} contains at most $\mathcal{O}(poly(N))$ quantum gates and $\{\mathrm{RZ} + \mathrm{CI}\}$ is a universal basis gate set, $\mathcal{O}(poly(N))$ RZ gates with a proper distribution over the classical inputs and $\mathcal{O}(poly(N))$ CI gates are sufficient to synthesis U_{BQP}. Taken together, the BQP circuit belongs to $\mathcal{F}$ when the number of $\{\mathrm{RZ} + \mathrm{CI}\}$ gates polynomially scale with N. However, according to the results of Lemma E.1, no algorithm can only use the measure-out data to learn this circuit within a polynomial time. This proves Theorem B.3.

SM F: Learnability of the proposed kernel-based ML model (Proof of Theorem 2)

This section provides the proof of Theorem 2, which analyzes how the prediction error of the proposed kernel-based ML model depends on the number of training examples n, the size of the quantum system N, and the dimension of classical inputs d. Recall that in the main text, the proposed state prediction model is

$$\hat{\sigma}_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \kappa_\Lambda\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) \tilde{\rho}_T(\boldsymbol{x}^{(i)}) \text{ with } \kappa_\Lambda\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) = \sum_{\boldsymbol{\omega}, \|\boldsymbol{\omega}\|_0 \leq \Lambda} 2^{\|\boldsymbol{\omega}\|_0} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(i)}) \in \mathbb{R}. \tag{F1}$$

What we intend to prove is the average discrepancy between $\mathrm{Tr}(\hat{\sigma}_n(\boldsymbol{x})O)$ and the ground truth $\mathrm{Tr}(\rho(\boldsymbol{x})O)$ when $\boldsymbol{x}$ is uniformly and randomly sampling from $[-\pi, \pi]^d$ and the local observable O is sampled from a prior distribution $\mathbb{D}_O$, i.e., $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} |\mathrm{Tr}(O\hat{\sigma}_n(\boldsymbol{x})) - \mathrm{Tr}(O\rho(\boldsymbol{x}))|^2$.

Before moving to proceed with the further analysis, let us exhibit the formal statement of Theorem 2.

Theorem (Restatement of Theorem 2)**.** *Following notations in the main text, consider a parametrized family of N-qubit states $\mathcal{Q}$ and a sum $O = \sum_{i=1} O_i$ of multiple local observables with $\sum_i \|O_i\|_\infty \leq B$ and the maximum locality of $\{O_i\}$ being K. Suppose $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \|\nabla_{\boldsymbol{x}} \mathrm{Tr}(\rho(\boldsymbol{x})O)\|_2^2 \leq C$. Then, let the dataset be $\mathcal{T}_{\mathsf{s}} = \{\boldsymbol{x}^{(i)} \to \tilde{\rho}(\boldsymbol{x}^{(i)})\}_{i=1}^n$ with $\boldsymbol{x}^{(i)} \sim \mathrm{Unif}[-\pi, \pi]^d$ and $n = |\mathfrak{C}(\Lambda)| \frac{2B^2 9^K}{\epsilon} \log\left(2 \cdot |\mathfrak{C}(\Lambda)|/\delta\right)$ with $\mathfrak{C}(\Lambda) = \{\boldsymbol{\omega} | \boldsymbol{\omega} \in \{0, \pm 1\}^d,\ s.t.\ \|\boldsymbol{\omega}\|_0 \leq \Lambda\}$. When the frequency is truncated to $\Lambda = 4C/\epsilon$, the state prediction model in Eq. (F1) achieves*

$$\mathbb{E}_{\mathcal{T}_{\mathsf{s}}}[\hat{\sigma}_n(\boldsymbol{x})] = \rho_\Lambda(\boldsymbol{x}) \tag{F2}$$

and with probability at least $1 - \delta$,

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} |\mathrm{Tr}(O\hat{\sigma}_n(\boldsymbol{x})) - \mathrm{Tr}(O\rho(\boldsymbol{x}))|^2 \leq \epsilon. \tag{F3}$$

To reach Theorem 2, we first use the triangle inequality to decouple the difference between the prediction and ground truth into the truncation error and the estimation error, i.e.,

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left[\left|\mathrm{Tr}\left(O\hat{\sigma}_n(\boldsymbol{x})\right) - \mathrm{Tr}\left(O\rho(\boldsymbol{x})\right)\right|^2\right] \tag{F4a}$$

$$\leq \left(\sqrt{\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left[\left|\mathrm{Tr}\left(O\rho_\Lambda(\boldsymbol{x})\right) - \mathrm{Tr}(O\rho(\boldsymbol{x}))\right|^2\right]} + \sqrt{\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left[\left|\mathrm{Tr}\left(O\sigma_n(\boldsymbol{x})\right) - \mathrm{Tr}(O\rho_\Lambda(\boldsymbol{x}))\right|^2\right]}\right)^2. \tag{F4b}$$

After decoupling, we then separately derive the upper bound of these two terms, where the relevant results are encapsulated in the following two lemmas whose proofs are given in the subsequent two subsections.

Lemma F.1 (Truncation error of ρ_Λ)**.** *Following notations in Theorem 2, assuming $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \|\nabla_{\boldsymbol{x}} \mathrm{Tr}(\rho(\boldsymbol{x})O)\|_2^2 \leq C$, the truncation error induced by removing high-frequency terms of ρ under the trigonometric expansion with $\|\boldsymbol{\omega}\|_0 \leq \Lambda$ is upper bounded by*

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} |\mathrm{Tr}(O\rho_\Lambda(\boldsymbol{x}) - \mathrm{Tr}(O\rho(\boldsymbol{x}))|^2 \leq \frac{C}{\Lambda}. \tag{F5}$$

Lemma F.2 (Estimation error of $\hat{\sigma}_n$)**.** *Following notations in Theorem 2, with probability at least $1-\delta$, the estimation error induced by finite training examples $\mathcal{T} = \{\tilde{\rho}(\boldsymbol{x}^{(i)})\}_{i=1}^n$ is upper bounded by*

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left[|\mathrm{Tr}(O\hat{\sigma}_n(\boldsymbol{x})) - \mathrm{Tr}(O\rho_\Lambda(\boldsymbol{x}))|^2\right] \leq |\mathfrak{C}(\Lambda)| \frac{1}{2n} B^2 9^K \log\left(\frac{2 \cdot |\mathfrak{C}(\Lambda)|}{\delta}\right), \tag{F6}$$

where $\mathfrak{C}(\Lambda) = \{\boldsymbol{\omega} | \boldsymbol{\omega} \in \{0, \pm 1\}^d,\ s.t.\ \|\boldsymbol{\omega}\|_0 \leq \Lambda\}$ refers to the set of truncated frequencies.

We are now ready to present the proof of Theorem 2.

Proof of Theorem 2. The difference between the prediction and ground truth can be obtained by integrating Lemmas F.1 and F.2 into Eq. (F4a). Mathematically, with probability at least $1 - \delta$, we have

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left[|\mathrm{Tr}(O\hat{\sigma}_n(\boldsymbol{x})) - \mathrm{Tr}(O\rho(\boldsymbol{x}))|^2\right] \leq \left(\sqrt{\frac{C}{\Lambda}} + \sqrt{|\mathfrak{C}(\Lambda)| \frac{1}{2n} B^2 9^K \log\left(\frac{2 \cdot |\mathfrak{C}(\Lambda)|}{\delta}\right)}\right)^2. \tag{F7}$$

To ensure the average prediction error is upper bounded by ϵ, it is sufficient to showcase when the inner two terms are upper bounded by $\sqrt{\epsilon}/2$. For the first term, the condition is satisfied when

$$\sqrt{\frac{C}{\Lambda}} \leq \frac{\sqrt{\epsilon}}{2} \Leftrightarrow \Lambda \geq \frac{4C}{\epsilon}. \tag{F8}$$

For the second term, we have

$$\sqrt{|\mathfrak{C}(\Lambda)|\frac{1}{2n}B^2 9^K \log\left(\frac{2\cdot|\mathfrak{C}(\Lambda)|}{\delta}\right)} \leq \frac{\sqrt{\epsilon}}{2} \Leftrightarrow n \geq |\mathfrak{C}(\Lambda)|\frac{2B^2 9^K}{\epsilon}\log\left(\frac{2\cdot|\mathfrak{C}(\Lambda)|}{\delta}\right). \tag{F9}$$

Taken together, with probability $1-\delta$, the prediction error is upper bounded by ϵ when the number of training examples satisfies

$$n \geq |\mathfrak{C}(4C/\epsilon)|\frac{2B^2 9^K}{\epsilon}\log\left(\frac{2\cdot|\mathfrak{C}(4C/\epsilon)|}{\delta}\right). \tag{F10}$$

□

1. Truncation error of the classical learning model (Proof of Lemma F.1)

Recall that under the trigonometric monomial expansion, the target state ρ with truncation and without truncation takes the form as

$$\rho_\Lambda = \sum_{\boldsymbol{\omega}\in\mathfrak{C}(\Lambda)} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\rho_{\boldsymbol{\omega}} \text{ and } \rho = \sum_{\boldsymbol{\omega}\in\mathfrak{C}(d)} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\rho_{\boldsymbol{\omega}}, \tag{F11}$$

respectively. The purpose of Lemma F.1 is to analyze the upper bound of the discrepancy between $\mathrm{Tr}(\rho_\Lambda O)$ and $\mathrm{Tr}(\rho O)$ induced by the truncation of high-frequency terms.

Proof of Lemma F.1. By adopting the explicit trigonometric monomial expansion of ρ and ρ_Λ in Eq. (F11), we have

$$\begin{aligned}
& \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left|\mathrm{Tr}(O\rho_\Lambda(\boldsymbol{x}) - \mathrm{Tr}(O\rho(\boldsymbol{x}))\right|^2 && \text{(F12a)}\\
= \; & \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left|\sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0>\Lambda} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\langle\langle\rho_{\boldsymbol{\omega}}|O\rangle\rangle\right|^2 && \text{(F12b)}\\
= \; & \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0>\Lambda}\sum_{\boldsymbol{\omega}',\|\boldsymbol{\omega}'\|_0>\Lambda} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\Phi_{\boldsymbol{\omega}'}(\boldsymbol{x})\langle\langle\rho_{\boldsymbol{\omega}}|O\rangle\rangle\langle\langle\rho_{\boldsymbol{\omega}'}|O\rangle\rangle && \text{(F12c)}\\
= \; & \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0>\Lambda}\sum_{\boldsymbol{\omega}',\|\boldsymbol{\omega}'\|_0>\Lambda} \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\Phi_{\boldsymbol{\omega}'}(\boldsymbol{x})\langle\langle\rho_{\boldsymbol{\omega}}|O\rangle\rangle\langle\langle\rho_{\boldsymbol{\omega}'}|O\rangle\rangle \mathsf{d}^d x && \text{(F12d)}\\
= \; & \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0>\Lambda}\sum_{\boldsymbol{\omega}',\|\boldsymbol{\omega}'\|_0>\Lambda} 2^{-\|\boldsymbol{\omega}\|_0}\delta_{\boldsymbol{\omega},\boldsymbol{\omega}'}\langle\langle\rho_{\boldsymbol{\omega}}|O\rangle\rangle\langle\langle\rho_{\boldsymbol{\omega}'}|O\rangle\rangle && \text{(F12e)}\\
= \; & \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0>\Lambda} 2^{-\|\boldsymbol{\omega}\|_0}|\langle\langle\rho_{\boldsymbol{\omega}}|O\rangle\rangle|^2 := \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0>\Lambda} 2^{-\|\boldsymbol{\omega}\|_0}\alpha_{\boldsymbol{\omega}}^2, && \text{(F12f)}
\end{aligned}$$

where the first three equalities follow a direct reformulation, Eq. (F12e) employs the orthogonality of basis functions in the trigonometric expansion, i.e.,

$$\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\Phi_{\boldsymbol{\omega}'}(\boldsymbol{x})\mathsf{d}^d x = 2^{-\|\boldsymbol{\omega}\|_0}\delta_{\boldsymbol{\omega},\boldsymbol{\omega}'}, \tag{F13}$$

and in the last equality we define $\alpha_{\boldsymbol{\omega}} \equiv \mathrm{Tr}(\rho_{\boldsymbol{\omega}}O)$ for clarification.

