

Learning k -Body Hamiltonians via Compressed Sensing

Muzhou Ma¹, Steven T. Flammia, John Preskill², and Yu Tong³

Abstract—We study the problem of learning a k -body Hamiltonian with M unknown Pauli terms that are not necessarily geometrically local. We propose a protocol that learns the Hamiltonian to precision ϵ with total evolution time $\mathcal{O}(M^{1/2+1/p}/\epsilon)$ up to logarithmic factors, where the error is quantified by the ℓ^p -distance between Pauli coefficients. Our learning protocol uses only single-qubit control operations and a GHZ state initial state, is non-adaptive, is robust against SPAM errors, and performs well even if M and k are not precisely known in advance or if the Hamiltonian is not exactly M -sparse. Methods from the classical theory of compressed sensing are used for efficiently identifying the M terms in the Hamiltonian from among all possible k -body Pauli operators. We also provide a lower bound on the total evolution time needed in this learning task, and we discuss the operational interpretations of the ℓ^1 and ℓ^2 error metrics. In contrast to most previous works, our learning protocol requires neither geometric locality nor any other relaxed locality conditions.

Index Terms—Hamiltonian learning, compressed sensing, quantum sensing, convex optimization.

I. INTRODUCTION

HAMILTONIAN learning is a fundamental problem in quantum physics, crucial for understanding and controlling quantum systems. Accurate models of Hamiltonians are

needed for tasks ranging from quantum simulation to quantum error correction, where precise knowledge of interactions allows for better optimization and control. It is also a fundamental problem in computer science, where the classically analogous task of learning undirected graphical models is central in many models of machine learning.

Learning Hamiltonians, particularly in large systems with many-body interactions, is complicated by the fact that dynamical access to the Hamiltonian naturally leads to entangled quantum states. Inferring a general unknown Hamiltonian from measurements in the computational basis is therefore nontrivial. Similarly, access via Gibbs states introduces challenges in state preparation and post-processing of measurement results. Prior work, which we review below, has nonetheless shown that an unknown Hamiltonian can be learned in polynomial time from its dynamics [1], [2], [3], [4], [5], [6] or from samples of a Gibbs state when the bipartite graph connecting qubits to interactions has bounded degree [1], [7], [8].

The classical theory of compressed sensing [9], [10], [11], [12], [13], [14], [15], [16] is a promising avenue for improving Hamiltonian learning and extending it to the setting where the interactions are sparse, or approximately sparse, in a known basis. It also has many applications in other types of learning tasks involving quantum states or processes [17], [18], [19], [20], [21], [22], [23], [24]. Classically, compressed sensing efficiently solves the following problem: given a linear measurement matrix $A : \mathbb{C}^n \rightarrow \mathbb{C}^m$ and an unknown vector $x \in \mathbb{C}^n$ which is promised to be s -sparse in a given basis, reconstruct x from $y = Ax$ when $m = \mathcal{O}(s \log n)$. At first glance this seems impossible for at least two reasons. First, the well-known Nyquist theorem states that learning a signal with bandwidth n requires measuring $\Omega(n)$ frequencies. This however does not take into account the sparsity of the underlying signal. But second, naively taking the sparsity into account quickly runs into fundamental computational bottlenecks, as finding the sparsest vector consistent with affine constraints is NP-hard [25].

Under surprisingly mild conditions on the measurement matrix A , compressed sensing overcomes these two obstacles. Moreover, the compressed-sensing estimate of the signal is robust to measurement noise due to sampling, robust to the signal being only “near” to a sparse signal, and is even efficient to compute via a reduction to an easy convex optimization problem. It is therefore natural to hope that compressed sensing can offer a powerful approach to reconstructing sparse Hamiltonians.

Received 27 August 2025; revised 10 April 2026; accepted 25 July 2026. Date of publication 6 August 2026; date of current version 17 September 2026. The work of Muzhou Ma was supported by Caltech Summer Undergraduate Research Fellowship (SURF). The work of John Preskill was supported by U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Grant DE-NA0003525 and Grant DE-SC0020290; in part by U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Quantum Systems Accelerator; and in part by the National Science Foundation under Grant PHY-1733907. The work of Yu Tong was supported by U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Grant DE-SC0020290. (Corresponding author: Muzhou Ma.)

Muzhou Ma was with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. He is now with the Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: mma2@caltech.edu).

Steven T. Flammia is with the Department of Computer Science, Virginia Tech, Alexandria, VA 22314 USA, and also with Phasecraft Inc., Washington, DC 20001 USA.

John Preskill is with the Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, CA 91125 USA, and also with the AWS Center for Quantum Computing, Pasadena, CA 91125 USA.

Yu Tong is with the Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, CA 91125 USA, and also with the Department of Mathematics and the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA.

Communicated by X. Wang, Associate Editor for Quantum.

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TIT.2026.3720996>.

Digital Object Identifier 10.1109/TIT.2026.3720996

A. Main Results

We consider the class of n -qubit k -body Hamiltonians, where all couplings are written in the Pauli basis and involve Paulis of weight at most k with $k = \mathcal{O}(1)$. Let $M \leq \sum_{l=0}^k 3^l \binom{n}{l} = \mathcal{O}(n^k)$ be the total number of nonzero couplings in the Hamiltonian. Write $H = \sum_{a=1}^M \mu_a P_a$ for the explicit unknown Hamiltonian, where the coefficients μ_a are assumed to be normalized to lie in the interval $[-1, 1]$. Our goal is to output the unknown Paulis P_a and estimates $\hat{\mu}_a$ of the coefficients that are accurate in ℓ^p error, defined as $\|\hat{\mu} - \mu\|_p = \left(\sum_{a=1}^M |\hat{\mu}_a - \mu_a|^p \right)^{1/p}$ for suitable p .

In what follows, we use $\tilde{\mathcal{O}}(f)$ to represent $\mathcal{O}(f \text{polylog}(n, M, 1/\delta, 1/\epsilon, 2^k))$ where δ is the failure probability and ϵ is the precision that we want to achieve.

Theorem 1 (Informal version of Theorem 9): There exists a learning protocol that uses $N = \tilde{\mathcal{O}}(M)$ independent non-adaptive experiments to learn, with probability at least $1 - \delta$, an estimate $\hat{\mu}$ of all coefficients with weight $\leq k$ that has ℓ^p error ($1 \leq p \leq 2$) at most ϵ , and uses a total evolution time

$$T = \tilde{\mathcal{O}}\left(\frac{M^{1/p+1/2}}{\epsilon}\right).$$

The protocol is robust to a constant amount of state preparation and measurement (SPAM) noise.

In the above, by *total evolution time*, we mean the total time needed to evolve with the unknown Hamiltonian in all experiments. Our method is also robust to modeling errors (i.e., the Hamiltonian may not be exactly k -local or have more than M terms), as discussed in Section V-E. The operational interpretations of the ℓ^p error metric for $p = 1$ and 2 are discussed in Section VII and correspond to notions of worst- and average-case error, respectively. It is notable that our algorithm nearly achieves Heisenberg-limited scaling, $\mathcal{O}(1/\epsilon)$, in the total evolution time. Also importantly, the classical computation required by our algorithm is polynomial. In fact, our estimator reduces to a convex optimization of the form

$$\underset{\mathbf{x}}{\text{minimize}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \alpha. \quad (1)$$

In Theorem 10, we give an explicit bound on the computational resources required to obtain an (approximately) optimal solution to Eq. (1). As our analysis of this optimization is self-contained (it does not rely on reducing Eq. 1 to a semidefinite program) and appears to be new, it may be of interest to a wider community.

We also prove a lower bound for the total evolution time for the case of ℓ^1 error.

Theorem 2 (Informal version of Theorem 11): Suppose a learning algorithm estimates $\hat{\mu}$ for the unknown Hamiltonian $H = \sum_{a=1}^M \mu_a P_a$ with $\|\hat{\mu} - \mu\|_p \leq \epsilon$ for $p \in [1, 2]$ in expectation value over experimental outcomes. Then

$$T = \Omega\left(\frac{M^{1/p}}{\epsilon \log(1/\gamma)}\right). \quad (2)$$

Here γ is the amount of measurement error (see (77)).

Interestingly, our lower bound holds for learning n -qubit Hamiltonians with the M terms *fixed* and having *known*

support. This naturally implies a lower bound for learning n -qubit M -term Hamiltonians without prior knowledge of which M terms are in the Hamiltonian. This suggests that parameter estimation, rather than support identification, may be the primary obstacle to improving the time complexity for Hamiltonian learning.

Remark 1: Comparing with the upper bound $\tilde{\mathcal{O}}(M^{1/2+1/p}/\epsilon)$, the gap is $\tilde{\mathcal{O}}(M^{1/2})$ for all $p \in [1, 2]$ simultaneously. We discuss the implications in the Discussion section and on how the lower bound is tight for the ℓ^1 -metric, as it should hold for the general case which is limited by some simpler cases, in contrast the upper bound provided by the algorithm if for general k -body Hamiltonians, which should work for the worst case.

B. Techniques

The lack of locality presents a considerable challenge to the Hamiltonian learning problem, as it prevents us from decoupling the learning problem into small problems that can be efficiently solved. We are therefore compelled to deal with the entire quantum system as a whole, and new techniques are needed to circumvent the exponential cost that usually comes from such problems. Compressed sensing utilizes sparsity to greatly reduce the complexity of reconstructing a high-dimensional vector, but prior to this work, it was far from clear how it can be applied to Hamiltonian learning. Although the Hamiltonian coefficients can be collected into a sparse vector, a generic observable does not depend on these coefficients linearly. The non-commutativity of the terms presents additional difficulties.

One key observation that led to this work is that we can apply compressed sensing to learn a completely commuting Hamiltonian, i.e., a Hamiltonian whose Pauli terms commute *on each qubit*. For such a Hamiltonian, the eigenvalue differences depend linearly on the coefficients, and we can estimate these differences efficiently as the eigenstates are all product states.

We then extend this method to general k -body Hamiltonians with M terms using Hamiltonian reshaping. Hamiltonian reshaping was used to decouple the dynamics and choose local bases in [5], but here we use it to obtain an effective Hamiltonian that is diagonal in a global basis set of our choosing. This effective Hamiltonian can then be learned using compressed sensing. However, this procedure will only allow us to learn the terms that are diagonal in the given basis, and multiple choices of bases are needed to cover all terms in the Hamiltonian. We then choose the basis randomly, and this allows us to cover all k -body terms with mild overhead. This idea comes from randomized measurement [26], [27], [28], but to the best of our knowledge this is the first time it has been applied to a dynamical setting.

C. Prior Work

Hamiltonian learning has been the focus of many previous works [1], [2], [3], [4], [5], [6], [17], [26], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47]. Most early works focused on achieving

the Heisenberg-limited scaling, i.e., estimating parameters of the Hamiltonian to precision ϵ with $\mathcal{O}(1/\epsilon)$ cost [48], for small quantum systems, or scalability for large quantum systems but without achieving the Heisenberg-limited scaling. Recently, [5] proposed a method to learn an n -qubit many-body Hamiltonian with total evolution time $\mathcal{O}(\log(n)/\epsilon)$, where ϵ is the ℓ^∞ -error on the coefficients, when the Hamiltonian is low-intersection, i.e., the number of Pauli terms acting on a qubit is a constant that is independent of the system size, and the terms in the Hamiltonian are known. This condition can be seen as a slight relaxation of geometric locality. The main technique used in [5] is Hamiltonian reshaping, which inserts random Pauli operators during the experiment to obtain an effective Hamiltonian that is easy to learn and preserves some of the coefficients. Later works obtained similar results for other control models [40] and in bosonic and fermionic scenarios [41], [42], [43]. These works all assume that the Hamiltonian terms are known and we only need to learn the coefficients.

The first work to consider achieving the Heisenberg-limited scaling for Hamiltonians with unknown terms is [45], the authors of which use a bootstrapping approach (as was also done in [40]) to gradually refine the estimation of the Hamiltonian. A rough estimate of the Hamiltonian is used in the time evolution to partially cancel out the actual Hamiltonian, so that a finer estimate of the residue can be obtained. This approach necessitates implementing multi-qubit operations during the time evolution which are adapted based on previous measurement results.

GHZ states and compressed sensing have been used previously to learn dephasing noise with long-range correlation and sparsity [17], but the dissipative nature of the noise makes it impossible to achieve the Heisenberg limit. Compared to the setup in [17], we also need to deal with the non-commutativity of the Hamiltonian terms.

Our learning protocol takes several important steps beyond prior works. First, it only applies single-qubit operations during time evolution, which is an enormous advantage for near-term implementation of algorithms that may have practical importance like Hamiltonian learning. Second, we do not require any locality assumptions beyond there being M terms of bounded strength, each of which is k -body. Lastly, our algorithm is completely non-adaptive. Compared to [45] in particular, the method in this work uses less total evolution time and number of experiments, which we will illustrate using several physically relevant examples below.

D. Applications

Sparse Sachdev-Ye model: We consider a generalization of the Sachdev-Ye (SY) Hamiltonian for qubit system [49]. In particular, we remove interaction between qubits independently to obtain a sparse SY model. The Hamiltonian takes the following form:

$$H_{SY} = \sum_{1 \leq i < j \leq n} \mu_{ij} (\sigma_i^x \sigma_j^x + \sigma_i^y \sigma_j^y + \sigma_i^z \sigma_j^z), \quad (3)$$

where the sum over i, j extends over n sites, $\mu_{ij} = \chi_{ij} J_{ij}$. The exchange constants J_{ij} are mutually uncorrelated and

selected i.i.d. standard Gaussian random variables. Random variables χ_{ij} are also i.i.d., with $\chi_{ij} = 1$ with probability p and $\chi_{ij} = 0$ with probability $1 - p$. The original SY model is recovered when $p = 1$. We denote the coefficient vector consisting of all μ_{ij} as μ . We are primarily interested in the case where $p = \Omega(1/n^\alpha)$ for $0 \leq \alpha < 1$.

We will first analyze how prior results can be applied to learn this Hamiltonian. The method in [5], based on Hamiltonian reshaping to decouple the dynamics, will have a total evolution time of $\exp(\mathcal{O}(n)/\epsilon)$ because each qubit is acted on by $\mathcal{O}(n)$ terms instead of $\mathcal{O}(1)$. References [1] and [2], based on the cluster expansion, will have unknown complexity due to the lack of Lieb-Robinson bound. Reference [6] will have a total evolution time $\tilde{\mathcal{O}}(M^{4/p} n \|H_{SY}\|^3 / \epsilon^4)$ for ℓ^p error to be below ϵ , where M is the number of terms in the Hamiltonian of sparse SY model as explicitly expressed in (4), and $\|H\| = \mathcal{O}(M)$. Reference [45] will have a total evolution time $\tilde{\mathcal{O}}(M^{1+1/p}/\epsilon)$ for ℓ^p error to be below ϵ , where M is the number of terms in the Hamiltonian of sparse SY model as explicitly expressed in (4), and $\|H\| = \mathcal{O}(M)$.

