

Efficient Learning for Linear Properties of Bounded-Gate Quantum Circuits

Corresponding Author: Dr Yuxuan Du

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This work investigates the sample and computational complexity of classically learning the linear properties of bounded-gate quantum circuits. This is an interesting, alternate approach to the problem raised in Reference [30], where it was studied for quantum learners. The main conclusion shows that the class of bounded-gate quantum circuits can be efficiently learned in terms of the sample complexity, i.e, the number of input-output samples extracted from the quantum circuit, that scales linearly with the dimension d of the input; however, at the cost of exponential computational complexity. While this result is interesting, I think it's important to compare the conclusions with that of [30]. Additionally, more details on the exponential computational complexity is also required.

Prima facie, it seems like while [30] infers the unknown quantum state from its finite copies, the classical learner studied in this work aims to approximate the linear property of the unknown quantum state defined through an appropriate observable. Consequently, the sample complexity for classical learner depends on the properties of the observable (which makes sense), but in addition, also depends on the number of bounded gates G which is scaled by the dimension d of the data. It would be great if the authors can shed some light in the Gd scaling in Theorem 1, and compare it with similar scaling (G scales with 2^n) in [30]. I think this would help to situate the framework better in the context of [30]. Furthermore, the kernel-based algorithm that the authors propose later distinguishes the two stages outlined above: first estimate the unknown quantum state via shadow tomography, and then use a kernel-based learning model for the linear property of the quantum state. The dimensional dependence is then controlled via truncating the frequencies of the kernel considered. Having said this, the algorithm is interesting and novel.

The main conclusions are supported by experimental evidence, with detailed analysis. Following Theorem 2, is it possible to identify any transition region based on λ , that can illustrate the transition from polynomial to exponential dependence?

The methodology seems to be correct to the best of my understanding. Importantly, the applications of the approach, including in variational quantum algorithms, that can do offline inference delineating the need to access quantum computers further validates the approach.

Some other minor comments:

1. In Theorem 1, it seems like from equation (2), the number of qubits N equals the dimension of the data. Is this true?
2. Typo in the statement of Theorem 1: the range of x values

(Remarks on code availability)

I checked the GitHub link attached, but could not file the code there.

Reviewer #2

(Remarks to the Author)

The authors introduce a quantum learning problem with clear separations in computational and sample complexities.

Specifically, they train a classical model from "quantum data" by mixing classical shadows techniques with a new trigonometric kernel, and they show that only "few" samples are needed for the model to generalise. A quantum machine is still required for training, to avoid exponential computational complexity, but not necessary for inference.

The closest study, as the authors correctly point out, is Ref.[30], but their algorithm requires a quantum machine for testing new data, while the new technique presented in this paper allows for purely classical inference. This is the key difference. However, it is not clear to me what are the main use cases for this learning approach. The numerical simulations presented in the main text, although compatible with the theory (which requires advanced and sophisticated techniques), are limited to toy use cases. The lack of convincing use cases is the main limiting factor of this paper. If the authors can convincingly discuss "practical" use cases (as they promise in the abstract) I'd be happy to reconsider my recommendation. For instance, I find the idea of pre-training VQA (briefly discussed in page 4), quite fascinating.

Other points:

1) the abstract is quite hard to read and, in my opinion, not up to the standards of Nat.Comm. From a first read, the problem looks contrived.

2) From Eq. 3 to Thm.1 included, I was puzzled by the role of the observable O . For instance, in Theorem 1 the authors say that $\tilde{f}_T$ is an estimation of $f(x, O)$, but for which O ? That is clarified later in the paper when they introduce classical shadows, but at this introductory stage the setting was very confusing. Moreover, if my understanding is correct, the function class Eq.(3) should contain the fact that O can be different (with bounded norm). I guess the authors wanted to avoid unnecessary details here, but the end result is actually harder to understand and not intuitive.

3) Eq. A10 (SM). If my understanding is correct, the matrix D_2 should be renamed to D_{-1} for consistency with the index of the Fourier frequencies

4) In appendix C (SM), the authors assume that results can be approximated by a normal distribution. Clearly it follows from central limit theorem, but is it possible to check when this approximation is accurate? Does it imply that the sample size must be larger than what is expected from Thm 1?

(Remarks on code availability)

there is no code, the github repository is empty

Reviewer #3

(Remarks to the Author)

The authors study the learnability of quantum circuits as a function of the number of RZ gates, assuming that the observer has classical access to the state. They prove several results, on both the sample and computational complexity of the task. They provide a lower bound that the sample complexity must be at least linear in d , by providing a simple example of circuits that require this many samples to learn. They also provide an upper bound using a fairly standard counting approach. Finally, they show that, even though the sample complexity is efficient, the computational complexity may not be, by adapting an example from another recent work. To address the computational complexity, the authors introduce a kernel method based on the Fourier transform of the signal with respect to the circuit parameters. They show that when gradients of the circuit are small, the Fourier transform can be truncated to contain only low-weight frequency vectors, which enables efficient learning. They provide discussions and numerical evidence that their approach works well in several cases.

I enjoyed reading the authors' work and believe that it is a valuable and high-quality contribution to our knowledge on learning quantum circuits. At the same time, I feel that the results of the work are not *especially* surprising, and the methods are not sufficiently novel, to warrant publication in Nature Communications. To elaborate on the first point: The fact that the sample complexity grows linearly in d is fairly similar to existing work (Ref 30) on the sample complexity as a function of the number of two-qubit gates. It is nice to know how this extends when we separate the Clifford from non-Clifford gates, but not too surprising that it works out to a similar scaling. The computational complexity is a more interesting question. In this regard, I believe that the authors truncated frequency method may be useful, but it also bears an extremely close similarity to existing classical algorithm for the same class of circuits (e.g. the LOWESA algorithm or several more recent papers on Pauli path simulations). Thus, I view this contribution more as an adaptation of an existing approach to a new context, rather than a new method in itself. Moreover, a downside of this approach is that its success is linked to the fact the quantum circuit can be classically simulated, using the same truncated frequency expansion. This seems like it will limit the impact of the authors approach in applications.

As a small comment on presentation, I would have liked to see a bit more discussion on the proof ideas inside the main text. I understand the authors have many results to present in a short space, but it is nice to give people an idea of how the results are shown without needing to dig through the Appendix.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I would like to thank the authors for the detailed response to my comments. The response is very convincing, and I appreciate the depth of explanation provided. I believe the modifications made have significantly improved the understanding of the contributions of the work. I thus recommend acceptance of this work.

(Remarks on code availability)

The code is self-contained and pretty easy to follow.

Reviewer #2

(Remarks to the Author)

I thank the authors for their careful and detailed response to my questions and those of the other reviewers. My main criticism was the lack of 'practical' applications, and the authors have now performed new numerical experiments showing the advantage (in some cases) of pretraining VQE using their method. These preliminary results maybe appropriate at this stage, though more thorough experiments will be required, possibly in follow-up studies, to better assess the performance of this method. For instance, it would be interesting to compare, either separately or jointly, both the computational and measurement overhead of the two approaches. I appreciate this might take some effort and might be beyond the scope of the current manuscript. Since the authors have also thoroughly responded to my other criticisms, in my opinion the manuscript may be publishable in Nat Comm

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

I thank the authors for their detailed and thoughtful response. I have read the response carefully. I have been convinced that the technical innovations required (1) proving the sample complexity bounds for classical learners, and (2) using the low-frequency approximation for learning as opposed to classical simulation, are significant and that neither (1) nor (2) follow immediately from previous work. At the same time, I maintain my stance that these are fairly plausible and not-too-surprising results from an intuitive point of view. This is just my personal perspective, and I expect it to be different from reader to reader.

I will slightly revise my recommendation upwards after reading the response of the authors. However, I would still lean against publication in Nature Communications if it were solely up to me.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to reviewers for the manuscript “Efficient Learning for Linear Properties of Bounded-Gate Quantum Circuits”

We would like to express our sincere thanks to all three reviewers for their valuable time and constructive comments to improve the quality of our manuscript. In this response letter, we have addressed all of the comments raised by each reviewer in the order in which they were presented. All changes appeared in the revised manuscript are highlighted with the teal color.

Reviewer 1

Reviewer Point P 1.1 — This work investigates the sample and computational complexity of classically learning the linear properties of bounded-gate quantum circuits. This is an interesting, alternate approach to the problem raised in Reference [30], where it was studied for quantum learners. The main conclusion shows that the class of bounded-gate quantum circuits can be efficiently learned in terms of the sample complexity, i.e, the number of input-output samples extracted from the quantum circuit, that scales linearly with the dimension d of the input; however, at the cost of exponential computational complexity. While this result is interesting, I think it’s important to compare the conclusions with that of [30]. Additionally, more details on the exponential computational complexity is also required.

Reply: We thank the reviewer for the positive affirmation of our manuscript and appreciate the reviewer’s thoughtful feedback. Regarding the relation with Ref. [1] (Reference [30] in the original manuscript), at a high level, these two works focus on **different tasks** and are rooted on **distinct learning paradigms**. In this regard, they are **complementary rather than competitive**, which collectively **enrich quantum learning theory** and **deepen our understanding of the capabilities and limitations of learning bounded-gate quantum circuits**. For clarity, in the remainder of the response, we first elucidate the complementary nature of these two works, followed by providing additional details about the exponential computational complexity.

Complementation with Ref. [1]. For comprehensive, we summarize the theoretical results achieved by Ref. [1] and our work in Table R1. As summarized in the table, Ref. [1] investigates the computational complexities of **quantum learners** across three learning tasks:

Task (i) *learning bounded-gate unitaries*;

Task (ii) *learning bounded-gate quantum states*;

Task (iii) *learning $\mathbb{R}$ -valued classical functions with 1-Lipschitz continuity*.

Different from Ref. [1], we explore the computational complexities of **classical learners** for a new task:

Task (iv) *predicting the expectation values of bounded-gate quantum circuit for a given observable*.

The achieved theoretical results under *different learning tasks* (i.e., Tasks (i)-(iv)) and *different learning paradigms* (i.e., quantum learners versus classical learners) are complementary, providing the following **two crucial insights** in the context of the learnability of bounded-gate quantum circuits.

Table R1: **Summarization of theoretical results achieved by Ref. [1] and our work.** The terms ‘Q-learner’ and ‘C-learner’ refer to the quantum and classical learning paradigms, respectively. Besides, the terms ‘Samp’ and ‘Comp’ refer to the abbreviations of ‘sample complexity’ and ‘computational complexity’, respectively. C-function refers to the classical function with ν being the number of variables in the classical function (Task (iii)). The notations N and ϵ denote the number of qubits and the prediction error. In Ref. [1], the explored bounded-gate circuit is realized by $\hat{G}$ two-qubit gates, while in our work the explored bounded-gate circuit is composed of d RZ gates and $G - d$ Clifford gates. The notations B and T refer to the norm of the observable and the snapshots when collecting each training example, respectively. The markers $\checkmark$, $\times$, $?$ separately represent ‘yes’, ‘no’, and ‘unknown’.

	Learn unitary (Task (i))	Learn state (Task (ii))	Learn C-function (Task (iii))	Learn observable (Task (iv))
Q-learner	Theorems 4&6, Ref. [1] Samp: $\begin{cases} \tilde{O}(\hat{G} \min\{\frac{1}{\epsilon^2}, \frac{\sqrt{2^N}}{\epsilon}\}) \\ \Omega(\frac{\hat{G}}{\epsilon}) \end{cases}$ Comp: $e^{\Omega(\min\{\hat{G}, N\})}$ Algorithm: $\checkmark$	Theorems 1&2, Ref. [1] Samp: $\tilde{O}(\frac{\hat{G}}{\epsilon^2})$ Comp: $e^{\Omega(\min\{\hat{G}, N\})}$ Algorithm: $\checkmark$	Theorem 7, Ref [1] Samp: $\Omega(\frac{1}{\epsilon^\nu})$ Comp: $\tilde{\Omega}(\frac{1}{\epsilon^\nu})$ Algorithm: $\times$	$?$
C-learner	$?$	$?$	$\times$	Theorem 1, this work Samp: $\begin{cases} \tilde{O}(\frac{B^2 d + B^2 N G}{\epsilon}) \\ \tilde{\Omega}(\frac{(1-\epsilon)d}{\epsilon T}) \end{cases}$ Comp: $O(\exp(N, d))$ Algorithm: $\checkmark$ (Theorem 2)

Before moving on to discuss the two insights, let us emphasize **the necessity of exploring both classical and quantum learners**. Recall that one of the major differences between Ref. [1] and our work lies in the employed learning paradigms. To be concrete, Ref. [1] focuses on the **quantum learner**, which may implement the learning model on quantum computers (e.g., quantum neural networks) and require quantum resources during **both the training and inference stages**. In contrast, our work centers on the **classical learner**, which implements the learning model on classical computers, requiring quantum resources **only during the data collection stage**. This means that both the training and inference stages for classical learners are conducted entirely on classical hardware. A comprehensive analysis about the computational complexities of both quantum and classical learners can **guide various practical applications** and **indicate the utility of quantum computers**. On the one side, given the foreseeable scarcity of quantum resources, if classical learners can achieve **comparable performance** to quantum learners for a specific task, they offer undeniable benefits in terms of resource efficiency. On the other side, if quantum learners demonstrate **provable computational advantages** over classical learners for a given task, this would establish the practical utility of quantum computing for that particular application. For example, a recent study [2] provides a compelling example of how computational capabilities differ between classical and quantum learners, demonstrating that quantum learners can achieve exponential runtime speedups in predicting the properties of ground states.

We now detail the two insights by combining Ref. [1] and our results.

Insight I: Computational complexity analysis. Based on the achieved computational complexities presented in Table R1, we draw the following key observations in the context of learning bounded-gate quantum circuits.

- Observation I. Let us revisit the achieved **sample complexity** in Tasks (i), (ii), and (iv) summarized in Table R1. That is, despite the decreasing levels of information gain—from learning an unknown unitary, to an unknown state, and finally to unknown expectation values, we observe a consistent

scaling behavior for both classical and quantum learners.

In particular, Theorems 1 and 4 in Ref. [1] demonstrate that when learning a circuit or a pure state composed of $\hat{G}$ two-qubit gates, $\Theta(\hat{G})$ training examples are sufficient and necessary for a **quantum learner** to ensure a bounded prediction error. Similarly, Theorem 1 in our work establishes that when predicting the expectation values of a bounded-gate circuit composed of d RZ gates and $G - d$ Clifford gates, $\Theta(d)$ training examples are sufficient and necessary for a classical learner to achieve a bounded prediction error.

We would like to note that although Ref. [1] and our work use different descriptions of bounded-gate circuits, **the results achieved in our study can be directly extended to their framework**. This is because our description of bounded-gate circuits can be efficiently transformed into their setting with only a **constant overhead**. More precisely, **a bounded-gate quantum circuit composed of $\hat{G}$ two-qubit gates can be implemented using at most $d = 24\hat{G}$ RZ gates and $G - d = 35\hat{G}$ Clifford gates to implement this circuit**. An intuition is provided in Fig. R1. Ref. [3] has proved that three CNOT gates, associated with eight single-qubit gates, are necessary and sufficient in order to implement an arbitrary unitary transformation of two qubits. Meanwhile, any single qubit gate can be realized by three RZ gates and four single-qubit Clifford gates [4]. Taken together, $3 \times 8 = 24$ RZ gates and $4 \times 8 + 3 = 35$ Clifford gates can implement an arbitrary unitary transformation of two qubits. Since there are $\hat{G}$ two-qubit gates, $d = 24\hat{G}$ RZ gates and $G - d = 35\hat{G}$ Clifford gates are sufficient to implement this circuit.

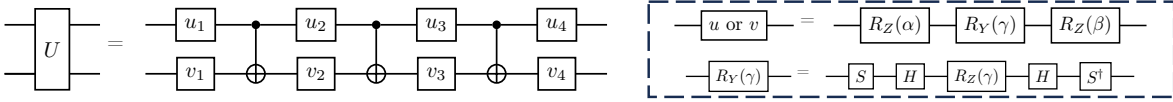


Figure R1: **Decomposition of an arbitrary two-qubit gate into RZ and Clifford gates.** The dashed box highlights the implementation of the single-qubit gates u_* and v_* with $* \in \{1, 2, 3, 4\}$ by RZ gates and single-qubit Clifford gates.

According to the result of Theorem 1 in our work, when a classical learner is applied to predict the expectation value of a bounded-gate quantum circuit composed of $\hat{G}$ two-qubit gates, $\tilde{O}((B^2\hat{G} + B^2N\hat{G})/\epsilon)$ and $\tilde{\Omega}((1 - \epsilon)\hat{G}/(\epsilon T))$ training examples are sufficient and necessary to achieve a bounded prediction error. **This result is consistent with the findings for Tasks (i) and (ii) achieved in Ref. [1], both of which display a linear-in- $\hat{G}$ sample complexity.**

The linear scaling behavior with respect to the number of tunable gates across different learners (i.e., quantum versus classical) and tasks (i.e., Tasks (i), (ii), and (iv)) suggests **a consistent linear-in- d (or linear-in- $\hat{G}$) sample complexity for classical learners in Tasks (i)&(ii) and for quantum learners in Task (iv)**. This conjecture is supported by two key facts. First, covering and packing nets, as well as (quantum) hypothesis selection, are robust tools for analyzing the sample complexity of learning bounded-gate quantum circuits, often providing tight bounds [5, 6, 7, 8, 9]. Second, the cardinality of these nets is intrinsically dependent on the number of gates (d , G , or $\hat{G}$), which leads to similar results across tasks, despite variations in the intended information gain.