The remainder of the proof uses the assumption of the norm of the gradients of the expectation, i.e., $\|\nabla_{\boldsymbol{x}}\mathrm{Tr}(\rho(\boldsymbol{x})O)\|_2^2 \leq C$, to derive the upper bound of Eq. (F12f). To do so, we first derive the explicit form of

the gradient under the trigonometric monomial expansion. That is, the gradient of $\mathrm{Tr}(\rho(\boldsymbol{x})O)$ with respect to $\boldsymbol{x}$ is a d-dimensional vector, i.e.,

$$
\begin{aligned}
& \nabla_{\boldsymbol{x}} \mathrm{Tr}(\rho(\boldsymbol{x})O) && \text{(F14a)}\\
= \;& \nabla_{\boldsymbol{x}} \sum_{\boldsymbol{\omega}} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{Tr}(\rho_{\boldsymbol{\omega}} O) \\
= \;& \sum_{\boldsymbol{\omega}} \nabla_{\boldsymbol{x}} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \alpha_{\boldsymbol{\omega}} && \text{(F14b)}\\
= \;& \Big[\sum_{\boldsymbol{\omega}} \Psi_{\boldsymbol{\omega}_1}(\boldsymbol{x}_1) \Phi_{\boldsymbol{\omega}_{2:d}}(\boldsymbol{x}_{2:d}) \alpha_{\boldsymbol{\omega}}, \cdots, \sum_{\boldsymbol{\omega}} \Phi_{\boldsymbol{\omega}_{1:i-1}}(\boldsymbol{x}_{1:i-1}) \Psi_{\boldsymbol{\omega}_i}(\boldsymbol{x}_i) \Phi_{\boldsymbol{\omega}_{i+1:d}}(\boldsymbol{x}_{i+1:d}) \alpha_{\boldsymbol{\omega}}, && \text{(F14c)}\\
& \cdots, \sum_{\boldsymbol{\omega}} \Psi_{\boldsymbol{\omega}_{1:d-1}}(\boldsymbol{x}_{1:d-1}) \Psi_{\boldsymbol{\omega}_d}(\boldsymbol{x}_d) \alpha_{\boldsymbol{\omega}} \Big]^\top \in \mathbb{R}^d, && \text{(F14d)}
\end{aligned}
$$

where $\Phi_{\boldsymbol{\omega}_{a:b}}(\boldsymbol{x}_{a:b}) \equiv \prod_{i=a}^{b} \Phi_{\boldsymbol{\omega}_i}(\boldsymbol{x}_i)$, and the derivative with respect to the i-th entry is

$$
\Psi_{\boldsymbol{\omega}_i}(\boldsymbol{x}_i) \equiv \nabla_{\boldsymbol{x}_i} \Phi_{\boldsymbol{\omega}_i}(\boldsymbol{x}_i) = \begin{cases} 0 & \text{if } \boldsymbol{\omega}_i = 0 \\ -\sin(\boldsymbol{x}_i) & \text{if } \boldsymbol{\omega}_i = 1 \\ \cos(\boldsymbol{x}_i) & \text{if } \boldsymbol{\omega}_i = -1 \end{cases}. \tag{F15}
$$

By making use of the above formula, the expectation value of $\|\nabla_{\boldsymbol{x}} \mathrm{Tr}(\rho(\boldsymbol{x})O)\|_2^2$ over $\boldsymbol{x} \sim \mathrm{Unif}[-\pi, \pi]^d$ satisfies

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{x} \sim [-\pi,\pi]^d} \|\nabla_{\boldsymbol{x}} \mathrm{Tr}(\rho(\boldsymbol{x})O)\|_2^2 && \text{(F16a)}\\
= \;& \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \Bigg[\Bigg(\sum_{\boldsymbol{\omega}} \Psi_{\boldsymbol{\omega}_1}(\boldsymbol{x}_1) \Phi_{\boldsymbol{\omega}_{2:d}}(\boldsymbol{x}_{2:d}) \alpha_{\boldsymbol{\omega}} \sum_{\boldsymbol{\omega}'} \Psi_{\boldsymbol{\omega}'_1}(\boldsymbol{x}_1) \Phi_{\boldsymbol{\omega}'_{2:d}}(\boldsymbol{x}_{2:d}) \alpha_{\boldsymbol{\omega}'} \Bigg) + \cdots && \text{(F16b)}\\
& + \Bigg(\sum_{\boldsymbol{\omega}} \Phi_{\boldsymbol{\omega}_{1:i-1}}(\boldsymbol{x}_{1:i-1}) \Psi_{\boldsymbol{\omega}_i}(\boldsymbol{x}_i) \Phi_{\boldsymbol{\omega}_{i+1:d}}(\boldsymbol{x}_{i+1:d}) \alpha_{\boldsymbol{\omega}} \sum_{\boldsymbol{\omega}'} \Phi_{\boldsymbol{\omega}'_{1:i-1}}(\boldsymbol{x}_{1:i-1}) \Psi_{\boldsymbol{\omega}'_i}(\boldsymbol{x}_i) \Phi_{\boldsymbol{\omega}'_{i+1:d}}(\boldsymbol{x}_{i+1:d}) \alpha_{\boldsymbol{\omega}'} \Bigg) + \cdots && \text{(F16c)}\\
& + \Bigg(\sum_{\boldsymbol{\omega}} \Phi_{\boldsymbol{\omega}_{1:d-1}}(\boldsymbol{x}_{1:d-1}) \Psi_{\boldsymbol{\omega}_d}(\boldsymbol{x}_d) \alpha_{\boldsymbol{\omega}} \sum_{\boldsymbol{\omega}'} \Phi_{\boldsymbol{\omega}'_{1:d-1}}(\boldsymbol{x}_{1:d-1}) \Psi_{\boldsymbol{\omega}'_d}(\boldsymbol{x}_d) \alpha_{\boldsymbol{\omega}'} \Bigg) \Bigg] \mathsf{d}^d x && \text{(F16d)}\\
= \;& \frac{1}{(2\pi)^d} \Bigg[\pi \int_{[-\pi,\pi]^{d-1}} \Bigg(\sum_{\boldsymbol{\omega}_{2:d}} \Phi_{\boldsymbol{\omega}_{2:d}}(\boldsymbol{x}_{2:d}) \alpha_{1,\boldsymbol{\omega}_{2:d}} \sum_{\boldsymbol{\omega}'_{2:d}} (\boldsymbol{x}_1) \Phi_{\boldsymbol{\omega}'_{2:d}}(\boldsymbol{x}_{2:d}) \alpha_{1,\boldsymbol{\omega}'_{2:d}} \Bigg) \mathsf{d}^{d-1} x + \cdots && \text{(F16e)}\\
& + \pi \int_{[-\pi,\pi]^{d-1}} \Bigg(\sum_{\boldsymbol{\omega}_{2:d}} \Phi_{\boldsymbol{\omega}_{2:d}}(\boldsymbol{x}_{2:d}) \alpha_{-1,\boldsymbol{\omega}_{2:d}} \sum_{\boldsymbol{\omega}'_{2:d}} (\boldsymbol{x}_1) \Phi_{\boldsymbol{\omega}'_{2:d}}(\boldsymbol{x}_{2:d}) \alpha_{-1,\boldsymbol{\omega}'_{2:d}} \Bigg) \mathsf{d}^{d-1} x + \cdots && \text{(F16f)}\\
& + \pi \int_{[-\pi,\pi]^{d-1}} \Bigg(\sum_{\boldsymbol{\omega}_{1:d-1}} \Phi_{\boldsymbol{\omega}_{1:d-1}}(\boldsymbol{x}_{1:d-1}) \alpha_{\boldsymbol{\omega}_{1:d-1},1} \sum_{\boldsymbol{\omega}'_{1:d-1}} \Phi_{\boldsymbol{\omega}'_{1:d-1}}(\boldsymbol{x}_{1:d-1}) \alpha_{\boldsymbol{\omega}'_{1:d-1},1} \Bigg) \mathsf{d}^{d-1} x && \text{(F16g)}\\
& + \pi \int_{[-\pi,\pi]^{d-1}} \Bigg(\sum_{\boldsymbol{\omega}_{1:d-1}} \Phi_{\boldsymbol{\omega}_{1:d-1}}(\boldsymbol{x}_{1:d-1}) \alpha_{\boldsymbol{\omega}_{1:d-1},-1} \sum_{\boldsymbol{\omega}'_{1:d-1}} \Phi_{\boldsymbol{\omega}'_{1:d-1}}(\boldsymbol{x}_{1:d-1}) \alpha_{\boldsymbol{\omega}'_{1:d-1},-1} \Bigg) \mathsf{d}^{d-1} x \Bigg] && \text{(F16h)}\\
= \;& \Bigg[\sum_{\boldsymbol{\omega}_{2:d}} 2^{-\|[1,\boldsymbol{\omega}_{2:d}]\|_0} |\alpha_{1,\boldsymbol{\omega}_{2:d}}|^2 + \sum_{\boldsymbol{\omega}_{2:d}} 2^{-\|[-1,\boldsymbol{\omega}_{2:d}]\|_0} |\alpha_{-1,\boldsymbol{\omega}_{2:d}}|^2 + \cdots && \text{(F16i)}\\
& + \sum_{\boldsymbol{\omega}_{1:d-1}} 2^{-\|[\boldsymbol{\omega}_{1:d-1},1]\|_0} |\alpha_{\boldsymbol{\omega}_{1:d-1},1}|^2 + \sum_{\boldsymbol{\omega}_{1:d-1}} 2^{-\|[\boldsymbol{\omega}_{1:d-1},-1]\|_0} |\alpha_{\boldsymbol{\omega}_{1:d-1},-1}|^2 \Bigg] := \bigstar, && \text{(F16j)}
\end{aligned}
$$

where Eqs. (F16e)-(F16h) are obtained by taking the expectation value on the gradient term $\Psi_{\boldsymbol{\omega}_i}$ for $\forall i \in [d]$, the quantity $\alpha_{\pm 1,\boldsymbol{\omega}_{2:d}}$ refers to $\mathrm{Tr}(\rho_{\pm 1,\boldsymbol{\omega}_{2:d}} O)$ with the frequency value at the 1-st position being ± 1 (the same rule applies to other terms such as $\alpha_{\boldsymbol{\omega}_{1:i},i+1,\boldsymbol{\omega}_{i+2:d}}$), and Eq. (F16j) is obtained by taking integral over each term in the bracket and the orthogonality of basis functions in $\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})$ and $\Phi_{\boldsymbol{\omega}'}(\boldsymbol{x})$ as shown in Eq. (F13).

We now use the reformulated $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]}\|\nabla_{\boldsymbol{x}}\operatorname{Tr}(\rho(\boldsymbol{x})O)\|_2^2$ to derive the upper bound of the truncation error. Namely, Eq. (F12f) can be reformulated as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}\left|\operatorname{Tr}(O\rho_\Lambda(\boldsymbol{x})-\operatorname{Tr}(O\rho(\boldsymbol{x}))\right|^2 && \text{(F17a)}\\
=&\sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0>\Lambda}2^{-\|\boldsymbol{\omega}\|_0}|\alpha_{\boldsymbol{\omega}}|^2 && \text{(F17b)}\\
=&\frac{1}{\Lambda}\left(\Lambda\sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0>\Lambda}2^{-\|\boldsymbol{\omega}\|_0}|\alpha_{\boldsymbol{\omega}}|^2\right) && \text{(F17c)}\\
\leq&\frac{\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]}\|\nabla_{\boldsymbol{x}}\operatorname{Tr}(\rho(\boldsymbol{x})O)\|_2^2}{\Lambda}, && \text{(F17d)}\\
\leq&\frac{C}{\Lambda}, && \text{(F17e)}
\end{aligned}$$

where the last second inequality is support by the result indicated below and the last inequality is supported by the assumption $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]}\|\nabla_{\boldsymbol{x}}\operatorname{Tr}(\rho(\boldsymbol{x})O)\|_2^2\leq C$.