We then analyze the cost of this task using the protocol in this work. The probability that the absolute value of any fixed coefficient exceeding ξ is $e^{-\Omega(\xi^2)}$, and consequently by a union bound, the probability of there existing one term in H_{SY} such that the absolute value of its coefficient is larger than ξ is $1 - n^2 e^{-\Omega(\xi^2)}$. Then we can choose $\xi = \mathcal{O}(\sqrt{\log(n/\delta_{SY})})$ to ensure that all coefficients are bounded by ξ with probability at least $1 - \delta_{SY}/2$. Therefore we can consider the rescaled Hamiltonian $\xi^{-1} H_{SY}$ whose coefficients are all bounded by 1 with $1 - \delta_{SY}/2$ probability. By the Chernoff-Hoeffding theorem, we can also see that the number of non-zero terms in the Hamiltonian is $n^2 p + \mathcal{O}(n \sqrt{p \log(1/\delta_{SY})})$ with probability at least $1 - \delta_{SY}/2$. Therefore with probability at least $1 - \delta_{SY}$, we only need to deal with a Hamiltonian with

$$M = n^2 p + \mathcal{O}(n \sqrt{p \log(1/\delta_{SY})}) \quad (4)$$

terms, each of which is bounded by 1. We need to ensure the ℓ^p -error in the end is ϵ/ξ to account for the rescaling. Based on the above, we use our main result Theorem 9 to arrive at the following result:

Corollary 1 (The sparse Sachdev-Ye model): To learn the Hamiltonian defined in (3) with an estimate $\hat{\mu}$ of all coefficients that have ℓ^p error ($1 \leq p \leq 2$) at most ϵ using the proposed algorithm with probability at least $(1 - \delta)(1 - \delta_{SY})$, the total evolution time complexity used is

$$T = \tilde{\mathcal{O}}\left(\frac{(n^2 p)^{1/p+1/2}}{\epsilon}\right).$$

Note that in the above we hide the $\text{polylog}(1/\delta_{SY})$ factor. Compared to previous works [1], [2], [5], [6], [45], our method has the advantage of guaranteed $\text{poly}(n)$ scaling and achieves the Heisenberg limit. Moreover, it can utilize sparsity to gain further advantage since the cost decreases with p .

Hamiltonians with power law interactions: For the power law interaction, the Hamiltonian (on a D -dimensional lattice) is

$$H = \sum_{i \leq j} H_{ij}, \quad (5)$$

where H_{ij} is supported on qubits i and j , and $\|H_{ij}\| \leq 1/(1 + d(i, j)^\alpha)$, where $d(i, j)$ is the Euclidean distance between sites i and j . We will consider two separate cases: $\alpha \leq D$ or $\alpha > D$.

For $\alpha \leq D$, which includes Hamiltonians of some interesting physical systems such as the Coulomb potential system ($\alpha = 1$, $D = 3$), [5] will have a total evolution time of $\exp(\mathcal{O}(n)/\epsilon)$ because each qubit is acted on by $\mathcal{O}(n)$ terms instead of $\mathcal{O}(1)$. References [1] and [2] will have an unknown complexity due to the lack of Lieb-Robinson bound [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66]. Reference [6] will have $\tilde{\mathcal{O}}(n^{(1+8/p)}\|H\|^3/\epsilon^4)$ for ℓ^p error to be below ϵ , where $\|H\| = \mathcal{O}(n^{2-\alpha/D})$. Reference [45] will have a total evolution time $\tilde{\mathcal{O}}(n^{(2/p+2)}/\epsilon)$ for ℓ^p error to be below ϵ .

For $\alpha > D$, where the decaying effect is more significant and the interactions tend to be more local in space compared to the $\alpha \leq D$ case, [5] will have a total evolution time of $\exp(\mathcal{O}(n)/\epsilon)$ because each qubit is acted on by $\mathcal{O}(n)$ terms instead of $\mathcal{O}(1)$, the same reason as in the $\alpha \leq D$ case. References [1] and [2] will have a total evolution time of $\tilde{\mathcal{O}}(n^{4/p}/\epsilon^2)$ for ℓ^p error to be below ϵ when there is a Lieb-Robinson bound [51] (each coefficient needs to be estimated to precision $\epsilon n^{-2/p}$), and an unknown complexity when there is not. Reference [6] will have $\tilde{\mathcal{O}}(n^{(1+8/p)}\|H\|^3/\epsilon^4)$ for ℓ^p error to be below ϵ , where $\|H\| = \mathcal{O}(n)$. Reference [45] will have a total evolution time $\tilde{\mathcal{O}}(\min(n^{(2/p+2)}/\epsilon, n^{2(1+\kappa)/p}/\epsilon^{(1+\kappa)}))$ for ℓ^p error to be below ϵ , with $\kappa = 2D/(\alpha + D)$.

However, with the protocol in this work, we can efficiently learn any Hamiltonian of α -power law interactions with the Heisenberg scaling regardless of α :

Corollary 2 (The power law interaction Hamiltonians): To learn a α -power law Hamiltonian of n -qubit on a D -dimensional lattice as defined in (5) with an estimate $\hat{\mu}$ of all coefficients that have ℓ^p error ($1 \leq p \leq 2$) at most ϵ using the proposed algorithm with probability at least $1 - \delta$, the total evolution time complexity used is

$$T = \tilde{\mathcal{O}}\left(\frac{n^{2/p+1}}{\epsilon}\right).$$

The total evolution time above directly follows from the fact that there are $\mathcal{O}(n^2)$ non-zero terms in the Hamiltonian.

We summarize the comparison of results in Table I.

II. OVERVIEW

Below we will provide an overview of the results and techniques in this work.

A. Background

Our results depend on several important tools developed in previous works. We will briefly introduce these tools here. Hamiltonian reshaping, proposed in [5] is an approach to reshape the Hamiltonian during time evolution into a new Hamiltonian H_{eff} that is easy to learn and also contains useful information about the original Hamiltonian.

In this work we will need to use the Hamiltonian reshaping technique to obtain an effective Hamiltonian that is diagonal with respect to a given single-qubit Pauli eigenbasis. More precisely, given a vector $\beta = (\beta_1, \beta_2, \dots, \beta_n)$, where $\beta_i \in \{x, y, z\}$,

we want the effective Hamiltonian to only contain Pauli terms that are either I or σ^{β_i} ($\sigma^{x,y,z}$ refers to the Pauli-X, Y, and Z matrices respectively) on the i th qubit. To achieve this, we apply, with an interval of τ , Pauli operators randomly drawn from the set $\mathcal{K}_\beta$ which is a qubit-wise abelian subgroup of the Pauli group on n qubits with $|\mathcal{K}_\beta| = 2^n$. This results in an effective Hamiltonian

$$H_{\text{eff}} = \frac{1}{2^n} \sum_{Q \in \mathcal{K}_\beta} Q H Q = \sum_{a: P_a \in \mathcal{K}_\beta} \mu_a P_a. \quad (6)$$

Note that because $\mathcal{K}_\beta$ is abelian, H_{eff} consists of commuting Pauli terms, making it easier to learn the coefficients, as will be discussed in Section V-C. For a more detailed discussion of the Hamiltonian reshaping technique see Section IV-A.

Compressed sensing provides a powerful tool to reconstruct a sparse high-dimensional vector from few measurements. This is relevant to the problem we are studying: the Hamiltonian coefficients form an M -sparse vector, but this vector is at the same time high-dimensional, as any of the $\mathcal{O}(n^k)$ k -local Pauli terms can be present.

Compressed sensing involves solving the ℓ^1 -minimization problem stated in (1). For the coefficient matrix $\mathbf{A}$, we will use what we call a *weight- k Hadamard matrix* (Definition 1), which is a $2^n \times D$ submatrix of the Walsh-Hadamard matrix of size $2^n \times 2^n$. Here $D = \sum_{l=0}^k \binom{n}{l}$ is the total number of bit-strings with Hamming weight at most k . In Section V-A, we will see that the transformation from coefficients to eigenvalues in the completely commuting Hamiltonian H_{eff} is described by this matrix. Therefore by solving the corresponding ℓ^1 minimization problem we can reconstruct the sparse coefficient vector with bounded error from few eigenvalue estimates (more precisely, differences between eigenvalues). For a more detailed discussion see Section IV-B. In particular we prove that the weight- k Hadamard matrix has the desired restricted isometry property (RIP).

B. The Hamiltonian Learning Protocol

We focus on learning a k -body Hamiltonian with M terms, which takes the form

$$H = \sum_{P \in \mathbb{P}_n} \mu_P P, \quad (7)$$

where $\mathbb{P}_n$ is the set of all Pauli matrices, $-1 \leq \mu_P \leq 1$, and $\mu_P \neq 0$ only for M Pauli terms P_a with $\text{wt}(P_a) \leq k$. Here we will briefly describe our protocol for learning this Hamiltonian.

First, the Hamiltonian reshaping procedure discussed in Section IV-A provides an effective Hamiltonian consisting of completely commuting Pauli terms. Section V-A contains a detailed discussion of the eigenvalues and eigenvectors of these Hamiltonians, which we will briefly summarize here. As stated in (6), the effective Hamiltonian takes the form $H_{\text{eff}} = \sum_{P \in \mathcal{K}_\beta} \mu_P P$, where

$$\mathcal{K}_\beta = \left\{ \bigotimes_{j=1}^n Q^j : Q^j = I \text{ or } \sigma^{\beta_j} \right\}, \quad (8)$$

TABLE I
TOTAL EVOLUTION TIMES OF DIFFERENT METHODS TO ACHIEVE ℓ^p ERROR ϵ

	Sparse SY model	Power law interactions $\alpha \leq D$	Power law interactions $\alpha > D$
[1]	unknown ¹	unknown ¹	$\tilde{\mathcal{O}}(n^{4/p}/\epsilon^2)$ (with L-R bound) unknown ¹ (without L-R bound)
[5]	$\exp(\mathcal{O}(n)/\epsilon)$	$\exp(\mathcal{O}(n)/\epsilon)$	$\exp(\mathcal{O}(n)/\epsilon)$
[2]	unknown ¹	unknown ¹	$\tilde{\mathcal{O}}(n^{4/p}/\epsilon^2)$ (with L-R bound) unknown ¹ (without L-R bound)
[6]	$\tilde{\mathcal{O}}\left(\frac{(n^2 p)^{3+4/p} n}{\epsilon^4}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^{7+8/p-3\alpha/D}}{\epsilon^4}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^{4+8/p}}{\epsilon^4}\right)$
[45] ³	$\tilde{\mathcal{O}}\left(\frac{(n^2 p)^{1/p+1}}{\epsilon}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^{2/p+2}}{\epsilon}\right)$	$\tilde{\mathcal{O}}\left(\min\left(\frac{n^{2/p+2}}{\epsilon}, \frac{n^{2(1+\kappa)/p}}{\epsilon^{1+\kappa}}\right)\right)^2$
This work	$\tilde{\mathcal{O}}\left(\frac{(n^2 p)^{1/p+1/2}}{\epsilon}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^{2/p+1}}{\epsilon}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^{2/p+1}}{\epsilon}\right)$

¹ Lack of Lieb-Robinson bound.

² $1 + \kappa = 1 + 2D/(\alpha + D) > 1$.

³ Using the method in [45, Theorem 5.1], the total numbers of experiments for these examples are $\tilde{\mathcal{O}}(n^4 p^2)$, $\tilde{\mathcal{O}}(n^4)$, and $\tilde{\mathcal{O}}(\min\{n^4, n^{4\kappa/p}/\epsilon^{2\kappa}\})$ respectively, compared to $\tilde{\mathcal{O}}(n^2 p)$, $\tilde{\mathcal{O}}(n^2)$, $\tilde{\mathcal{O}}(n^2)$ using the method in this work.

as given in (16). We note that within the set $\mathcal{K}_\beta$, each element can be represented by a bit-string $c \in \{0, 1\}^n$. This is because, for the i th qubit, the Pauli operator has a component on it that is either the identity or σ^{β_i} , and we choose $c_i = 0$ for the former and $c_i = 1$ for the latter. Based on the above, we can rewrite H_{eff} as $H_{\text{eff}} = \sum_{c \in \{0,1\}^n} \mu_c P_c$, where $P_c = \bigotimes_i (\sigma^{\beta_i})^{c_i}$, $c_i \in \{0, 1\}$ with the coefficient $\mu_c = \mu_P$ if $P_c = P$, and $\mu_c = 0$ otherwise. We then collect all coefficients μ_c into a 2^n -dimensional vector $\boldsymbol{\mu}_\beta$. Note that this vector contains at most M non-zero entries.

Each eigenvector can similarly be uniquely associated with a bit-string b , and a Pauli term P_c contributes to the corresponding eigenvalue by an amount of $(-1)^{c \cdot b}$. From the above observation, we can see that the relation between the eigenvalues and coefficients of H_{eff} can be written down in a more compact form:

$$\boldsymbol{\lambda} = \mathbf{H} \boldsymbol{\mu}_\beta, \quad (9)$$

where $\mathbf{H}$ is the $2^n \times 2^n$ Hadamard matrix defined in Definition 1. Therefore we can reconstruct the coefficients of the Hamiltonian by $\boldsymbol{\mu}_\beta = \mathbf{H}^{-1} \boldsymbol{\lambda} = \frac{1}{2^n} \mathbf{H} \boldsymbol{\lambda}$. This approach has the drawback that we will need to estimate all 2^n of the eigenvalues. While techniques such as “bins and peelings” for Walsh-Hadamard transforms with sparsity in the transform domain [67], [68] could theoretically reduce the computational cost, as used in [3] and [24], they cannot be used to achieve the Heisenberg-limited scaling and typically assume that the sparse support is randomly distributed across the entire vector space, rather than localized to the subspace with weight at most k .

We therefore need to consider an alternative approach, which is through compressed sensing. Using (9) to provide the constraint in (1) and minimizing the ℓ^1 norm of the coefficients, we will be able to recover the coefficients $\boldsymbol{\mu}_\beta$. Quantitatively, if H_{eff} contains M non-zero terms, then the

coefficients can be recovered with $\tilde{\mathcal{O}}(M)$ eigenvalue estimates. For a more detailed discussion see Section V-C.