- **Observation II.** We next turn to rethink the achieved **runtime complexities** in Tasks (i), (ii), and (iv) summarized in Table R1. An immediate observation is that, for all tasks, there are no efficient *quantum learners* for learning bounded-gate unitaries or quantum states [1, Theorems 2&6]. In addition, there are no efficient *classical learners* capable of accurately predicting the mean-value

space of arbitrary bounded-gate quantum circuits (Theorem 1 in our manuscript). These results are **complementary** and **offer valuable insights** into the learnability of bounded-gate quantum circuits.

First, the inherent limitations of quantum learners in learning bounded-gate unitaries or states strongly suggest that **there are no efficient classical learners capable of accomplishing Tasks (i) and (ii)**. This is because from an information-theoretic perspective, classical learners cannot extract much more information than quantum learners. Consequently, classical learners are unlikely to outperform quantum learners in these tasks.

Second, while neither quantum nor classical learners can efficiently learn arbitrary bounded-gate quantum circuits, it is natural to ask **whether specific conditions on the explored bounded-gate circuits could lead to the computational separations**, i.e., quantum learners can provide efficient solutions, whereas classical learners would still face computational hardness. Answering this question can enhance our understanding of the learnability of bounded-gate quantum circuits and help identify the practical utility of quantum learners. Ref. [2] offers initial insights into this question. Specifically, by adapting their proof analysis, a **computational separation can be reached** between classical and quantum learners when predicting **the mean-value space of BQP-type circuits**.

Insight II: Learning algorithm design. The computational hardness of learning bounded-gate quantum circuits in Tasks (i), (ii) and (iv) highlights the need to develop advanced learning algorithms with **provable computational efficiency under practical and realistic conditions**. Both Ref. [1] and our work make initial strides in this direction. Below, we summarize the advancements of each method and outline potential directions for future research.

- For Tasks (i) and (ii), Ref. [1] proposes **efficient learning algorithms for relatively restrictive scenarios** where the number of two-qubit gates $\hat{G}$ **scales logarithmically** with the number of qubits N , i.e., $\hat{G} \sim \mathcal{O}(\log(N))$. Conceptually, at the inference stage, given any new quantum circuit composed of $\mathcal{O}(\log(N))$ two-qubit gates, the proposed algorithm first identifies the non-trivial $\mathcal{O}(\log(N))$ qubits on which two-qubit gates are applied and then employs Pauli-based tomography on these qubits. The achieved result naturally raises two future research directions: (1) Are there efficient learning algorithms beyond the logarithmic scaling scenario? (2) Can classical learners achieve performance comparable to quantum learners in Tasks (i) and (ii)?
- For Task (iv), we devise an efficient **classical-shadow predictor** based on the kernel methods. This model ensures accurate predictions of the mean-value space for *any local observables* and *new rotation angles* of RZ gates, once the gradient norms are well-bounded. While a direct comparison may not be entirely fair, it is worth highlighting that **the conditions required for provable efficiency in our approach are much less restrictive than those in Ref. [1]**. Specifically, our protocol can efficiently learn a broad class of quantum circuits composed of a polynomial number of gates, as opposed to being limited to logarithmic scenarios. The achieved results, combined with the unsolved issues in Table R1, lead to two interesting questions: (1) Are there alternative efficient classical learning algorithms? (2) Is there a computational separation between quantum and classical learning algorithms for Task (iv)?

Remark. The two key insights presented above, along with the corresponding observations, highlight the **complementary roles** of Ref. [1] and our work in advancing both theoretical understanding and

algorithmic development related to the learnability of bounded-gate circuits. From the **theoretical perspective**, the computational complexities achieved in these works offer **valuable implications** for bounded-gate quantum circuits outlined in Table R1, shedding light on the computational challenges and possibilities in different scenarios. From the **algorithmic perspective**, the developed learning algorithms demonstrate distinct strengths: Ref. [1] focuses on **quantum learners**, establishing foundational methods for understanding quantum circuit with **logarithmic gate scaling**, while our work pioneers **classical learning approaches** with provable efficiency under more practical conditions. Together, these complementary strategies expand the landscape of learning algorithms for bounded-gate quantum circuits, paving the way for deeper exploration of their utility and limitations in quantum computing.

More details regarding the exponential computational complexity. We have followed the reviewer’s suggestion to provide more details regarding the exponential computational complexity. Recall that the computational complexity results presented in Theorem 1 are based on the findings of Ref. [2, Lemma 1, version 1]. During the review process, their manuscript was updated with a new theoretical result that achieves the same conclusion under a **more widely studied complexity-theoretic assumption** [2, Theorem 2, version 2]. To ensure clarity and comprehensiveness, we have incorporated these updates into the revised manuscript. Specifically, in the updated version, we include a proof sketch of Theorem 1 (including the exponential runtime complexity) in the Methods section and provide additional proof details in Supplementary Material E.

For completeness, we present the appended proof sketch for the exponential computational complexity. **The major modification compared to the initial version is the adoption of the assumption $\text{BQP} \not\subseteq \text{P/poly}$ to demonstrate the computational hardness.** Unlike the HeurP/poly complexity classes used in the original version, BQP is a more extensively studied complexity class for decision problems, offering a more intuitive and accessible framework for understanding the computational hardness.

Proof sketch. Recall Eq. (3) of our manuscript. That is, the explored concept class of the mean-value space of the bounded-gate circuits takes the form as

$$\mathcal{F} = \left\{ f(\mathbf{x}, O) = \text{Tr}(\rho(\mathbf{x})O) \mid U \in \text{Arc}(U, d, G) \right\}, \quad (1)$$

where $\mathbf{x} \in [-\pi, \pi]^d$, $\rho(\mathbf{x})$ is the quantum state prepared by the bounded-gate circuit $U(\mathbf{x})$, $\text{Arc}(U, d, G)$ refers to the set of all possible gate arrangements consisting of d RZ gates and $G - d$ CI gates, and O is an unknown local observable with the bounded norm. Intuitively, the **computational complexity** of the classical learner is characterized by **the required runtime cost** to identify a suitable hypothesis $h(\cdot)$ that achieves a low **average prediction error** relative to the target concept $f(\cdot)$ for $\forall f \in \mathcal{F}$, i.e.,

$$\mathbb{E}_{\mathbf{x}} |h(\mathbf{x}, O) - f(\mathbf{x}, O)|^2 \leq \epsilon. \quad (2)$$

To demonstrate the computational hardness of learning the mean-value space of bounded-gate circuits in Eq. (1), we leverage results from the BQP complexity class. Specifically, we construct a tailored concept class $\mathcal{F}$ that incorporates a concept f^* realized by the BQP circuit. Consequently, **if an efficient classical learner were to exist for this task, it would imply that any BQP language could be decided in P/poly—a scenario that is widely considered implausible.**

The implementation of the target concept f^* follows a similar manner with Ref. [2]. Let L be an arbitrary BQP language. Since $L \in \text{BQP}$, there exists a corresponding quantum circuit U^* which decides input bitstrings $\mathbf{x} \in \{0, 1\}^N$ correctly on average with respect to the data distribution. In particular,

define $O' = Z_1 \otimes \mathbb{I}_{2^{N-1}}$ as the observable, which measures the Pauli-Z operator on the first qubit. The target concept f^* is defined as

$$f^*(x, O') = \text{Tr}(O' \rho^*(x)), \quad \text{with} \quad \rho = U^* |x\rangle \langle x| U^{*\dagger}. \quad (3)$$

The measurement result $f^*(x, O')$, whether positive or negative, determines the decision regarding whether the input bitstrings x belong to the language L or not. As such, we **connect the problem of learning mean-value space of bounded-gate circuits with complexity theory**.

The remaining step to **complete the proof involves demonstrating that the quantum state $\rho^*(x)$ can be efficiently prepared using RZ and Clifford gates**, as the primary technical contribution in this context. According to the definition of BQP, the circuit U^* consists of at most a polynomial number of quantum gates with respect to the qubit count N . Given the universality of RZ and Clifford gates, any BQP circuit U^* and any bitstring x can be efficiently represented using RZ and Clifford gates with at most a constant linear overhead. Taken together, this completes the proof.

The revisions to our manuscript are summarized as follows.

Main text

1. (Lines 207-209, Page 3). We append the words “As detailed in SM J, these two works are complementary, which collectively reveal the capabilities and limitations of learning bounded-gate quantum circuits.”
2. (Lines 497-554, Page 7). We append the proof sketch regarding the exponential computational complexity in Methods, titled “Proof sketch of Theorem 1”.

Supplementary Material

1. (Lines 738-838, Pages 27-30). We append the subsection “SM J.1 complementation with Ref. [1]” to elucidate the complementary role between Ref. [1] and our work.
2. (Lines 465-525, Pages 15-16). We update and expand the proof related to the exponential runtime complexity in SM E.

Reviewer Point P 1.2 — Prima facie, it seems like while [30] infers the unknown quantum state from its finite copies, the classical learner studied in this work aims to approximate the linear property of the unknown quantum state defined through an appropriate observable. Consequently, the sample complexity for classical learner depends on the properties of the observable (which makes sense), but in addition, also depends on the number of bounded gates G which is scaled by the dimension d of the data. It would be great if the authors can shed some light in the Gd scaling in Theorem 1, and compare it with similar scaling (G scales with 2^n) in [30]. I think this would help to situate the framework better in the context of [30]. Furthermore, the kernel-based algorithm that the authors propose later distinguishes the two stages outlined above: first estimate the unknown quantum state via shadow tomography, and then use a kernel-based learning model for the linear property of the quantum state. The dimensional dependence is then controlled via truncating the frequencies of the kernel considered. Having said this, the algorithm is interesting and novel.

Reply: We appreciate the reviewer’s positive remark that “the algorithm is interesting and novel”. We agree with the reviewer’s suggestion to clarify the scaling behavior of the achieved sample complexity

and to compare our results with those of Ref. [1], thereby better situating our work in its broader context. For clarity, we address the reviewer’s comments iteratively in the following, providing detailed explanations for each point.

Comments: “Prima facie, ..., d of the data”. We thank the reviewer for the thoughtful observation regarding the relationship between our work and Ref. [1].

Comments: “It would be great if .. Gd scaling in Theorem 1,”. To provide an intuitive explanation for why the sample complexity of learning bounded-gate quantum circuits scales linearly with the number of RZ gates d and the total number of quantum gates G , it is essential to highlight a key principle in statistical learning theory [10]. Informally, under the PAC learning framework, the required sample complexity to infer a hypothesis $h(\cdot)$ that achieves a low prediction error with the target concept $f(\cdot)$ **logarithmically depends on the cardinality of the concept class $\mathcal{F}$** .

According to Eq. (1), the explored concept class of the mean-value space of the bounded-gate circuits takes the form as $\mathcal{F} = \{f(x, O) = \text{Tr}(\rho(x)O) | U \in \text{Arc}(U, d, G)\}$. As explained in the reply of Point P 1.1 (Observation I, Insight I), a commonly used tool to quantify the cardinality of $\mathcal{F}$ is the concept of covering and packing nets, broadly used in both classical and quantum machine learning [5, 6, 7, 8, 9]. In this regard, the problem is reduced to analyzing the cardinality of $\mathcal{F}$ through the packing net.

Following the definition of packing net (Definition 1 in Supplementary Material), let $(\mathcal{U}, \mathfrak{d})$ be a metric space. The subset $\mathcal{V}$ 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, the packing net captures the complexity of a concept class by discretizing it into small “balls” of a specified resolution. **A richer concept class requires more such balls to approximate the space effectively, which directly correlates to a higher sample complexity.** In our case, the richness of the concept class $\mathcal{F}$ is determined by the number of RZ gates d and the total number of quantum gates G , which explains the scaling behavior of the sample complexity.

- Upper bound $\tilde{O}(\frac{B^2 d + B^2 N G}{\epsilon})$. The linear dependence on G and d in the upper bound of the sample complexity stems from the quantification of the packing net of the concept class $\mathcal{F}$. Particularly, the cardinality of $\mathcal{F}$ is determined by all possible arrangements of d RZ gates and $G - d$ Clifford gates across N -qubit wires, which is inherently a *combinatorial problem*. Through analysis, we obtain that the upper bound of the packing number of $\mathcal{F}$, i.e., $\mathcal{M}(\mathcal{F}, \epsilon, |\cdot|) \leq \binom{G}{d} \cdot N^d \cdot 3^{G-d} \cdot \left(\frac{N}{2}\right)^{G-d}$. Then, according to the result of statistical learning theory, the required sample complexity logarithmically depends on $\mathcal{M}(\mathcal{F}, \epsilon, |\cdot|)$, leading to the linear scaling with d and G .
- Lower bound $\tilde{\Omega}(\frac{(1-\epsilon)d}{\epsilon T})$. Referencing the analysis in Supplementary Material C, we design a specific learning task (Eq. (C6) in the original version) and employ Fano’s lemma (Eq. (C15) in the original version) to establish the lower bound of the sample complexity, i.e., $\Pr[\mathbf{a}^* \neq \bar{\mathbf{a}}] \geq 1 - \frac{I(\mathbf{a}^*; \bar{\mathbf{a}}) + \log 2}{\log |\hat{\mathcal{F}}|}$.

Here $\hat{\mathcal{F}}$ denotes the concept class corresponding to the designed learning task, which involves a two-qubit circuit with d RZ gates and $G - d$ Clifford gates, $\mathbf{a}^*$ corresponds to the target concept, and $\bar{\mathbf{a}}$ corresponds to the inferred hypothesis. To achieve the lower bound, we independently quantify the upper bound of the mutual information $I(\mathbf{a}^*; \bar{\mathbf{a}})$ and the cardinality of the concept class $|\hat{\mathcal{F}}|$. The dependence on d arises from **the derivation of the cardinality of the concept class**, as $\hat{\mathcal{F}} = 2^d$.

Remark. To improve the readability of our manuscript, we have consolidated the above explanations into a proof sketch of Theorem 1, which is now included in the Methods section of the updated version.

Comments: “and compare it with similar scaling (G scales with 2^N) in Ref. [30]”. While both our work and Ref. [1] exhibit similar scaling behavior, we would like to clarify that these studies address **different tasks under distinct learning paradigms**, as summarized in Table R1. Consequently, the results of these two works are **complementary**, collectively advancing our understanding of the learnability of bounded-gate quantum circuits. For a detailed explanation, please refer to our reply to Point P 1.1.

In addition, the similar scaling behavior between our work and Ref. [1] arises from the use of covering/packing nets to analyze the sample complexity of learning bounded-gate quantum circuits (explained in Observation I, Insight I of the reply to Point P 1.1). Note that the cardinality of these nets is inherently determined by the number of quantum gates (d , G or $\hat{G}$), resulting in a linear-in- d (or linear-in- $\hat{G}$) sample complexity across tasks. Consequently, similar to the analysis in Ref. [1, Figure 1], predicting **the mean-value space with a classical learner** remains **sample-efficient** as long as the number of quantum gates scales at most polynomially with the number of qubits, i.e., $G \sim \mathcal{O}(\text{poly}(N))$.

Comments: “I think this would help to situate the framework better in the context of [30].” We appreciate the reviewer’s suggestion and have updated our manuscript to better align the language and context between our work and Ref. [1]. For completeness and self-consistency, we provide a brief overview here. Please refer to the reply of Point P 1.1 for a detailed explanation.

Our primary modifications include emphasizing the complementary relationship between these two works and their respective implications for quantum learning theory, as indicated in Table R1. Additionally, we have analytically derived a method to translate the bounded-gate quantum circuit composed of $\hat{G}$ two-qubit gates, as explored in Ref. [1], into our framework (RZ gates and Clifford gates). Importantly, as shown in Figure R1, we demonstrate that **this translation incurs only a constant overhead**, highlighting that **the results of both works are fully compatible**.

Comments: “Furthermore, the kernel-based algorithm ... interesting and novel”. We appreciate the reviewer’s positive feedback on the proposed kernel-based learning models for predicting the linear properties of quantum states. In the updated version, we have expanded our efforts to showcase the utility of these kernel-based methods in facilitating the pre-training of VQAs, as suggested by Reviewer #2. As detailed in the reply of Point P 2.2, we **conducted systematic simulations demonstrating the capability of pre-training multiple ansatzes up to 50 qubits**. The simulation results, which include applications to TFIM and H_5 molecules, validate the superiority of our proposed methods compared to other learning approaches such as multi-layer perceptrons and random Fourier feature regression models. We hope these additional results further enhance the quality of our manuscript and highlight the broad applicability of the proposed protocols.

The revisions to our manuscript are summarized as follows.

Main text

1. (Lines 497-554, Page 7). We have included the proof sketch of Theorem 1 in the Methods section, where we provide an explanation of the scaling behavior of G and d in both the upper and lower bounds.
2. (Lines 389-423, Page 6). We have included new simulation results demonstrating the use of the proposed model to pretrain a 50-qubit VQE for estimating the ground-state energy of the Transverse-Field Ising Model, as shown in Fig. 3(b), titled “[Pretraining performance of the proposed ML model on a 50-qubit TFIM](#)”.

Supplementary Material

1. (Lines 738-838, Pages 27-30). We have added a new Section J.1 to compare the similar scaling discussed in Ref. [30], incorporating all the explanations provided in our response.
2. (Lines 1129-1333, Pages 39-47). Additional numerical simulations showcasing the use of the proposed ML model for pretraining VQAs, along with comparisons to other learning models, have been included in SM K.4.

Reviewer Point P 1.3 — The main conclusions are supported by experimental evidence, with detailed analysis. Following Theorem 2, is it possible to identify any transition region based on Λ , that can illustrate the transition from polynomial to exponential dependence?

Reply: Thank you for the insightful comment. The required Λ in the proposed ML model is **task-dependent, influenced by factors such as the desired accuracy ϵ , the employed ansatz $U(\mathbf{x})$, and the target observable O** . To illustrate this claim, in what follows, we first provide a brief overview of how the computational complexity of the proposed learning model depends on Λ followed by a systematic numerical analysis to demonstrate the problem-dependent behavior of the required Λ .