The proof of Eq. (F17d). According to Eq. (F16j), to reach Eq. (F17d), it is equivalent to proving that for $\forall\Lambda\in[d]$,

$$\Lambda\sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0>\Lambda}2^{-\|\boldsymbol{\omega}\|_0}|\alpha_{\boldsymbol{\omega}}|^2\leq\bigstar. \tag{F18}$$

Note that depending on the number of non-zero entries of the frequency (i.e., $\|\boldsymbol{\omega}\|_0$), the left and right hand sides in Eq. (F18) can be decomposed into $d-\Lambda$ parts and $d-1$ parts, respectively. Mathematically, the left hand side of Eq. (F18) yields

$$\Lambda\left(\sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0=\Lambda+1}2^{-\|\boldsymbol{\omega}\|_0}|\alpha_{\boldsymbol{\omega}}|^2+\sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0=\Lambda+2}2^{-\|\boldsymbol{\omega}\|_0}|\alpha_{\boldsymbol{\omega}}|^2+\cdots+\sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0=d}2^{-\|\boldsymbol{\omega}\|_0}|\alpha_{\boldsymbol{\omega}}|^2\right). \tag{F19}$$

Besides, the right hand side of Eq. (F18), i.e., $\bigstar$, can be rewritten as

$$\begin{aligned}
\bigstar=&\left[\left(\sum_{\boldsymbol{\omega}_{2:d},\|\boldsymbol{\omega}_{2:d}\|_0=0}2^{-\|[1,\boldsymbol{\omega}_{2:d}]\|_0}|\alpha_{1,\boldsymbol{\omega}_{2:d}}|^2+\cdots+\sum_{\boldsymbol{\omega}_{2:d},\|\boldsymbol{\omega}_{2:d}\|=d-1}2^{-\|[1,\boldsymbol{\omega}_{2:d}]\|_0}|\alpha_{1,\boldsymbol{\omega}_{2:d}}|^2\right) && \text{(F20)}\\
&+\cdots && \text{(F21)}\\
&+\left(\sum_{\boldsymbol{\omega}_{1:d-1},\|\boldsymbol{\omega}_{1:d-1}\|_0=1}2^{-\|[\boldsymbol{\omega}_{1:d-1},-1]\|_0}|\alpha_{\boldsymbol{\omega}_{1:d-1},-1}|^2+\cdots+\sum_{\boldsymbol{\omega}_{1:d-1},\|\boldsymbol{\omega}_{1:d-1}\|_0=d-1}2^{-\|[\boldsymbol{\omega}_{1:d-1},-1]\|_0}|\alpha_{\boldsymbol{\omega}_{1:d-1},-1}|^2\right)\Bigg] && \text{(F22)}\\
=&\left(\sum_{\boldsymbol{\omega}_{2:d},\|\boldsymbol{\omega}_{2:d}\|_0=0}2^{-\|[1,\boldsymbol{\omega}_{2:d}]\|_0}|\alpha_{1,\boldsymbol{\omega}_{2:d}}|^2+\cdots\sum_{\boldsymbol{\omega}_{1:d-1},\|\boldsymbol{\omega}_{1:d-1}\|_0=0}2^{-\|[\boldsymbol{\omega}_{1:d-1},-1]\|_0}|\alpha_{\boldsymbol{\omega}_{1:d-1},-1}|^2\right) && \text{(F23)}\\
&+\cdots && \text{(F24)}\\
&+\left(\sum_{\boldsymbol{\omega}_{2:d},\|\boldsymbol{\omega}_{2:d}\|_0=d-1}2^{-\|[1,\boldsymbol{\omega}_{2:d}]\|_0}|\alpha_{1,\boldsymbol{\omega}_{2:d}}|^2+\cdots+\sum_{\boldsymbol{\omega}_{1:d-1},\|\boldsymbol{\omega}_{1:d-1}\|_0=d-1}2^{-\|[\boldsymbol{\omega}_{1:d-1},-1]\|_0}|\alpha_{\boldsymbol{\omega}_{1:d-1},-1}|^2\right), && \text{(F25)}
\end{aligned}$$

where the second equality is acquired by rearranging, as in every bracket, all frequencies have the same number of non-zero entries.

In conjunction with Eqs. (F19) and (F25), to achieve Eq. (F18), we need to demonstrate that for any $\Lambda<\Lambda'<d$, the following relation is satisfied, i.e.,

$$\Lambda\sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0=\Lambda'}2^{-\|\boldsymbol{\omega}\|_0}|\alpha_{\boldsymbol{\omega}}|^2\leq\sum_{\boldsymbol{\omega}_{2:d},\|\boldsymbol{\omega}_{2:d}\|_0=\Lambda'-1}2^{-\|[1,\boldsymbol{\omega}_{2:d}]\|_0}|\alpha_{1,\boldsymbol{\omega}_{2:d}}|^2+\cdots+\sum_{\boldsymbol{\omega}_{1:d-1},\|\boldsymbol{\omega}_{1:d-1}\|_0=\Lambda'-1}2^{-\|[\boldsymbol{\omega}_{1:d-1},-1]\|_0}|\alpha_{\boldsymbol{\omega}_{1:d-1},-1}|^2. \tag{F26}$$

To achieve this goal, we next prove that the right hand side in Eq. (F26) equals to

$$\Lambda'\sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0=\Lambda'}2^{-\|\boldsymbol{\omega}\|_0}|\alpha_{\boldsymbol{\omega}}|^2,\quad\text{with }\Lambda<\Lambda'<d. \tag{F27}$$

Recall that in Eq. (F27), each frequency $\boldsymbol{\omega} \in \{0, \pm 1\}^d$ with $\|\boldsymbol{\omega}\|_0 = \Lambda'$ appears Λ' times and the total number of frequencies is

$$\Lambda \cdot \left| \left\{ \boldsymbol{\omega} \in \{0, \pm 1\}^d \middle| \|\boldsymbol{\omega}\|_0 = \Lambda' \right\} \right| = \Lambda \cdot \left| \binom{d}{d-\Lambda'} \right| \cdot 2^{\Lambda'} = \Lambda' \frac{d!}{(d-\Lambda')!\Lambda'!} 2^{\Lambda'} = \frac{d!}{(d-\Lambda')!(\Lambda'-1)!} 2^{\Lambda'}. \tag{F28}$$

In addition, the total number of frequencies in the right hand side of Eq. (F26) is

$$\binom{d-1}{d-\Lambda'-1} 2^{\Lambda'} + \cdots + \binom{d-1}{d-\Lambda'-1} 2^{\Lambda'} = d \cdot \binom{d-1}{d-\Lambda'-1} 2^{\Lambda'} = d \frac{(d-1)!}{(d-\Lambda')(\Lambda'-1)} 2^{\Lambda'} = \frac{d!}{(d-\Lambda')(\Lambda'-1)} 2^{\Lambda'}. \tag{F29}$$

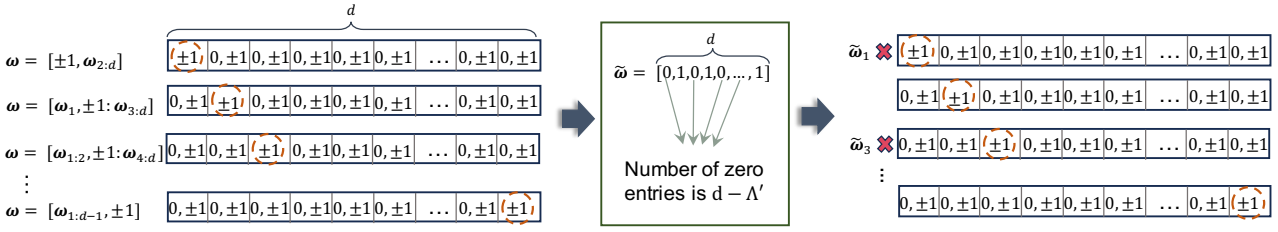

Figure F.5: **A visual interpretation about the appearance times of $\tilde{\omega}$ in the right hand side of Eq. (F26)**.

Combining with the above two equations, we know that the total number of frequencies in Eq. (F27) and the right hand side of Eq. (F26) is the same. Accordingly, their equivalence can be reached if we can show that each frequency $\boldsymbol{\omega} \in \{0, \pm 1\}^d$ with $\|\boldsymbol{\omega}\|_0 = \Lambda'$ appears Λ' times in the right hand side of Eq. (F26). This is indeed the case, as the visual interpretation is shown in Fig. F.5. Denote that the specified frequency as $\tilde{\boldsymbol{\omega}}$ with $\|\tilde{\boldsymbol{\omega}}\|_0 = \Lambda'$. In other words, there are $d - \Lambda'$ entries whose value is 0. This property allows us to determining how many times of $\tilde{\boldsymbol{\omega}}$ appearing in the right side of Eq. (F26). In particular, as shown in the left panel of Fig. F.5, the right hand side of Eq. (F26) can be divided into d groups, depending on the location of ± 1. Note that among these d groups, every entry with the value 0 in $\tilde{\boldsymbol{\omega}}$ precludes one group and only $d - (d - \Lambda') = \Lambda'$ feasible groups are preserved, as shown in the right panel of Fig. F.5. Moreover, for each feasible group, $\tilde{\boldsymbol{\omega}}$ can only appear once, because of the orthogonality of different frequencies. Consequently, we obtain that each frequency $\tilde{\boldsymbol{\omega}} \in \{0, \pm 1\}^d$ with $\|\tilde{\boldsymbol{\omega}}\|_0 = \Lambda'$ appears Λ' times, indicating that the right hand side in Eq. (F26) equals to Eq. (F27).

In other words, we achieve

$$\bigstar = \sum_{\Lambda'=1}^{d} \Lambda' \sum_{\boldsymbol{\omega}, \|\boldsymbol{\omega}\|_0 = \Lambda'} 2^{-\|\boldsymbol{\omega}\|_0} |\alpha_{\boldsymbol{\omega}}|^2 \tag{F30}$$

$$\Rightarrow \quad \bigstar = \sum_{\boldsymbol{\omega}, \|\boldsymbol{\omega}\|_0 > \Lambda} \|\boldsymbol{\omega}\|_0 2^{-\|\boldsymbol{\omega}\|_0} |\alpha_{\boldsymbol{\omega}}|^2 + \sum_{\boldsymbol{\omega}, \Lambda \geq \|\boldsymbol{\omega}\|_0 \geq 1} \|\boldsymbol{\omega}\|_0 2^{-\|\boldsymbol{\omega}\|_0} |\alpha_{\boldsymbol{\omega}}|^2 \tag{F31}$$

$$\Rightarrow \quad \bigstar \geq \sum_{\boldsymbol{\omega}, \|\boldsymbol{\omega}\|_0 > \Lambda} \|\boldsymbol{\omega}\|_0 2^{-\|\boldsymbol{\omega}\|_0} |\alpha_{\boldsymbol{\omega}}|^2 \tag{F32}$$

$$\Rightarrow \quad \bigstar \geq \Lambda \sum_{\boldsymbol{\omega}, \|\boldsymbol{\omega}\|_0 > \Lambda} 2^{-\|\boldsymbol{\omega}\|_0} |\alpha_{\boldsymbol{\omega}}|^2. \tag{F33}$$

Taken together, the truncation error can be upper bounded by the averaged gradient norm of the specified circuit. □

2. Estimation error of the classical learning model (Proof of Lemma F.2)

The core of the proof is to show that the state prediction model $\hat{\sigma}_n(\boldsymbol{x})$ in Eq. (F1) is equal to the trigonometric expansion of the truncated target quantum state $\rho_\Lambda(\boldsymbol{x})$ if we take the expectation over the training data, which includes the randomness from the sampled inputs $\boldsymbol{x}^{(i)}$ and classical shadow. With this relation in mind, we can quantify the estimation error of the proposed state prediction model $\hat{\sigma}_n(\boldsymbol{x})$ when calculating the expectation value of an unseen state under the specified observable O, i.e., $\mathbb{E}_{\boldsymbol{x} \sim [-\pi, \pi]^d} [|\operatorname{Tr}(O\hat{\sigma}_n(\boldsymbol{x})) - \operatorname{Tr}(O\rho_\Lambda(\boldsymbol{x}))|^2]$.

Proof of Lemma F.2. We first prove the equivalence between the expectation of the classical representations and the target quantum state, i.e., $\mathbb{E}_{\mathcal{T}}[\hat{\sigma}_n(\boldsymbol{x})] = \rho_\Lambda(\boldsymbol{x})$. Following the explicit form of $\hat{\sigma}_n(\boldsymbol{x})$ in Eq. (F1), we obtain

$$\begin{aligned}
\mathbb{E}_{\mathcal{T}}[\hat{\sigma}_n(\boldsymbol{x})] = &\quad \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{x}^{(i)}\sim[-\pi,\pi]^d}\left[\kappa_\Lambda\left(\boldsymbol{x},\boldsymbol{x}^{(i)}\right)\right]\mathbb{E}_{s_1^{(\boldsymbol{x}^{(i)})},\ldots,s_N^{(\boldsymbol{x}^{(i)})}}\tilde{\rho}_1(\boldsymbol{x}^{(i)}) && \text{(F34a)}\\
= &\quad \mathbb{E}_{\boldsymbol{x}^{(1)}\sim[-\pi,\pi]^d}\left[\kappa_\Lambda\left(\boldsymbol{x},\boldsymbol{x}^{(1)}\right)\right]\rho(\boldsymbol{x}^{(1)}) && \text{(F34b)}\\
= &\quad \mathbb{E}_{\boldsymbol{x}^{(1)}\sim[-\pi,\pi]^d}\sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0\le\Lambda}2^{\|\boldsymbol{\omega}\|_0}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(1)})\rho(\boldsymbol{x}^{(1)}) && \text{(F34c)}\\
= &\quad \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0\le\Lambda}\Phi_{\omega}(\boldsymbol{x})2^{\|\boldsymbol{\omega}\|_0}\mathbb{E}_{\boldsymbol{x}^{(1)}\sim[-\pi,\pi]^d}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(1)})\rho(\boldsymbol{x}^{(1)}) && \text{(F34d)}\\
= &\quad \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0\le\Lambda}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})2^{\|\boldsymbol{\omega}\|_0}\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(1)})\rho(\boldsymbol{x}^{(1)})\mathsf{d}^d x^{(1)} && \text{(F34e)}\\
= &\quad \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0\le\Lambda}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})2^{\|\boldsymbol{\omega}\|_0}\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(1)})\sum_{\boldsymbol{\omega}'}\Phi_{\boldsymbol{\omega}'}(\boldsymbol{x})\rho_{\boldsymbol{\omega}}\mathsf{d}^d x^{(1)} && \text{(F34f)}\\
= &\quad \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0\le\Lambda}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})2^{\|\boldsymbol{\omega}\|_0}\sum_{\boldsymbol{\omega}'}\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(1)})\Phi_{\boldsymbol{\omega}'}(\boldsymbol{x}^{(1)})\rho_{\boldsymbol{\omega}}\mathsf{d}^m x^{(1)} && \text{(F34g)}\\
= &\quad \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|_0\le\Lambda}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\rho_{\boldsymbol{\omega}} && \text{(F34h)}\\
= &\quad \rho_\Lambda(\boldsymbol{x}), && \text{(F34i)}
\end{aligned}$$

where $s_j^{(\boldsymbol{x}^{(i)})}$ denotes the randomized measurement outcome for the j-th qubit for the state $\rho(\boldsymbol{x}^{(i)})$ with $s_j^{(\boldsymbol{x}^{(i)})} \in \{|0\rangle, |1\rangle, |\pm\rangle, |\pm\imath\rangle\}$, Eq. (F34b) uses the fact that each $\boldsymbol{x}^{(i)}$ is sampled independently and uniformly from $[-\pi,\pi]^d$, Eq. (F34c) employs the explicit formula of the truncated trigonometric monomial kernel κ_Λ, Eq. (F34f) adopts trigonometric expansion of the quantum state, i.e., $\rho(\boldsymbol{x}^{(i)}) = \sum_{\boldsymbol{\omega}}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(i)})\rho_{\boldsymbol{\omega}}$, and Eq. (F34h) comes from the orthogonality of basis functions in Eq. (F13).