To complete our description of the learning protocol for H_{eff} , we only need to describe the experimental setup to learn the eigenvalue corresponding to a given eigenvector, which is a product state. We call such an experiment a phase estimation experiment. It consists of the following steps: (1) Prepare the initial state $\frac{1}{\sqrt{2}}(|0\rangle_\beta + |b\rangle_\beta)$ as in (39). (2) Let the system evolve for time t while applying random Pauli operators from $\mathcal{K}_\beta$ (defined in (8)) with interval τ . (3) Measure the observables X_β^b or Y_β^b (by measuring every qubit in either σ^x , σ^y , or σ^z basis) as defined in (41) and (44) to obtain a ± 1 outcome.

With these experiments we can estimate eigenvalue difference $\lambda_b - \lambda_0$ through the robust frequency estimation protocol introduced in Section IV-C. In particular, these estimations have Heisenberg-limited scaling and are SPAM-robust. We will further discuss how we estimate $\lambda_b - \lambda_0$ and the complexity of the number of independent non-adaptive experiments and total evolution time in Section V-B. Combined with compressed sensing, we can now learn all coefficients in H_{eff} efficiently and with Heisenberg-limited scaling. Theorem 5 also tells us that in order to keep the ℓ^1 -error of the coefficient estimates below ϵ we need to estimate the randomly sampled eigenvalues to precision $\tilde{\mathcal{O}}(\epsilon M^{-1/2})$, which results in a total evolution time of $\tilde{\mathcal{O}}(M^{1/2}/\epsilon)$ for eigenvalue estimate using robust phase estimation according to Theorem 6. The total evolution time needed for learning H_{eff} with ℓ^1 -error at most ϵ is therefore $\tilde{\mathcal{O}}(M^{1.5}/\epsilon)$.

We can now learn H_{eff} , whose terms are all contained in the set $\mathcal{K}_\beta$ as defined in (8). In other words, for each Pauli term P in H_{eff} , its component on the j th qubit is either I or σ^{β_j} where $\beta_j \in \{x, y, z\}$. We will next focus on the way to choose a set of β so that all the terms in the Hamiltonians are covered with large probability. We can show that uniformly randomly choosing β is a good strategy. Since each Pauli term P in the

Hamiltonian is k -body, for a uniformly randomly generated β , $P \in \mathcal{K}_\beta$ with probability at least 3^{-k} . It therefore takes $L = 3^k \log(M/\delta_{\text{basis}})$ samples of β to ensure that all the M terms are included in at least one of the L instances with probability at least $1 - \delta_{\text{basis}}$. This ensures that each term P is learned via one of the H_{eff} . Taking this overhead L into account, we then arrive at our main result (using the notation of $\mathcal{K}_\beta$ as defined in (8) and $\mathbb{P}_n^{(k)}$ for the set of Pauli operators on at most k qubits as defined in (15)):

Theorem 3 (*Learning a k -body Hamiltonian containing M terms*): We assume that the quantum system is evolving under a k -body Hamiltonian with M terms (Definition 2). With N_{exp} independent non-adaptive (β, b_j, t_j, τ) -phase estimation experiments (Definition 3), $j = 1, 2, \dots, N_{\text{exp}}$, with the SPAM error (Definition 4) satisfying $\epsilon_{\text{SPAM}} \leq 1/(3\sqrt{2})$, we can obtain estimates $\hat{\mu}_P$ for every $P \in \mathbb{P}_n^{(k)}$ such that, with probability at least $1 - \delta$

$$\left(\sum_{P \in \mathbb{P}_n^{(k)} \setminus \{I\}} |\hat{\mu}_P - \mu_P|^p \right)^{1/p} \leq \epsilon, \quad (10)$$

for $1 \leq p \leq 2$. In the above N_{exp} , $\{t_j\}$, and τ satisfy

$$N_{\text{exp}} = \tilde{\mathcal{O}}(3^k M),$$

$$T = \sum_j t_j = \tilde{\mathcal{O}}\left(\frac{9^k M^{1/p+1/2}}{\epsilon}\right),$$

and $\tau = \Omega(3^{-k} \epsilon / (M^{1/p+3/2} \log(M/\delta)))$.

This theorem is stated as Theorem 9 later in the text. Although $k = O(1)$ throughout, we keep factors $2^{O(k)}$ explicit to help inform the reader on some k dependence. More precise results are in Remark 3.

We also discuss the robustness of our method to modeling errors, i.e., the Hamiltonian may contain more than M terms and may not be exactly k -body, in Section V-E. The computational complexity of solving the ℓ^1 -minimization problem in (1) is discussed in Section V-F. Operational interpretations of the ℓ^1 - and ℓ^2 -error metrics are discussed in Section VII.

C. Lower Bounds

Besides the learning protocol we also provide a lower bound for the Hamiltonian learning task we consider. Specifically we consider the dependence on the tolerated ℓ^1 -error ϵ and the number of Pauli terms M . The adaptive experiments are modeled as in [69], where all experiments form a tree such that the outcome of the experiment at a vertex determines which child leaf to move to, thus determining the next experiment to perform. Our main tool for the total evolution time lower bound is Assouad's lemma [70], which provides a lower bound on the achievable ℓ^1 -error given two prerequisites: (1) an estimate of how hard it is to distinguish two output probability distributions if they come from two Hamiltonians that differ slightly, and (2) a lower bound on the penalty in ℓ^1 -error if such a pair of Hamiltonians are not correctly distinguished. The second prerequisite is easy to fulfill, as discussed in the proof of Theorem 11. For the first prerequisite, we follow [5] to capture the difficulty of correctly distinguishing the output

probability distributions by induction on the tree of adaptive experiments. We show that for a series of adaptive experiments with Hamiltonians H and H' respectively, if each of them involve a SPAM error of order γ , then the total variation distance between the output distributions is upper bounded by $1 - \gamma^{\|H-H'\|T}$. Combining this with Assouad's lemma allows us to prove the lower bound: If a learning algorithm can learn the coefficients of an n -qubit Hamiltonian $H = \sum_{a=1}^M \mu_a P_a$ with any unknown parameters $|\mu_a| \leq 1$, i.e., after multiple rounds of noisy experiments, the algorithm can estimate any $\mu = (\mu_1, \dots, \mu_M)$ to ϵ_1 -error in the ℓ^1 -norm in expectation value averaged over experimental outcomes, then,

$$T \geq \frac{M}{\epsilon_1 e \log(1/\gamma)}. \quad (11)$$

Details of the proofs are in Section VI and Appendix D.

III. NOTATIONS

In this work we use $\|\mathbf{x}\|_p$ to denote the ℓ^p -norm of vector $\mathbf{x}$. For a bit string b , we use $|b|$ to denote its Hamming weight. The Pauli matrices are denoted by $\sigma^x, \sigma^y, \sigma^z$. We use the following notation to denote the Pauli eigenstates:

$$\begin{aligned} |1, z\rangle &= |0\rangle, \quad |-1, z\rangle = |1\rangle, \quad |1, x\rangle = |+\rangle, \quad |-1, x\rangle = |-\rangle, \\ |1, y\rangle &= \frac{1}{\sqrt{2}}(|0\rangle + i|1\rangle), \quad |-1, y\rangle = \frac{1}{\sqrt{2}}(|0\rangle - i|1\rangle). \end{aligned} \quad (12)$$

We denote the set of all n -qubit Pauli matrices by $\mathbb{P}_n$:

$$\mathbb{P}_n = \left\{ \bigotimes_{i=1}^n P_i : P_i = I, \sigma^x, \sigma^y, \text{ or } \sigma^z \right\}. \quad (13)$$

For each $P = \bigotimes_{i=1}^n P_i \in \mathbb{P}_n$, we define its weight by

$$\text{wt}(P) = |\{i : P_i \neq I\}|. \quad (14)$$

We denote the set of Pauli operators that are at most k -body by

$$\mathbb{P}_n^{(k)} = \{P \in \mathbb{P}_n : \text{wt}(P) \leq k\}. \quad (15)$$

IV. BACKGROUND

A. Hamiltonian Reshaping

The basic idea behind [5] is that we can reshape the Hamiltonian during time evolution into a new Hamiltonian that is easy to learn and also contains useful information about the original Hamiltonian. In this section we will provide a brief introduction to this technique.

Hamiltonian reshaping is done by doing random unitary transformations to the time evolution operator corresponding to the unknown Hamiltonian and obtaining the effective Hamiltonian H_{eff} . At each time step we apply a unitary transformation that is randomly sampled from an ensemble, which we will describe later in this section. As a result the system evolves under a random Hamiltonian transformed under the unitary operation. We assume that these unitary transformations are applied instantaneously. The same analysis as the one underlying the randomized Hamiltonian simulation algorithm known as qDRIFT [71], [72], [73] then tells us that the effective Hamiltonian H_{eff} is the ensemble average of the random Hamiltonians. This approach is similar in spirit to

dynamical decoupling [74], [75], which applies unitary operations in a deterministic manner. However, the randomness in our protocol is helpful as it allows us to efficiently apply random unitaries from a potentially exponentially large set.

In this work we will need to use the Hamiltonian reshaping technique to obtain an effective Hamiltonian that is diagonal with respect to a given single-qubit Pauli eigenbasis. More precisely, given a vector $\beta = (\beta_1, \beta_2, \dots, \beta_n)$, where $\beta_i \in \{x, y, z\}$, we want the effective Hamiltonian to only contain Pauli terms that are either I or σ^{β_i} on the i th qubit. To achieve this, we apply, with an interval of τ , Pauli operators randomly drawn from the set

$$\mathcal{K}_\beta = \left\{ \bigotimes_{j=1}^n Q^j : Q^j = I \text{ or } \sigma^{\beta_j} \right\}. \quad (16)$$

Here we can see that $|\mathcal{K}_\beta| = 2^n$. $\mathcal{K}_\beta$ here is an abelian subgroup of the Pauli group on n qubits. Also, τ needs to be sufficiently small as will be analyzed in Theorem 4. More precisely, the evolution of the quantum system is described by

$$Q_r e^{-iH\tau} Q_r \dots Q_2 e^{-iH\tau} Q_2 Q_1 e^{-iH\tau} Q_1, \quad (17)$$

where Q_k , $k = 1, 2, \dots, r$, is uniformly randomly drawn from the set $\mathcal{K}_\beta$. In one time step of length τ , the quantum state evolves as

$$\begin{aligned} \rho &\mapsto \frac{1}{2^n} \sum_{Q \in \mathcal{K}_\beta} Q e^{-iH\tau} Q \rho Q e^{iH\tau} Q \\ &= \rho - \frac{1}{2^n} \sum_{Q \in \mathcal{K}_\beta} i\tau [QH Q, \rho] + \mathcal{O}(\tau^2) \\ &= \rho - i\tau [H_{\text{eff}}, \rho] + \mathcal{O}(\tau^2) = e^{-iH_{\text{eff}}\tau} \rho e^{iH_{\text{eff}}\tau} + \mathcal{O}(\tau^2), \end{aligned} \quad (18)$$

where H_{eff} , the effective Hamiltonian, is

$$H_{\text{eff}} = \frac{1}{2^n} \sum_{Q \in \mathcal{K}_\beta} Q H Q. \quad (19)$$

In Appendix A we will provide a bound on how far the actual dynamics deviate from the dynamics induced by the effective Hamiltonian.

The above can be seen as a linear transformation of the Hamiltonian H . It is therefore reasonable to consider what is the effect of the above transformation on each Pauli term in the Hamiltonian. For a Pauli term P_a , we note that there are two possible outcomes:

Lemma 1: Let P be a Pauli operator and let $\mathcal{K}_\beta$ be as defined in (16). Then

$$\frac{1}{2^n} \sum_{Q \in \mathcal{K}_\beta} Q P Q = \begin{cases} P, & \text{if } P \in C(\mathcal{K}_\beta), \\ 0, & \text{if } P \notin C(\mathcal{K}_\beta), \end{cases} \quad (20)$$

where $C(\mathcal{K}_\beta)$ denotes the centralizer of the subgroup $\mathcal{K}_\beta$.

Proof: If $P \in C(\mathcal{K}_\beta)$, then $Q P Q = P$ for all $Q \in \mathcal{K}_\beta$, and averaging over all Q therefore yields P . If $P \notin C(\mathcal{K}_\beta)$, then there exists $Q_0 \in \mathcal{K}_\beta$ such that $P Q_0 = -Q_0 P$ (because two Pauli matrices either commute or anti-commute). Consider the mapping $\phi : \mathcal{K}_\beta \rightarrow \mathcal{K}_\beta$ defined by $\phi(Q) = Q_0 Q$. This is a bijection from $\mathcal{K}_\beta$ to itself, and it can be readily checked that if Q commutes

with P , then $\phi(Q)$ anti-commutes with P ; if Q anti-commutes with P , then $\phi(Q)$ commutes with P . Consequently $Q P Q = P$ for half of all $Q \in \mathcal{K}_\beta$ and $Q P Q = -P$ for the other half. Thus taking the average yields 0. $\square$

The above lemma allows us to select which Pauli terms we want to preserve in the Hamiltonian and which ones we want to discard. Because $C(\mathcal{K}_\beta) = \mathcal{K}_\beta$, we have

$$H_{\text{eff}} = \sum_{a: P_a \in \mathcal{K}_\beta} \mu_a P_a. \quad (21)$$

Note that because $\mathcal{K}_\beta$ is abelian, H_{eff} consists of commuting Pauli terms, making it easier to learn the coefficients, as will be discussed in Section V-C. Also the coefficients μ_a of P_a that are in $\mathcal{K}_\beta$ are preserved in this effective Hamiltonian. We therefore need to choose a set of β so that each Pauli term of H is contained in at least one of the H_{eff} corresponding to a β . We will discuss the way to do this in Section V-D.

While Lemma 1 concerns the uniform average over all 2^n elements of a set of commuting Pauli operators, in our learning protocol we will randomly sample from this set. To assess the protocol's accuracy, we will use the following theorem.