The relation between the computational cost and Λ . Let us first recall the mathematical form of the proposed ML model, i.e.,

$$h_s(\mathbf{x}, O) = \frac{1}{n} \sum_{i=1}^n \kappa_\Lambda(\mathbf{x}, \mathbf{x}^{(i)}) g(\mathbf{x}^{(i)}, O), \quad (4)$$

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

$$\kappa_\Lambda(\mathbf{x}, \mathbf{x}^{(i)}) = \sum_{\boldsymbol{\omega}, \|\boldsymbol{\omega}\|_0 \leq \Lambda} 2^{\|\boldsymbol{\omega}\|_0} \Phi_{\boldsymbol{\omega}}(\mathbf{x}) \Phi_{\boldsymbol{\omega}}(\mathbf{x}^{(i)}) \in \mathbb{R},$$

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

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

As shown in Theorem 2 of the main text, the quantity Λ impacts the **computational complexity** of the proposed ML model $h_s(\mathbf{x}, O)$ by determining both **the required number of training examples**

n and **the dimensionality of the trigonometric monomial feature basis** (denoted by p). More concisely, denote the *truncated frequency set* as

$$\mathfrak{C}(\Lambda) = \{\omega | \omega \in \{0, \pm 1\}^d, \text{ s.t. } \|\omega\|_0 \leq \Lambda\}, \quad (5)$$

the proposed ML model achieves ϵ -prediction error when the number of training examples and the dimensionality of the trigonometric monomial feature basis (i.e., cardinality of the truncated frequency set) satisfy

$$n = |\mathfrak{C}(\Lambda)| 2B^2 9^K \epsilon^{-1} \log(2|\mathfrak{C}(\Lambda)|/\delta) \quad \text{and} \quad p = |\mathfrak{C}(\Lambda)|, \quad \text{respectively.} \quad (6)$$

Furthermore, we clarify the relation between the quantity Λ , the average gradient norm C , and the desired prediction error ϵ . That is, when the desired error is ϵ and the average gradient norm is upper bounded by $\mathbb{E}_{\mathbf{x} \sim [-\pi, \pi]^d} \|\nabla_{\mathbf{x}} \text{Tr}(\rho(\mathbf{x})O)\|_2^2 \leq C$, it is sufficient to set

$$\Lambda = 4C/\epsilon \Leftrightarrow \mathfrak{C}(4C/\epsilon). \quad (7)$$

In this context, understanding the scaling behavior of Λ , is equivalent to analyzing the scaling of the cardinality of the truncated frequency set $|\mathfrak{C}(4C/\epsilon)|$. Given that the upper bound of the gradient norm C and the desired ϵ vary depending on the *quantum circuits, the observables, and tasks*, we argue that the required Λ in the proposed ML model is inherently **task-dependent**.

In what follows, we conduct a systematic analysis to heuristically understand the behavior of Λ , or equivalently $|\mathfrak{C}(4C/\epsilon)|$.

The scaling behavior of $\mathfrak{C}(\Lambda)$. We begin by examining how the cardinality of the truncated frequency set $\mathfrak{C}(\Lambda)$ in Eq. (5) scales with the input dimensions d (i.e., the number of RZ gates) and the truncation thresholds Λ . Mathematically, we have

$$p \equiv |\mathfrak{C}(\Lambda)| = \sum_{k=0}^{\Lambda} \binom{d}{k} 2^k. \quad (8)$$

The scaling behavior is visualized in Fig. R2(a), where p exhibits *two distinct phases*: in Phase I, for small truncation values Λ , p scales exponentially with both Λ and the input dimension d ; in Phase II, as Λ increases further, the growth of p begins to saturate and eventually converges to a fixed value. Notably, such a scaling imposes a substantial computational burden for large d and Λ . As shown in the plot, when both d and Λ are large, the computational overhead becomes unaffordable, e.g., when $d = 50$ and $\Lambda = 10$, p is above 10^{12} .

The relation between Λ and the gradient norm C . The above analysis indicates that the practical efficiency of the proposed ML model lies in the **small- Λ regime**, especially for the large d . According to Eq. (7), this is equivalent to requesting a small value about the ratio $4C/\epsilon$. In this regard, the problem reduces to determining **how the average gradient norm C scales in practice**.

We perform the following numerical simulations to investigate the scaling behavior of the average gradient norm with $C = \mathbb{E}_{\mathbf{x} \sim [-\pi, \pi]^d} \|\nabla_{\mathbf{x}} \text{Tr}(\rho(\mathbf{x})O)\|_2^2$. Specifically, the class of states $\{\rho(\mathbf{x})\}$ is prepared by applying an N -qubit hardware-efficient ansatz $U(\mathbf{x})$ to an N -qubit product state $|+\rangle^{\otimes N}$ with $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$, i.e.,

$$\left\{ \rho(\mathbf{x}) = U(\mathbf{x})(|+\rangle \langle +|)^{\otimes N} U(\mathbf{x})^\dagger | \mathbf{x} \in [-\pi, \pi]^d \right\}$$

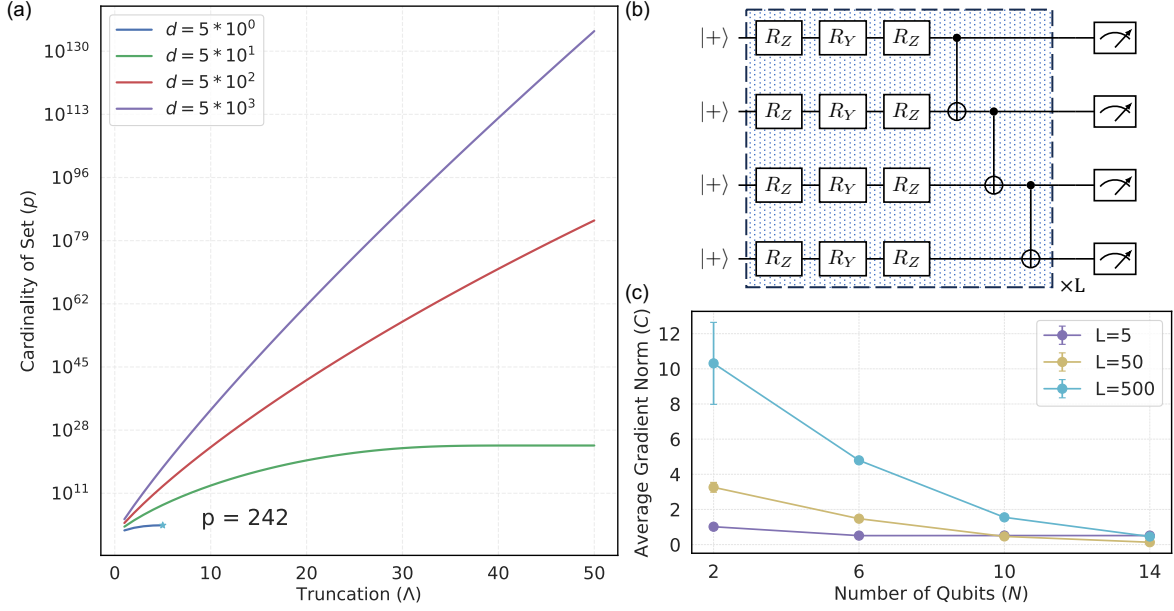


Figure R2: **The problem-dependent behavior of the truncation threshold Λ .** (a) SCALING BEHAVIOR OF p . The x-axis and y-axis represent the truncation thresholds and the cardinality of the truncation set $p \equiv |\mathfrak{C}(\Lambda)|$, respectively. The symbol ‘ $\star$ ’ marks that the cardinality of the truncation set is $p = 242$ when $d = 5$ and $\Lambda = 5$. (b) QUANTUM CIRCUIT FOR AVERAGE GRADIENT NORM COMPUTATION. The panel illustrates the quantum circuits used for computing the average gradient norm C when the number of qubits is $N = 4$. The dashed box highlights the implementation of hardware-efficient ansatz and the notation ‘ L ’ denotes the number of layers. (c) SCALING BEHAVIOR OF C . The simulation results of the average gradient norm C with varied qubit count N and layer number L . The vertical bar refers to the variance of the gradient norm.

and

$$U(\mathbf{x}) = \prod_{l=1}^L \left(\bigotimes_{j=1}^N RZ(\mathbf{x}_{j*(l-1)}) RY(\mathbf{x}_{j*(l-1)+1}) RZ(\mathbf{x}_{j*(l-1)+2}) \right) u_e, \quad (9)$$

where L denotes the number of layers, u_e refers to the entanglement layer consisting of N CNOT gates applied sequentially to adjacent qubits, and the dimension of the input yields

$$d = 3 * N * L.$$

The observable interacted with $\rho(\mathbf{x})$ is fixed to be $O = (X_1 + Y_1 + Z_1)/3$, where X_1 (or Y_1, Z_1) refers to apply the Pauli operator X (or Y, Z) to the first qubit. An illustration of the employed quantum circuit is depicted in Fig. R2(b).

The hyper-parameter settings are as follows. The number of qubits N varies from 2 to 14, and for each specified qubit count, the number of layers L is varied from 5 to 500. For each configuration, we uniformly and randomly sample the input 500 times to estimate the average gradient norm C and its variance.

The scaling behavior of the average gradient norm is illustrated in Fig. R2(c). An important observation is that, for all layer numbers L , **the average gradient norm C converges to a small value with the negligible variance as the qubit count N increases.** When $N = 14$, the value of C with

$L=50$ is around 0.13 ± 10^{-6} . Although the concrete value of C depends on the employed ansatz $U(x)$ and the qubit count N , the obtained simulation results support a reasonable conjecture that $C \ll 1$ as the number of qubits increases.

The above findings validate the potential of the proposed ML model to predict interested properties of large-qubit circuits. According to the relation $\Lambda = 4C/\epsilon$ indicated in Eq. (7) and the scaling behavior of Λ exhibited in Fig. R2(a), we conclude that **the primary factor influencing the applicability of the proposed ML model is the error threshold ϵ** . That is, when the given task requires high-precision predictions such that $\Lambda = 4C/\epsilon$ yields a very large constant, the proposed ML model will encounter computational issues; otherwise, the proposed ML model can be efficient.

Remark. In our response to Point P 2.2, we have conducted systematic simulations using the proposed ML model to pre-train VQAs with up to $N = 50$ qubits under varied ansatz settings. We hope these newly added simulation results further demonstrate the utility of the proposed learning model and clarify that **a small value of Λ typically ensures good prediction performance**, addressing the reviewer’s concerns.

Main text

1. (Lines 275-276, Page 4) We append the words “See SM K for heuristic evidence.”

Supplementary Material

1. (Lines 905-938, Pages 31-33) We append a new subsection titled “SM K.1 Heuristic evidence about the truncation value Λ ”, which covers all heuristic evidence explained in this reply, to highlight that a small value of Λ typically ensures good prediction performance.

Reviewer Point P 1.4 — The methodology seems to be correct to the best of my understanding. Importantly, the applications of the approach, including in variational quantum algorithms, that can do offline inference delineating the need to access quantum computers further validates the approach.

Reply: We thank the reviewer for the positive affirmation of our proposal. To better emphasize the **practical applications** enabled by the proposed algorithm, the revised manuscript includes additional simulation results showcasing how our model facilitates **the pre-training of VQAs** across various tasks, scaling up to 50 qubits. For a detailed explanation of these results, please refer to our response to Point P 2.2. We hope these newly added contents further underscore the practical significance of our work.

Some other minor comments

Reviewer Point P 1.5 — In Theorem 1, it seems like from equation (2), the number of qubits N equals the dimension of the data. Is this true?

Reply: We thank the reviewer for the comment. In fact, **the dimension of the data corresponds to the number of RZ gates** (denoted by d), not the number of qubits N . Thus, the two quantities, i.e., the dimension of the data and the number of qubits, are generally **independent**, except in specific cases such that an N -qubit circuit contains exactly $d = N$ RZ gates.

To clarify this point and address the reviewer’s concern, we have updated the relevant text in the manuscript as follows.

Main text

1. (Lines 123-125, Page 2) an N -qubit $U(x)$ can always be expressed as ... where d is the number of RZ gates, and u_e refers to the unitary composed of CI gates and the identity gate $\mathbb{I}_2$ with $CI = \{H, S, CNOT\}$.

Reviewer Point P 1.6 — Typo in the statement of Theorem 1: the range of x values.

Reply: We thank the reviewer for bringing this issue to our attention. We have corrected the typo in the statement of Theorem 1 as follows.

Main text

1. (Line 177, Page 3) ..., a dataset containing n training examples with $x^{(i)} \in [-\pi, \pi]^d$.

Reviewer Point P 1.7 — I checked the GitHub link attached, but could not file the code there.

Reply: We have followed the reviewer’s suggestion to make all codes related to this work open source. The source code for all numerical simulations is now available at [Github Repository](#).

Reviewer 2

Reviewer Point P 2.1 — The authors introduce a quantum learning problem with clear separations in computational and sample complexities. Specifically, they train a classical model from “quantum data” by mixing classical shadows techniques with a new trigonometric kernel, and they show that only “few” samples are needed for the model to generalise. A quantum machine is still required for training, to avoid exponential computational complexity, but not necessary for inference.

Reply: We sincerely appreciate the reviewer’s effort in reviewing our work and providing a clear and accurate summary of our contributions.

Reviewer Point P 2.2 — The closest study, as the authors correctly point out, is Ref.[30], but their algorithm requires a quantum machine for testing new data, while the new technique presented in this paper allows for purely classical inference. This is the key difference. However, it is not clear to me what are the main use cases for this learning approach. The numerical simulations presented in the main text, although compatible with the theory (which requires advanced and sophisticated techniques), are limited to toy use cases. The lack of convincing use cases is the main limiting factor of this paper. If the authors can convincingly discuss “practical” use cases (as they promise in the abstract) I’d be happy to reconsider my recommendation. For instance, I find the idea of pre-training VQA (briefly discussed in page 4), quite fascinating.

Reply: Thanks for the valuable comments. We agree with the reviewer’s perspective that demonstrating the pre-training of VQAs on larger-scale and more use cases would further improve the quality and practical relevance of our manuscript. To address this, in the updated version, we have included use-case-level demonstrations showcasing how our approach can be applied to pretrain *variational quantum eigensolver* (VQE) [11] effectively.

To ensure clarity, we have organized the remainder of this reply as follows. We first present numerical simulation results that demonstrate how our method effectively reduces quantum resource requirements for pre-training VQE, with **varied qubit counts (up to 50 qubits) and different families of Hamiltonians of practical relevance**. Notably, these results also highlight the superiority of our proposed methods compared to other benchmark learning approaches. We then conclude this reply with a **high-level summary** of the utility and broad applicability of our approach.

Practical Applications in pretraining VQEs.

In the following, we begin by briefly recapping the implementation details of pre-training VQAs for self-consistency. Then, we present numerical simulations across varied settings to comprehensively demonstrate how the proposed machine learning (ML) model enhances VQEs by reducing quantum resource demands.

Algorithmic implementation. The proposed pre-training strategy is summarized in Alg. 1 (Page 29 in the original Supplementary Material). Specifically, for a given quantum device and its corresponding ansatz $U(\mathbf{x})$, we first build a training dataset $\mathcal{T}_s$ by randomly generating inputs $\mathbf{x}$ and collecting the corresponding classical shadows $\tilde{\rho}_T(\mathbf{x})$ with T shots. Once the dataset is prepared, it is used to train the ML model $h_s(\mathbf{x})$ with the specified observable. The ML model $h_s(\mathbf{x})$ serves as a **surrogate** to optimize VQAs **entirely on the classical side**, removing the need for direct implementation on the specified quantum device.

Algorithm 1: Optimizing variational quantum algorithms in an offline manner

1. (Dataset construction) Randomly generate classical inputs $\mathbf{x} \sim [-\pi, \pi]^d$, feed it to the specified N -qubit quantum device to obtain the pre-measured state $U(\mathbf{x})|0\rangle^{\otimes N}$, and apply Pauli-based classical shadow with T shots to constitute a single training example $(\mathbf{x}, \tilde{\rho}_T(\mathbf{x}))$;
 2. (Dataset construction) Repeat the above procedure n times to construct the training dataset $\mathcal{T}_s$;
 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.
-

When the task is employing *variational quantum eigensolver* (VQE) to find the minimum energy of a given N -qubit Hamiltonian H , the detailed procedure for Step 3 in Alg. 1 is outlined below.

Recall that the mathematical description of VQE takes the form as

$$\min_{\mathbf{x} \in [-\pi, \pi]^d} \text{Tr}(HU(\mathbf{x})\rho_0U(\mathbf{x})^\dagger). \quad (10)$$

This optimization task can be achieved by the gradient descent method, i.e., the update of trainable parameters at the m -th iteration is

$$\mathbf{x}^{(m)} = \mathbf{x}^{(m-1)} - \eta \nabla_{\mathbf{x}} \text{Tr}(HU(\mathbf{x})\rho_0U(\mathbf{x})^\dagger), \quad (11)$$

where η refers to the learning rate. **Conventionally, quantum computers are necessary to calculate the gradients** $\nabla_{\mathbf{x}} \text{Tr}(HU(\mathbf{x})\rho_0U(\mathbf{x})^\dagger)$ through the parameter shift rule [12]. Accordingly, for M iterations, the total number of measurements (or equivalently the total queries to quantum device) is

$$\mathcal{Q} = 2 \times d \times M \times T \times L, \quad (12)$$

where T refers to snapshots for each circuit configuration and L refers to the number of Pauli terms in the Hamiltonian H . In this context, the **required quantum resources** become **increasingly expensive** and even prohibitive for very deep circuits (large d), extended iteration counts (large M), large-scale quantum chemical problems ($L \sim \mathcal{O}(N^2)$), or scenarios demanding low measurement errors (large T).