To complete the proof, we next move to analyze the estimation error $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}[|\operatorname{Tr}(O\hat{\sigma}_n(\boldsymbol{x})) - \operatorname{Tr}(O\rho_\Lambda(\boldsymbol{x}))|^2]$. Define

$$\tilde{A}_{\boldsymbol{\omega}} = \frac{1}{n}\sum_{i=1}^{n}2^{\|\boldsymbol{\omega}\|_0}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(i)})\operatorname{Tr}(\tilde{\rho}_1(\boldsymbol{x}^{(i)})O) - \operatorname{Tr}(\rho_{\boldsymbol{\omega}}O). \tag{F35}$$

By making use of the trigonometric expansion of ρ_Λ and the explicit formalism of $\hat{\sigma}_n(\boldsymbol{x})$, the estimation error can be reformulated as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}\left[\left|\operatorname{Tr}(O\hat{\sigma}_n(\boldsymbol{x})) - \operatorname{Tr}(O\rho_\Lambda(\boldsymbol{x}))\right|^2\right] && \text{(F36a)}\\
= &\quad \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\kappa_\Lambda(\boldsymbol{x},\boldsymbol{x}^{(i)})\operatorname{Tr}(\tilde{\rho}_1(\boldsymbol{x}^{(i)})O) - \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|\le\Lambda}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\operatorname{Tr}(\rho_{\boldsymbol{\omega}}O)\right|^2\right] && \text{(F36b)}\\
= &\quad \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}\left[\left|\sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|\le\Lambda}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\left(\frac{1}{n}\sum_{i=1}^{n}2^{\|\boldsymbol{\omega}\|_0}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(i)})\operatorname{Tr}(\tilde{\rho}_1(\boldsymbol{x}^{(i)})O) - \operatorname{Tr}(\rho_{\boldsymbol{\omega}}O)\right)\right|^2\right] && \text{(F36c)}\\
= &\quad \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}\left[\sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|\le\Lambda}\sum_{\boldsymbol{\omega}',\|\boldsymbol{\omega}'\|\le\Lambda}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\Phi_{\boldsymbol{\omega}'}(\boldsymbol{x})\tilde{A}_{\boldsymbol{\omega}}\tilde{A}_{\boldsymbol{\omega}'}\right] && \text{(F36d)}\\
= &\quad \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|\le\Lambda}\sum_{\boldsymbol{\omega}',\|\boldsymbol{\omega}'\|\le\Lambda}\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\Phi_{\boldsymbol{\omega}'}(\boldsymbol{x})\mathsf{d}^d x\tilde{A}_{\boldsymbol{\omega}}\tilde{A}_{\boldsymbol{\omega}'} && \text{(F36e)}\\
= &\quad \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|\le\Lambda}2^{-\|\boldsymbol{\omega}\|_0}\tilde{A}_{\boldsymbol{\omega}}^2 && \text{(F36f)}
\end{aligned}$$

$$= \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|\leq\Lambda} 2^{-\|\boldsymbol{\omega}\|_0} \left| \frac{1}{n}\sum_{i=1}^{n} 2^{\|\boldsymbol{\omega}\|_0} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(i)}) \operatorname{Tr}(\tilde{\rho}_1(\boldsymbol{x}^{(i)})O) - \operatorname{Tr}(\rho_{\boldsymbol{\omega}}O) \right|^2 \tag{F36g}$$

$$\equiv \sum_{\boldsymbol{\omega},\|\boldsymbol{\omega}\|\leq\Lambda} |\tilde{D}_{\boldsymbol{\omega}}(\mathcal{T})|^2, \tag{F36h}$$

where Eq. (F36d) adopts the explicit form of the trigonometric monomial kernel κ_Λ, Eq. (F36d) employs the definition $\tilde{A}_{\boldsymbol{\omega}}$ in Eq. (F35), and Eq. (F36f) comes from the evaluation of the orthogonality of basis functions in Eq. (F13).

Such reformulation suggests that the derivation of the upper bound of $\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}[|\operatorname{Tr}(O\hat{\sigma}_n(\boldsymbol{x})) - \operatorname{Tr}(O\rho(\boldsymbol{x}))|^2]$ is reduced to deriving the upper bound of $\tilde{D}_{\boldsymbol{\omega}}(\mathcal{T})$ for $\forall\boldsymbol{\omega}\in\mathfrak{C}(\Lambda)$. To do so, we rewrite the term $\operatorname{Tr}(\rho_{\boldsymbol{\omega}}O)$ in $\tilde{D}_{\boldsymbol{\omega}}(\mathcal{T})$ as the trigonometric monomial expansion of $\rho(\boldsymbol{x})$, i.e.,

$$\operatorname{Tr}(\rho_{\boldsymbol{\omega}}O) = 2^{\|\boldsymbol{\omega}\|_0}\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \operatorname{Tr}(\rho(\boldsymbol{x})O)\mathsf{d}^d x \tag{F37a}$$

$$= 2^{\|\boldsymbol{\omega}\|_0}\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \operatorname{Tr}(\rho(\boldsymbol{x})O) \tag{F37b}$$

$$= 2^{\|\boldsymbol{\omega}\|_0}\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\mathbb{E}_{s_1^{\boldsymbol{x}},\ldots,s_N^{\boldsymbol{x}}} \operatorname{Tr}(\tilde{\rho}_1(\boldsymbol{x})O). \tag{F37c}$$

Accordingly, the quantity $|\tilde{D}_{\boldsymbol{\omega}}(\mathcal{T})|^2$ yields

$$|\tilde{D}_{\boldsymbol{\omega}}(\mathcal{T})|^2 = \left| \frac{1}{n}\sum_{i=1}^{n} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(i)}) \operatorname{Tr}(\tilde{\rho}_1(\boldsymbol{x}^{(i)})O) - \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\mathbb{E}_{s_1^{\boldsymbol{x}},\ldots,s_N^{\boldsymbol{x}}} \operatorname{Tr}(\tilde{\rho}_1(\boldsymbol{x})O) \right|^2. \tag{F38}$$

This formalism hints that we can use Hoeffding's inequality to bound $\tilde{D}_{\boldsymbol{\omega}}(\mathcal{T})$. Recall that the requirement of applying Hoeffding's inequality is ensuring the expectation value is bounded. In our case, we have

$$|\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \operatorname{Tr}(\tilde{\rho}_1(\boldsymbol{x})O)| \leq |\operatorname{Tr}(\tilde{\rho}_1(\boldsymbol{x})O)| \tag{F39a}$$

$$\leq \|O\|_\infty \|\tilde{\rho}_1(\boldsymbol{x})\|_1 \tag{F39b}$$

$$= 3^K B, \tag{F39c}$$

where the first inequality uses the Cauchy–Schwarz inequality and $|\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})|\leq 1$, the second inequality adopts the von Neumann's trace inequality with Hölder's inequality, and the last inequality exploits $\|\tilde{\rho}_1(\boldsymbol{x})\|_1 \leq 3^K$ [79, Eq. (F37)] and the condition $\|O\|_\infty \leq B$.

The bounded expectation term enables us to use Hoeffding's inequality to attain the following result, i.e.,

$$\Pr\left[\tilde{D}_{\boldsymbol{\omega}}(\mathcal{T})^2 \geq \tau^2\right] = \Pr\left[\tilde{D}_{\boldsymbol{\omega}}(\mathcal{T}) \geq \tau\right] \leq 2\exp\left(-\frac{2n\tau^2}{B^2 9^K}\right). \tag{F40}$$

Denote the set of truncated frequencies as $\mathfrak{C}(\Lambda) = \{\boldsymbol{\omega}|\boldsymbol{\omega}\in\{0,\pm 1\}^d,\ s.t.\ \|\boldsymbol{\omega}\|_0\leq\Lambda\}$. This leads to

$$\Pr\left[\sum_{\boldsymbol{\omega}\in\mathfrak{C}(\Lambda)} |D_{\boldsymbol{\omega}}(\mathcal{T})|^2 \geq |\mathfrak{C}(\Lambda)|\tau^2\right] \leq \sum_{\boldsymbol{\omega}\in\mathfrak{C}(\Lambda)} \Pr\left[|D_{\boldsymbol{\omega}}(\mathcal{T})|^2 \geq \tau^2\right] \leq |\mathfrak{C}(\Lambda)|\cdot 2\exp\left(-\frac{2n\tau^2}{B^2 9^K}\right). \tag{F41}$$

Let the right hand side be δ. We have

$$\tau = \sqrt{\frac{1}{2n}B^2 9^K \log\left(\frac{2\cdot|\mathfrak{C}(\Lambda)|}{\delta}\right)}. \tag{F42}$$

This concludes the proof as with probability at least $1-\delta$, the mean-square error between the prediction and the ground truth taken over the randomness of the sampled inputs and classical shadow is upper bounded by

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d}\left[\left|\operatorname{Tr}(O\sigma_n^{(1)}(\boldsymbol{x})) - \operatorname{Tr}(O\rho(\boldsymbol{x}))\right|^2\right] \leq |\mathfrak{C}(\Lambda)|\frac{1}{2n}B^2 9^K \log\left(\frac{2\cdot|\mathfrak{C}(\Lambda)|}{\delta}\right). \tag{F43}$$

□

SM G: Computational time for training and prediction

Here we analyze the computational cost of the proposed ML model. For clarity, here we separately analyze the computational time of our proposal required in the training and inference.

Training time. Recall that the training procedure of our model amounts to loading the collected training dataset $\mathcal{T}_{\mathsf{s}} = \{\boldsymbol{x}^{(i)}, \tilde{\rho}_T(\boldsymbol{x}^{(i)})\}_{i=1}^n$ to the classical memory. According to the explanation of classical shadow in SM A 1, the required computation cost to store and load the N-qubit state $\tilde{\rho}_T(\boldsymbol{x}^{(i)})$ with T snapshots is $\mathcal{O}(NT)$. Simultaneously, the computation cost to store and load the classical control $\boldsymbol{x}^{(i)}$ is $\mathcal{O}(d)$. Combining these facts with the result of Theorem 2 such that the required total number of training examples n of our model, the computation cost to load the dataset $\mathcal{T}_{\mathsf{s}}$ is

$$\mathcal{O}(nNT) = \mathcal{O}\left(NT|\mathfrak{C}(\Lambda)|\frac{B^2 9^K}{\epsilon}\log\left(\frac{2\cdot|\mathfrak{C}(\Lambda)|}{\delta}\right)\right) = \widetilde{\mathcal{O}}\left(\frac{B^2 9^K NT|\mathfrak{C}(4C/\epsilon)|}{\epsilon}\right). \tag{G1}$$

As discussed in the main text, when C is bounded (i.e., $C \sim \mathcal{O}(1/\mathrm{poly}(N))$ or $C \sim \mathcal{O}(1/\exp(N))$) or d is a small constant in many practical scenarios, the cardinality $|\mathfrak{C}(4C/\epsilon)|$ polynomially scales with N and d, and thus training our model is computationally efficient.