Theorem 4: Let β be a length- n string of x, y, z , and let $\mathcal{K}_\beta$ be as defined in (16). Let U be the random unitary defined in (17). Let $V = e^{-iH_{\text{eff}}t}$ for H_{eff} given in (21), and $t = r\tau$. We define the quantum channels $\mathcal{U}$ and $\mathcal{V}$ be

$$\mathcal{U}(\rho) = \mathbb{E}[U \rho U^\dagger], \quad \mathcal{V}(\rho) = V \rho V^\dagger.$$

Then

$$\|\mathcal{U} - \mathcal{V}\|_\diamond \leq \frac{4M^2 t^2}{r}.$$

Here the norm is the diamond norm (or completely-bounded norm), given in terms of the Schatten L_1 norm by $\|\mathcal{U}\|_\diamond := \max_{X: \|X\|_1 \leq 1} \|(\mathcal{U} \otimes \mathbb{1}_n)(X)\|_1$. The proof can be found in Appendix A, where we analyze how the error in each segment of time-evolution contributes to the total diamond distance between the quantum channels $\mathcal{U}$ and $\mathcal{V}$. The error in each segment is analyzed using a Taylor expansion.

B. Compressed Sensing

The task of learning a sparse coefficient vector is studied in the field of compressed sensing, which is related to learning the M -sparse coefficient vector in the Hamiltonian learning scenario. Compressed sensing provides a powerful tool to reconstruct a sparse high-dimensional vector from few measurements. This is relevant to the problem we are studying: the Hamiltonian coefficients form an M -sparse vector, but this vector is at the same time high-dimensional, as any of the $\mathcal{O}(n^k)$ k -local Pauli terms can be present.

The basic setup of compressed sensing is as follows: we want to reconstruct a vector $\mathbf{x} \in \mathbb{C}^D$ based on observations $\mathbf{y} \in \mathbb{C}^\Gamma$ of the form

$$\mathbf{y} = \mathbf{A} \mathbf{x} + \mathbf{e}, \quad (22)$$

where $\mathbf{A}$ is a $\Gamma \times D$ matrix and $\mathbf{e} \in \mathbb{C}^\Gamma$ represents (typically small) noise in the observation. When $\Gamma < D$, we will not be able to uniquely determine $\mathbf{x}$ in general. However, below we will use $\mathbf{A}$ that has the *restricted isometry*

property (RIP), which will be discussed in detail in Appendix B. The restricted isometry property of $\mathbf{A}$ guarantees that the ℓ^2 -norm of any sparse vector $\mathbf{x}$ will be approximately preserved after applying $\mathbf{A}$. This intuitively implies that we can have enough information to reconstruct $\mathbf{x}$ given access to $\mathbf{Ax}$. A matrix satisfying the RIP is nearly an isometry when acting on sufficiently sparse vectors, and thus the vector we want to reconstruct is not distorted beyond recognition when transformed through this matrix.

We further assume that $\mathbf{x}$ is M -sparse and solve the following ℓ^1 -minimization problem,

$$\underset{\mathbf{z} \in \mathbb{C}^D}{\text{minimize}} \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{Az} - \mathbf{y}\|_2 \leq \eta \sqrt{\Gamma}, \quad (23)$$

with Γ being sufficiently large but potentially small compared to D . The solution $\mathbf{x}^\# \in \mathbb{C}^D$ is guaranteed to be close to the actual vector $\mathbf{x}$ if each entry of the error vector $\mathbf{e}$ is at most η in absolute value. The optimization problem (23) can be solved efficiently as we show in Section V-F.

In this work, we will focus on what we call a *weight- k Hadamard matrix*, which we define as follows:

Definition 1 (Weight- k Hadamard matrix $\mathbf{H}^{(k)}$): Let $\mathbf{H}$ be the Hadamard matrix of size $2^n \times 2^n$, i.e.,

$$\mathbf{H} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\otimes n}.$$

For each $i = 1, 2, \dots, 2^n$, we let b_i be the n -bit string representing the integer $i - 1$. Then the weight- k Hadamard matrix, denoted as $\mathbf{H}^{(k)}$, is the sub-matrix of $\mathbf{H}$ consisting of every i th column of $\mathbf{H}$ where $|b_i| \leq k$. Here $|b_i|$ is the Hamming weight of the bit string b_i .

One can readily see that $\mathbf{H}^{(k)}$ is of size $2^n \times D$ where $D = \sum_{l=0}^k \binom{n}{l}$. Below we will use this matrix for our compressed sensing task. Before we state the result we will first introduce a notation: we define for a vector $\mathbf{x}$ and integers $s, p \geq 0$,

$$\sigma_s(\mathbf{x})_p = \min_{\mathbf{y}: \|\mathbf{y}\|_0 \leq s} \|\mathbf{x} - \mathbf{y}\|_p. \quad (24)$$

In other words $\sigma_s(\mathbf{x})_p$ is the smallest error in the p -norm one can achieve when approximating $\mathbf{x}$ with an s -sparse vector.

Theorem 5 (Compressed sensing with the weight- k Hadamard matrix): Let $D = \sum_{l=0}^k \binom{n}{l}$. Let $\mathbf{A} \in \mathbb{C}^{\Gamma \times D}$ be a matrix whose rows are independently randomly sampled from the rows of $\mathbf{H}^{(k)}$ with replacement. Let $\delta_{CS} \in (0, 1)$ and let $M > 0$ be an integer. If

$$\Gamma \geq CM \max\{\ln^2(M) \ln(M \ln(D)) \ln(D), \ln(1/\delta_{CS})\}, \quad (25)$$

then, with probability at least $1 - \delta_{CS}$, the following statement holds for every $\mathbf{x} \in \mathbb{C}^D$. Let noisy samples $\mathbf{y} = \mathbf{Ax} + \mathbf{e}$ be given with $\|\mathbf{e}\|_2 \leq \eta \sqrt{\Gamma}$, and let $\mathbf{x}^\#$ be the solution of the ℓ^1 -minimization problem (23). Then

$$\|\mathbf{x} - \mathbf{x}^\#\|_p \leq \frac{C_1}{M^{1-1/p}} \sigma_M(\mathbf{x})_1 + C_2 M^{1/p-1/2} \eta, \quad 1 \leq p \leq 2. \quad (26)$$

$\sigma_M(\mathbf{x})_1$ is defined as in (24). All constants $C, C_1, C_2 > 0$ are universal.

The proof of this theorem is provided in Appendix B. From the above theorem we can see that, if $\mathbf{x}$ is itself M -sparse, then $\sigma_M(\mathbf{x})_1 = 0$, and the ℓ^1 -minimization allows us to estimate $\mathbf{x}$ to within $C_2 M^{1/p-1/2} \eta$ error in the ℓ^p -norm.

C. Robust Frequency Estimation

Taking advantage of the Hamiltonian reshaping procedure, we will be able to identify eigenvectors of the effective Hamiltonian, and as a part of the protocol we want to estimate the corresponding eigenvalue. The estimation needs to be performed in a noise-robust way in order to achieve Heisenberg-limited scaling and SPAM robustness. The robust frequency estimation protocol that we introduce in this section helps us achieve this goal. This protocol follows the same idea as the robust phase estimation protocol [76], but is modified to obtain a good confidence interval rather than minimizing the mean-squared error.

From the phase estimation experiments to be introduced in Section V-B, we will generate signals fluctuating around $\cos(\theta t)$ and $\sin(\theta t)$ where θ corresponds to the eigenvalue we want to estimate. The goal is therefore to reconstruct θ based on noisy estimates of $\cos(\theta t)$ and $\sin(\theta t)$. Through the robust frequency estimation protocol, we can accomplish this task with the cost stated in the following theorem:

Theorem 6 (Robust frequency estimation): Let $\theta \in [-A, A]$. Let $X(t)$ and $Y(t)$ be independent random variables satisfying

$$\begin{aligned} |X(t) - \cos(\theta t)| &< 1/\sqrt{2}, \text{ with probability at least } 2/3, \\ |Y(t) - \sin(\theta t)| &< 1/\sqrt{2}, \text{ with probability at least } 2/3. \end{aligned} \quad (27)$$

Then with K independent non-adaptive¹ samples $X(t_1), X(t_2), \dots, X(t_K)$ and $Y(t_1), Y(t_2), \dots, Y(t_K)$, $t_j \geq 0$, for

$$K = \mathcal{O}(\log(A/\epsilon)(\log(1/q) + \log \log(A/\epsilon))), \quad (28)$$

$$T = \sum_{j=1}^K t_j = \mathcal{O}((1/\epsilon)(\log(1/q) + \log \log(A/\epsilon))),$$

$$\max_j t_j = \mathcal{O}(1/\epsilon), \quad (29)$$

we can obtain a random variable $\hat{\theta}$ such that

$$\Pr[|\hat{\theta} - \theta| > \epsilon] \leq q. \quad (30)$$

For a detailed proof see Appendix D. The main idea is to reduce the estimation problem into a sequence of decision problems with binary outputs in a way similar to bisection. Solving each of these decision problems updates our knowledge of θ , and in the process we obtain a smaller and smaller interval containing the exact value. We stop the process when the interval is small enough to provide sufficient accuracy.

V. THE HAMILTONIAN LEARNING PROTOCOL

The Hamiltonians we will focus on in this work are defined as follows

Definition 2 (k -body Hamiltonians with M terms): A k -body Hamiltonians with M terms takes the form

$$H = \sum_{P \in \mathbb{P}_n} \mu_P P, \quad (31)$$

where $\mathbb{P}_n$ is the set of all Pauli matrices defined in (13), $-1 \leq \mu_P \leq 1$, and $\mu_P \neq 0$ only for M Pauli terms P_a with $\text{wt}(P_a) \leq k$.

¹By “non-adaptive” we mean that the choice of each t_j does not depend on the value of $X(t_{j'})$ or $Y(t_{j'})$ for any $j' < j$.

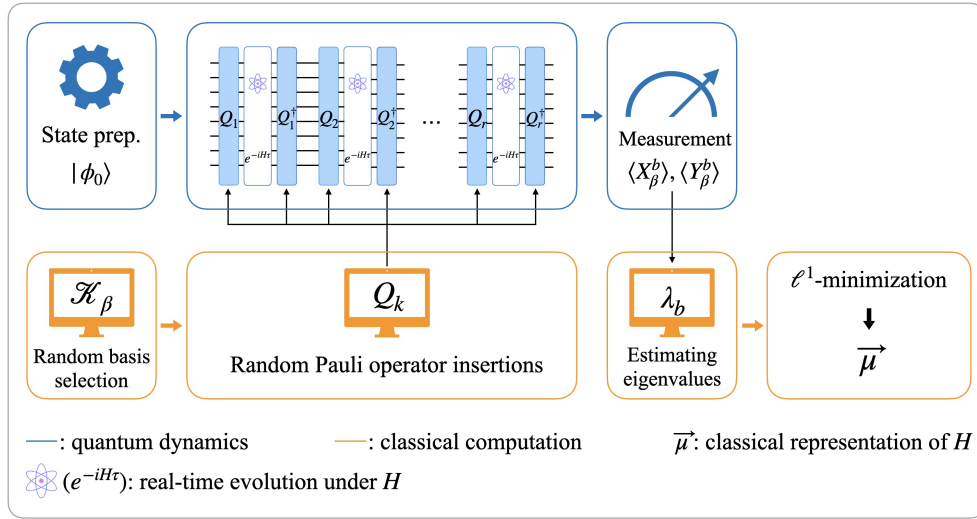


Fig. 1. **The learning protocol.** Quantum processes are illustrated in blue blocks, and classical processes are illustrated in orange blocks. State preparation and measurement (SPAM) errors are taken into account in the experiments.

A. Completely Commuting Hamiltonians

The Hamiltonian reshaping procedure discussed in Section IV-A provides an effective Hamiltonian consisting of completely commuting Pauli terms, by which we mean that the Pauli terms commute with each other when acting on any one of the qubits. One can readily verify that the Pauli terms in $\mathcal{K}_\beta$ (16) satisfy this requirement. We call these Hamiltonians *completely commuting Hamiltonians*. Equivalently, we can assign an σ^x , σ^y , or σ^z basis to each qubit, thus forming a basis for the entire Hilbert space, and then require the Hamiltonian to be diagonal in this basis.

The effective Hamiltonian from Hamiltonian reshaping takes the form

$$H_{\text{eff}} = \sum_{P \in \mathcal{K}_\beta} \mu_P P, \quad (32)$$

as shown in (21). The set $\mathcal{K}_\beta$ of completely commuting Pauli terms is specified by the basis given through $\beta = (\beta_1, \beta_2, \dots, \beta_n)$, as defined in (16). Based on the above, we can rewrite H_{eff} in another way:

$$H_{\text{eff}} = \sum_{c \in \{0,1\}^n} \mu_c P_c, \quad (33)$$

where

$$P_c = \bigotimes_i (\sigma^{\beta_i})^{c_i}, \quad c_i \in \{0, 1\}, \quad (34)$$

the coefficient $\mu_c = \mu_P$ if $P_c = P$, and $\mu_c = 0$ otherwise. We then collect all coefficients μ_c into a 2^n -dimensional vector μ_β . Note that this vector contains at most M non-zero entries.

The effective Hamiltonian H_{eff} has the nice property that its eigenvalues and eigenstates can be written down explicitly. For an n -bit string $b = (b_0, b_1, \dots, b_n)$ we define

$$|b\rangle_\beta = \bigotimes_{i=1}^n |(-1)^{b_i}, \beta_i\rangle. \quad (35)$$

The Pauli eigenstates $|\pm 1, \beta\rangle$ are defined as in (12). Then the above $|b\rangle_\beta$ is an eigenstate of H_{eff} . We can write down its corresponding eigenvalue, which is

$$\lambda_b = \sum_{c \in \{0,1\}^n} \mu_c (-1)^{c \cdot b}, \quad (36)$$

where $c \cdot b = c_1 b_1 + c_2 b_2 + \dots + c_n b_n \pmod{2}$. Collecting all eigenvalues λ_b into a 2^n -dimensional vector λ , we can then write down the relation between the eigenvalues and coefficients in a more compact form:

$$\lambda = \mathbf{H} \mu_\beta, \quad (37)$$

where $\mathbf{H}$ is the $2^n \times 2^n$ Hadamard matrix defined in Definition 1. Therefore we can reconstruct the coefficients of the Hamiltonian by

$$\mu_\beta = \mathbf{H}^{-1} \lambda = \frac{1}{2^n} \mathbf{H} \lambda. \quad (38)$$

This approach has the drawback that we will need to estimate all 2^n of the eigenvalues. While techniques such as “bins and peelings” for Walsh-Hadamard transforms with sparsity in the transform domain [67], [68] could theoretically reduce the computational cost, as used in [3] and [24], they cannot be used to achieve the Heisenberg-limited scaling and typically assume that the sparse support is randomly distributed across the entire vector space, rather than localized to the subspace with weight at most k . We therefore need to consider an alternative approach in Section V-C.