The proposed ML model $h_s(\mathbf{x})$ can substantially reduce the computational overhead of querying quantum devices. Recall that the proposed ML model $h_s(\mathbf{x})$ in Eq. (4) works as a **classical shadow state predictor**, where given any observable H with the bounded norm, $h_s(\mathbf{x}, H)$ approximates the expectation value $\text{Tr}(HU(\mathbf{x})\rho_0U(\mathbf{x})^\dagger)$ for $\forall \mathbf{x} \in [-\pi, \pi]^d$. In this regard, we can **pretrain corresponding VQEs** by optimizing the parameters $\mathbf{x}$ that minimizes the **surrogate loss**, i.e.,

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x} \in [-\pi, \pi]^d} h_s(\mathbf{x}, H). \quad (13)$$

The resulting parameters $\hat{\mathbf{x}}$ can then be used as **initialization** for executing the corresponding VQEs on a quantum device, allowing for further fine-tuning.

Remark. In the original manuscript, we adopted a zero-order gradient optimizer to minimize the surrogate loss $h_s(\mathbf{x}, H)$ in Eq. (13). This choice was made because the explored toy models had very small dimensions, making this approach sufficient to achieve good performance. Notably, the proposed model $h_s(\mathbf{x}, H)$ is **purely classical and fully differentiable**. Consequently, auto-differentiation tools can be utilized to estimate the minima $\hat{\mathbf{x}}$ using **higher-order optimization methods** with better performance, as we will demonstrate subsequently.

Numerical results. We next conduct numerical investigations to evaluate the performance of the proposed ML model in pretraining various VQE tasks of practical interest. For clarity and completeness, this section is organized into four parts: *Part (I) Problem setup*, *Part (II) Benchmark methods*, *Part (III) Evaluation metrics*, and *Part (IV) Analysis of the simulation results*.

Part (I) Problems setup. We focus on employing VQE to estimate the ground-state energies of **two families of Hamiltonians**, which are

- ① *One-dimensional Transverse-Field Ising models (TFIMs).* The mathematical definition of the family of N -qubit (one-dimensional) TFIMs is

$$\mathcal{H}_{\text{TFIM}} = \left\{ H_{\text{TFIM}}(J; h) = \sum_{i=1}^{N-1} J Z_i Z_{i+1} + h \sum_{i=1}^N X_i \mid J, h \in \mathbb{R} \right\}, \quad (14)$$

where J and h refer to the coupling strength and the external field applied in a transverse direction, respectively.

- ② *H₅ molecules.* The families of H₅ molecules, denoted by $\mathcal{H}_{\text{H}_5}$, describe H₅ molecules with different bond lengths R , i.e.,

$$\mathcal{H}_{\text{H}_5} = \left\{ H_{\text{H}_5}(R), R \in \mathbb{R}_+ \right\}. \quad (15)$$

Under the Jordan-Wigner Hamiltonian representation, each Hamiltonian $H_{H_5}(R)$ corresponds to a 10-qubit Hamiltonian composed of 444 weighted Pauli operators. For simplicity, we do not provide the explicit forms of $H_{H_5}(R)$ here.

Remark. Both families of Hamiltonians are widely studied in the literature, garnering attention in quantum many-body physics and quantum chemistry, spanning algorithmic developments to experimental investigations. In particular, The TFIM serves as a prototypical model for understanding quantum phase transitions, non-equilibrium dynamics, and critical phenomena in condensed matter systems [13]. Its simplicity, coupled with its rich physics, makes it a cornerstone for testing quantum algorithms [14, 15] and benchmarking quantum simulators [16, 17]. In addition, the hydrogen molecules (e.g., H_2 and H_5) represent small yet nontrivial systems that capture essential features of electron correlation and quantum entanglement in molecular systems [18]. It is widely used to benchmark quantum computational methods in solving quantum chemistry problems with near-term quantum devices [19, 20].

Part (II) Benchmark methods. To attain a comprehensive understanding about the capabilities of the proposed ML model, we compare its performance with **three other types of classical learning surrogates** in pretraining VQEs. Additionally, we use **conventional VQEs** as a benchmark to evaluate the discrepancy between the *optimized results* and the *exact ground-state energy*. Below, we outline the fundamental mechanisms of each approach.

- *Method I: Conventional VQE.* The implementation of VQE mainly follows the rules introduced in Eq. (11). The only difference is that we set the number of measurements to infinity (i.e., $T \rightarrow \infty$) and compute the gradients using the *automatic differentiation method*, i.e., the update of trainable parameters is completed by the Adam optimizer [21].

Remark. The infinite measurements can eliminate the influence of estimation errors, allowing us to **isolate and examine how the employed ansatz $U(x)$ affects the performance**, i.e., the discrepancy between the optimized results of VQE and the exact ground-state energy.

- *Method II: The proposed ML model.* The first surrogate applied to the pre-training VQE is the proposed ML model defined in Eq. (4), i.e.,

$$h_s(x, H) = \frac{1}{n} \sum_{i=1}^n \kappa_\Lambda(x, x^{(i)}) g(x^{(i)}, H),$$

where $\kappa_\Lambda(x, x^{(i)})$ is the truncated trigonometric monomial kernel introduced in Eq. (5) and $g(x^{(i)}, H) = \text{Tr}(\tilde{\rho}_T(x^{(i)})H)$ refers to the shadow estimation of $\text{Tr}(\rho(x^{(i)})H)$ with T snapshots for each training example in $\mathcal{T}_s$ (Step 2 in Alg. 1).

Given access to the proposed ML model $h_s(x, H)$, the initialized parameters of $U(x)$ take the form as

$$\hat{x} = \arg \min_x h_s(x, H). \quad (16)$$

To find the parameters minimizing the surrogate loss $h_s(x, H)$, we employ Adam optimizer, an advanced gradient-descent-based method.

- *Method III: kernel ridge regression.* We identify that when the employed ansatz $U(x)$ consists of d RZ gates and an arbitrary number of Clifford gates, the estimated energy in VQE can be **reformulated into an inner-product form under the Pauli transfer matrix (PTM) representations**, i.e.,

$$\text{Tr}(HU(x)\rho_0U(x)^\dagger) = \langle \mathbf{w}^*, \vec{\Phi}(x) \rangle, \quad (17)$$

where $\vec{\Phi}(\mathbf{x}) = \{\Phi_{\omega}(\mathbf{x})\}_{\omega}$ is the 3^d -dimensional vector with $\omega \in \{0, \pm 1\}^d$ and $\Phi_{\omega}(\mathbf{x})$ being defined in Eq. (5), and the optimal parameters $\mathbf{w}^* = \{2^{\|\omega\|_0} \text{Tr}(\rho_{\omega} \mathbf{H})\}_{\omega}$ is also a 3^d -dimensional vector with ρ_{ω} being parameter-independent matrix.

The inner-product reformulation above enables us to **utilize kernel ridge regression (KRR) [10] to construct a learning surrogate**. Specifically, given the training dataset $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$ with $y^{(i)} = \text{Tr}(\mathbf{H}U(\mathbf{x}^{(i)})\rho_0U(\mathbf{x}^{(i)})^\dagger)$ in the ideal case or $y^{(i)} = \text{Tr}(\tilde{\rho}_T(\mathbf{x}^{(i)})\mathbf{H})$ in the shadow case for $\forall i \in [n]$, we can use a KRR model $h_{\text{KRR}}(\mathbf{x}; \mathbf{w}) = \langle \mathbf{w}, \vec{\Phi}(\mathbf{x}) \rangle$ to minimize the empirical risk, i.e.,

$$\min_{\mathbf{w}} \frac{1}{2} \sum_{i=1}^n \left(h_{\text{KRR}}(\mathbf{x}^{(i)}; \mathbf{w}) - y^{(i)} \right)^2 + \frac{\alpha}{2} \|\mathbf{w}\|_2, \quad (18)$$

where α is a hyper-parameter. Under the *dual representation* [10], the closed-form solution of the above minimization problem is

$$h_{\text{KRR}}(\mathbf{x}; \hat{\mathbf{w}}) = \mathbf{k}(\mathbf{x}) \cdot (K + \alpha \mathbb{I}_n)^{-1} \mathbf{y}, \quad (19)$$

where $\hat{\mathbf{w}}$ refers to the optimized parameters, $\mathbf{y} = [y^{(1)}, \dots, y^{(n)}]$ is the label vector, $K \in \mathbb{R}^{n \times n}$ refers to the kernel matrix with

$$K_{ij} \equiv k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \langle \vec{\Phi}(\mathbf{x}^{(i)}), \vec{\Phi}(\mathbf{x}^{(j)}) \rangle = \prod_{m=1}^d (1 + 2 \cos(\mathbf{x}_m^{(i)} - \mathbf{x}_m^{(j)})), \quad \forall i, j \in [n], \quad (20)$$

and $\mathbf{k}(\mathbf{x})$ is an n -dimensional vector whose i -th entry is $\mathbf{k}(\mathbf{x})_i = \langle \vec{\Phi}(\mathbf{x}), \vec{\Phi}(\mathbf{x}^{(i)}) \rangle = \prod_{m=1}^d (1 + 2 \cos(\mathbf{x}_m^{(i)} - \mathbf{x}_m))$.

Given access to the optimized learning model $h_{\text{KRR}}(\mathbf{x}; \hat{\mathbf{w}})$ in Eq. (19), the pre-training VQE amounts to solving the following minimization problem,

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} h_{\text{KRR}}(\mathbf{x}, \hat{\mathbf{w}}). \quad (21)$$

The optimization is completed by the Adam optimizer. After that, the optimized result serves as the initialized parameters of $U(\mathbf{x})$ for the corresponding VQE task.

• **Method IV: Random Fourier features.** The computational efficiency of KRR highly depends on the number of training examples n . When the number of training examples n becomes large, the space and time cost of KRR scales with $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$, respectively. To address the computational bottleneck, **an alternative is employing random Fourier features (RFF) techniques** to learn the given dataset [22]. More specifically, when the target kernel is *shift-invariant*, i.e., $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}')$, the KRR model $h_{\text{KRR}}(\mathbf{x}; \mathbf{w})$ can be estimated by **RFF-based regression model**, i.e.,

$$h_{\text{RFF}}(\mathbf{x}, \mathbf{w}) = \langle \mathbf{w}, \vec{\Psi}(\mathbf{x}) \rangle, \quad (22)$$

where $\vec{\Psi}(\mathbf{x}) = [\cos(\frac{\gamma}{d'} \langle \boldsymbol{\theta}_1, \mathbf{x} + b_1 \rangle), \sin(\frac{\gamma}{d'} \langle \boldsymbol{\theta}_1, \mathbf{x} + b_1 \rangle), \dots, \cos(\frac{\gamma}{d'} \langle \boldsymbol{\theta}_{d'}, \mathbf{x} + b_{d'} \rangle), \sin(\frac{\gamma}{d'} \langle \boldsymbol{\theta}_{d'}, \mathbf{x} + b_{d'} \rangle)]^\top$ is the RFF with d' dimensions, γ and d' are tunable hyperparameters, $\boldsymbol{\theta}_i$ for $\forall i \in [d']$ are d' -dimensional vectors sampled from a multivariate standard normal distribution, and b_i for $\forall i \in [d']$ is scalar sampled from the uniform distribution ranging from $-\pi$ to π .

Following Eq. (20), an immediate observation is that the kernel under consideration is *shift-invariant*, justifying the use of an RFF-based regression model as a surrogate. Specifically, given the training dataset $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$, we train the RFF-based regression model by minimizing the empirical risk

$$\min_{\mathbf{w}} \frac{1}{2} \sum_{i=1}^n \left(h_{\text{RFF}}(\mathbf{x}^{(i)}; \mathbf{w}) - y^{(i)} \right)^2 + \frac{\alpha}{2} \|\mathbf{w}\|_2, \quad (23)$$

where α is a hyper-parameter.

Denote the optimized parameters as $\hat{\mathbf{w}}$. As before, we use Adam optimizer to locate the solution that minimizes the surrogate loss, i.e.,

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} h_{\text{RFF}}(\mathbf{x}, \hat{\mathbf{w}}). \quad (24)$$

The obtained estimate serves as the initialization parameters of $U(\mathbf{x})$ for the corresponding VQEs.

• *Method V: Multi-layer Perceptron.* We last exploit the performance of deep learning protocols for VQE's parameter initialization. For this purpose, we adopt multi-layer perceptron (MLP) as the deep learning protocol [23], whose mathematical expression is

$$h_{\text{MLP}}(\mathbf{x}; \mathbf{W}) = W_L \text{ReLU}(\dots \text{ReLU}(W_2 \text{ReLU}(W_1 \mathbf{x}))), \quad (25)$$

where $\mathbf{W} = \{W_i\}_{i=1}^L$ refers to L trainable weight matrices and the size of W_i for $\forall i \in [L]$ is adjustable. Given the training dataset $\mathcal{T}_s = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$, we train MLP by minimizing the empirical risk

$$\min_{\mathbf{W}} \frac{1}{2} \sum_{i=1}^n \left(h_{\text{MLP}}(\mathbf{x}^{(i)}; \mathbf{w}) - y^{(i)} \right)^2 + \frac{\alpha}{2} \sum_i \|W_i\|_F, \quad (26)$$

where α is a hyper-parameter.

Given access to the optimized MLP $h_{\text{MLP}}(\mathbf{x}, \hat{\mathbf{W}})$, the initialization parameters for $U(\mathbf{x})$ are acquired via solving

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} h_{\text{MLP}}(\mathbf{x}, \hat{\mathbf{W}}). \quad (27)$$

For fair comparison, the optimization of MLP is also completed by the Adam optimizer.

Remark. While the employed MLP lacks provable guarantees, benchmarking its performance is crucial given the growing success of deep learning in quantum system learning tasks [24, 25]. For completeness and comprehensiveness, it is important to evaluate the MLP on bounded-gate circuits and compare its performance with ML-based methods.

Part (III) Metrics. For comprehensive, we employ two metrics to evaluate the performance of different benchmarking methods.

- ① *Prediction error.* Our first metric concerns the ability of the trained learning models to predict unseen examples. Denote the test dataset as $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n_{\text{test}}}$ such that the test examples are sampled from the same distribution with the training examples. The prediction error refers to the mean square error on the test dataset, i.e.,

$$R_{\text{predict}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(h_*(\mathbf{x}) - y^{(i)} \right)^2, \quad (28)$$

where $h_* \in \{h_s, h_{\text{KRR}}, h_{\text{RFF}}, h_{\text{MLP}}\}$.

- ② *Absolute energy difference.* The second metric compares the difference between the estimated energy $\hat{E}_{\text{VQE}}$ returned by VQE and the exact minimum ground state energy E_{Exact} of the given Hamiltonian, i.e.,

$$R_{\text{abs}} = \left| \hat{E}_{\text{VQE}} - E_{\text{Exact}} \right|. \quad (29)$$

The estimated energy is defined by $\hat{E}_{\text{VQE}} = \text{Tr}(H U(\hat{\mathbf{x}}) \rho_0 U(\hat{\mathbf{x}})^\dagger)$. Given different parameters $\hat{\mathbf{x}}$ returned by varied learning surrogates with $h_* \in \{h_s, h_{\text{KRR}}, h_{\text{RFF}}, h_{\text{MLP}}\}$, we can evaluate how these surrogates advance the parameter initialization by comparing R_{abs} .

Part (IV) Analysis of simulation results. We proceed with numerical simulations to showcase the capabilities of the proposed ML model in pretraining VQAs. To achieve this, we first employ the *five benchmark methods* to minimize the energy of the *two families Hamiltonians* $\mathcal{H}_{\text{TFIM}}$ and $\mathcal{H}_{\text{H}_5}$ introduced earlier, setting the number of qubits to $N = 10$. This scale allows for **a systematic evaluation of different approaches through ablation studies**. Subsequently, we demonstrate the application of the proposed ML approach to minimize the TFIM Hamiltonian with $N = 50$, highlighting its **scalability and effectiveness**.

① *Ablation study with $N = 10$.* For clarity, we **individually present and explain the simulation results of pre-training VQEs for TFIM and the H_5 molecule** introduced in Part I.

◆ *Results of TFIM.* The detailed settings of TFIMs are as follows. We investigate six TFIMs in $\mathcal{H}_{\text{TFIM}}$. Following notations in Eq. (14), these Hamiltonians are constructed by fixing $h = -0.5$ and varying the coupling strength with $J \in \{-0.5, -0.3, -0.1, 0.1, 0.3, 0.5\}$.

Moreover, Hamiltonian-informed ansatz is adopted to estimate their ground state energies, i.e.,

$$U(\mathbf{x}) = \prod_{i=1}^{N-1} \exp\left(-i \frac{\mathbf{x}_i}{2} (Z_i \otimes Z_{i+1})\right) \prod_{i=N}^{2N-1} \left(-i \frac{\mathbf{x}_i}{2} (X_i)\right) \text{Ha}^{\otimes N}, \quad (30)$$

where i refers to the imaginary unit and Ha refers to the Hadamard gate. As illustrated in Fig. R3, optimal parameters for this ansatz exist to minimize the six Hamiltonians under investigation. In particular, for each setting, we perform five repetitions with different initial parameters to collect statistical results, using a learning rate of 0.8 for the Adam optimizer. Fig. R3(a) illustrates the statistical performance of VQEs. That is, for each configuration, the exact ground-state energy can be accurately estimated after 200 iterations with *well-initialized parameters*. For clarity, we visualize the training dynamics of the VQE for the top-performing setting in Fig. R3(b). An immediate observation is that across all settings, the absolute error between the estimated energy and the exact value remains small, which are 0.24, 0.04, 5.06×10^{-4} , 4.92×10^{-4} , 0.04, and 0.24 for $J = -0.5, -0.3, -0.1, 0.1, 0.3, 0.5$, respectively.