Inference (prediction) time. Suppose that the observable O is constituted by multiple local observables with a bounded norm, i.e., $O = \sum_{i=1}^q O_i$ and $\sum_l \|O_i\| \leq B$, and the maximum locality of $\{O_i\}$ is K. Following the definition of our model in Eq. (F1), given a new input $\boldsymbol{x}$ and the observable O, the prediction yields

$$h_{\mathsf{s}} \equiv \mathrm{Tr}(O\hat{\sigma}_n(\boldsymbol{x})) = \frac{1}{n}\sum_{i=1}^n \kappa_\Lambda\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right)\mathrm{Tr}\left(O\tilde{\rho}_T(\boldsymbol{x}^{(i)})\right). \tag{G2}$$

In this regard, the evaluation involves summing over the assessment of each training example $(\boldsymbol{x}^{(i)}, \tilde{\rho}_T(\boldsymbol{x}^{(i)}))$ for $\forall i \in [n]$, and the evaluation of each training example can be further decomposed into two components. That is, the first component is classically computing the shadow estimation $\mathrm{Tr}(O\tilde{\rho}_T(\boldsymbol{x}^{(i)}))$; the second component is calculating the kernel function $\kappa_\Lambda(\boldsymbol{x}, \boldsymbol{x}^{(i)})$ for $\forall i \in [n]$. According to the runtime complexity of shadow estimation elucidated in SM A 1, when $K \sim \mathcal{O}(1)$, the computation of each $\mathrm{Tr}(O_i\tilde{\rho}_T(\boldsymbol{x}^{(i)}))$ can be completed in $\mathcal{O}(T)$ time after storing the classical shadow in the classical memory. In this regard, the computation cost to complete the first part is

$$\mathcal{O}(Tq). \tag{G3}$$

Moreover, based on the explicit definition of the kernel $\kappa_\Lambda(\cdot,\cdot)$, the computation cost of evaluating $\kappa_\Lambda(\boldsymbol{x}, \boldsymbol{x}^{(i)})$ is

$$\mathcal{O}(|\mathfrak{C}(4C/\epsilon)|). \tag{G4}$$

In conjunction with the computation cost of each example and the total number of training examples n, the required predicting time of our proposal is

$$\mathcal{O}(n(Tq + |\mathfrak{C}(4C/\epsilon)|)) \leq \widetilde{\mathcal{O}}\left(\frac{TqB^2 9^K|\mathfrak{C}(4C/\epsilon)|^2}{\epsilon}\right). \tag{G5}$$

In conclusion, when C is bounded, the whole procedure of our proposal (encompassing the training and predicting) is both computational and memory efficient, which is upper bounded by

$$\widetilde{\mathcal{O}}\left(\frac{TNqB^2 9^K|\mathfrak{C}(4C/\epsilon)|^2}{\epsilon}\right). \tag{G6}$$

SM H: Classical prediction model with the full expansion (Proof of Corollary 1)

This section comprises two parts. In SM H 1, we elaborate on the implementation details of the protocol introduced in Corollary 1. Then, in SM H 2, we present the proof details of Corollary 1.

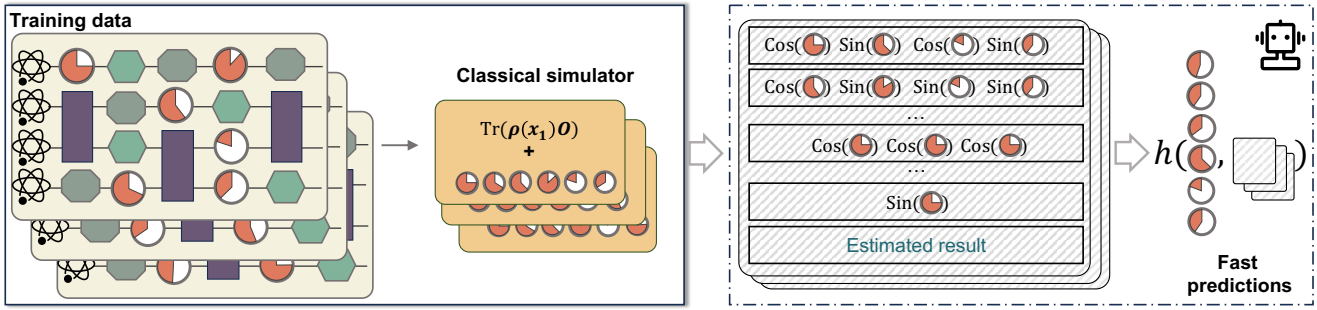

Figure H.6: **Purely classical learning model with the full expansion.** The pipeline is similar to the one presented in Fig. 1 of the main text. The only difference is the way of collecting training data, where the circuit layout is known by the learner and the expectation value $\mathrm{Tr}(\rho(bx)O)$ should be efficiently calculated by the employed classical simulators.

1. Implementation of the classical prediction model

The implementation of the protocol introduced in Corollary 1 is visualized in Fig. H.6. The only difference with the one introduced in the main text is the way of collecting training data. Specifically, when the tunable quantum circuit architecture formed by RZ + CI is known to the learner, the classical simulators (e.g., tensor network simulators, LOWESA, and near Clifford circuits simulators) are employed to collect the training data

$$\mathcal{T}_{\mathsf{c}} = \left\{ \boldsymbol{x}^{(i)} \to \mathrm{Tr}\left(\rho(\boldsymbol{x}^{(i)})O\right) \right\}_{i=1}^{n}. \tag{H1}$$

Once the training dataset is prepared, the learner uses it to form the kernel-based prediction model, i.e., given a new input $\boldsymbol{x}$ and an observable O, it takes the form as

$$h_{\mathsf{c}}(\boldsymbol{x}, O) = \frac{1}{n}\sum_{i=1}^{n} \kappa\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) \mathrm{Tr}(\rho(\boldsymbol{x}^{(i)})O) \text{ with } \kappa(\boldsymbol{x}, \boldsymbol{x}^{(i)}) = \sum_{\boldsymbol{\omega}, \|\boldsymbol{\omega}\|_0 \le d} 2^{\|\boldsymbol{\omega}\|_0} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(i)}) \in \mathbb{R}. \tag{H2}$$

Note that this protocol is relatively more restrictive than the one introduced in the main text, as the dimension d is required to be sufficiently low with the full expansion and the expectation value under the observable O should be efficiently calculated by the exploited classical simulators.

2. Proof of Corollary 1

The formal statement of Corollary 1 is as follows.

Corollary (Restatement of Corollary 1)**.** *Following notations in the main text, consider a parametrized family of N-qubit states $\mathcal{Q}$ and $\|O\|_\infty \le B$ Then, let the dataset be $\mathcal{T}_{\mathsf{c}} = \{\boldsymbol{x}^{(i)} \to \mathrm{Tr}(\rho(\boldsymbol{x}^{(i)})O)\}_{i=1}^{n}$ with $\boldsymbol{x}^{(i)} \sim \mathrm{Unif}[-\pi, \pi]^d$ and $n = 3^d B^2 \log\left(\frac{2\cdot 3^d}{\delta}\right)(2\epsilon)^{-1}$. Then with probability at least $1-\delta$, the average prediction error of the prediction model in Eq. (H2) satisfies*

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left| h_{\mathsf{c}}(\boldsymbol{x}, O) - \mathrm{Tr}(O\rho(\boldsymbol{x})) \right|^2 \le \epsilon. \tag{H3}$$

Proof of Corollary 1. Since the full expansion is adopted, the truncation error is eliminated. We only need to quantify the estimation error induced by the finite training examples. The discrepancy can be reformulated as

$$\mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left[\left| h_{\mathsf{c}}(\boldsymbol{x}, O) - \mathrm{Tr}(O\rho(\boldsymbol{x})) \right|^2 \right] \tag{H4a}$$

$$= \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left[\left| \sum_{\boldsymbol{\omega}} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \left(\frac{1}{n}\sum_{i=1}^{n} 2^{\|\boldsymbol{\omega}\|_0} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(i)}) \mathrm{Tr}(\rho(\boldsymbol{x}^{(i)})O) - \mathrm{Tr}(\rho_{\boldsymbol{\omega}} O) \right) \right|^2 \right] \tag{H4b}$$

$$= \mathbb{E}_{\boldsymbol{x}\sim[-\pi,\pi]^d} \left[\sum_{\boldsymbol{\omega}} \sum_{\boldsymbol{\omega}'} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \Phi_{\boldsymbol{\omega}'}(\boldsymbol{x}) A_{\boldsymbol{\omega}} A_{\boldsymbol{\omega}'} \right] \tag{H4c}$$

$$= \sum_{\boldsymbol{\omega}} \sum_{\boldsymbol{\omega}'} \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \Phi_{\boldsymbol{\omega}'}(\boldsymbol{x}) \mathrm{d}^d x A_{\boldsymbol{\omega}} A_{\boldsymbol{\omega}'} \tag{H4d}$$

$$= \sum_{\boldsymbol{\omega}} 2^{-\|\boldsymbol{\omega}\|_0} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) A_{\boldsymbol{\omega}}^2 \tag{H4e}$$

$$= \sum_{\boldsymbol{\omega}} 2^{-\|\boldsymbol{\omega}\|_0} \left| \frac{1}{n} \sum_{i=1}^{n} 2^{\|\boldsymbol{\omega}\|_0} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(i)}) \operatorname{Tr}(\rho(\boldsymbol{x}^{(i)}) O) - \operatorname{Tr}(\rho_{\boldsymbol{\omega}} O) \right|^2 \tag{H4f}$$

$$\equiv \sum_{\boldsymbol{\omega}} |D_{\boldsymbol{\omega}}(\mathcal{T})|^2, \tag{H4g}$$

where Eq. (H4f) comes from the evaluation of the orthogonality of basis functions in Eq. (F13). Recall the analysis in the proof of Theorem 1, the quantity $|D_{\boldsymbol{\omega}}(\mathcal{T})|^2$ yields

$$|D_{\boldsymbol{\omega}}(\mathcal{T})|^2 \quad = \left| \frac{1}{n} \sum_{i=1}^{n} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(i)}) \operatorname{Tr}\left(\rho(\boldsymbol{x}^{(i)}) O\right) - \mathbb{E}_{\boldsymbol{x} \sim [-\pi,\pi]^d} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \operatorname{Tr}(\rho(\boldsymbol{x}) O) \right|^2. \tag{H5}$$

This formalism hints that we can use Hoeffding's inequality to bound $D_{\boldsymbol{\omega}}(\mathcal{T})$. In this case, we have

$$|\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \operatorname{Tr}(\rho(\boldsymbol{x}) O)| \leq |\operatorname{Tr}(\rho(\boldsymbol{x}) O)| \leq \|O\|_\infty \|\rho(\boldsymbol{x})\|_1 = B, \tag{H6}$$

where the first inequality uses the Cauchy–Schwarz inequality and $|\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})| \leq 1$, the second inequality adopts von Neumann's trace inequality with Hölder's inequality, and the last inequality exploits $\|\rho(\boldsymbol{x})\|_1 = 1$ and $\|O\|_\infty \leq B$.

The bounded expectation term enables us to use Hoeffding's inequality to attain the following result, i.e.,

$$\Pr[D_{\boldsymbol{\omega}}(\mathcal{T})^2 \geq \tau^2] = \Pr[D_{\boldsymbol{\omega}}(\mathcal{T}) \geq \tau] \leq 2 \exp\left(-\frac{2n\tau^2}{B^2}\right). \tag{H7}$$

This leads to

$$\Pr\left[\sum_{\boldsymbol{\omega} \in \mathfrak{C}(d)} |D_{\boldsymbol{\omega}}(\mathcal{T})|^2 \geq |\mathfrak{C}(d)| \tau^2\right] \leq \sum_{\boldsymbol{\omega} \in \mathfrak{C}(d)} \Pr\left[|D_{\boldsymbol{\omega}}(\mathcal{T})|^2 \geq \tau^2\right] \leq 3^d \cdot 2 \exp\left(-\frac{2n\tau^2}{B^2}\right). \tag{H8}$$

Let the right hand side be δ. We have

$$\tau = \sqrt{\frac{1}{2n} B^2 \log\left(\frac{2 \cdot 3^d}{\delta}\right)}. \tag{H9}$$

This concludes the proof as with probability at least $1 - \delta$, the mean-square error between the prediction and the ground truth taken over the randomness of sampled inputs and classical shadow is upper bounded by

$$\mathbb{E}_{\boldsymbol{x} \sim [-\pi,\pi]^d} \left[\left| \operatorname{Tr}(O \hat{\sigma}_n^{(\mathrm{c})}(\boldsymbol{x})) - \operatorname{Tr}(O \rho(\boldsymbol{x})) \right|^2 \right] \leq 3^d \frac{1}{2n} B^2 \log\left(\frac{2 \cdot 3^d}{\delta}\right). \tag{H10}$$

Let the right-hand side be equal to the tolerant error ϵ. We have

$$n = \frac{3^d B^2 \log\left(\frac{2 \cdot 3^d}{\delta}\right)}{2\epsilon}. \tag{H11}$$

□

SM I: Learning bounded-gate quantum circuit with CI gates and parameterized multi-qubit gates

In the main text, we primarily elaborate on the application of the proposed ML model in predicting bounded-gate quantum circuits consisting of RZ and CI gates. However, our proposal is adaptable and can be readily extended to predict bounded-gate quantum circuits composed of alternative basis gate sets. To exemplify this flexibility, in this section, we illustrate how our proposal and the associated theoretical findings (Theorem 2) can be effectively expanded

to encompass a wider context, specifically bounded-gate circuits incorporating CI gates alongside parameterized multi-qubit gates generated by arbitrary Pauli strings.