B. The Experimental Setup

As illustrated in Fig 1, in an experiment we will start from an initial state

$$\begin{aligned} |\phi_0\rangle &= \frac{1}{\sqrt{2}} (|0\rangle_\beta + |b\rangle_\beta) \\ &= \frac{1}{\sqrt{2}} \left(\bigotimes_{i=1}^n |1, \beta_i\rangle + \bigotimes_{i=1}^n |(-1)^{b_i}, \beta_i\rangle \right), \end{aligned} \quad (39)$$

where $|b\rangle_\beta$ is defined in (35), and the bit string $b \neq 0$. This state can be prepared from a $|b|$ -qubit GHZ state by applying

σ^x on some of the qubits. The GHZ state can be prepared in constant depth using measurements and feedback [77], [78]. Because b is uniformly sampled from all bit-strings of length n , the average size of the GHZ state that needs to be prepared is $n/2$.

Remark 2: Note that if one restricts sampling to smaller entangled states, for example for GHZ states with bit-strings with $\mathbb{E}|b| \leq \bar{w}$ for some $\bar{w} < n/2$, reducing the required expected GHZ size to $\bar{w}$ qubits. However, this changes the row distribution from uniform over all 2^n rows to uniform over a subsets of rows. However, the BOS orthonormality (with $K = 1$), which is required for establishing the RIP property in the compressed sensing protocol, of the columns of $\mathbf{H}^{(k)}$ holds with respect to the full 2^n -row uniform measure and does not automatically extend to the restricted measure.

One way to avoid using entangled states is to estimate $\lambda_0 - \lambda_b = \sum_u \lambda_{b_u} - \lambda_{b_{u+1}}$, where each b_u and b_{u+1} differ on only one bit, and all $\{b_u\}$ form a path on the boolean hypercube linking 0 to b . In this approach we only need to use states that are superpositions of $|b_u\rangle_\beta$ and $|b_{u+1}\rangle_\beta$, which are unentangled product states. The drawback is that one needs to increase the number of experiments, which result in an $\mathcal{O}(n)$ overhead in the total cost. It is unclear whether this is the optimal way to restrict to unentangled states, and we leave a rigorous treatment of the optimal tradeoff as future work.

We then let the system evolve while applying random Pauli operators with interval τ for time t to perform Hamiltonian reshaping as discussed in (17). Because $|0\rangle_\beta$ and $|b\rangle_\beta$ are both eigenstates of the effective Hamiltonian H_{eff} corresponding to eigenvalues λ_0 and λ_b (defined in (36)), as discussed in Section V-A, at time t we will approximately obtain the state

$$|\phi_t\rangle = \frac{1}{\sqrt{2}}(e^{-i\lambda_0 t}|0\rangle_\beta + e^{-i\lambda_b t}|b\rangle_\beta). \quad (40)$$

In the end, we then measure the observable

$$X_\beta^b = \bigotimes_{i=1}^n Q_i, \quad (41)$$

where each Q_i is chosen to be a Pauli operator such that

$$Q_i |1, \beta_i\rangle = |(-1)^{b_i}, \beta_i\rangle, \quad Q_i |(-1)^{b_i}, \beta_i\rangle = |1, \beta_i\rangle. \quad (42)$$

Such Q_i operators can be constructed as follows: in the case of $b_i = 1$, if $\beta_i = z$ then $Q_i = \sigma^x$, and if $\beta_i = y$ or x then $Q_i = \sigma^z$; in the case of $b_i = 0$ then $Q_i = I$. This observable therefore satisfies

$$X_\beta^b |0\rangle_\beta = |b\rangle_\beta, \quad X_\beta^b |b\rangle_\beta = |0\rangle_\beta.$$

The expectation value will be

$$\langle \phi_t | X_\beta^b | \phi_t \rangle = \cos((\lambda_b - \lambda_0)t). \quad (43)$$

Note that in order to measure X_β^b we only need to measure each individual Q_i to obtain a ± 1 outcome and then multiply them together.

Similarly, we can measure the observable

$$Y_\beta^b = \bigotimes_{i=1}^n Q_i, \quad (44)$$

where the Q_i operators are chosen slightly differently. For the first i for which $b_i = 1$, we choose Q_i so that

$$Q_i |1, \beta_i\rangle = i |(-1)^{b_i}, \beta_i\rangle, \quad Q_i |(-1)^{b_i}, \beta_i\rangle = -i |1, \beta_i\rangle. \quad (45)$$

Note that we can choose $Q_i = \sigma^y$ if $\beta_i = z$, $Q_i = \sigma^x$ if $\beta_i = y$, and $Q_i = -\sigma^y$ if $\beta_i = x$ to satisfy this requirement. For all other i we choose Q_i to satisfy (42). The resulting operator Y_β^b satisfies

$$Y_\beta^b |0\rangle_\beta = i |b\rangle_\beta, \quad Y_\beta^b |b\rangle_\beta = -i |0\rangle_\beta.$$

Therefore the expectation value is

$$\langle \phi_t | Y_\beta^b | \phi_t \rangle = -\sin((\lambda_b - \lambda_0)t). \quad (46)$$

We summarize the above experiment in the following definition

Definition 3 (Phase estimation experiment): We call the procedure below a (β, b, t, τ) -phase estimation experiment:

- 1) Prepare the initial state $\frac{1}{\sqrt{2}}(|0\rangle_\beta + |b\rangle_\beta)$ as in (39).
- 2) Let the system evolve for time t while applying random Pauli operators from $\mathcal{K}_\beta$ (defined in (16)) with interval τ .
- 3) Measure the observables X_β^b or Y_β^b (by measuring every qubit in either σ^x , σ^y , or σ^z basis) as defined above to obtain a ± 1 outcome.

The goal of the above experiment is to estimate $\lambda_b - \lambda_0$, for which we also need to use the robust frequency estimation protocol introduced in Section IV-C. Below we will discuss how this is done, and analyze the effect of the Hamiltonian reshaping error and the state preparation and measurement (SPAM) error.

We model the SPAM error as follows:

Definition 4: The preparation of the initial state involves an error channel $\mathcal{E}_{\text{prep}}$ applied after the ideal state preparation channel, and the measurement involves an error channel $\mathcal{E}_{\text{meas}}$ applied before the ideal measurement channel. We assume that

$$\|\mathcal{E}_{\text{prep}} - \mathcal{I}\|_\diamond + \|\mathcal{E}_{\text{meas}} - \mathcal{I}\|_\diamond \leq \epsilon_{\text{SPAM}}.$$

Through a (β, b, t, τ) -phase estimation experiment, if we measure X_β^b in the end, we will obtain a random variable $s^x(t) \in \{\pm 1\}$. If we had the exact state $|\phi_t\rangle$ as defined in (40), then we would have $\mathbb{E}[s^x(t)] = \cos((\lambda_b - \lambda_0)t)$. However, due to the Hamiltonian reshaping error and the SPAM error, we have

$$|\mathbb{E}[s^x(t)] - \cos((\lambda_b - \lambda_0)t)| \leq \frac{4M^2 t^2}{r} + \epsilon_{\text{SPAM}}. \quad (47)$$

Notice that the first term on the right-hand side comes from the Hamiltonian reshaping error, where we set $t = r\tau$ as in Theorem 4. Moreover the variance of $s^x(t)$ is at most 1 because it can only take values ± 1 . We assume that $\epsilon_{\text{SPAM}} \leq 1/(3\sqrt{2})$, and choose r to be $r = \mathcal{O}(M^2 t^2)$ so that

$$\frac{4M^2 t^2}{r} < \frac{1}{3\sqrt{2}}.$$

Note that τ and r are related through $\tau = t/r$. Then we have

$$|\mathbb{E}[s^x(t)] - \cos((\lambda_b - \lambda_0)t)| < \frac{2}{3\sqrt{2}}.$$

We then take 54 independent samples of $s^x(t)$ and average them, denoting the sample average by $X(t)$. By Chebyshev's inequality, this ensures

$$\begin{aligned} \Pr[|X(t) - \mathbb{E}[s^x(t)]| \geq 1/(3\sqrt{2})] \\ = \Pr[|X(t) - \mathbb{E}[X(t)]| \geq 1/(3\sqrt{2})] \\ \leq \frac{1}{54 \times 1/(3\sqrt{2})^2} = \frac{1}{3}. \end{aligned}$$

Therefore combining the above with (47) we have

$$\Pr[|X(t) - \cos((\lambda_t - \lambda_0)t)| \geq 1/\sqrt{2}] \leq 1/3.$$

This guarantees estimating $\cos((\lambda_t - \lambda_0)t)$ to a constant $1/\sqrt{2}$ accuracy with at least $2/3$ probability, which gives us the $X(t)$ required in the robust frequency estimation protocol in Theorem 6. The $Y(t)$ in Theorem 6 can be similarly obtained. We also note that because $|\lambda_b - \lambda_0| \leq 2M$, we can set $A = 2M$ in Theorem 6. Therefore we can state the following for the phase estimation experiment (using the notation of $\mathcal{K}_\beta$ as defined in (16) and $\mathbb{P}_n^{(k)}$ for the set of Pauli operators on at most k qubits as defined in (15)):

Theorem 7: We assume that the quantum system is evolving under a k -body Hamiltonian with M terms (Definition 2). With N_{exp} independent non-adaptive $(\beta, b_j, t_j, \tau_j)$ -phase estimation experiments (Definition 3), $j = 1, 2, \dots, N_{\text{exp}}$, with the SPAM error (Definition 4) satisfying $\epsilon_{\text{SPAM}} \leq 1/(3\sqrt{2})$, we can obtain an estimate $\hat{\theta}$ such that

$$\Pr[|\hat{\theta} - (\lambda_b - \lambda_0)| \geq \eta] \leq q,$$

where λ_0 and λ_b (defined in (36)) are eigenvalues of the effective Hamiltonian H_{eff} defined in (33). In the above N_{exp} , $\{t_j\}$, and $\{\tau_j\}$ satisfy

$$N_{\text{exp}} = \mathcal{O}(\log(M/\eta)(\log(1/q) + \log \log(M/\eta))),$$

$$T = \sum_{j=1}^{N_{\text{exp}}} t_j = \mathcal{O}\left(\frac{1}{\eta}(\log(1/q) + \log \log(M/\eta))\right),$$

$$\max_j t_j = \mathcal{O}(1/\eta), \quad (48)$$

and $\tau_j = \Omega(1/(M^2 t_j))$. $D = \sum_{l=0}^k \binom{n}{l} = \Theta(n^k)$.

C. Learning a Completely Commuting Hamiltonian

In the last section we have shown how to estimate eigenvalue differences $\lambda_b - \lambda_0$ through phase estimation experiments. Up to fixing a global phase, we have the eigenvalues λ_b (we will discuss the global phase issue in more detail later in this section). These eigenvalues can be used to recover the Hamiltonian coefficients through an inverse Hadamard transform as described in (38), but a naive approach would not be efficient as there are 2^n different eigenvalues to estimate. Note that we have not yet used the information that the coefficient vector μ_β contains at most M non-zero entries, and compressed sensing is a tool that utilizes exactly this information.

Before we apply compressed sensing, we first note that all the nonzero entries of μ_β of the unknown k -body Hamiltonian

must (by assumption) correspond to k -body Pauli terms, which already helps us reduce the size of the linear system. Let $\mu_\beta^{(k)}$ be the D -dimensional vector consisting of the entries of μ that correspond to k -body Pauli terms, where $D = \sum_{l=0}^k \binom{n}{l}$. Then we can write (37) as

$$\lambda = \mathbf{H}^{(k)} \mu_\beta^{(k)},$$

where $\mathbf{H}^{(k)}$ is the $2^n \times D$ weight- k Hadamard matrix defined in Definition 1.

As is commonly done in compressed sensing, we generate Γ independent samples of the rows of $\mathbf{H}^{(k)}$, use those rows to form a matrix $\mathbf{A}$, and use the corresponding entries of λ to form a vector $\mathbf{y}$. Then we solve the ℓ^1 -minimization problem in (23), where η is the estimation error upper bound for each entry of $\mathbf{y}$. When Γ satisfies (25), then by Theorem 5 the solution of the ℓ^1 -minimization problem (23) is guaranteed to be close to the actual coefficients. More precisely, denoting the solution by $\hat{\mu}_\beta^{(k)}$, we have $\|\hat{\mu}_\beta^{(k)} - \mu_\beta^{(k)}\|_p \leq C_2 M^{1/p-1/2} \eta$ for $1 \leq p \leq 2$.

In the above we assumed that we can estimate eigenvalues λ_b for any $b \in \{0, 1\}^n$ to precision η . This is impossible given the ambiguity coming from the global phase. However, we will explain why this does not affect our ability to learn the Hamiltonian. Note that the phase estimation experiment allows us to estimate $\lambda_b - \lambda_0$. We can therefore fix the global phase by first assuming $\lambda_0 = 0$ to obtain λ_b . This can be done by shifting the effective Hamiltonian, so that we are in fact estimating the eigenvalues of

$$\tilde{H}_{\text{eff}} = \left(-\sum_{c \neq 0} \mu_c\right) I + \sum_{c \neq 0} \mu_c P_c.$$

Therefore the above procedure guarantees that we can learn all the coefficients μ_c accurately. Note that these μ_c correspond to all Pauli terms that are k -body and are in $\mathcal{K}_\beta$. Therefore we have obtained a set of estimates $\hat{\mu}_P$ for all $P \in \mathcal{K}_\beta$ and $\text{wt}(P) \leq k$ that satisfies

$$\left(\sum_{P \in \mathcal{K}_\beta: 1 \leq \text{wt}(P) \leq k} |\hat{\mu}_P - \mu_P|^p\right)^{1/p} \leq C M^{1/p-1/2} \eta, \quad (49)$$

for $1 \leq p \leq 2$ and a universal constant C .