The process of pretraining VQEs by classical learning surrogates follows the steps outlined in Alg. 1. Specifically, in Steps 1&2, we construct the training dataset $\mathcal{T}_s$ by randomly sampling $\mathbf{x} \in [-\pi, \pi]^{19}$ from a uniform distribution and inputting these values into the ansatz $U(\mathbf{x})$ in Eq. (30) to generate classical shadows (when $N = 10$, the dimension of the input $\mathbf{x}$ is $d = 19$). The snapshot for each quantum state $\rho(\mathbf{x})$ is fixed to be $T = 100$ and the number of trainable examples is set $n = 7500$.

Once the dataset $\mathcal{T}_s$ is collected, we move to Step 3 of Alg. 1. That is, the dataset $\mathcal{T}_s$ is used to pretrain VQEs by the four classical surrogates introduced in Part (II) to estimate the ground-state energies of the six explored TFIMs. The hyper-parameter settings of each learning surrogate are as follows.

- **Proposed ML model.** For all configurations, we fix the truncation value as $\Lambda = 2$ and the number of training examples as $n = 7500$ to obtain the corresponding learning surrogate $h_s(\mathbf{x}; \text{H})$ following Eq. (4). In the procedure of minimizing the surrogate loss as defined in Eq. (16), we set the learning rate of Adam optimizer as 0.01 and the total number of iterations as 500.
- **Kernel regression model.** For all configurations, we set the number of training examples as $n = 100$ to avoid the memory issue. The learning models h_{KRR} are trained following Eq. (19). Once the trained models are available under different J values, we follow Eq. (21) to minimize their corresponding surrogate losses. As with the proposed ML model, we set the learning rate of Adam optimizer as 0.01 and the total number of iterations as 500.

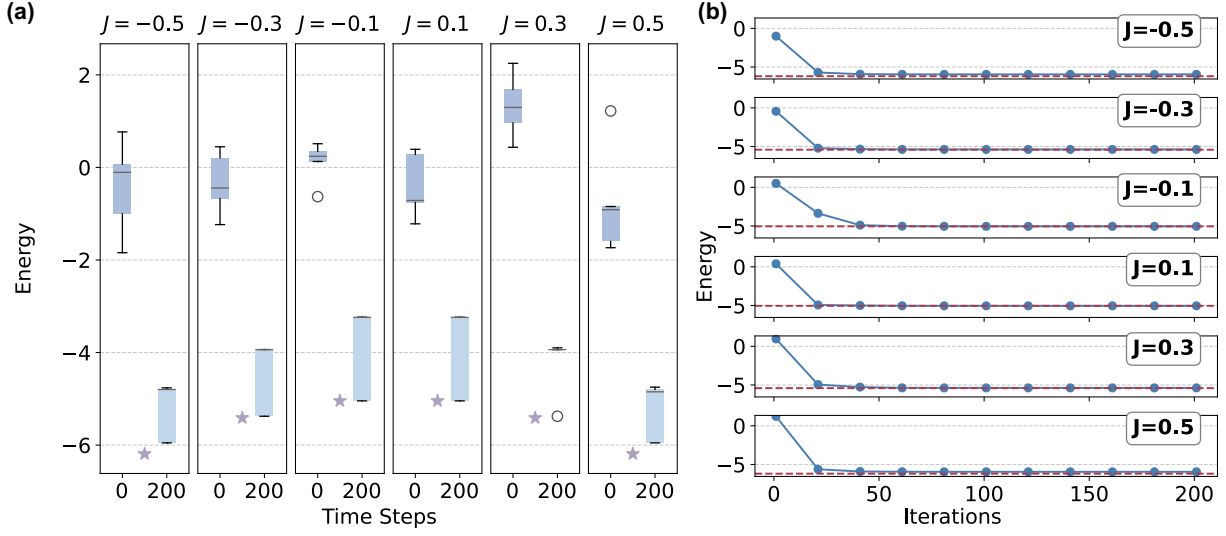


Figure R3: Performance of conventional VQEs on the family of TFIMs with $N = 10$. (a) STATISTICAL PERFORMANCE OF CONVENTIONAL VQES. The box plots show the distribution of the estimated energies at the first iteration and after 200 iterations for each setting, with J values ranging from -0.5 to 0.5 . Each box represents the interquartile range (IQR), with the horizontal line inside indicating the median energy. The whiskers extend to the most extreme data points within 1.5 times the IQR, and outliers beyond this range are shown as individual circles. The purple stars represent the exact ground-state energies. (b) TRAINING DYNAMICS OF THE BEST VQES. The plots show the convergence of the estimated energies toward the exact ground-state energies (dashed red lines) over 200 iterations for the top-performing instance in each setting.

- **Random Fourier Feature.** For all configurations, we set the dimension of random Fourier features as $d' = 800$ in Eq. (22), to match the feature dimension of the proposed learning models $h_s(\mathbf{x}; \mathbf{H})$. We follow Eq. (23) to h_{RFF} under different J values. The number of training examples is set as $n = 7500$ and the hyper-parameter is fixed to be $\alpha = 1$. In the procedure of minimizing the surrogate loss as defined in Eq. (24), we follow the same routine with $h_s(\mathbf{x}; \mathbf{H})$, where the learning rate of Adam optimizer is 0.01 and the total number of iterations is 500.
- **MLP.** For all configurations, the exploited $h_{\text{MLP}}(\mathbf{x}, \mathbf{W})$ in Eq. (25) refers to a 2-hidden-layer MLP, where the dimension of the input layer, the first hidden layer, the second hidden layer, and the output layer is 19, 32, 8, and 1, respectively. The total number of trainable parameters of MLP is 872, comparable to those used in the proposed learning model and KRR. Moreover, each hidden layer is followed by a ReLU activation function. We use $n = 7500$ training examples to optimize MLPs under varied J values are as follows. The learning rate of the Adam optimizer is set as 0.001. The batch size is set as 200. The hyper-parameter α is set as 0.0001. Once $h_{\text{MLP}}(\mathbf{x}, \hat{\mathbf{W}})$ is accessible, we follow the surrogate loss in Eq. (27) to pretrain VQEs. As with the proposed ML model, we set the learning rate of Adam optimizer as 0.01 and the total number of iterations as 500.

Remark. To assess the impact of noisy labels caused by the finite snapshots of classical shadows in each training example, we also benchmark the performance of the employed learning surrogate under the assumption of an infinite number of shots for each configuration, i.e., $T \rightarrow \infty$, $\tilde{\rho}_T(\mathbf{x}) = \rho(\mathbf{x})$,

and $y^{(i)} = \text{Tr}(\rho(x^{(i)})H)$. We repeat each configuration described above five times to collect statistical results.

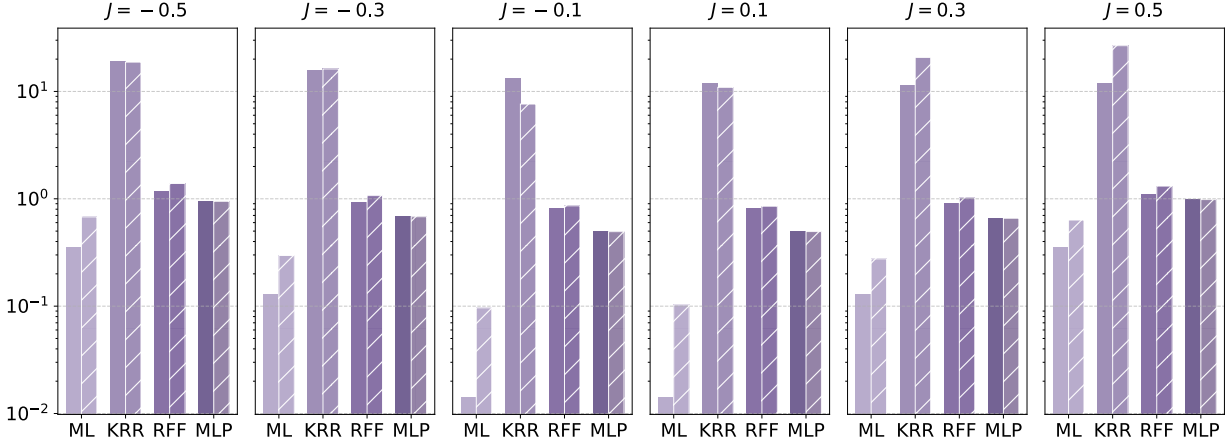


Figure R4: **Prediction error of four learning surrogates for the family of TFIMs with $N = 10$.** Each subplot corresponds to the simulation results for a specific value of J in the TFIM. Solid bars represent the prediction error under ideal conditions ($T \rightarrow \infty$ in the training dataset $\mathcal{T}_s$), while shadowed bars with a “/” pattern denote the noisy case ($T = 100$). The four surrogates are labeled as ML (the proposed ML model), KRR (Kernel ridge regression), RFF (random Fourier features), and MLP (Multi-Layer Perceptron), respectively.

The predictor error of the employed four learning surrogates (i.e., the first metric R_{predict} in Eq. (28)) is demonstrated in Fig. R4, which is evaluated by 500 test examples. For all subplots, the proposed ML models outperform the rest three learning surrogates under both ideal (highlighted by the solid bar) and noisy (highlighted by the shadowed bar) scenarios. Particularly, the prediction error of the proposed ML model under the ideal (noisy) setting achieves 0.36 (0.68), 0.13 (0.30), 0.01 (0.10), 0.01 (0.10), 0.13 (0.28), and 0.36 (0.63) when $J = -0.5, -0.3, -0.1, 0.1, 0.3, 0.5$, respectively. Another observation is that while the noisy labels caused by finite snapshots (i.e., $T = 100$) can slightly degrade prediction performance of the proposed ML model, the impact is not substantial, with a maximum degradation of no more than 0.32 across all settings.

Next, we use the optimized learning surrogates to pretrain VQEs for the six explored TFIMs. The estimated energies, both with and without pretraining by the four learning surrogates, are illustrated in Fig. R5. An immediate observation is that compared to other learning surrogates, **the estimated energy obtained from the proposed ML model is greatly closer to the exact result in both ideal and noisy settings**. Notably, unlike the proposed ML model, the performance of VQEs does not improve after pretraining with other learning surrogates and, in some cases, may **even degrade**. These contrastive outcomes further validate the effectiveness of our proposed learning protocol.

Another key observation from the simulation results is that the performance of the proposed ML model remains comparable between the ideal and noisy settings, as illustrated in Figs. R5(a) and (b), respectively. This indicates that **the noisy labels introduced by the finite snapshots of classical shadows only subtly effect the performance of the proposed ML model**. Furthermore, the results suggest that **a constant number of snapshots** (e.g., $T = 100$) is sufficient to achieve strong performance, **aligning with the findings of Theorem 2**.

Finally, we compare the performance of the conventional VQE (using infinite measurements) with the

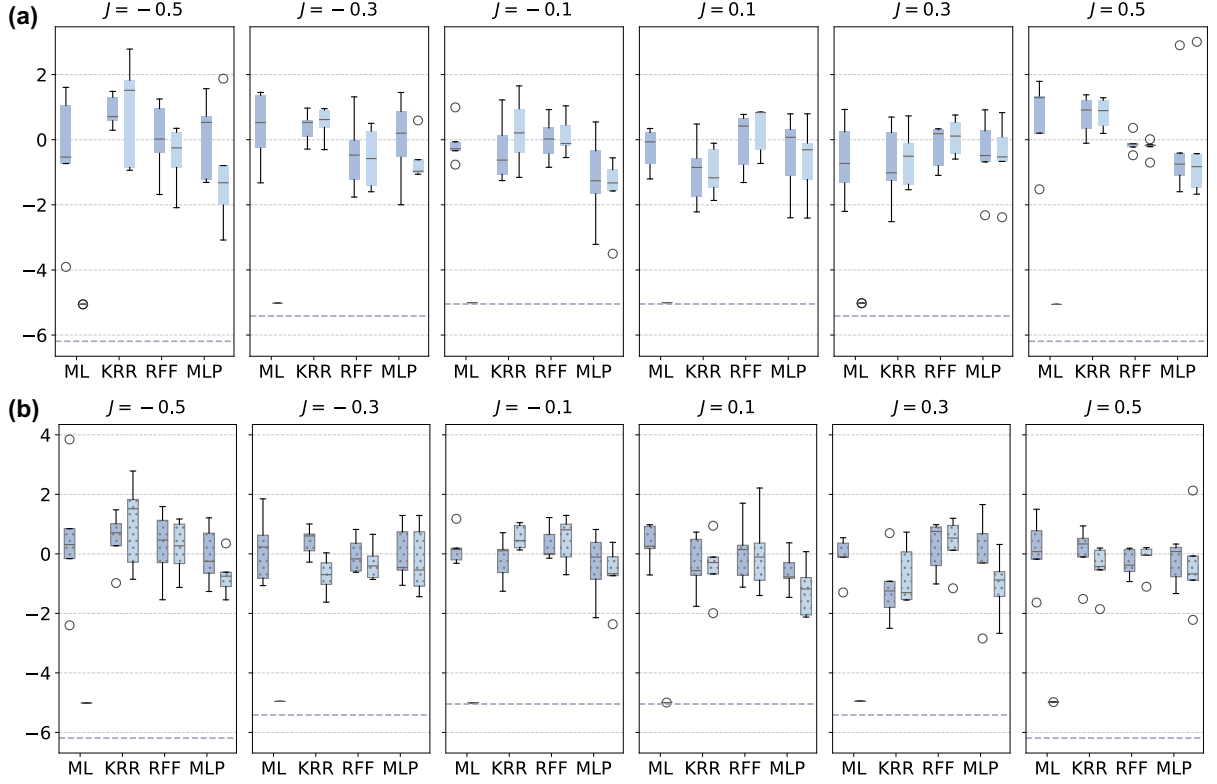


Figure R5: **Pretraining performance of different learning surrogates on the family of TFIMs with $N = 10$.** The upper panel (a) and the lower panel separately exhibit the simulation results under the ideal and noisy settings ($T \rightarrow \infty$ and $T = 100$). Each subplot corresponds to a specific J value in the TFIM, and the box plots depict the estimated ground-state energy distributions for four learning surrogates. For each surrogate, the estimated energies without pre-training and after pretraining are visualized by the deep and light blue colors, respectively. Dashed horizontal lines represent the exact ground-state energies.

Table R2: **Comparison between conventional VQEs and the proposed ML model on a family of TFIMs with $N = 10$.** The table presents the estimated ground state energies for various coupling strengths J , as obtained from the conventional VQE and the proposed ML model under both ideal and noisy conditions (labeled by ‘Pretrain (ideal)’ and ‘Pretrain (shadow)’, respectively). The notation ‘ $a \pm b$ ’ indicates that the mean value is a and the variance is b .

	$J = -0.5$	$J = -0.3$	$J = -0.1$	$J = 0.1$	$J = 0.3$	$J = 0.5$
VQE	-5.252 ± 0.327	-4.512 ± 0.498	-3.960 ± 0.784	-3.961 ± 0.784	-4.213 ± 0.339	-5.26 ± 0.32
Pretrain (ideal)	-5.056 ± 0.000	-5.020 ± 0.000	-5.002 ± 0.000	-5.002 ± 0.000	-5.021 ± 0.000	-5.057 ± 0.000
Pretrain (shadow)	-5.012 ± 0.000	-4.959 ± 0.000	-5.002 ± 0.000	-4.996 ± 0.000	-4.951 ± 0.000	-4.981 ± 0.000

proposed ML model. The statistical results are summarized in Table R2. The first key observation is that **the proposed ML model, in both ideal and noisy settings, generally achieves energy estimates that are comparable to or better than those of the conventional VQE**, particularly at coupling strengths $J \in \{-0.3, -0.1, 0.1, 0.3\}$. The second key observation is **the consistency and stability of the proposed ML model’s performance across both settings**, especially at intermediate coupling

The simulation results in Fig. R6 validate that optimal parameters for the ansatz in Eq. (31) exist to minimize the six Hamiltonians under investigation. In particular, for each setting, we perform five repetitions with different initial parameters to collect statistical results, using a learning rate of 0.8 for the Adam optimizer. Fig. R6(a) illustrates the statistical performance of VQEs. That is, for each configuration, the exact ground-state energy can be accurately estimated after 200 iterations, especially for the small bond lengths. For clarity, we visualize the training dynamics of the VQE for the top-performing setting in Fig. R6(b). Specially, across all settings, the absolute error between the estimated energy and the exact value remains small, which are 0.04, 0.04, 0.05, 0.06, 0.07, and 0.08 for $R = 0.5\text{\AA}, 0.6\text{\AA}, 0.7\text{\AA}, 0.8\text{\AA}, 0.9\text{\AA}, 1.0\text{\AA}$, respectively.

Similar to the task of TFIMs, the process of pretraining VQEs by classical learning surrogates follows the steps outlined in Alg. 1. In Steps 1&2, we construct the training dataset $\mathcal{T}_s$ by randomly sampling $\mathbf{x} \in [-\pi, \pi]^{14}$ from a uniform distribution and inputting these values into the ansatz $U(\mathbf{x})$ in Eq. (31) to generate classical shadows. The snapshot for each quantum state $\rho(\mathbf{x})$ is fixed to be $T = 100$ and the number of trainable examples is set $n = 7500$. Once the dataset $\mathcal{T}_s$ is collected, we move to Step 3 of Alg. 1, where the four classical surrogates introduced in Part (II) employ $\mathcal{T}_s$ to pretrain VQEs. The hyperparameter settings for each learning surrogate are identical to those used in the TFIM tasks.