For ease of notations, given an N-qubit quantum circuit, we define the j-th parameterized gate generated by the Pauli string $P_{\boldsymbol{a}_j} \in \{\mathbb{I}, X, Y, Z\}^N$ with $\boldsymbol{a}_j \in \{0,1,2,3\}^N$ as

$$\mathrm{RP}(\boldsymbol{x}_j) = \exp\left(-\imath\frac{\boldsymbol{x}_j}{2}P\right) \equiv \cos\left(\frac{\boldsymbol{x}_j}{2}\right)\mathbb{I}_{2^N} + \imath\sin\left(\frac{\boldsymbol{x}_j}{2}\right)P. \tag{I1}$$

Then the bounded-gate quantum circuit takes the form as

$$U(\boldsymbol{x}) = \prod_{j=1}^{d}(\mathrm{RP}_{\boldsymbol{a}_j}(\boldsymbol{x}_j)u_e), \tag{I2}$$

where $u_e \in \mathrm{CI}$ refers to an arbitrary Clifford gate or their combinations. In what follows, we demonstrate how the proposed ML model effectively predicts the incoherent dynamics of bounded-gate circuits in Eq. (I1), with the resulting prediction error aligning with the findings outlined in Theorem 2.

Recall the implementation of the proposed ML model and the proof of Theorem 2 presented in SM F. In brief, if the pre-measured state under the Pauli-basis expansion can be expressed as the form

$$\rho(\boldsymbol{x}) = U(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N}U(\boldsymbol{x})^\dagger = \sum_{\boldsymbol{\omega}}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\langle\langle 0|\,\mathfrak{U}_{\boldsymbol{\omega}}^\dagger \equiv \sum_{\boldsymbol{\omega}}\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\rho_{\boldsymbol{\omega}}, \tag{I3}$$

where the trigonometric monomial basis is $\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) = \prod_{i=1}^{d}\begin{cases}1 & \text{if } \boldsymbol{\omega}_i = 0\\ \cos(\boldsymbol{x}_i) & \text{if } \boldsymbol{\omega}_i = 1\\ \sin(\boldsymbol{x}_i) & \text{if } \boldsymbol{\omega}_i = -1\end{cases}$, then we can form the state prediction model in Eq. (F1), i.e.,

$$\hat{\sigma}_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\kappa\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right)\tilde{\rho}_T(\boldsymbol{x}^{(i)}), \quad \text{with } \kappa(\boldsymbol{x}, \boldsymbol{x}^{(i)}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d, \|\boldsymbol{k}\|_2\leq\Lambda}\cos\left(\pi\boldsymbol{k}\cdot\left(\boldsymbol{x} - \boldsymbol{x}^{(i)}\right)\right) \in \mathbb{R}, \tag{I4}$$

and the analysis in Theorem 2 applies. In this respect, the key aspect of extending our proposal from $\{\mathrm{RZ} + \mathrm{CI}\}$ gate set to $\{\mathrm{RP}_{\boldsymbol{a}_j} + \mathrm{CI}\}$ gate set involves demonstrating that each parameterized gate $\mathrm{RP}_{\boldsymbol{a}_j}$ with $j \in [d]$ can be represented in the form specified by Eq. (A10), i.e.,

$$\mathfrak{R}_{\mathrm{P}_{\boldsymbol{a}_j}}(\boldsymbol{x}_j) = D_0' + \cos(\boldsymbol{x}_j)D_1' + \sin(\boldsymbol{x}_j)D_2', \tag{I5}$$

where D_0', D_1', and D_2' are three constant matrices with the size $2^N \times 2^N$.

We next prove that the form in Eq. (I5) is satisfied for any Pauli string $P_{\boldsymbol{a}_j}$. For clarity, in the following analysis, we denote $P_{\boldsymbol{a}_j}$ and $\mathrm{RP}_{\boldsymbol{a}_j}(\boldsymbol{x}_j)$ as P_j and $\mathrm{RP}_j(x)$, respectively. According to the definition of PTM, we have

$$[\mathfrak{R}_{\mathrm{P}_{\boldsymbol{a}_j}}(\boldsymbol{x}_j)]_{ik} = \mathrm{Tr}\left(\mathrm{RP}_j(x)P_i\,\mathrm{RP}_j(-x)P_k\right). \tag{I6}$$

Since P_i, P_j, and P_k are Pauli strings, any two of them must commute or anti-commute to each other. By exploiting this property, we next separately analyze the value of $[\mathfrak{R}_{\mathrm{P}_{\boldsymbol{a}_j}}(\boldsymbol{x}_j)]_{ik}$ for $\forall i, k \in [4^d]$.

Case I: $[P_i, P_j] = 0$ *or* $[P_k, P_j] = 0$. In this case, Eq. (I6) can be simplified to

$$[\mathfrak{R}_{\mathrm{P}_{\boldsymbol{a}_j}}(\boldsymbol{x}_j)]_{ik} = \mathrm{Tr}(P_iP_k) = \delta_{ij}. \tag{I7}$$

In other words, for the indices $i, k \in [4^d]$ whose corresponding Pauli strings commute with P_j, the corresponding PTM entry is a constant, which is zero if $i \neq k$ and 1 if $i = k$. According to the formula in Eq. (I5), the relevant entries, whose indices satisfy $[P_i, P_j] = 0$ or $[P_k, P_j] = 0$, belong to D_0'.

Case II: $\{P_i, P_j\} = 0$ *or* $\{P_k, P_j\} = 0$. In this case, Eq. (I6) can be simplified to

$$[\mathfrak{R}_{\mathrm{P}_{\boldsymbol{a}_j}}(\boldsymbol{x}_j)]_{ik} = \mathrm{Tr}(\mathrm{RP}_j(2x)P_iP_k) = \cos(x)\,\mathrm{Tr}(P_iP_k) + \imath\sin(x)\,\mathrm{Tr}(P_jP_iP_k) = \cos(x)\delta_{ij} + \imath\sin(x)\,\mathrm{Tr}(P_jP_iP_k), \tag{I8}$$

where the second equality employs $\mathrm{RP}_j(2x) = \cos(x)\mathbb{I}_{2^N} + \imath\sin(x)P_j$. In other words, for the indices $i, k \in [4^d]$ whose corresponding Pauli strings anticommute with P_j, (i) when $i = k$, the corresponding PTM entry only contributes to D_1' in Eq. (I5); (ii) when $i \neq k$ and $P_jP_iP_k = \mathbb{I}_{2^N}$, the corresponding PTM entry only contributes to D_2' in Eq. (I5); (iii) else, the corresponding PTM entry is zero, which only contributes to D_0'.

Taken together, for any pair indices (i, k), the corresponding entry only contributes to one of D_0', D_1', and D_2', indicating the relation in Eq. (I5). As a result, our proposal and Theorem 2 applies to the basis gate set $\{\mathrm{RP}_{\boldsymbol{a}_j}, \mathrm{CI}\}$.

SM J: More numerical simulations

In this section, we demonstrate more simulation details omitted in the main text. Specifically, in SM J 1, we provide more simulation details about predicting properties of rotational N-qubit GHZ states. Then, in SM J 2, we illustrate more simulation results about the global Hamiltonian simulation task. Last, in SM J 3, we demonstrate how to use the proposed ML model to enhance variational quantum algorithms, including variational quantum Eigen-solver and quantum neural networks.

1. Numerical simulations of N-qubit rotational GHZ states

Dataset construction of N-qubit rotational GHZ states. The mathematical form of the N-qubit rotational GHZ states is

$$|\mathrm{GHZ}(\boldsymbol{x})\rangle = \left(\mathrm{RY}_1(\boldsymbol{x}_1)\otimes \mathrm{RY}_{N/2}(\boldsymbol{x}_2)\otimes \mathrm{RY}_N(\boldsymbol{x}_3)\right)\frac{|0\cdots 0\rangle + |1\cdots 1\rangle}{\sqrt{2}}. \tag{J1}$$

The circuit implementation of this class of quantum states is visualized in Fig. J.7(a).

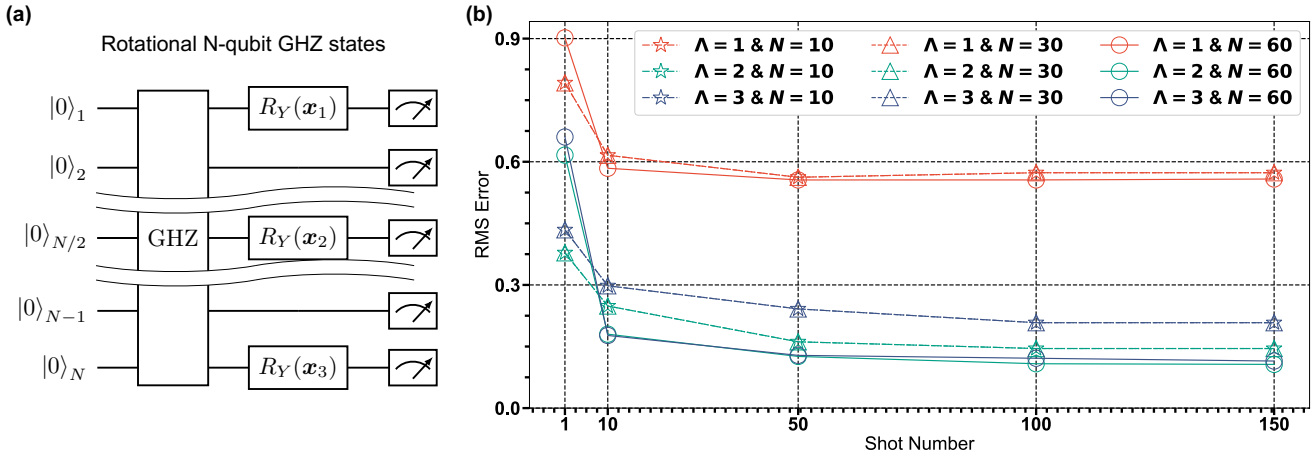

Figure J.7: **More simulation results of rotational N-qubit GHZ states**. (a) Circuit implementation of the rotational N-qubit GHZ states. The three RY gates apply to the first qubit, the $(N/2)$-th qubit, and the last qubit. (b) Prediction error. The root mean squared (RMS) error of the trained ML model with varied truncation Λ, the shot number T, and the qubit count N.

In both subtasks presented in the main text, the classical shadow of each example $(\boldsymbol{x}^{(i)}, \tilde{\rho}_T(\boldsymbol{x}^{(i)}))$ in the training dataset $\mathcal{T}$ are acquired by PastaQ library [113]. In the first subtask of two-point correlation estimation, the exact value is obtained by matrix product operators (MPO) provided by PastaQ [113]. In the second subtask of predicting expectation values on $Z_1 \otimes Z_N$, the accurate results yield

$$\begin{aligned} & \langle \mathrm{GHZ}(\boldsymbol{x})|Z_1\otimes Z_N|\mathrm{GHZ}(\boldsymbol{x})\rangle \\ = \; & -\sin(\boldsymbol{x}_1)\sin(\boldsymbol{x}_2)\cos(\boldsymbol{x}_3) - \cos(\boldsymbol{x}_1)\sin(\boldsymbol{x}_2)\sin(\boldsymbol{x}_3) + \sin(\boldsymbol{x}_1)\cos(\boldsymbol{x}_2)\sin(\boldsymbol{x}_3) + \cos(\boldsymbol{x}_1)\cos(\boldsymbol{x}_2)\cos(\boldsymbol{x}_3). \end{aligned} \tag{J2}$$

Hyper-parameter settings. The random seed to collect training examples and test examples is set as 1234 and 123, respectively. The hyper-parameter of MPO used to calculate the exact values of the two-point correlation of rotational GHZ states is as follows. The cutoff value is set as 10^{-8} and the max dimension is set as 50.

Cardinality of the frequency set $\mathfrak{C}(\Lambda)$. In the main text, we adopt three settings of the maximum frequency length, i.e., $\Lambda = 1, 2, 3$, to evaluate the performance of the proposed ML model. The corresponding cardinality of the frequency set is $|\mathfrak{C}(\Lambda = 1)| = 7$, $|\mathfrak{C}(\Lambda = 2)| = 19$, and $|\mathfrak{C}(\Lambda = 3)| = 27$, respectively. The similar performance between $\Lambda = 2$ and $\Lambda = 3$ (full expansion) in Fig. 2 indicates that truncating the high-frequency terms does not apparently affect the capability of the proposed ML model.

Prediction error versus varied number of qubits N. We append more simulation results about how the prediction error of the proposed ML model depends on the shot number T, the truncation Λ, and the qubit count

N. To be specific, we fix the number of training examples to be $n = 500$ and collect these training examples under different qubit counts N, where the maximum shot number is set as $T = 150$. Fig. J.7(b) visualizes the root mean squared (RMS) prediction error under different settings on 10 test examples. The achieved results indicate that the performance of the proposed ML model is dominated by the truncation number Λ (i.e., the dimension of classical controls d) and not sensitive to the number of qubits, which echoes our theoretical analysis. More precisely, when $\Lambda \geq 2$ and $T \geq 100$, the prediction error attains a very low value for both $N = 10, 30, 60$.