Multiplying the overhead of compressed sensing as described in (25) by the number of samples N_{exp} and total evolution time T described in Theorem 7, we arrive at the following theorem (using the notation of $\mathcal{K}_\beta$ as defined in (16) and $\mathbb{P}_n^{(k)}$ for the set of Pauli operators on at most k qubits as defined in (15)):

Theorem 8 (Learning the coefficients of terms in $\mathcal{K}_\beta$): We assume that the quantum system is evolving under a k -body Hamiltonian with M terms (Definition 2). With N_{exp} independent non-adaptive $(\beta, b_j, t_j, \tau_j)$ -phase estimation experiments (Definition 3), $j = 1, 2, \dots, N_{\text{exp}}$, with the SPAM error (Definition 4) satisfying $\epsilon_{\text{SPAM}} \leq 1/(3\sqrt{2})$, we can obtain estimates $\hat{\mu}_P$ for every $P \in \mathcal{K}_\beta \cap \mathbb{P}_n^{(k)}$, such that (49) holds with probability at least $1 - \delta_{\text{CS}} - \Gamma q$. In the above N_{exp} , $\{t_j\}$, and $\{\tau_j\}$ satisfy

$$N_{\text{exp}} = \mathcal{O}(\Gamma \log(M/\eta)(\log(1/q) + \log \log(M/\eta))),$$

$$T = \sum_{j=1}^{N_{\text{exp}}} t_j = \mathcal{O} \left(\frac{\Gamma}{\eta} (\log(1/q) + \log \log(M/\eta)) \right),$$

$$\max_j t_j = \mathcal{O}(1/\eta), \quad (50)$$

$$\Gamma = \mathcal{O} \left(M(\ln^2(M) \ln(M \ln(D)) \ln(D) + \ln(1/\delta_{\text{CS}})) \right), \quad (51)$$

and $\tau_j = \Omega(1/(M^2 t_j))$. $D = \sum_{l=0}^k \binom{n}{l} = \Theta(n^k)$.

The success probability in the theorem comes from considering all possible ways the protocol may fail and then taking the union bound: it may fail because the sub-sampled matrix in compressed sensing does not have the RIP property, which contributes a failure probability of δ_{CS} , and it may fail because any one of the eigenvalue estimations is not accurate enough, which contributes a failure probability of Γq . Note that the failure probability can be made small with little overhead, as the dependence on $1/q$ and $1/\delta_{\text{CS}}$ are both logarithmic.

D. Randomized Basis Selection

In the previous section we have discussed how to learn all the coefficients corresponding to Hamiltonian terms contained in the set $\mathcal{K}_\beta$ (defined in (16)). In this section, we will focus on the way to choose a set of $\mathcal{K}_\beta$ so that all the terms in the Hamiltonians are covered with large probability. We will show that uniformly randomly choosing $\mathcal{K}_\beta$ is a good strategy.

To see this, we observe that by uniformly randomly sampling a $\beta \in \{x, y, z\}^n$, for a fixed Pauli operator $P \in \mathbb{P}_n$,

$$\Pr[P \in \mathcal{K}_\beta] = 3^{-\text{wt}(P)} \geq 3^{-k}.$$

Therefore, each time we learn a randomly sample β , there is at least 3^{-k} probability that we will learn the coefficient μ_P for $\text{wt}(P) \leq k$. If we generate L independent samples of β , then the probability of a term P being not covered in any of the L instances is at most $(1 - 3^{-k})^L$. By the union bound, to ensure that all the M terms are included in at least one of the L instances with probability at least $1 - \delta_{\text{basis}}$, we need

$$M(1 - 3^{-k})^L \leq M \exp(3^{-k}L) \leq \delta_{\text{basis}}. \quad (52)$$

It is thus sufficient to choose $L = 3^k \log(M/\delta_{\text{basis}})$. Taking this overhead L into account, we then arrive at our main result (using the notation of $\mathcal{K}_\beta$ as defined in (16) and $\mathbb{P}_n^{(k)}$ for the set of Pauli operators on at most k qubits as defined in (15)):

Theorem 9 (Learning a k -body Hamiltonian containing M terms): We assume that the quantum system is evolving under a k -body Hamiltonian with M terms (Definition 2). With N_{exp} independent non-adaptive (β, b_j, t_j, τ) -phase estimation experiments (Definition 3), $j = 1, 2, \dots, N_{\text{exp}}$, with the SPAM error (Definition 4) satisfying $\epsilon_{\text{SPAM}} \leq 1/(3\sqrt{2})$, we can obtain estimates $\hat{\mu}_P$ for every $P \in \mathbb{P}_n^{(k)}$ such that, with probability at least $1 - \delta$

$$\left(\sum_{P \in \mathbb{P}_n^{(k)} \setminus \{I\}} |\hat{\mu}_P - \mu_P|^p \right)^{1/p} \leq \epsilon, \quad (53)$$

for $1 \leq p \leq 2$. In the above N_{exp} , $\{t_j\}$, and τ satisfy

$$N_{\text{exp}} = \tilde{\mathcal{O}}(3^k M),$$

$$T = \sum_j t_j = \tilde{\mathcal{O}} \left(\frac{9^k M^{1/p+1/2}}{\epsilon} \right),$$

and $\tau = \Omega(3^{-k} \epsilon / (M^{1/p+3/2} \log(M/\delta)))$.

In the above, we choose τ to be the same for all experiments for simplicity. One can also choose τ differently for each experiment to reduce the total number of Pauli operators that are needed, as is done in the remark below. We are primarily concerned with the cases with $p = 1$ and $p = 2$, which have specific operational interpretations that we will discuss in Section VII, but our result holds for all $1 \leq p \leq 2$.

Remark 3: In the above we have used the $\tilde{\mathcal{O}}$ -notation to hide polylogarithmic factors. Here we provide a more precise statement. We constructed a protocol that with probability at least $1 - L\delta_{\text{CS}} - L\Gamma q - \delta_{\text{basis}}$, can obtain estimates $\hat{\mu}_P$ for $P \in \mathbb{P}_n^{(k)}$ such that

$$\left(\sum_{P \in \mathbb{P}_n^{(k)} \setminus \{I\}} |\hat{\mu}_P - \mu_P|^p \right)^{1/p} \leq CLM^{1/p-1/2} \eta,$$

for $1 \leq p \leq 2$ and a positive universal constant C . This protocol uses

$$N_{\text{exp}} = \mathcal{O}(L\Gamma \log(M/\eta)(\log(1/q) + \log \log(M/\eta))),$$

experiments and

$$T = \sum_{j=1}^{N_{\text{exp}}} t_j = \mathcal{O} \left(\frac{L\Gamma}{\eta} (\log(1/q) + \log \log(M/\eta)) \right),$$

$$\max_j t_j = \mathcal{O}(1/\eta), \quad (54)$$

total evolution time. In the above $\{t_j\}$, $\{\tau_j\}$, Γ , and L satisfy

$$\Gamma = \mathcal{O} \left(M(\ln^2(M) \ln(M \ln(D)) \ln(D) + \ln(1/\delta_{\text{CS}})) \right), \quad (55)$$

$$L = \mathcal{O} \left(3^k \log(M/\delta_{\text{basis}}) \right), \quad (56)$$

and $\tau_j = \Omega(1/(M^2 t_j))$. $D = \sum_{l=0}^k \binom{n}{l} = \Theta(n^k)$.

E. Robustness to Modeling Errors

In Theorem 5 we have shown that the Hamiltonian learning protocol is robust against SPAM error. In this section we will discuss the robustness against modeling errors. More precisely, we will ask two questions: (1) What if the Hamiltonian contains more than M terms? (2) What if the terms in the Hamiltonian are not exactly k -body? We will show that our learning protocol can still work if these two types of modeling errors occur.

The robustness against the first type of modeling error, i.e., that the Hamiltonian contains more than M terms, but all terms are still k -body, is a direct consequence of Theorem 5, in which an additional error term $\sigma_M(\mathbf{x})_1$ (defined in (24)) is included to account for the error of approximating a potentially non-sparse vector with an M -sparse one. In the context of Theorem 5, this means an extra term is included in the estimation error for the coefficients, which gives us

$$\left(\sum_{P \in \mathbb{P}_n^{(k)} \setminus \{I\}} |\hat{\mu}_P - \mu_P|^p \right)^{1/p} \leq \frac{C3^k \sigma_M(\mu)_1 \log(M/\delta)}{M^{1-1/p}} + \epsilon, \quad (57)$$

for some universal constant C , where ϵ is the error bound without modeling errors as defined in (53), and μ is the vector consisting of μ_P for all $P \in \mathbb{P}_n \setminus \{I\}$. Here we have used the fact that the best M -sparse approximation error for terms in $\mathcal{K}_\beta$ must be smaller than or equal to that for the whole coefficient vector. From the above we can see that if the Hamiltonian can be well-approximated by one with M k -body terms, then we can still obtain an accurate estimate for its leading coefficients.

For the second type of modeling error, we will first focus on the procedure of learning terms in $\mathcal{K}_\beta$ described in Section V-C. The coefficients for all terms in $\mathcal{K}_\beta$ are collected into a vector μ_β . Those that are also k -body are collected into $\mu_\beta^{(k)}$, while the rest form a vector $\bar{\mu}_\beta^{(k)}$. Here μ_β contains $D = \sum_{l=0}^k \binom{n}{l}$ entries.

We denote by $\tilde{\mathbf{A}}$ the submatrix of the original Walsh-Hadamard matrix $\mathbf{H}$ consisting of the Γ rows corresponding to the randomly sampled rows of A in Theorem 5. We can write $\tilde{\mathbf{A}}$ as $\tilde{\mathbf{A}} = [\mathbf{A} \ \mathbf{B}]$, where $\mathbf{A} \in \mathbb{R}^{\Gamma \times D}$ is the submatrix correspond to rows of the weight- k Hadamard matrix as defined in Definition 1, and $\mathbf{B} \in \mathbb{R}^{\Gamma \times (2^n - D)}$. Then the eigenvalue estimates $\mathbf{y}$ and the coefficients are related through

$$\mathbf{y} = \tilde{\mathbf{A}}\mu_\beta + \mathbf{e} = \mathbf{A}\mu_\beta^{(k)} + \mathbf{B}\bar{\mu}_\beta^{(k)} + \mathbf{e}, \quad (58)$$

where $\mathbf{e}$ is the vector consisting of errors on each eigenvalue estimate, and using the phase estimation experiments described in Section V-B we can ensure that each entry of $\mathbf{e}$ is at most η in absolute value, thus ensuring $\|\mathbf{e}\| \leq \sqrt{\Gamma}\eta$. We therefore have

$$\|\mathbf{y} - \mathbf{A}\mu_\beta^{(k)}\|_2 \leq \|\mathbf{B}\bar{\mu}_\beta^{(k)}\|_2 + \sqrt{\Gamma}\eta. \quad (59)$$

Since each entry of $\mathbf{B}$ is ± 1 , we can prove that $\|\mathbf{B}\mathbf{w}\|_2 \leq \sqrt{\Gamma}\|\mathbf{w}\|_1$ for any $(2^n - D)$ -dimensional vector $\mathbf{w}$, and therefore

$$\|\mathbf{y} - \mathbf{A}\mu_\beta^{(k)}\|_2 \leq \sqrt{\Gamma}(\eta + \|\bar{\mu}_\beta^{(k)}\|_1). \quad (60)$$

This motivates us to solve a modified ℓ^1 -minimization problem

$$\underset{\mathbf{z} \in \mathbb{C}^D}{\text{minimize}} \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2 \leq \sqrt{\Gamma}\tilde{\eta}, \quad (61)$$

where $\tilde{\eta}$ is any number such that $\tilde{\eta} \geq \eta + \|\bar{\mu}_\beta^{(k)}\|_1$. Solving this ℓ^1 -minimization problem yields a solution, which we denote by $\hat{\mu}_\beta^{(k)}$. By Theorem 5 we have

$$\|\hat{\mu}_\beta^{(k)} - \mu_\beta^{(k)}\|_p \leq \frac{C_1}{M^{1-1/p}} \sigma_M(\mu_\beta^{(k)})_1 + C_2 M^{1/p-1/2} \tilde{\eta}. \quad (62)$$

We let $\epsilon^{(k)}$ be a known upper bound of $\|\bar{\mu}_\beta^{(k)}\|_1$, and considering all $3^k \log(M/\delta)$ randomly sampled indices β , we can therefore produce estimates $\hat{\mu}_P$ for all k -body coefficients satisfying (63), shown at the bottom of the page, for universal constants $C_1, C_2 > 0$, with the same total evolution time as given in Theorem 9. Note that this error bound include both types of modeling errors.

F. Computational Complexity

In this section we will show that the classical post-processing takes time that is polynomial in n and $\log(1/\epsilon_1)$ where n is the number of qubits and ϵ_1 is the allowed ℓ^1 -error on the coefficients. The classical post-processing comprises solving $\text{poly}(n, \log(1/\epsilon_1))$ ℓ^1 -minimization problems as described in (23), which we restate here:

$$\underset{\mathbf{z} \in \mathbb{C}^D}{\text{minimize}} \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta \sqrt{\Gamma}.$$

Here $\mathbf{A}$ is a $\Gamma \times D$ submatrix of the Hadamard matrix. An important feature that we will use later is that all entries of $\mathbf{A}$ are at most 1 in absolute value (in fact they are all ± 1). Moreover, the admissible set of this optimization problem is non-empty, since the exact coefficient vector $\mathbf{x}$ that we want to recover satisfies $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \sqrt{\Gamma}\eta$, and $\mathbf{x}$ is M -sparse whose entries are bounded by 1 in absolute value. We only need to show that approximately solving each of these ℓ^1 -minimization problems takes time $\text{poly}(n, \log(1/\epsilon_1))$.

We will first demonstrate how this problem can be reformulated as a semidefinite programming (SDP) problem to build intuition. Because certain SDPs are pathological (for example, the optimal solution is exponentially large), this does not immediately imply that the solution is efficient. We therefore provide a self-contained analysis of the computational complexity by examining its dual (67), and we rigorously prove that the classical computation requires weak polynomial time.

1) *Reformulation to an SDP:* We first make the substitution

$$\mathbf{z} = \mathbf{z}^+ - \mathbf{z}^-, \quad \mathbf{v} = \mathbf{A}(\mathbf{z}^+ - \mathbf{z}^-) - \mathbf{y},$$

where $\mathbf{z}^\pm = (z_1^\pm, z_2^\pm, \dots, z_D^\pm)$, and they are related to the $\mathbf{z}$ in (23) via $\mathbf{z} = \mathbf{z}^+ - \mathbf{z}^-$ and $z_i^\pm = \max\{z_i, 0\}$, $z_i^- = -\min\{0, z_i\}$. Then (23) can be written as:

$$\begin{aligned} &\text{minimize } \mathbf{z}^+ + \mathbf{z}^- \\ &\text{subject to: } \|\mathbf{v}\|_2 \leq \eta \sqrt{\Gamma}, \mathbf{z}^+ \geq 0, \quad \mathbf{z}^- \geq 0. \end{aligned} \quad (64)$$

Introduce the block matrix

$$\mathbf{W} = \begin{pmatrix} \eta^2 \Gamma & \mathbf{v}^* \\ \mathbf{v} & \mathbf{I} \end{pmatrix}, \quad (65)$$

where the first entry is a scalar, and the lower right block is an identity matrix. Then $\mathbf{W} \geq 0$ is equivalent to $\|\mathbf{v}\|_2 \leq \eta \sqrt{\Gamma}$ by a standard inequality on such block matrices.