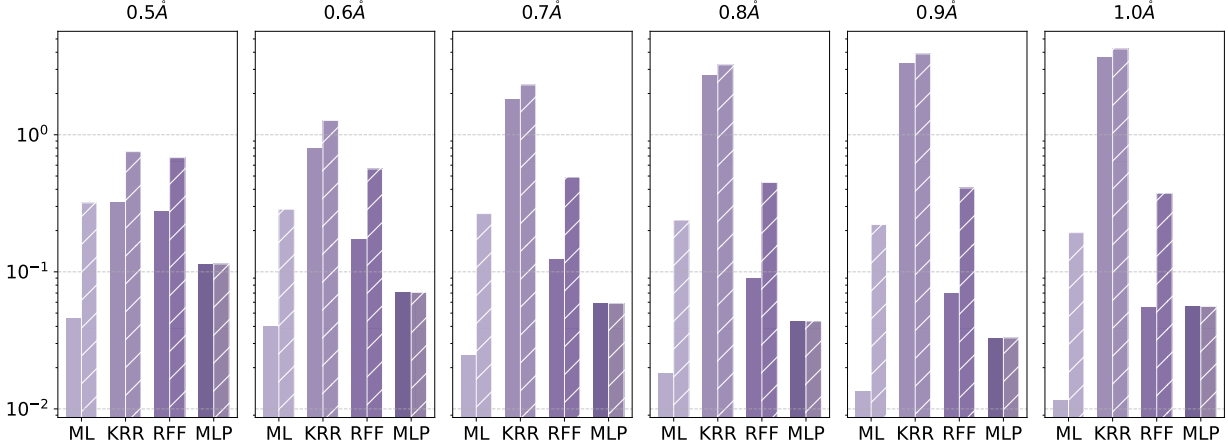


Figure R7: **Prediction error of four learning surrogates for the family of H_5 molecules with $N = 10$.** Each subplot corresponds to the simulation results for a specific bond length R . The solid bars, shaded bars, and x-axis labels follow the same representations as those in Fig. R4.

The predictor error R_{predict} of the employed four learning surrogates is shown in Fig. R7, which is evaluated by 500 test examples. For all explored bond lengths, the proposed ML model outperforms the other three learning surrogates under the ideal setting. Additionally, under the noisy setting, the proposed ML model performs slightly worse than MLP but remains superior to the other two learning surrogates. More specifically, the prediction error of the proposed ML model under the ideal (noisy) setting achieves 0.046 (0.319), 0.040 (0.286), 0.025 (0.266), 0.018 (0.239), 0.013 (0.222), and 0.012 (0.193) when $R = 0.5\text{\AA}, 0.6\text{\AA}, 0.7\text{\AA}, 0.8\text{\AA}, 0.9\text{\AA}, 1.0\text{\AA}$, respectively. The increased discrepancy between the ideal and shadow results, compared to the TFIM task, arises from the presence of numerous global Pauli terms in the H_5 molecules, which substantially amplify the Pauli-based shadow estimation error.

We proceed to use the optimized four learning surrogates to pretrain VQEs for the six explored H_5 molecules. The estimated energies, both with and without pretraining by the four learning surrogates, are illustrated in Fig. R8. The first key observation is that, **under both ideal and noisy conditions, the**

parameters returned by the proposed ML model substantially improve estimation accuracy after pretraining. In contrast, other learning surrogates fail to reduce the estimation error through pretraining and, in some cases, even lead to degraded performance. Furthermore, while the shadowed dataset increases the test prediction error of the proposed ML model discussed in Fig. R7, it **notably improves the estimation of the ground-state energy compared to both the ideal case and MLP**, particularly when the bond length is $R = 0.8\text{\AA}$. These observations align with the results of Theorem 2 in the main text.

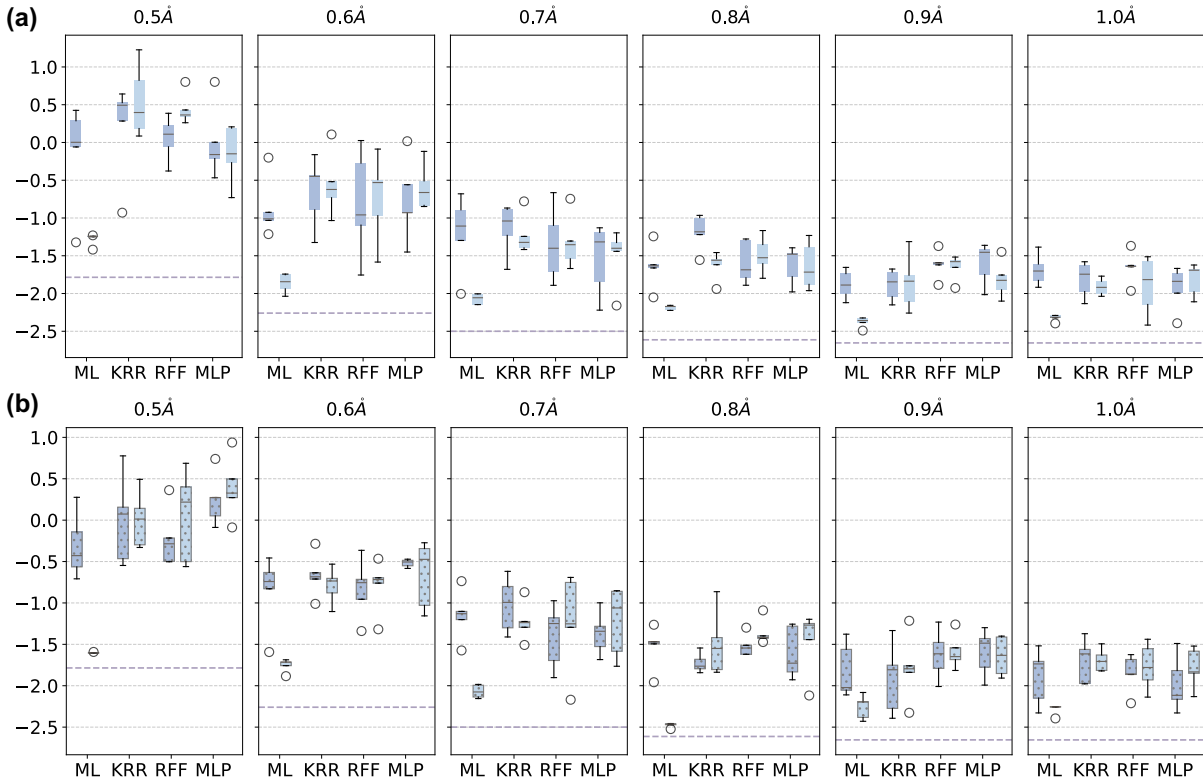


Figure R8: **Pretraining performance of four learning surrogates on the family of H_5 molecules with $N = 10$.** The upper panel (a) and the lower panel separately exhibit the simulation results under the ideal and noisy settings ($T \rightarrow \infty$ and $T = 100$). Each subplot corresponds to a specific bond length R in $H_{H_5}(R)$, and the box plots depict the estimated ground-state energy distributions for four learning surrogates. For each surrogate, the estimated energies without pre-training and after pretraining are visualized by the deep and light blue colors, respectively. Dashed horizontal lines represent the exact ground-state energies.

Finally, we compare the performance of the conventional VQE (using infinite measurements) with the proposed ML model. The statistical results are summarized in Table R2. A key observation is that the proposed ML model, under both ideal and noisy settings, generally produces energy estimates that are comparable to but slightly less accurate than those of the conventional VQE. This slight inferiority may be attributed to the increased gradient norm when the Hamiltonian involves more than 400 Pauli terms, which necessitates more training examples and larger truncation values to achieve improved performance.

② *Simulation results of 50-qubit TFIM.* We now turn to evaluate the performance of the proposed

Table R3: **Comparison between conventional VQEs and the proposed ML model on a family of H_5 with $N = 10$.** The table presents the estimated ground state energies for various coupling strengths J , as obtained from the conventional VQE and the proposed ML model under both ideal and noisy conditions (labeled by ‘Pretrain (ideal)’ and ‘Pretrain (shadow)’, respectively). The notation ‘ $a \pm b$ ’ indicates that the mean value is a and the variance is b .

	$R = 0.5\text{\AA}$	$R = 0.6\text{\AA}$	$R = 0.7\text{\AA}$	$R = 0.8\text{\AA}$	$R = 0.9\text{\AA}$	$R = 1.0\text{\AA}$
VQE	-1.747 ± 0.000	-2.219 ± 0.000	-2.45 ± 0.000	-2.555 ± 0.000	-2.584 ± 0.000	-2.571 ± 0.000
Pretrain (ideal)	-1.276 ± 0.005	-1.862 ± 0.013	-2.072 ± 0.004	-2.189 ± 0.001	-2.376 ± 0.004	-2.322 ± 0.002
Pretrain (shadow)	-1.601 ± 0.000	-1.752 ± 0.005	-2.069 ± 0.001	-2.475 ± 0.001	-2.257 ± 0.017	-2.283 ± 0.003

ML model in the large-scale scenario. Specifically, we focus on the TFIM defined in Eq. (14) with the number of qubits set to $N = 50$, where the hyperparameters of TFIMs are fixed to be $J = -0.2$ and $h = -0.5$. Moreover, the Hamiltonian-informed ansatz $U(x)$ defined in Eq. (30) is employed to implement VQE with the initialized state being the zero product state $|0\rangle^{\otimes 50}$.

As in the previous tasks, the pretraining process mainly follows the steps outlined in Alg. 1. That is, in Steps 1&2, we construct the training dataset $\mathcal{T}_s$ by randomly sampling $x \in [-\pi, \pi]^{99}$ from a uniform distribution and inputting these values into the ansatz $U(x)$ to generate classical shadows. The snapshot for each quantum state $\rho(x)$ is fixed to be $T = 300$ and the number of trainable examples is set as $n = 1500$. Once the dataset $\mathcal{T}_s$ is collected, we move to Step 3 of Alg. 1, where the proposed ML model employs $\mathcal{T}_s$ to pretrain VQEs. Similar to the 10-qubit task, the Adam optimizer is used to minimize the surrogate loss, with a learning rate of 0.01 and a total of 500 iterations.

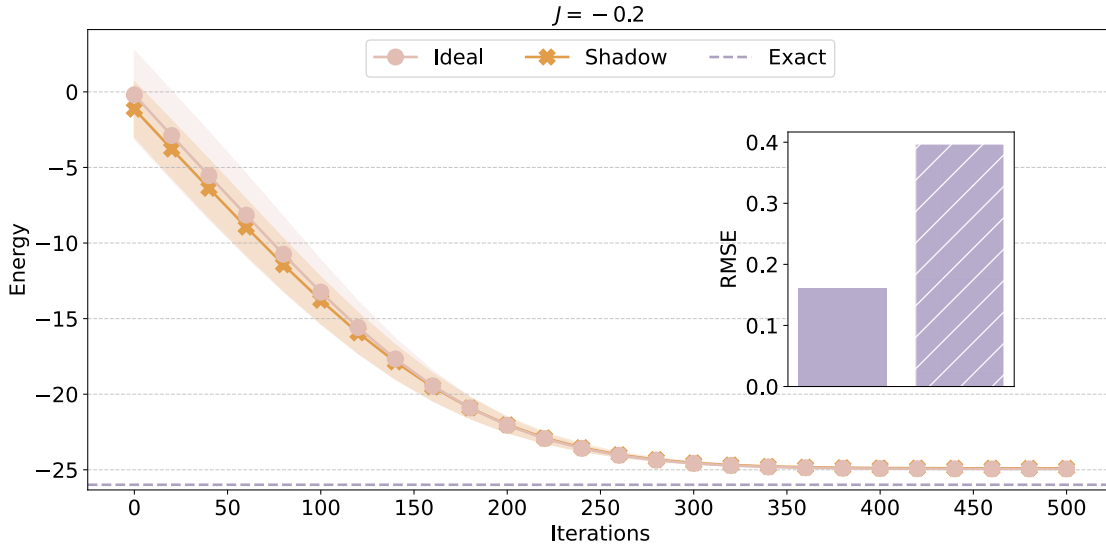


Figure R9: **Pretraining performance of the proposed ML model on a 50-qubit TFIM.** The inset presents the prediction error R_{predict} of the proposed ML model for both the ideal and shadowed cases, i.e., $T \rightarrow \infty$ and $T = 300$. The main plot depicts the training dynamics of the proposed ML model over 500 iterations under ideal (circle markers) and shadowed (star markers) settings. The shaded areas represent the variance, while the dashed line indicates the exact ground-state energy.

The simulation results are illustrated in Fig. R6. The inset illustrates the prediction error R_{predict} of

the proposed ML model h_s , evaluated using 500 test examples sampled from the same data distribution as the training examples. While the prediction error for the shadow dataset is slightly higher than that for the ideal dataset, both remain remarkably low, demonstrating the success of the learning process.

The main plot illustrates the optimization dynamics of the proposed ML model during the minimization of the surrogate loss in Eq. (16). Notably, for both the noisy and ideal cases, **the estimated energies after pretraining are comparable and closely approximate the optimal results**, which are -24.95 , -24.90 , and -25.99 for the ideal dataset, shadow dataset, and the exact value, respectively. Furthermore, for both the noisy and ideal cases, **the variance diminishes as the number of iterations increases**. These results highlight **the effectiveness of the proposed ML model in large-scale scenarios**, showcasing its potential to facilitate a wide range of practical use-case tasks.

Overview of broad applicability.

We conclude this reply by offering an overview of the broad applicability of the proposed ML model. For clarity, we first explain the rationale for focusing on toy examples in the original manuscript. We then elaborate on the potential of the proposed ML model for pretraining VQAs and explore its broader applicability to other quantum tasks.

Intention of the original manuscript. In the original manuscript, **our primary focus was to establish the theoretical foundation of the proposed ML model** and provide proof-of-principle demonstrations using toy examples. These examples were intended to illustrate the feasibility of the model in specific scenarios. To address the reviewer’s concerns, the revised manuscript incorporates additional large-scale examples, highlighting the model’s potential to address practical use-case tasks.

Pretraining VQAs. Beyond the newly added simulations on pretraining VQEs, we would like to emphasize that **there are ongoing research directions related to pretraining VQAs**, particularly in refining the third step of Alg. 1. For example, how to further enhance the performance of classical learning surrogates when only a limited number of interactions with quantum systems are permitted. Developing more advanced techniques for this step will further enhance the efficiency and applicability of the proposed framework, making it suitable for a broader range of quantum computing challenges.

Applicability to other tasks. Recall that the offline learning paradigm explored in this work is motivated by its potential to **significantly reduce quantum resource consumption**—a critical consideration given the scarcity of quantum resources in the foreseeable future. In this regard, the proposed ML model extends its utility beyond pretraining VQAs to **support various resource-intensive tasks**, such as enabling Hamiltonian simulation through Trotter decomposition, as outlined in the main text (Fig. 3(a)) and Supplementary Material K.3.

All in all, the proposed ML model not only provides a robust framework for pretraining VQAs but also holds promise for addressing diverse and resource-intensive quantum computing tasks, paving the way for broader advancements in the field.

We provide a brief summary of all the modifications made in the revised manuscript below.

Main text

1. (Lines 389-423, Page 6) We have included new simulation results demonstrating the use of the proposed model to pretrain a 50-qubit VQE for estimating the ground-state energy of the Transverse-Field Ising Model, as shown in Fig. 3(b), titled “[Pretraining performance of the proposed ML model on a 50-qubit TFIM](#)”.

Supplementary Material

1. (Lines 1129-1333, Page 39-47) Additional numerical simulations showcasing the use of the proposed ML model for pretraining VQAs, along with comparisons to other learning models, have been included in SM K.4, titled “[More simulation results of enhanced variational quantum algorithms by pretraining](#)”.

Other points

Reviewer Point P 2.3 — 1) The abstract is quite hard to read and, in my opinion, not up to the standards of Nat.Comm. From a first read, the problem looks contrived.

Reply: Thanks for the comment. To improve the clarity and readability of our manuscript, we have followed the reviewer's advice to update the abstract as follows.

Main text

1. (Lines 6-30, Page 1) [Efficiently predicting the behavior of large-scale quantum circuits is essential for advancing quantum computing and understanding quantum systems. Despite its importance, it remains elusive whether classical methods can effectively infer linear properties of quantum circuits from measured data generated under varying input configurations. Here we address this knowledge gap by analyzing the learnability of bounded-gate quantum circuits, showing that achieving a small prediction error requires sample complexity scaling linearly with the number of tunable gates, while runtime cost may scale exponentially. To address this challenge, we devise a learning framework that can balance accuracy and efficiency in many practical scenarios by leveraging classical shadow and compact representations of input configurations. Our proposal advances the exploration of quantum algorithms with practical utilities and enables resource-efficient quantum system certification. Numerical simulations validate our proposal across diverse applications, including quantum information processing, Hamiltonian simulation, and pretraining variational quantum algorithms, scaling up to 60 qubits.](#)

Reviewer Point P 2.4 — 2) From Eq. 3 to Thm.1 included, I was puzzled by the role of the observable O . For instance, in Theorem 1 the authors say that $\tilde{f}_T$ is an estimation of $f(x, O)$, but for which O ? That is clarified later in the paper when they introduce classical shadows, but at this introductory stage the setting was very confusing. Moreover, if my understanding is correct, the function class Eq.(3) should contain the fact that O can be different (with bounded norm). I guess the authors wanted to avoid unnecessary details here, but the end result is actually harder to understand and not intuitive.

Reply: We thank the reviewer for pointing out this issue. The reviewer's understanding of the role of the observable O is correct. Indeed, our intention in the original manuscript was to avoid unnecessary details at this introductory stage, but we recognize that this approach may have caused confusion. To address the reviewer's concerns and enhance the readability of the manuscript, we have made the following revisions.