2. More details of synthetic global Hamiltonian simulation

Hyper-parameter settings. In the main text, we adopt three settings of the maximum frequency length, i.e., $\Lambda = 1, 2, 3, 4$, to evaluate the performance of the proposed ML model. The corresponding cardinality of the frequency set is $|\mathfrak{C}(\Lambda = 1)| = 61$, $\mathfrak{C}|(\Lambda = 2)| = 1801$, $|\mathfrak{C}(\Lambda = 3)| = 472761$, and respectively. When evaluating the statistical performance of the proposed ML model with $n < 50000$, we sample a subset from the whole dataset using different random seeds. The random seeds for all relevant simulations are set as 12345, 22222, 33333, 44444, and 55555, respectively.

Standard error for the simulation results in Fig. 3(b). For ease of visualization, Fig. 3(b) only demonstrates the averaged RMS prediction error of our model. For completeness, here we append the corresponding standard error in Table I. Particularly, for all settings with $\Lambda \in \{1, 2, 3, 4\}$, the averaged RMS prediction error and the corresponding standard error continuously decrease with the increased number of training examples.

Table I: **Simulation results of predicting the magnetization of states evolved by the 60-qubit global Hamiltonian.**

	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
$\Lambda = 1$	0.0424 ± 0.0222	0.0210 ± 0.0126	0.0130 ± 0.0109	0.0152 ± 0.0117	0.0143 ± 0.0088
$\Lambda = 2$	0.1556 ± 0.0959	0.1862 ± 0.0933	0.0729 ± 0.0472	0.0729 ± 0.0472	0.037 ± 0.0293
$\Lambda = 3$	0.3244 ± 0.3473	0.3736 ± 0.3249	0.4012 ± 0.2055	0.3502 ± 0.1357	0.2631 ± 0.1003
$\Lambda = 4$	2.1227 ± 1.061	0.9519 ± 0.7148	0.9678 ± 0.6590	0.7926 ± 0.4887	0.6432 ± 0.5222

The role of shot number T. Here we conduct numerical simulations to explore how the shot number T influences the prediction error. Specifically, we fix the number of training examples as $n = 500$ and the truncation value as $\Lambda = 1$, but vary the shot number from $T = 50$ to $T = 00$. Each setting is repeated five times to collect the statistical results, where the randomness stems from distilling different training datasets from 600 training examples. The simulation results are summarized in Table. II, where the averaged prediction error does not decrease with the increased number of measurements. This phenomenon echoes our theoretical analysis, suggesting that once the shot number exceeds a certain threshold, it does not heavily affect the performance of the proposed model.

Table II: **Simulation results of predicting the magnetization of states evolved by the 60-qubit global Hamiltonian with varied shot numbers**.

	$T = 50$	$T = 500$
$n = 500$ & $\Lambda = 1$	0.0130 ± 0.0109	0.0251 ± 0.0038

3. More simulation results on enhanced variational quantum algorithms

As shown in SM A 3, a major application of the proposed ML model is enhancing variational quantum algorithms by substantially reducing the quantum resource demands. In this subsection, we provide a comprehensive explanation of the corresponding algorithmic implementation and then proceed to numerical simulations that demonstrate the efficacy of our approach.

Algorithmic implementation. Recall that contemporary quantum devices encounter limitations such as connectivity, gate fidelities, and coherence time. To overcome these constraints, experimentalists typically adopt hardware-efficient ansatzes (HEAs) constructed from the native gate set from the specified quantum device to implement various variational quantum algorithms [114, 115]. Essentially, these HEAs are tailored to individual quantum devices, maintaining consistent gate layouts while adjusting parameters for various computational tasks, as delineated by the

framework formulated in Eq. (3) of the main text. This intrinsic relation warrants the use of the proposed ML model to enhance plenty of variational quantum algorithms associated with device-specific HEAs.

The learning framework is summarized in Alg. 1. For a given quantum device and its corresponding HEA, the learner first constructs the training dataset following the outlined procedure in the main text. Once the dataset is prepared, the proposed ML model serves as a surrogate to optimize variational quantum algorithms in which the predictions correspond to the gradient information, obviating the need for direct implementation on the specified quantum device. Since the optimization process (i.e., the third step) is conducted entirely on classical processors, it

Algorithm 1: Optimizing variational quantum algorithms in an offline manner

1. (Dataset construction) Randomly generate classical controls $\boldsymbol{x} \sim [-\pi, \pi]^d$, feed it to the specified N-qubit quantum device to obtain the pre-measured state $U(\boldsymbol{x})\left|0\right\rangle^{\otimes N}$, and apply Pauli-based classical shadow with T shots to constitute a single training example $(\boldsymbol{x}, \tilde{\rho}_T(\boldsymbol{x}))$;
2. (Dataset construction) Repeat the above procedure n times to construct the training dataset $\mathcal{T}_{\mathsf{s}}$;
3. (Downstream task optimization) Formalize the prediction model h_{s} following Eq. (6) and use it to optimize various downstream tasks without access to the quantum processor.

significantly reduces the overhead associated with accessing sparse quantum devices in the contemporary era.

In what follows, we demonstrate how our approach supports two crucial classes of variational quantum algorithms: the variational quantum Eigen-solvers (VQEs) and quantum neural networks (QNNs). Note that the learning scheme presented below can be readily extended to wide applications covered by variational quantum algorithms.

Ground state energy estimation by VQE. Here we first briefly recap the mechanism of VQE when applied to estimate the ground state energy of a specified Hamiltonian. Given an N-qubit Hamiltonian $\mathfrak{H} \in \mathbb{C}^{2^N \times 2^N}$, the ground state energy estimation aims to find its minimum eigenvalue, i.e.,

$$\mathfrak{E}^* = \min_{\left|\psi\right\rangle \in \mathbb{C}^{2^N}} \left\langle\psi\right|\mathfrak{H}\left|\psi\right\rangle. \tag{J3}$$

To estimate the ground state energy $\mathfrak{E}^*$, VQE adopts an ansatz $W(\boldsymbol{\theta})$ to prepare a variational quantum state $\left|\psi(\boldsymbol{\theta})\right\rangle = W(\boldsymbol{\theta})\left|\psi_0\right\rangle^{\otimes N}$ with a fixed input state $\left|\psi_0\right\rangle$. The trainable parameters $\boldsymbol{\theta}$ are optimized by minimizing the loss function $\mathcal{L}(\boldsymbol{\theta}, \mathfrak{H}) = \mathrm{Tr}\left(\left(\left|\psi_0\right\rangle\left\langle\psi_0\right|\right)W(\boldsymbol{\theta})^\dagger \mathfrak{H} W(\boldsymbol{\theta})\right)$. The optimization of VQE follows an iterative manner, i.e., the classical optimizer continuously leverages the output of the quantum circuits to update $\boldsymbol{\theta}$ and the update rule is

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \frac{\partial \mathcal{L}(\boldsymbol{\theta}^{(t)}, \mathfrak{H})}{\partial \boldsymbol{\theta}}, \tag{J4}$$

where η refers to the learning rate. The first-order gradient in the equation can be obtained by parameter shift rule [116]. Mathematically, the derivative with respect to the k-th parameter $\boldsymbol{x}_k$ for $\forall k \in [d]$ is

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}^{(t)}, \mathfrak{H})}{\partial \boldsymbol{\theta}_k} = \frac{1}{2\sin\alpha}\left[\mathrm{Tr}\left(\left(\left|\psi_0\right\rangle\left\langle\psi_0\right|\right)W(\boldsymbol{\theta}^{(t,+)})\mathfrak{H}W(\boldsymbol{\theta}^{(t,+)})^\dagger\right) - \mathrm{Tr}\left(\left(\left|\psi_0\right\rangle\left\langle\psi_0\right|\right)W(\boldsymbol{\theta}^{(t,-)})\mathfrak{H}W(\boldsymbol{\theta}^{(t,-)})^\dagger\right)\right], \tag{J5}$$

where $\boldsymbol{\theta}^{(t,\pm)} = \boldsymbol{\theta}^{(t)} \pm \alpha \boldsymbol{e}_k$, $\boldsymbol{e}_k$ is the unit vector along the $\boldsymbol{\theta}_k$ axis, and α can be any real number but the multiple of π because of the diverging denominator.

Supported by the proposed ML model, the optimization of $\boldsymbol{\theta}$ can be entirely carried out on classical processors. More precisely, define $W(\boldsymbol{\theta}) = U(\boldsymbol{x})$ (see Fig. J.8 for the visual interpretation), the two terms of the derivatives in Eq. (J5) can be predicted by h_{s} in Alg. 1 and the optimized parameters can be obtained in an offline manner. For convenience, in the subsequent context, we dub the VQEs optimized by the proposed ML model as *Offline-VQE*. An alternative scenario involves utilizing the optimized parameters of Offline-VQE as effective initial parameters for running VQEs on real quantum devices. In either case, the offline optimization enabled by the proposed ML model dramatically reduces the demand for quantum resources, a critical advantage given the scarcity of available quantum processors in the current landscape.

Binary classification by QNN. Denote the input space as $\mathcal{Z}$, the binary label (class) space as $\mathcal{Y} = \{0, 1\}$, and the training dataset as $\mathcal{D} = \{(\boldsymbol{z}^{(i)}, y^{(i)})\}_{i=1}^n$ with samples drawn i.i.d. from an unknown probability distribution $\mathbb{D}$ on $\mathcal{Z} \times \mathcal{Y}$. The purpose of a binary classification algorithm is using $\mathcal{D}$ to infer a hypothesis (a.k.a., a binary classifier) $g_{\mathcal{D}} : \mathcal{Z} \to \mathbb{R}$ from the hypothesis space to separate training examples from binary classes. This is equivalent to identifying an optimal hypothesis minimizing the expected risk $\mathsf{R}(g) = \mathbb{E}_{(\boldsymbol{z},\boldsymbol{y})\sim\mathbb{D}}[\ell(g(\boldsymbol{z}), y)]$, where $\ell(\cdot,\cdot)$ is the per-sample loss and for clarity we specify it as the square error [117]. Unfortunately, the inaccessible distribution $\mathbb{D}$ forbids

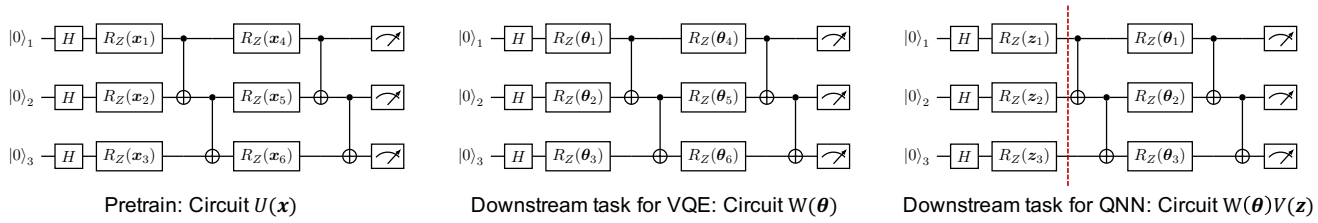

Figure J.8: **A toy model of optimizing variational quantum algorithms in an offline manner**. Left panel: In the pre-training state, the learner collects training data from a 3-qubit device with a hardware-efficient ansatz (HEA) $U(\boldsymbol{x})$, which contains $d = 6$ tunable parameters. Then the learner can use the collected data to train an ML model introduced in the main text to enhance various downstream tasks. Middle panel: The optimization of variational quantum Eigen-solvers (VQEs) associated with the ansatz $W(\boldsymbol{\theta})$ can be accomplished by the proposed ML model without the access to the quantum resources. Right panel: The optimization of quantum neural networks (QNNs) associated with the encoding circuit $V(\boldsymbol{z})$ and the trainable circuit $W(\boldsymbol{\theta})$ can be accomplished by the proposed ML model without the access to the quantum resources.

us to assess the expected risk directly. In practice, the binary classification algorithm alternatively learns an empirical classifier $\hat{h}$ as the global minimizer of the (regularized) loss function

$$\mathcal{L}(g, \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(g(\boldsymbol{z}^{(i)}) - y^{(i)}\right)^2 + \mathfrak{E}(g), \tag{J6}$$

where $\mathfrak{E}(g)$ is an optional regularizer. Given an unseen example $\boldsymbol{z}'$, its predicted label is 0 if $g(\boldsymbol{z}') < 0.5$; otherwise, its predicted label is 1.

When QNN is employed to implement the binary classifier, the hypothesis g is realized by variational quantum circuits followed by a predefined measurement operator Π_0 [118, 119]. The mathematical expression of the hypothesis space for the binary quantum classifier (BiQC) is

$$\mathcal{G} = \left\{ g(\boldsymbol{z}, \boldsymbol{\theta}) = \mathrm{Tr}\left(V(\boldsymbol{z})(|0\rangle\langle 0|)^{\otimes N}V(\boldsymbol{z})^\dagger W(\boldsymbol{\theta})^\dagger \Pi_0 W(\boldsymbol{\theta})\right) \middle| \boldsymbol{\theta} \in \Theta \right\}, \tag{J7}$$

where $W(\boldsymbol{\theta})$ is the trainable circuit and $V(\boldsymbol{z})$ is another variational circuit that encodes the training example $\boldsymbol{z}^{(i)}$ into the quantum state $\rho(\boldsymbol{z}^{(i)})$. As with VQE, the optimization of trainable parameters $\boldsymbol{\theta}$ can be completed by the gradient descent methods (e.g., stochastic gradient descent or batch gradient descent) and the derivatives can be acquired by the parameter shift rule given in Eq. (J5).