The optimization becomes

$$\begin{aligned} &\text{minimize } \mathbf{z}_+ + \mathbf{z}_- \\ &\text{subject to: } \mathbf{z}_+ \geq 0, \quad \mathbf{z}_- \geq 0, \\ &\quad \mathbf{v} = \mathbf{A}(\mathbf{z}_+ - \mathbf{z}_-) - \mathbf{y}, \\ &\quad \mathbf{W} \geq 0, \quad \mathbf{W} = \begin{pmatrix} \eta^2 \Gamma & \mathbf{v}^* \\ \mathbf{v} & \mathbf{I} \end{pmatrix}, \end{aligned} \quad (66)$$

where the objective is linear and all the constraints are positive semi-definite constraints, meaning that this is an SDP.

$$\left(\sum_{P \in \mathbb{P}_n^{(k)} \setminus \{I\}} |\hat{\mu}_P - \mu_P|^p \right)^{1/p} \leq \epsilon + \frac{C_1 3^k \sigma_M(\mu)_1 \log(M/\delta)}{M^{1-1/p}} + \frac{C_2 3^k \epsilon^{(k)}}{M^{1/2-1/p}}, \quad (63)$$

In practice, semidefinite programs can be solved efficiently. However, to rigorously prove that (23) is solvable in polynomial time, a more careful analysis of the optimization error and the properties of the constraints, as in the following section, is required within the context of the Hamiltonian learning scenario we are considering.

2) *Self-Contained Analysis*: For the ℓ^1 -minimization problem in (23), we consider a related problem in the following form

$$\text{minimize } \|\mathbf{Az} - \mathbf{y}\|_2 \quad \text{subject to } \|\mathbf{z}\|_1 \leq \xi. \quad (67)$$

This is equivalent to a convex quadratic program,

$$\begin{aligned} & \text{minimize } \frac{1}{2} \|\mathbf{A}(\mathbf{z}^+ - \mathbf{z}^-) - \mathbf{y}\|_2^2, \\ & \text{subject to } \sum_i (z_i^+ + z_i^-) \leq \xi, \quad \mathbf{z}^+ \geq \mathbf{0}, \quad \mathbf{z}^- \geq \mathbf{0}, \end{aligned} \quad (68)$$

where $\mathbf{z}^\pm = (z_1^\pm, z_2^\pm, \dots, z_D^\pm)$, and they are related to the $\mathbf{z}$ in (67) via $\mathbf{z} = \mathbf{z}^+ - \mathbf{z}^-$ as before. This convex quadratic program can be solved in time $\text{poly}(\Gamma, D, \log(1/\epsilon'))$ so that the $\|\mathbf{Az} - \mathbf{y}\|_2$ obtained is ϵ' away from the optimal value [79].

We will first study the properties of the optimization problem (67) and discuss how it helps us solve the original ℓ^1 minimization problem (23). We denote by $\mathbf{z}(\xi)$ the solution to (67) (which may not be unique, but any one of the solutions can be used) and by $f(\xi)$ the corresponding value of the objective function, i.e.

$$f(\xi) = \|\mathbf{Az}(\xi) - \mathbf{y}\|_2.$$

We note that when ξ is large enough, the constraint in (67) will have no effect, and the optimization problem is solved by any $\mathbf{z}$ that minimizes $\|\mathbf{Az} - \mathbf{y}\|_2$, which is equivalent to $\mathbf{A}^\dagger \mathbf{Az} = \mathbf{A}^\dagger \mathbf{y}$. The smallest value for ξ for which this happens is

$$\xi_0 = \min\{\|\mathbf{z}\|_1 : \mathbf{A}^\dagger \mathbf{Az} = \mathbf{A}^\dagger \mathbf{y}\}. \quad (69)$$

Lemma 2: Let $\mathbf{A}^+$ be the Moore–Penrose pseudoinverse of $\mathbf{A}$.

- (i) If $\xi \geq \xi_0$, then $f(\xi) = \sqrt{\mathbf{y}^\dagger \mathbf{y} - \mathbf{y}^\dagger \mathbf{A} \mathbf{A}^+ \mathbf{y}}$.
- (ii) If $0 \leq \xi \leq \xi_0$, then $\|\mathbf{z}(\xi)\|_1 = \xi$ and $f(\xi)$ is a strictly decreasing function of ξ .
- (iii) f is Lipschitz continuous with Lipschitz constant $\sqrt{\Gamma}$. More precisely, for any $0 \leq \xi_1 \leq \xi_2$, we have

$$f(\xi_2) \leq f(\xi_1) \leq f(\xi_2) + \sqrt{\Gamma}(\xi_2 - \xi_1).$$

Proof: For (i), we note that the constraint in (67) no longer has any effect and we are solving an unconstrained quadratic optimization problem. Solving this unconstrained quadratic optimization problem readily yields the expression for $f(\xi)$, which does not depend on ξ .

For (ii), if $\|\mathbf{z}(\xi)\|_1 < \xi$, then there exists an open neighborhood U of $\mathbf{z}(\xi)$ such that $U \subset \{\mathbf{z} : \|\mathbf{z}\|_1 \leq \xi\}$. Because $\mathbf{z}(\xi)$ minimizes $\|\mathbf{Az} - \mathbf{y}\|_2^2$ within the open neighborhood U , the gradient of this objective function must be $\mathbf{0}$ at $\mathbf{z}(\xi)$. This implies $\mathbf{A}^\dagger \mathbf{Az}(\xi) = \mathbf{A}^\dagger \mathbf{y}$. By (69), we have $\|\mathbf{z}(\xi)\|_1 \geq \xi_0$. Therefore $\xi_0 \leq \|\mathbf{z}(\xi)\|_1 < \xi$, which contradicts $\xi \leq \xi_0$. This proves $\|\mathbf{z}(\xi)\|_1 = \xi$.

Next we will prove that $f(\xi)$ is strictly decreasing for $0 \leq \xi \leq \xi_0$. If for $\xi_1, \xi_2 \in [0, \xi_0]$ we have $\xi_1 < \xi_2$

and $f(\xi_1) \leq f(\xi_2)$, then $\|\mathbf{Az}(\xi_1) - \mathbf{y}\|_2 \leq \|\mathbf{Az}(\xi_2) - \mathbf{y}\|_2$, which implies that $\mathbf{z}(\xi_1)$ is a solution to the optimization problem (67) with $\xi = \xi_2$. Note that $\mathbf{z}(\xi_1)$ is in the interior of the admissible set of the this optimization problem, and therefore the gradient argument in the preceding paragraph applies, giving us $\mathbf{A}^\dagger \mathbf{Az}(\xi_1) = \mathbf{A}^\dagger \mathbf{y}$, which results in the same contradiction. This proves that $f(\xi_1) > f(\xi_2)$.

For (iii), we only need to prove that $f(\xi_1) \leq f(\xi_2) + \sqrt{\Gamma}(\xi_2 - \xi_1)$ for $\xi_1 \leq \xi_2 \leq \xi_0$. We define $\mathbf{z}' = (\xi_1/\xi_2)\mathbf{z}(\xi_2)$. Then by (ii) we have $\|\mathbf{z}' - \mathbf{z}_2\|_1 = (1 - \xi_1/\xi_2)\|\mathbf{z}(\xi_2)\|_1 = \xi_2 - \xi_1$. Also, $\|\mathbf{z}'\| = \xi_1$ and $\mathbf{z}'$ is therefore in the admissible set of (67) with $\xi = \xi_1$. Consequently

$$\begin{aligned} f(\xi_1) & \leq \|\mathbf{Az}' - \mathbf{y}\|_2 \leq \|\mathbf{Az}(\xi_2) - \mathbf{y}\|_2 + \|\mathbf{Az}(\xi_2) - \mathbf{Az}'\|_2 \\ & = f(\xi_2) + \|\mathbf{A}(\mathbf{z}(\xi_2) - \mathbf{z}')\|_2. \end{aligned} \quad (70)$$

Note that $\|\mathbf{A}(\mathbf{z}(\xi_2) - \mathbf{z}')\|_2 \leq \|\mathbf{A}\|_{1 \rightarrow 2} \|\mathbf{z}(\xi_2) - \mathbf{z}'\|_1 \leq \|\mathbf{A}\|_{1 \rightarrow 2} (\xi_2 - \xi_1)$, where

$$\|\mathbf{A}\|_{1 \rightarrow 2} := \max_{\|\mathbf{z}\|_1 \leq 1} \|\mathbf{Az}\|_2.$$

Because all entries of $\mathbf{A}$ are ± 1 , we therefore have $\|\mathbf{A}\|_{1 \rightarrow 2} \leq \sqrt{\Gamma}$. Hence $\|\mathbf{A}(\mathbf{z}(\xi_2) - \mathbf{z}')\|_2 \leq \sqrt{\Gamma}(\xi_2 - \xi_1)$. This combined with (70) then results in the desired inequality. $\square$

We then show that the above properties of $f(\xi)$ allow us to solve (23) by finding the right ξ . We consider the following cases for (23): In the first case, $\sqrt{\Gamma}\eta < \sqrt{\mathbf{y}^\dagger \mathbf{y} - \mathbf{y}^\dagger \mathbf{A} \mathbf{A}^+ \mathbf{y}} = f(\xi_0)$. In this scenario, the admissible set is empty, but we do not need to worry about this scenario because in our setup the exact coefficient vector is always in the admissible set. In the second case, $\sqrt{\Gamma}\eta = f(\xi_0)$. The inequality constraint is then equivalent to $\mathbf{A}^\dagger \mathbf{Az} = \mathbf{A}^\dagger \mathbf{y}$, an equality constraint. Therefore (23) can be reduced to a linear program and solved in weak polynomial time. Therefore, the only case we need to examine closely is the third case, where $\sqrt{\Gamma}\eta > \sqrt{\mathbf{y}^\dagger \mathbf{y} - \mathbf{y}^\dagger \mathbf{A} \mathbf{A}^+ \mathbf{y}}$.

Lemma 3: We assume $\sqrt{\Gamma}\eta > \sqrt{\mathbf{y}^\dagger \mathbf{y} - \mathbf{y}^\dagger \mathbf{A} \mathbf{A}^+ \mathbf{y}}$. If there exists $\mathbf{x}$ such that $\|\mathbf{Ax} - \mathbf{y}\|_2 \leq \sqrt{\Gamma}\eta$, and $\|\mathbf{y}\|_2 > \sqrt{\Gamma}\eta$, then there exists a unique $\xi^* \in [0, \min\{\|\mathbf{x}\|_1, \xi_0\}]$ such that $f(\xi^*) = \sqrt{\Gamma}\eta$, and $\mathbf{z}(\xi^*)$ is a solution of (23).

Algorithm 1 Solving the ℓ^1 -Minimization Problem (23)

Input: $\mathbf{A} \in \mathbb{C}^{\Gamma \times D}, \mathbf{y} \in \mathbb{C}^\Gamma, M \in \mathbb{N}_+, \eta > 0$, error tolerance $\nu > 0$.

$\xi_L \leftarrow 0, \xi_R \leftarrow M$;

while $\xi_R - \xi_L > 2\nu$ **do**

Solve the convex quadratic program (68) with $\xi = (\xi_L + \xi_R)/2$ to obtain the minimum $f(\xi)^2/2$;

if $f(\xi) > \sqrt{\Gamma}\eta$ **then**

$\xi_L \leftarrow (\xi_L + \xi_R)/2$;

else

$\xi_R \leftarrow (\xi_L + \xi_R)/2$;

end if

end while

$\tilde{\xi} = (\xi_L + \xi_R)/2$;

Solve the convex quadratic program (68) with $\xi = \tilde{\xi}$ to obtain the solution $\tilde{\mathbf{x}}^\#$;

Output: $\tilde{\mathbf{x}}^\#$.

Proof: Because $f(\|\mathbf{x}\|_1) \leq \|\mathbf{Ax} - \mathbf{y}\|_2 \leq \sqrt{\Gamma}\eta$, and $f(0) = \|\mathbf{y}\|_2$, we have

$$\max\{f(\|\mathbf{x}\|_1), f(\xi_0)\} \leq \sqrt{\Gamma}\eta < f(0),$$

where we have used Lemma 2 (i) which states that $f(\xi_0) = \sqrt{\mathbf{y}^\dagger \mathbf{y} - \mathbf{y}^\dagger \mathbf{A} \mathbf{A}^\dagger \mathbf{y}}$. Because f is continuous and strictly decreasing in $[0, \xi_0]$, we therefore have a unique ξ^* such that $f(\xi^*) = \sqrt{\Gamma}\eta$, and

$$0 < \xi^* \leq \min\{\|\mathbf{x}\|_1, \xi_0\}.$$

□

We then show that solving the optimization problem (67) with the ξ^* in Lemma 3 yields the solution to the ℓ^1 -minimization problem (23).

Lemma 4: With the same assumptions as Lemma 3, $\mathbf{z}(\xi^*)$ solves the ℓ^1 -minimization problem (23).

Proof: First note that because $f(\xi^*) = \sqrt{\Gamma}\eta$, $\mathbf{z}(\xi^*)$ is in the admissible set of (23). If there exists $\mathbf{z}'$ in the admissible set of (23) such that

$$\|\mathbf{z}'\|_1 < \|\mathbf{z}(\xi^*)\|_1 = \xi^*, \quad (71)$$

where we have used $0 < \xi^* \leq \xi_0$ and Lemma 2 (ii) to show $\|\mathbf{z}(\xi^*)\|_1 = \xi^*$, then

$$f(\|\mathbf{z}'\|_1) \leq \|\mathbf{Az}' - \mathbf{y}\|_2 \leq \sqrt{\Gamma}\eta = f(\xi^*),$$

implying $\|\mathbf{z}'\|_1 \geq \xi^*$ by Lemma 2 (ii), which contradicts (71). □

Therefore solving the ℓ^1 -minimization problem (23) reduces to finding ξ^* , which can be done through a bisection procedure as described in Algorithm V-F2. Each bisection step involves solving (67) for a certain ξ , which is equivalent to the convex quadratic program (68). In the bisection process, we start from an interval $[0, M]$, in each iteration compute $f(\xi)$ where ξ is the mid-point of the interval, thereby deciding whether ξ^* is in the right half of the interval or the left half. This allows us to get a new interval with half the size but still contains ξ^* , thereby proceeding to the next iteration. When the interval size is 2ν we can take the mid-point $\tilde{\xi}^*$ as the estimate for ξ^* , which ensures

$$|\tilde{\xi}^* - \xi^*| \leq \nu.$$

The total number of iterations needed is at most $\lceil \log_2(M/\nu) \rceil$.