Main text

1. (Lines 137-155, Page 2) To avoid confusion, we have removed redundant descriptions of the observable at the beginning of the problem setup and concentrated on elucidating the general purpose of classical learning paradigm. The revised words are:
 "We denote the total number of gates and CI gates in $U(\mathbf{x})$ as G and $G - d$, respectively. When $\rho(\mathbf{x})$ is measured by a known observable O , its incoherent dynamics are described the mean-value space, i.e., $f(\mathbf{x}, O) = \text{Tr}(\rho(\mathbf{x})O)$ with $\mathbf{x} \in [-\pi, \pi]^d$. The purpose of most classical learners is to obtain a learning model $h(\mathbf{x}, O)$ to predict $f(\mathbf{x}, O)$. The general paradigm is shown in Fig. 1(a). To accommodate the constraints of modern quantum devices, the training data are exclusively gathered by incoherent measurements. As such, the classical learner feeds $\mathbf{x}$ into the quantum circuit and receives $\tilde{f}_T(\mathbf{x}, O)$, as the estimation of $f(\mathbf{x}, O)$ incurred by T finite shots. By repeating this procedure with varied inputs, the classical learner constructs a training dataset $\mathcal{T} = \{\mathbf{x}^{(i)}, \tilde{f}_T(\mathbf{x}, O)\}_{i=1}^n$. Besides, the classical learner is kept unaware of the circuit layout details, except for the gate count G and the dimension of classical inputs d . Last, the inferred model $h(\mathbf{x}, O)$ should be offline, where the predictions for new inputs $\mathbf{x}$ is performed solely on the classical side."
2. (Lines 156-169, Page 3) After outlining the basic mechanism, we provide the mathematical formulation of the learning problem and specify the conditions for the quantum observables under consideration. The revised words are:
 "According to the above paradigm, the concept class of the mean-value space of bounded-gate circuit is $\mathcal{F} = \left\{ f(\mathbf{x}, O) = \text{Tr}(\rho(\mathbf{x})O) \middle| U \in \text{Arc}(U, d, G) \right\}$, where $\mathbf{x} \in [-\pi, \pi]^d$ and $\text{Arc}(U, d, G)$ refers to the set of all possible gate arrangements consisting of d RZ gates and $G - d$ CI gates. The hardness of learning $\mathcal{F}$ by classical learners is explored by separately assessing the required sample complexity and runtime complexity of training a classical model $h(\mathbf{x}, O)$ that attains a low *average* prediction error with $f(\mathbf{x}, O)$. Mathematically, the average performance of the learner is expected to satisfy $\mathbb{E}_{\mathbf{x} \sim \mathbb{D}_{\mathcal{X}}} |h(\mathbf{x}, O) - f(\mathbf{x}, O)|^2 \leq \epsilon$, where the classical input $\mathbf{x}$ is sampled from the distribution $\mathbb{D}_{\mathcal{X}}$. Here we suppose O consists of q local observables with a B -bounded norm, i.e., $O = \sum_{i=1}^q O_i$ and $\sum_l \|O_l\|_{\infty} \leq B$, and the maximum locality of $\{O_i\}$ is K . "

Reviewer Point P 2.5 — 3) Eq. A10 (SM). If my understanding is correct, the matrix D_2 should be renamed to D_{-1} for consistency with the index of the Fourier frequencies.

Reply: We thank the reviewer for pointing out this typo. We have fixed this issue in the revised version.

Supplementary Material

1. (Line 80, Page 3) ..., the PTM representation of $\mathfrak{R}_Z(x_j)$ gate is

$$\mathfrak{R}_Z(x_j) = \dots = D_0 + \cos(x_j)D_1 + \sin(x_j)D_{-1},$$

$$\text{where } \dots D_{-1} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Reviewer Point P 2.6 — 4) In appendix C (SM), the authors assume that results can be approximated by a normal distribution. Clearly it follows from central limit theorem, but is it possible to check when this approximation is accurate? Does it imply that the sample size must be larger than what is expected from Thm 1?

Reply: Thanks for the comment. To account for the role of the number of measurements T in the lower bound, Theorem 1 adopts the assumption that the binomial distribution of measurement outcomes can be approximated by a normal distribution. We would like to highlight that this assumption is **widely utilized in quantum computation and quantum information theory** [26, 27, 28, 29].

Regarding the concern about the accuracy of this approximation, the required sample size T depends on the probability p in the binomial distribution $B(T, p)$ and is independent of the qubit count N . A commonly used criterion is $Tp > 5$ and $T(1 - p) > 5$, e.g., when $p = 0.5$, $T = 10$ is sufficient [30]. Consequently, for most values of p , the required sample size T is relatively small and does not compromise the generality or validity of the results presented in Theorem 1.

We would like to emphasize that **the assumption can be omitted if the primary focus is solely on the number of RZ gates d** , without considering the measurement cost T . In this case, the upper bound of the mutual information $I(\mathbf{a}^*; \bar{\mathbf{a}})$ of the designed learning task (i.e., Eq. (C18) in Supplementary Material C) can be effectively derived using Holevo's theorem [31, 32], rather than relying on the KL divergence between the two normal distributions formed by the measurement outcomes. In this context, a similar lower bound of the sample complexity can still be obtained with a linear dependence on d .

To address the reviewer's concerns, we have made the following revisions.

Supplementary Material

1. (Lines 331-332, Page 10) This assumption is widely utilized in quantum computation and quantum information theory [26, 27, 28, 29].
2. (Lines 334-337, Page 10) Note that the assumption can be omitted if the primary focus is solely on the number of RZ gates d , without considering the measurement cost T . In this case, the upper bound of the mutual information $I(\mathbf{a}^*; \bar{\mathbf{a}})$ can be effectively derived using Holevo's theorem [31, 32], where a similar lower bound of the sample complexity can still be obtained with a linear dependence on d .

Reviewer Point P 2.7 — there is no code, the github repository is empty.

Reply: We have followed the reviewer's suggestion to make all codes related to this work open source. The source code for all numerical simulations is now available at [Github Repository](#).

Reviewer 3

Reviewer Point P 3.1 — The authors study the learnability of quantum circuits as a function of the number of RZ gates, assuming that the observer has classical access to the state. They prove several results, on both the sample and computational complexity of the task. They provide a lower bound that the sample complexity must be at least linear in d , by providing a simple example of circuits that require this many samples to learn. They also provide an upper bound using a fairly standard counting approach. Finally, they show that, even though the sample complexity is efficient, the computational complexity may not be, by adapting an example from another recent work. To address the computational complexity, the authors introduce a kernel method based on the Fourier transform of the signal with respect to the circuit parameters. They show that when gradients of the circuit are small, the Fourier transform can be truncated to contain only low-weight frequency vectors, which enables efficient learning. They provide discussions and numerical evidence that their approach works well in several cases.

I enjoyed reading the authors' work and believe that it is a valuable and high-quality contribution to our knowledge on learning quantum circuits.

Reply: We sincerely appreciate the reviewer's thoughtful and accurate summary, as well as the positive feedback on our work.

Reviewer Point P 3.2 — At the same time, I feel that the results of the work are not *especially* surprising, and the methods are not sufficiently novel, to warrant publication in Nature Communications.

Reply: We thank reviewer for the comment. While we respectfully disagree with the reviewer's perspective that our results lack sufficient novelty for publication in Nature Communications, we appreciate the opportunity to address these concerns.

We believe the **reviewer's misunderstanding regarding the novelty and contributions of our work arises from confusion about the main contributions and research focus of Ref. [1]** (Ref. [30] in the original manuscript), Pauli path simulators, and our study. In fact, **none of the prior works** address whether **classical learners** can efficiently predict the linear properties of bounded-gate quantum circuits or how to design an efficient algorithm capable of predicting a broad class of such circuits. In other words, prior works and our study consider **different settings** and collectively advance the understanding of learning bounded-gate quantum circuits.

- From a theoretical perspective, our work focuses on exploring the computational complexity of predicting the mean values of bounded-gate quantum circuits using **classical learners**, while Ref. [1] investigates the computational complexity of predicting bounded-gate states and unitaries using **quantum learners**. Together, these results significantly **enrich quantum learning theory** and deepen our understanding of the learnability of bounded-gate quantum circuits. Please refer to the reply of Point P 3.3 for elaborations.
- From an algorithmic perspective, the algorithm proposed in Ref. [1] is designed for predicting target unitaries or states, but it is limited in scope as it applies only to circuits with **a logarithmic number of quantum gates** relative to the number of qubits. Unlike Ref. [1], both our algorithm and Pauli path simulators aim to predict the mean values of bounded-gate quantum circuits. Remarkably, our proposal **differs from Pauli path simulators by addressing different scenarios**, which

broaden the range of applications in this field. For instance, our algorithm outperforms Pauli path simulators in cases where the circuit contains numerous Clifford gates or when the goal is to predict the mean values of a bounded-gate circuit with unknown noise. Please refer to the reply of Point P 3.4 for details.

We would like to further refer to the positive comments of Reviewers #1 and #2 to further emphasize the contributions of our work. Specifically, Reviewer #1 noted “This is an interesting, alternate approach to the problem raised in Reference [30], where it was studied for quantum learners.” Similarly, Reviewer #2 highlighted “The closest study, ..., is Ref.[30], but their algorithm requires a quantum machine for testing new data, while the new technique presented in this paper allows for purely classical inference.”

In the detailed responses below, we **systematically address each point raised by the reviewer, providing clarifications and additional insights**. We hope this will convincingly demonstrate the novelty, importance, and broad applicability of our work, reaffirming its suitability for publication in Nature Communications.

Reviewer Point P 3.3 — To elaborate on the first point: The fact that the sample complexity grows linearly in d is fairly similar to existing work (Ref 30) on the sample complexity as a function of the number of two-qubit gates. It is nice to know how this extends when we separate the Clifford from non-Clifford gates, but not too surprising that it works out to a similar scaling.

Reply: We thank the reviewer for the comment and the opportunity to clarify the contributions of our work. While the linear scaling of sample complexity with respect to d may appear similar to the results in Ref. [1], we emphasize that **our work extends beyond this observation** by addressing **different settings** and **introducing new insights** into the learnability of bounded-gate quantum circuits.

Table R4: **Summarization of theoretical results achieved by Ref. [1] and our work** (duplicate of Table R1). The terms ‘Q-learner’ and ‘C-learner’ refer to the quantum and classical learning paradigms, respectively. Besides, the terms ‘Samp’ and ‘Comp’ refer to the abbreviations of ‘sample complexity’ and ‘computational complexity’, respectively. C-function refers to the classical function with ν being the number of variables in the classical function (Task (iii)). The notations N and ϵ denote the number of qubits and the prediction error. In Ref. [1], the explored bounded-gate circuit is realized by $\hat{G}$ two-qubit gates, while in our work the explored bounded-gate circuit is composed of d RZ gates and $G - d$ Clifford gates. The notations B and T refer to the norm of the observable and the snapshots when collecting each training example, respectively. The markers ✓, ✗, ? separately represent ‘yes’, ‘no’, and ‘unknown’.

	Learn unitary (Task (i))	Learn state (Task (ii))	Learn C-function (Task (iii))	Learn observable (Task (iv))
Q-learner	Theorems 4&6, Ref. [1] Samp: $\begin{cases} \tilde{O}(\hat{G} \min\{\frac{1}{\epsilon^2}, \frac{\sqrt{2N}}{\epsilon}\}) \\ \Omega(\frac{\hat{G}}{\epsilon}) \end{cases}$ Comp: $e^{\Omega(\min\{\hat{G}, N\})}$ Algorithm: ✓	Theorems 1&2, Ref. [1] Samp: $\tilde{O}(\frac{\hat{G}}{\epsilon^2})$ Comp: $e^{\Omega(\min\{\hat{G}, N\})}$ Algorithm: ✓	Theorem 7, Ref [1] Samp: $\Omega(\frac{1}{\epsilon^\nu})$ Comp: $\tilde{\Omega}(\frac{1}{\epsilon^\nu})$ Algorithm: ✗	?
C-learner	?	?	✗	Theorem 1, this work Samp: $\begin{cases} \tilde{O}(\frac{B^2 d + B^2 N G}{\epsilon}) \\ \tilde{\Omega}(\frac{(1-\epsilon)d}{\epsilon T}) \end{cases}$ Comp: $O(\exp(N, d))$ Algorithm: ✓(Theorem 2)

Complementary to Ref. [1]. For clarity, let us first **distinguish the different settings** explored in Ref. [1] and our work. A detailed explanation about the complementary role is provided in the reply to Point P 1.1. For convenience, we **summarize the key results** here and include Table R1, relabeled as Table R4 in this reply.

As shown in Table R4, Ref. [1] focuses on using **quantum learners** to predict **bounded-gate pure states and unitaries**, whereas our work centers on using **classical learners** to predict the **mean values of bounded-gate quantum circuits**. Given these distinct objectives, the results of the two works are **not directly comparable**. An immediate observation is that our work is the **first** to analyze the learnability of bounded-gate quantum circuits by classical learners and to propose efficient classical algorithms for predicting a broad class of bounded-gate quantum circuits.

We would also like to address the reviewer’s characterization of the similar scaling behavior between our results and those in Ref. [1] as “*not too surprising*.” **On the contrary, the similar scaling actually provides crucial new insights into the learnability of bounded-gate quantum circuits.** Again, as elaborated in the Reply to Point P 1.1, it is essential to explore both classical and quantum learners to fully understand the computational complexities of learning bounded-gate quantum circuits. This is because the relevant results have critical implications for practical applications and the utility of quantum computers:

- For the same task, if classical learners can achieve **similar performance** with that of quantum learners, they offer undeniable advantages in terms of resource efficiency.
- For the same task, if quantum learners exhibit provable **computational advantages** over classical learners, it underscores the practical utility of quantum computing for that specific application.

Thus, exploring the interplay between classical and quantum learners is not only insightful but also foundational to advancing quantum learning theory and its applications. For example, a recent study [2] provides a compelling example of how computational capabilities differ between classical and quantum learners, verifying the necessity of quantum learners to achieve exponential runtime speedups in predicting the properties of ground states.

By combining the computational complexities achieved in Ref. [1] and our work under different settings, we reach **two key observations about the learnability of bounded-gate quantum circuits** (please refer to the section titled ‘*Insight I: Computational complexity analysis*’ in Point P 1.1 for the detailed explanations).

Observation I. Despite the **decreasing levels of information gain** from learning an unknown unitary, to an unknown state, and finally to unknown expectation values, we observe **a consistent scaling behavior for both classical and quantum learners**. This linear scaling behavior with respect to the number of tunable gates across different learners (i.e., quantum versus classical) and tasks (i.e., learning bounded-gate unitaries, states, and expectation values) suggests **a consistent linear-in- d (or linear-in- $\hat{G}$) sample complexity for classical learners in tasks of predicting unitaries and states generated by bounded quantum gates, and for quantum learners in the task of predicting expectation values of bounded-gate quantum circuits**.

Observation II. The computational challenges faced by quantum learners in learning bounded-gate unitaries and pure states, combined with the computational limitations of classical learners in predicting expectation values of bounded-gate quantum circuits, offer valuable insights into the unresolved questions highlighted in Table R4. For example, the limitations of quantum learners

in learning bounded-gate unitaries and pure states imply that **classical learners cannot achieve efficiency in these tasks either**. This is because from an information-theoretic perspective, classical learners cannot extract more useful information than quantum learners.

Technical contributions. In addition to highlighting the complementary roles of Ref. [1] and our work, we would like to restate our technical contributions. Due to the **distinct learning paradigms** (classical versus quantum learners) and **tasks** (predicting expectation values versus unitaries or pure states), our analysis **necessitated the development of novel techniques** to reach the computational complexity. For instance, to derive the sample complexity lower bound, we designed a tailored learning task and employed advanced statistical learning tools to quantify how the number of measurements T impacts the prediction performance. These results may hold independent interests and contribute broadly to the field.

Implications. Lastly, we underscore that the sample efficiency and computational hardness of learning mean values of bounded-gate quantum circuits by classical learners strongly motivated the development of an efficient kernel method applicable across a broad range of scenarios (i.e., Theorem 2 in our manuscript). This differs from Ref. [1], which primarily focuses on theoretical scaling. Instead, our work extends to practical applications, such as enabling resource-efficient pretraining for VQAs. This practical utility is further demonstrated through the additional simulations (up to 50 qubits) presented in the Reply to Point P 2.2.

We provide a brief summary of all the modifications made in the revised manuscript as follows.

Supplementary Material

1. (Lines 738-838, Pages 27-30) All explanations related to the complementation between Ref. [1] and our work have been appended to SM J.1, titled “**SM J.1 complementation with Ref. [1]**”.
2. (Lines 1129-1333, Pages 39-47) Additional numerical simulations showcasing the potentials of the proposed ML model for pretraining VQAs, along with comparisons to other learning models, have been included in SM K.4..

Reviewer Point P 3.4 — The computational complexity is a more interesting question. In this regard, I believe that the authors truncated frequency method may be useful, but it also bears an extremely close similarity to existing classical algorithm for the same class of circuits (e.g. the LOWESA algorithm or several more recent papers on Pauli path simulations). Thus, I view this contribution more as an adaptation of an existing approach to a new context, rather than a new method in itself. Moreover, a downside of this approach is that its success is linked to the fact the quantum circuit can be classically simulated, using the same truncated frequency expansion. This seems like it will limit the impact of the authors approach in applications.

Reply: We thank the reviewer for the comment. While we respectfully disagree with the perspective that our contribution is merely an adaptation of existing approaches, we appreciate the opportunity to clarify the novelty and impact of our work. Below, we address the reviewer’s concerns by **first elaborating on the key differences between the problem explored in our work and existing classical algorithms such as LOWESA**. We then highlight that **this fundamental difference** emphasizes the **unique benefits and advancements our protocol** offered in many practically relevant scenarios. Finally, we

present **numerical evidence** that demonstrates the superiority and broader applicability of our proposed method compared to LOWESA.