Supported by the proposed ML model, the optimization of $\boldsymbol{\theta}$ for BQC can be entirely carried out on classical processors. According to Alg. 1, the classical control $\boldsymbol{x}$ should be divided into parts, where the first part refers to the training example $\boldsymbol{z}$ and the second part refers to the trainable parameters $\boldsymbol{\theta}$ in BiQCs. In addition, the circuit architecture of $U(\boldsymbol{x})$ amounts to the combination of $W(\boldsymbol{\theta})$ and $V(\boldsymbol{z})$ (see Fig. J.8 for the visual interpretation). In this way, for an arbitrary input $\boldsymbol{x}$ and trainable parameters $\boldsymbol{\theta}$, the derivatives $\partial\mathcal{L}/\partial\boldsymbol{\theta}_k$, which is formed by $g(\boldsymbol{z}, \boldsymbol{\theta}^{(t)})$ and $g(\boldsymbol{z}, \boldsymbol{\theta}^{(t,\pm)})$, can be estimated by the proposed ML model without accessing quantum resources. Due to this offline property, we call BiQCs optimized by the proposed ML model as *Offline-BiQCs*.

Remark. The approach of leveraging the proposed ML model to propel BiQCs can be readily extended to more complex scenarios. Specifically, the proposed ML model can enhance the optimization of multi-class quantum classifiers [118] and quantum regression models [120], significantly reducing the quantum resource overhead. Moreover, it can also improve quantum neural networks with diverse architectures, such as data re-uploading strategy [121] and quantum convolutional neural networks [122, 123].

Numerical results. To exhibit the effectiveness of our proposal for enhancing VQEs and QNNs, we employ it to estimate the ground state of the Transverse-field Ising (TFI) Hamiltonian and to classify a synthetic binary dataset, respectively. For both tasks, we consider the same quantum devices comprising $N = 3$ qubits with a linear qubit connectivity, and the identical ansatz, i.e.,

$$U(\boldsymbol{x}) = \prod_{l=1}^{3}\left(\mathrm{CNOT}_{2,3}\,\mathrm{CNOT}_{1,2}\bigotimes_{i=1}^{3}\mathrm{RY}(\boldsymbol{x}_{i+3(l-1)})\right), \tag{J8}$$

where $\mathrm{CNOT}_{a,b}$ denotes applying CNOT gate to a-th and b-th qubits. This ansatz contains in total $d = 9$ classical controls.

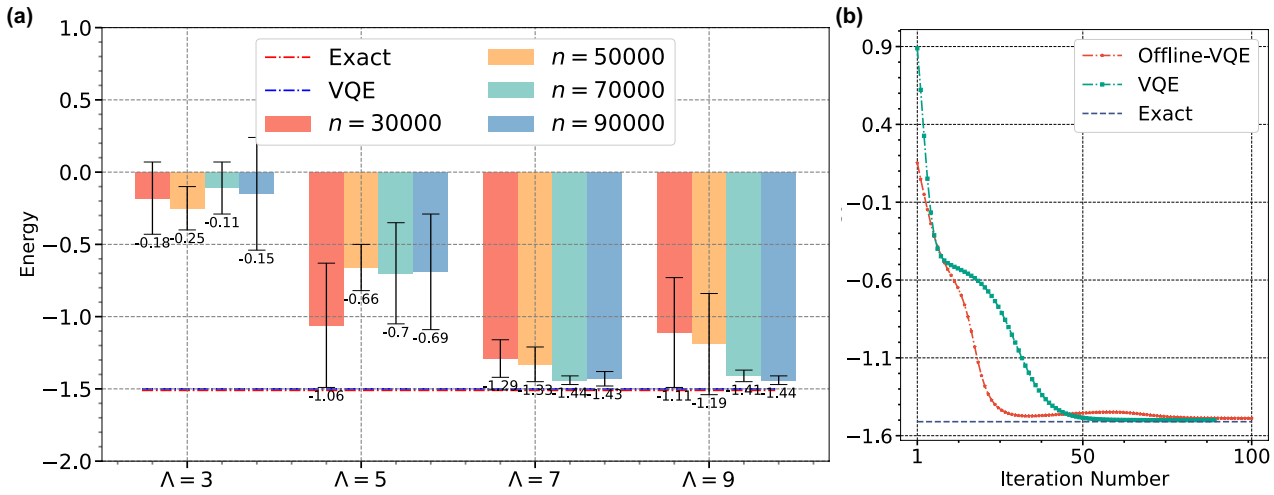

Figure J.9: **Simulation results for ground state energy estimation of TFI Hamiltonian**. (a) The estimated energy of Offline-VQE under different hyper-parameter settings. The label $n = a$ refers that the number of training data is a. The label 'Exact' denotes the accurate ground state energy of the employed TFI Hamiltonian. The blue dotted line highlights the estimated result of traditional VQE with infinite shots. The vertical bar reflects the standard error of Offline-VQE in each setting. (b) The optimization process of Top-1 Offline-VQE and conventional VQE.

Pre-training. At the stage of dataset construction, we collect in total 90000 training examples and the shot number of each training example is set as $T = 1000$. The random seed to generate such a dataset is set as 123. Once the data collection is completed, we form the kernel-based learning model to accomplish the following two downstream tasks on the classical side.

Downstream task I: ground state energy estimation of TFI model. The mathematical form of the exploited 3-qubit one-dimensional (1D) TFI Hamiltonian is $H_{\text{TFI}} = -0.1(Z_1Z_2 + Z_2Z_3) + 0.5(X_1 + X_2 + X_3)$. The ground state energy of H_{TFI} is $\mathfrak{E}^*_{\text{TFI}} = -1.51$Ha.

In numerical simulations, we adopt different hyper-parameter settings to evaluate the performance of Offline-VQE. In particular, we vary the number of training examples as $n \in \{30000, 50000, 70000, 90000\}$ and set the truncation frequency as $\Lambda \in \{3, 5, 7, 9\}$. The maximum iteration is set as 200. The initial parameters are uniformly sampled from the range $[-\pi, \pi]^9$. Each setting is repeated 5 times to obtain the statistical results. For comprehensive, we employ the conventional VQE with the same ansatz, the same initial parameters, the same optimizer, and infinite measurements as the benchmark.

The achieved results are depicted in Fig. J.9. Fig. J.9(a) demonstrates the performance of Offline-VQE under various settings. A key observation is that with the increased frequency truncation and number of training examples, the estimated ground state energy of Offline-VQE converges to the exact result. Namely, when $\Lambda \geq 7$ and $n \geq 70000$, the estimation error is less than 0.1Ha. Besides, Fig. J.9(b) compares the top-1 Offline-VQE (i.e., $\Lambda = 7$ and $n = 90000$) with conventional VQE. The achieved results exhibit a similar convergence rate during the optimization, where both of them converge to the near-optimal value after 50 iterations. These results validate the potential of Offline-VQE in advancing conventional VQEs.

Downstream task II: binary synthetic data classification. The construction rule of the binary synthetic dataset follows the study [115]. That is, we randomly and uniformly sample classical data $\boldsymbol{z}$ in the range $[-\pi, \pi]^3$ and then embed them into a 3-qubit quantum circuit $V(\boldsymbol{z}) = \text{CNOT}_{2,3}\,\text{CNOT}_{1,2}\bigotimes_{j=1}^{3}\text{RY}(\boldsymbol{z}_j)$. Then, we evolve the quantum state $V(\boldsymbol{z})\,|0\rangle^{\otimes 3}$ by the unitary $W(\boldsymbol{\theta}^*) = \prod_{l=1}^{2}\left(\text{CNOT}_{2,3}\,\text{CNOT}_{1,2}\bigotimes_{j=1}^{3}\text{RY}(\boldsymbol{\theta}^*_{j+3(l-1)})\right)$ with the fixed parameters $\boldsymbol{\theta}^*$ and measuring the evolved state under the observable $X \otimes \mathbb{I}_4$. For the input $\boldsymbol{z}^{(i)}$, its label is defined as

$$y^{(i)} = \text{sign}\left(\text{Tr}\left(W(\boldsymbol{\theta}^*)V(\boldsymbol{z}^{(i)})(|0\rangle\langle 0|)^{\otimes 3}V(\boldsymbol{z}^{(i)})^\dagger W(\boldsymbol{\theta}^*)^\dagger\,(X \otimes \mathbb{I}_4)\right)\right), \tag{J9}$$

where $\text{sign}(\cdot)$ denotes the sign function. We collect 500 positive samples and 500 negative samples and split them into the training set and the test set with a ratio of 0.2.

In numerical simulations, we adopt different hyper-parameter settings to evaluate the performance of Offline-BiQC. The varied settings are analogous to those employed in the offline-VQE, where the number of training examples is set as $n \in \{30000, 50000, 70000\}$ and the truncation frequency is set as $\Lambda \in \{3, 5, 7, 9\}$. The initial parameters are

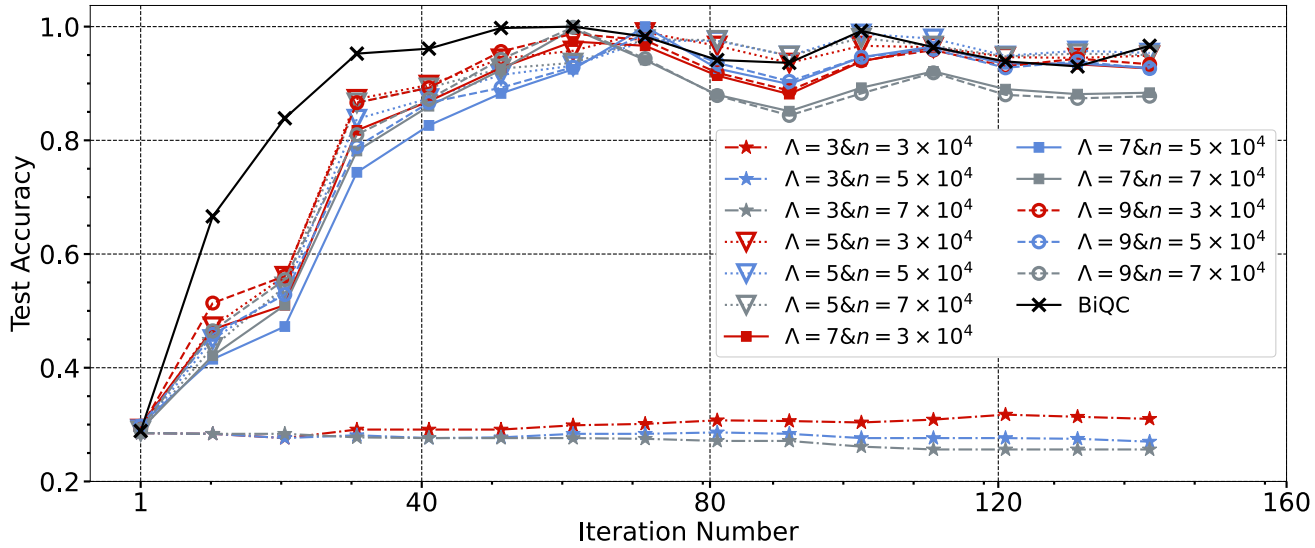

Figure J.10: **Simulation results of Offline-BiQC**. The figure depicts the test accuracy of Offline-BiQC under varied settings. The label '$\Lambda = a\&n = b$' denotes that the truncation value is a and the number of training examples used to form the classical representations is b. The label 'BiQC' refers to the conventional BiQC optimized in an online manner.

uniformly sampled from the range $[-\pi, \pi]^6$. The learning rate and the maximum iteration are set as 0.5 and 140, respectively. For comprehensive, we employ the conventional BiQC with the same encoding and training circuits, the same initial parameters, the same optimizer, and infinite measurements as the benchmark.

The simulation results are shown in Fig. J.10. When $\Lambda = 3$, the test accuracy of Offline-BiQC is lower than 40% no matter how the number of training examples n is. These results reflect the crucial role of the truncation value in warranting the performance of Offline-BiQC. In addition, Offline-BiQC attains a superior performance with an increased truncation value. Namely, when $\Lambda \geq 5$ and $n \geq 30000$, its performance is comparable with conventional BiQC, where both of them attain a test accuracy above 85%. These results validate the capability of Offline-BiQC in advancing the study of quantum neural networks by greatly reducing the quantum resources demand. Another phenomenon is that Offline-BiQC and BiQC encounter the oscillated test accuracy after 80 iterations. This is mainly caused by the employed large learning rate. Adopting a smaller learning rate, associated with the learning rate decay strategy, can mitigate this issue.