Difference between our work and existing methods. The primary distinction between our work and existing classical algorithms lies in their *respective purposes*. Specifically, **our work focuses on developing a shadow state predictor** to predict the mean values of a bounded-gate quantum circuit for *new rotation angles and observables*. In contrast, **existing classical algorithms aim to efficiently simulate the mean values** of a bounded-gate quantum circuit for varying rotation angles while *keeping the observables fixed*. This fundamental difference—**prediction versus estimation**—necessitates distinct theoretical analyses and underscores the complementary roles of these two approaches. In the following, we elaborate on how the introduction of the *Pauli transfer matrix* (PTM) and truncation methods within the learning paradigm enables our work to achieve several novel insights that are not attainable with existing classical algorithms.

- The employment of PTM. PTM is a widely used tool in quantum computing, with applications ranging from gate set tomography [33] to quantum learning theory [34] and the analysis of quantum circuit dynamics [35, 36, 37, 38]. In the context of simulating the mean values of bounded-gate quantum circuits, PTM is typically combined with Pauli path simulations to compute the contribution of each path [35, 36]. More specifically, given an N -qubit quantum circuit $U(\mathbf{x})$ composed of d RZ gates and $G - d$ Clifford gates, the expectation value of the given observable O under the PTM representation yields

$$\text{Tr}(\rho(\mathbf{x})O) = \text{Tr}(U(\mathbf{x})(|0\rangle\langle 0|)^{\otimes N}U(\mathbf{x})^\dagger O) = \sum_{\omega} \Phi_{\omega}(\mathbf{x}) \text{Tr}(\rho_{\omega}O), \quad (32)$$

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

$$\Phi_{\omega}(\mathbf{x}) = \prod_{i=1}^d \begin{cases} 1 & \text{if } \omega_i = 0 \\ \cos(\mathbf{x}_i) & \text{if } \omega_i = 1 \\ \sin(\mathbf{x}_i) & \text{if } \omega_i = -1 \end{cases}, \quad (33)$$

and $\text{Tr}(\rho_{\omega}O)$ refers to the expectation value of the *purely-Clifford circuit* for the path indexed by ω . For Pauli path simulators, the expectation value for a portion of $\{\omega\}$ should be **explicitly computed on the classical hardware**.

However, our work adopts **a fundamentally different approach to leveraging PTM**. Instead of computing nontrivial $\text{Tr}(\rho_{\omega}O)$ like Pauli path simulators, we utilize PTM to derive the explicit formula of the bounded-gate quantum state with respect to the input $\mathbf{x}$, i.e.,

$$\rho(\mathbf{x}) = U(\mathbf{x})(|0\rangle\langle 0|)^{\otimes N}U(\mathbf{x})^\dagger = \sum_{\omega} \Phi_{\omega}(\mathbf{x})\rho_{\omega}, \quad (34)$$

where $\Phi_{\omega}(\mathbf{x})$ and ρ_{ω} follow the same definitions in Eq. (32). This explicit form inspires us to design the shadow state predictor, i.e.,

$$\hat{\sigma}_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \kappa_{\Lambda}(\mathbf{x}, \mathbf{x}^{(i)}) \tilde{\rho}_T(\mathbf{x}^{(i)}) \text{ with } \kappa_{\Lambda}(\mathbf{x}, \mathbf{x}^{(i)}) = \sum_{\omega, \|\omega\|_0 \leq \Lambda} 2^{\|\omega\|_0} \Phi_{\omega}(\mathbf{x}) \Phi_{\omega}(\mathbf{x}^{(i)}), \quad (35)$$

where Λ denotes the truncation value, $\{\mathbf{x}^{(i)}\}_{i=1}^n$ refers to the n training examples, and $\tilde{\rho}_T(\mathbf{x}^{(i)})$ denotes the classical shadow of the state $\rho(\mathbf{x}^{(i)})$ with T snapshots.

In the proof of Theorem 2 (Lemma F.2 in the manuscript), we prove that given any new input x , the expectation value of the shadow state predictor $\hat{\sigma}_n(x)$ matches the truncated target quantum state $\rho_\Lambda(x)$ for any given input x , i.e.,

$$\mathbb{E}[\hat{\sigma}_n(x)] = \rho_\Lambda(x) \equiv \sum_{\omega, \|\omega\|_0 \leq \Lambda} \Phi_\omega(x) \rho_\omega.$$

This result represents a key technical contribution of our work, setting our approach apart from the use of PTM in Pauli path simulators. Specifically, our work leverages PTM to derive the explicit formula of the concept class with respect to the input x and subsequently design the shadow state prediction model $\hat{\sigma}_n(x)$, whereas Pauli path simulators such as LOWESA utilize PTM to compute values for each nontrivial path.

- Truncation strategy. The low-weight truncation scheme, as with PTM, is also a widely utilized strategy in quantum computing. However, **the approach to truncation and the elements being truncated differ substantially across applications**, leading to distinct theoretical results and heuristic performances. For Pauli path simulators, low-weight truncation is applied to the frequencies ω (e.g., LOWESA) or the weights of Pauli operators to achieve computational efficiency while maintaining a low **estimation error**. In contrast, in our work, low-weight truncation is applied to the frequencies to ensure computational efficiency while **minimizing the prediction error**. This fundamental distinction in purpose underscores that **prior results on Pauli path simulators offer no direct implications for learning protocols**, necessitating an entirely new theoretical framework.

Recall that Theorem 1 in our manuscript establishes that **no efficient learning model can predict all bounded-gate quantum circuits**. Consequently, the most critical challenge lies in **identifying specific classes of bounded-gate quantum circuits** that can be **efficiently learned**—a question that classical simulators do not need to address. To tackle this, we establish a connection between the **gradient norm** of bounded-gate circuits and **the truncation error**. Specifically, as shown in Lemma F.1 in our manuscript, we prove that when the gradient norm of a given quantum circuit is well-bounded, the proposed learning model can achieve efficient predictions with low prediction error. This novel result, which is absent in LOWESA, highlights the **uniqueness of our approach** and opens new avenues for developing efficient learning models tailored to predicting the mean values of bounded-gate quantum circuits.

Unique benefits of our learning model in two scenarios. As explained previously, our method focuses on the **shadow state prediction**, while LOWESA centers on **Heisenberg simulations**. These differences define the complementary strengths of the two approaches, with our model demonstrating notable advantages in the following two scenarios.

- Circuits with many Clifford gates. Classical Pauli path simulators, such as LOWESA, must explicitly simulate Clifford gates, which, although theoretically efficient ($\mathcal{O}(N^2)$ complexity with N being the qubit count), can be **computationally expensive in terms of wall-clock time**. In contrast, our learning model is **insensitive to the number of Clifford gates**, as these gates do not affect the learning process. We validate this advantage through numerical simulations subsequently, demonstrating superior performance in this regime.

- (ii) Circuits with unknown system noise. Classical simulators, including both tensor network simulators and Pauli path simulators, rely on assumptions about noise strength when simulating the dynamics of noisy quantum circuits. However, these assumptions generally **do not align with the actual noise present in the system**. In contrast, a learning-based approach (such as the extension of our model) can **automatically capture and adapt to inherent system noise** by leveraging the dataset collected from the target quantum system, offering a more flexible and accurate alternative.

Numerical evidence. To further substantiate the advantages of our approach in Scenario (i) explained above, we conduct the numerical simulations to compare the wall-clock runtime of our method and LOWESA using open-source implementations.

The problem setup is as follows. The number of qubits is fixed to be $N = 32$. The employed ansatz $U(x)$ is an adaptation of the Trotterized transverse-field Ising Hamiltonian circuit, designed with a corresponding bricklayer topology. A visualization of this circuit is provided in Fig. R10. Specifically, the ansatz is designed with 50,000 layers. All angles of the R_{ZZ} gates and R_X gates are set to zero, except for the 16th R_{ZZ} gate in the first layer, which is assigned arbitrary angles. As such, the explored ansatz is tailored to include **a single tunable parameter**. The initialized state is $|0\rangle^{\otimes N}$ and the observable employed takes the form as $O = \mathbb{I}^{\otimes 15} \otimes Z \otimes \mathbb{I}^{\otimes 16}$.

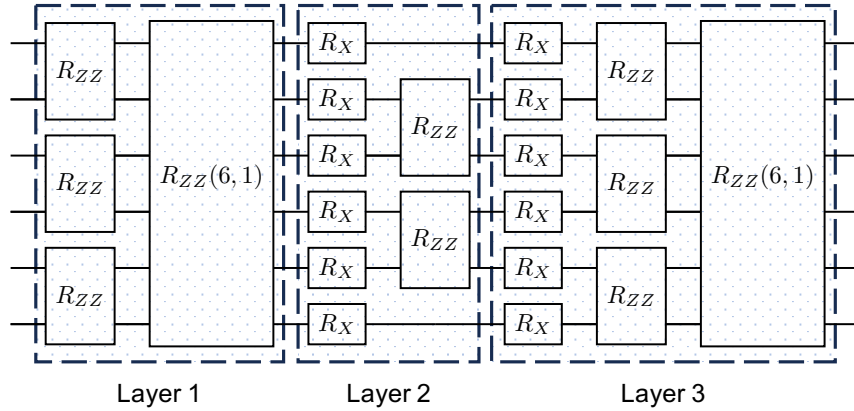


Figure R10: **The employed ansatz with the bricklayer topology.** The figure visualizes the implementation of the employed ansatz $U(x)$ with three layers and $N = 6$ qubits. Each layer consists of alternating R_{ZZ} and R_X gates. The R_{ZZ} gate labeled $R_{ZZ}(6,1)$ in the odd layers connects the 6-th and 1st qubits. For circuits with more than three layers, Layers 2 and 3 are alternated to construct deeper circuits while preserving the bricklayer topology. Specifically, the gate arrangement in Layer 2 is repeated for all even-numbered layers, while the gate arrangement in Layer 3 is repeated for all odd-numbered layers.

The hyperparameter settings for LOWESA are as follows. The implementation is based on the open-source PauliPropagation.jl package available on GitHub. The truncation threshold for the weights is set to 5, and the minimum coefficient magnitude is set to 4. We provide 10 input angles, i.e., $x \in \{-\pi, -0.8\pi, -0.6\pi, -0.4\pi, -0.2\pi, 0, 0.2\pi, 0.4\pi, 0.6\pi, 0.8\pi\}$, to the ansatz $U(x)$ and use LOWESA to estimate the corresponding expectation values.

The hyperparameter settings of the proposed learning model h_s in Eq. (4) are as follows. The number of training examples is set as 1000. For simplicity, we consider the ideal scenario such that the infinite measurements are adopted to construct the dataset. Given access to the training dataset, we

train the learning model by setting the truncation threshold Λ as 2. Once training is completed, we use the same 10 input angles with LOWESA as test examples to evaluate the prediction performance and inference time.

Table R5: **Performance and wall-clock time of LOWESA and the proposed learning model.** The notation ‘ $a \pm b$ ’ indicates that the mean value is a and the variance is b . The wall-clock time for the proposed learning approach focuses on the inference stage.

	LOWESA	Our work
Estimation (Prediction) error	0 ± 0.00	0 ± 0.0002
Wall-clock time (seconds)	2.003 ± 0.0107	$8.6808 \times 10^{-5} \pm 2.202 \times 10^{-8}$

The performance and computational efficiency between LOWESA and the proposed learning model are summarized in Table R5. Regarding the prediction (estimation) error, both approaches achieve comparable accuracy, with the proposed method showing a slight variance of 0.0002. However, in the wall-clock time, **our proposed learning model demonstrates a dramatic reduction in inference time compared to LOWESA**, achieving a reduction by **approximately 2.3×10^4 times**. This distinction underscores a scenario where the learning model surpasses classical simulators, emphasizing their complementary roles.

To improve the clarity and readability of our manuscript, we have conducted the following modifications in the updated version.

Main text

1. (Lines 331-332, Page 5) We append the words “ **In SM J, we show how our proposal differs from Pauli path simulators**”.

Supplementary Material

1. (Lines 839-896, Pages 30-31) All explanations in this reply are appended to SM J.2 titled “**Connections and differences with Pauli path simulators**”.
2. (Lines 1335-1354, Pages 47-48) All newly added numerical simulations are appended to SM K.5 titled “**Comparison with LOWESA**”.

Reviewer Point P 3.5 — As a small comment on presentation, I would have liked to see a bit more discussion on the proof ideas inside the main text. I understand the authors have many results to present in a short space, but it is nice to give people an idea of how the results are shown without needing to dig through the Appendix.

Reply: Thanks for the comment. We have followed the reviewer’s suggestion to include more key insights from the proofs in the main text to enhance the clarity of the theoretical results. To manage space constraints, these newly added details have been incorporated into the Methods section.

The key modifications are summarized below.

1. (Lines 497-579, Page 7) We append the proof sketches of Theorem 1 and Theorem 2 in Methods.

References

- [1] 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.
- [2] Riccardo Molteni, Casper Gyurik, and Vedran Dunjko. Exponential quantum advantages in learning quantum observables from classical data. *arXiv preprint arXiv:2405.02027*, 2024.
- [3] G. Vidal and C. M. Dawson. Universal quantum circuit for two-qubit transformations with three controlled-not gates. *Phys. Rev. A*, 69:010301, Jan 2004.
- [4] Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge University Press, 2010.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] Xinbiao Wang, Yuxuan Du, Zhuozhuo Tu, Yong Luo, Xiao Yuan, and Dacheng Tao. Transition role of entangled data in quantum machine learning. *Nature Communications*, 15(1):3716, 2024.
- [10] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- [11] 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.
- [12] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3):032331, 2019.

- [13] Amit Dutta, Gabriel Aeppli, Bikas K. Chakrabarti, Uma Divakaran, Thomas F. Rosenbaum, and Diptiman Sen. *Quantum Phase Transitions in Transverse Field Spin Models: From Statistical Physics to Quantum Information*. Cambridge University Press, 2015.
- [14] James Dborin, Vinul Wimalaweera, Fergus Barratt, Eric Ostby, Thomas E O’Brien, and Andrew G Green. Simulating groundstate and dynamical quantum phase transitions on a superconducting quantum computer. *Nature Communications*, 13(1):5977, 2022.
- [15] Elijah Pelofske, Andreas Bärtshi, and Stephan Eidenbenz. Short-depth qaoa circuits and quantum annealing on higher-order ising models. *npj Quantum Information*, 10(1):30, 2024.
- [16] Alexander Miessen, Daniel J Egger, Ivano Tavernelli, and Guglielmo Mazzola. Benchmarking digital quantum simulations above hundreds of qubits using quantum critical dynamics. *PRX Quantum*, 5(4):040320, 2024.
- [17] 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.
- [18] Sam McArdle, Suguru Endo, Alán Aspuru-Guzik, Simon C Benjamin, and Xiao Yuan. Quantum computational chemistry. *Reviews of Modern Physics*, 92(1):015003, 2020.
- [19] James I Colless, Vinay V Ramasesh, Dar Dahlen, Machiel S Blok, Mollie E Kimchi-Schwartz, Jarrod R McClean, Jonathan Carter, Wibe A de Jong, and Irfan Siddiqi. Computation of molecular spectra on a quantum processor with an error-resilient algorithm. *Physical Review X*, 8(1):011021, 2018.
- [20] Pei Liu, Ruixia Wang, Jing-Ning Zhang, Yingshan Zhang, Xiaoxia Cai, Huikai Xu, Zhiyuan Li, Jiaxiu Han, Xuegang Li, Guangming Xue, et al. Performing su (d) operations and rudimentary algorithms in a superconducting transmon qudit for d= 3 and d= 4. *Physical Review X*, 13(2):021028, 2023.
- [21] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [22] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.
- [23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [24] 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.
- [25] 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.
- [26] Ke Li. Second-order asymptotics for quantum hypothesis testing. *The Annals of Statistics*, 42(1):171 – 189, 2014.

- [27] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, 2016.
- [28] Giacomo Torlai, Guglielmo Mazzola, Giuseppe Carleo, and Antonio Mezzacapo. Precise measurement of quantum observables with neural-network estimators. *Physical Review Research*, 2(2):022060, 2020.
- [29] Alistair WR Smith, Kiran E Khosla, Chris N Self, and MS Kim. Qubit readout error mitigation with bit-flip averaging. *Science advances*, 7(47):eabi8009, 2021.
- [30] George EP Box, William H Hunter, Stuart Hunter, et al. *Statistics for experimenters*, volume 664. John Wiley and sons New York, 1978.
- [31] Alexander Semenovich Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problemy Peredachi Informatsii*, 9(3):3–11, 1973.
- [32] Ingemar Bengtsson and Karol Życzkowski. *Geometry of quantum states: an introduction to quantum entanglement*. Cambridge university press, 2017.
- [33] Daniel Greenbaum. Introduction to quantum gate set tomography. *arXiv preprint arXiv:1509.02921*, 2015.
- [34] Matthias C Caro. Learning quantum processes and hamiltonians via the pauli transfer matrix. *ACM Transactions on Quantum Computing*, 5(2):1–53, 2024.
- [35] Manuel S Rudolph, Enrico Fontana, Zoë Holmes, and Lukasz Cincio. Classical surrogate simulation of quantum systems with lowesa. *arXiv preprint arXiv:2308.09109*, 2023.
- [36] 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.
- [37] Guillermo González-García, J Ignacio Cirac, and Rahul Trivedi. Pauli path simulations of noisy quantum circuits beyond average case. *arXiv preprint arXiv:2407.16068*, 2024.
- [38] Thomas Schuster, Chao Yin, Xun Gao, and Norman Y Yao. A polynomial-time classical algorithm for noisy quantum circuits. *arXiv preprint arXiv:2407.12768*, 2024.