However, our ultimate goal is to recover the original coefficient vector $\mathbf{x}$, and we need to know how small ν needs to be in order for the approximate solution $\tilde{\mathbf{x}}^\#$ (obtained by solving (68) with $\xi = \tilde{\xi}^*$) to be close to $\mathbf{x}$. Next we will provide an answer to this question.

Note that solving (68) with $\xi = \tilde{\xi}^*$ is equivalent to solving the ℓ^1 -minimization problem (23) but with η replaced with some $\tilde{\eta}$. This $\tilde{\eta}$ can be identified by $f(\tilde{\xi}^*) = \sqrt{\Gamma}\tilde{\eta}$. Therefore by the Lipschitz continuity of f (Lemma 2 (iii)) we have

$$|\sqrt{\Gamma}\tilde{\eta} - \sqrt{\Gamma}\eta| = |f(\tilde{\xi}^*) - f(\xi^*)| \leq \sqrt{\Gamma}|\tilde{\xi}^* - \xi^*|,$$

which leads to

$$|\tilde{\eta} - \eta| \leq |\tilde{\xi}^* - \xi^*| \leq \nu.$$

From the above we can see that in the end we obtain an approximate solution $\tilde{\mathbf{x}}^\#$ that corresponds to the exact solution

of (23) but with η replaced with some $\tilde{\eta}$ satisfying $\tilde{\eta} \leq \eta + \nu$. We can therefore use Theorem 5 to obtain the following guarantee about the closeness to $\mathbf{x}$:

Theorem 10 (Compressed sensing with approximate solution): Let $D = \sum_{l=0}^k \binom{n}{l}$. Let $\mathbf{A} \in \mathbb{C}^{\Gamma \times D}$ be a matrix whose rows are independently randomly sampled from the rows of $\mathbf{H}^{(k)}$ with replacement. Let $\delta_{\text{CS}} \in (0, 1)$ and let $M > 0$ be an integer. If

$$\Gamma \geq CM \max\{\ln^2(M) \ln(M \ln(D)) \ln(D), \ln(1/\delta_{\text{CS}})\}, \quad (72)$$

then, with probability at least $1 - \delta_{\text{CS}}$, the following statement holds for every M -sparse $\mathbf{x} \in \mathbb{C}^D$ whose entries are bounded by 1 in absolute value. Let noisy samples $\mathbf{y} = \mathbf{Ax} + \mathbf{e}$ be given with $\|\mathbf{e}\|_2 \leq \eta \sqrt{\Gamma}$, and let $\tilde{\mathbf{x}}^\#$ be the approximate solution of the ℓ^1 -minimization problem (23) produced by Algorithm V-F2 with the error tolerance on ξ^* chosen to be $\nu > 0$. Then

$$\left\| \mathbf{x} - \tilde{\mathbf{x}}^\# \right\|_p \leq C_1 M^{1/p-1/2} (\eta + \nu), \quad 1 \leq p \leq 2. \quad (73)$$

Both constants C, C_1 are universal. The approximate solution $\tilde{\mathbf{x}}^\#$ is obtained through Algorithm V-F.2 which runs in time $\text{poly}(n, \Gamma, \log(1/\nu))$.

VI. LOWER BOUNDS

In this section we provide a lower bound for the Hamiltonian learning task we consider. Specifically we consider the dependence on the accuracy ϵ and the number of Pauli terms M . The adaptive experiments are modeled as in [69], where all experiments form a tree such that the outcome of the experiment at a vertex determines which child leaf to move to, thus determining the next experiment to perform. Our main tool for the total evolution time lower bound is Assouad's lemma [70], which provides a lower bound on the achievable ℓ^1 -error given two prerequisites: (1) an estimate of how hard it is to distinguish two output probability distributions if they come from two Hamiltonians that differ slightly, and (2) a lower bound on the penalty in ℓ^1 -error if such a pair of Hamiltonians are not correctly distinguished. The second prerequisite is easy to fulfill, as discussed in the proof of Theorem 11. For the first prerequisite, we will follow [5] to capture the difficulty of correctly distinguishing the output probability distributions by induction on the tree of adaptive experiments. In the following, we first characterize the model of quantum learning experiments following [69]. Details of the proofs are in Appendix C.

A. Model of Quantum Experiments

We consider the scenario of learning the unknown Hamiltonian H from dynamics. Consider the time evolution operator of H parameterized by time t :

$$U(t) = e^{-iHt}, \quad (74)$$

A learning agent can have access to $U(t)$ by quantum experiments without knowing H . The learning algorithm includes the scenario where each quantum experiment is chosen adaptively based on past measurement outcomes. Following [69], we first define a single experiment:

Definition 5 (An ideal experiment): Given a n -qubit unitary $U(t) = e^{-iHt}$ parameterized by time t , with H being the unknown Hamiltonian. A single ideal experiment $E^{(0)}$ is specified by:

1. an arbitrary n' -qubit initial state $|\psi_0\rangle \in \mathbb{C}^{2^{n'}}$ with an integer $n' \geq n$,
2. an arbitrary POVM $\mathcal{F} = \{M_i\}_i$ on an n' -qubit system,
3. an n' -qubit unitary of the following form,

$$U_{\text{exp}} = U_{K+1} (U(t_K) \otimes I) U_K \dots U_2 (U(t_1) \otimes I) U_1, \quad (75)$$

for some arbitrary integer K , arbitrary evolution times $t_1, \dots, t_K \in \mathbb{R}$, and arbitrary n' -qubit unitaries $U_1, \dots, U_K, U_{K+1}$. Here I is the identity unitary on $n' - n$ qubits subsystem.

A single run of $E^{(0)}$ returns an outcome from performing the POVM $\mathcal{F} = \{M_i\}_i$ on the state $U|\psi_0\rangle$. The evolution time of the experiment is defined as

$$t(E^{(0)}) := \sum_{k=1}^K t_k. \quad (76)$$

Next, we will take noise into account. A small SPAM error is unavoidable in any quantum experiment, and being robust against it is an important consideration in useful algorithms for characterizing and benchmarking quantum systems [28], [69], [76], [80], [81], [82], [83]. We model the SPAM error by a global depolarizing channel

$$\mathcal{D} : \rho \mapsto \mathcal{D}(\rho) = (1 - \gamma)\rho + \frac{\gamma \text{tr}(\rho)}{2^{n'}} I, \quad (77)$$

which is applied to the system right before the POVM $\mathcal{F}$. This is equivalent to changing the POVM to

$$\mathcal{F}^{(\gamma)} = \{(1 - \gamma)M_i + \gamma \text{tr}(M_i) (I/2^{n'})\}_i. \quad (78)$$

The noise strength is measured by the constant γ , which can take any value in $(0, 1)$.

In proving the lower bound, we are allowed to only consider SPAM error in the form of the global depolarizing noise applied in the way described above, because any algorithm that is robust to SPAM error in general must be robust to this type of SPAM error in particular. We will then define a single experiment with measurement noise as follows:

Definition 6 (An experiment with measurement noise): A single experiment $E^{(\gamma)}$ with measurement noise $\gamma \in (0, 1)$ is identical to a single ideal experiment E^0 except for the POVM, which is $\mathcal{F}^{(\gamma)}$ given in (78) instead of $\mathcal{F} = \{M_i\}_i$.

Note that a lower bound for learning with measurement noise naturally also holds for learning with both state-preparation and measurement errors, since the latter is a more difficult task. Therefore for simplicity we only consider measurement noise in this section.

With the definition of a single experiment with measurement noise, we can now formally define a learning algorithm with total evolution time T as follows, where we again adopt the definition in [5].

Definition 7 (Learning algorithm with bounded total evolution time): Consider $T > 0, 0.5 > \gamma > 0$. A learning algorithm with total evolution time T and measurement noise γ can

obtain measurement outcomes from an arbitrary number of experiments $E_1^{(\gamma)}, E_2^{(\gamma)}, \dots$ as long as

$$\sum_i t(E_i^{(\gamma)}) \leq T \quad (79)$$

The parameters specifying each experiment $E_i^{(\gamma)}$ can depend on the measurement outcomes from previous experiments $E_1^{(\gamma)}, \dots, E_{i-1}^{(\gamma)}$.

B. Lower Bound of the Total Evolution Time in Hamiltonian Learning

With the definition of the learning algorithm and the possible sets of experiments, we now present a fundamental lower bound on the total evolution time for any learning algorithm that tries to learn an unknown n -qubit Hamiltonian consisting of M Pauli terms from real-time evolution. The theorem is stated as follows:

Theorem 11: Given integers n and M , real numbers $\epsilon, \delta \in (0, 1)$, and a set $\{P_1, P_2, \dots, P_M\} \in \mathbb{P}_n \setminus \{I\}$ representing the M n -qubit Pauli terms in the Hamiltonian, we consider any learning algorithm with a total evolution time T and a constant measurement noise $\gamma \in (0, 0.5)$ as defined in Definition 7. Suppose that such a learning algorithm satisfies that for an n -qubit Hamiltonian $H = \sum_{a=1}^M \mu_a P_a$ with any unknown parameters $|\mu_a| \leq 1$, after multiple rounds of noisy experiments, the algorithm can estimate any $\mu = (\mu_1, \dots, \mu_M)$ to ϵ_p -error in the ℓ^p -norm for $p \in [1, 2]$ in expectation value averaged over experimental outcomes. Then,

$$T \geq \frac{M^{1/p}}{2\epsilon_p e \log(1/\gamma)}. \quad (80)$$

Notice that the lower bound in Theorem 11 is for learning n -qubit Hamiltonians with fixed M terms. It naturally implies a lower bound for learning n -qubit M -term Hamiltonians without prior knowledge of which M terms are in the Hamiltonian. Moreover, by restricting the terms in the Hamiltonian to be k -body, this lower bound also holds for the k -body, M -term, n -qubit Hamiltonian learning problem we discussed in the previous sections.

Remark 4: As two special cases of Theorem 11 that are related to the operational interpretations that would be further discussed in Section VII, the lower bounds for considering the ℓ^1 - and ℓ^2 -norm of the error are

$$T^{(1)} \geq \frac{M}{2\epsilon_1 e \log(1/\gamma)}, \quad T^{(2)} \geq \frac{\sqrt{M}}{2\epsilon_2 e \log(1/\gamma)}. \quad (81)$$

We begin by considering how well a single noisy experiment can distinguish two Hamiltonians.

Lemma 5 (TV for one experiment, Lemma 30 [5]): Consider a noisy experiment as defined in Definition 6. Let $p(i)$ and $p'(i)$ be the probability of obtaining the measurement outcome i by performing the POVM $\mathcal{F}^{(\gamma)}$ when the unknown Hamiltonian is H and H' respectively. Then

$$\text{TV}(p, p') \leq (1 - \gamma) \min(\|H - H'\| \cdot t(E^{(\gamma)}), 1), \quad (82)$$

where $t(E^{(\gamma)}) = \sum_{k=1}^K |t_k|$ is the total evolution time in this single experiment $E^{(\gamma)}$. Here TV denotes the total variation distance between probability distributions.

Proof: We define

$$\begin{aligned} |\psi\rangle &= U_{K+1}(U(t_K) \otimes I) U_K \dots U_2(U(t_1) \otimes I) U_1 |\psi_0\rangle, \\ |\psi'\rangle &= U_{K+1}(U'(t_K) \otimes I) U_K \dots U_2(U'(t_1) \otimes I) U_1 |\psi_0\rangle \end{aligned} \quad (83)$$

By the triangle inequality and telescoping sum, we have the following upper bound on the trace distance,

$$\begin{aligned} |||\psi\rangle\langle\psi| - |\psi'\rangle\langle\psi'| |||_1 &\leq \sum_{k=1}^K \|U(t_k) - U'(t_k)\|_0 \\ &\leq 2\|H - H'\| \cdot \sum_{k=1}^K |t_k| \\ &= 2\|H - H'\| \cdot t(E^{(\gamma)}), \end{aligned} \quad (84)$$

where we have used Lemma 9 in the second inequality. Because the ideal POVM measurement $\mathcal{F} = \{M_i\}_i$ is a quantum channel, we have

$$\begin{aligned} &\frac{1}{2} \sum_i |\langle\psi|M_i|\psi\rangle - \langle\psi'|M_i|\psi'\rangle| \\ &\leq \frac{1}{2} |||\psi\rangle\langle\psi| - |\psi'\rangle\langle\psi'| |||_1 \\ &\leq \min(\|H - H'\| \cdot t(E^{(\gamma)}), 1). \end{aligned} \quad (85)$$

Now we take measurement noise into account. Since the noisy POVM $\mathcal{F}^{(\gamma)} = \{\tilde{M}_i\}$ where

$$\tilde{M}_i = (1 - \gamma)M_i + \gamma \text{tr}(M_i) (I/2^{N'}),$$

the total variation distance between the measurement outcome distribution is

$$\begin{aligned} \text{TV}(p, p') &= \frac{1}{2} \sum_i |\langle\psi|\tilde{M}_i|\psi\rangle - \langle\psi'|\tilde{M}_i|\psi'\rangle| \\ &= \frac{1}{2} (1 - \gamma) \sum_i |\langle\psi|M_i|\psi\rangle - \langle\psi'|M_i|\psi'\rangle| \\ &\leq (1 - \gamma) \min(\|H - H'\| \cdot t(E^{(\gamma)}), 1). \end{aligned} \quad (86)$$

□

In order to include possible adaptivity in the choice of experiments, we consider the rooted tree representation $\mathcal{T}$ described in [84] and [85], which was also used in [5]. Each node in the tree corresponds to the sequence of measurement outcomes the algorithm has seen so far. We can also think of the node as the memory state of the algorithm. At each node v , the algorithm runs a single experiment $E_v^{(\gamma)}$ as defined in Definition 6 with the measurement noise specified by γ and evolution time specified by T_v .

For each experiment $E_v^{(\gamma)}$, it produces a measurement outcome $i \in 1, 2, \dots, L_v$. The outcome will determine which child node to move the algorithm to from node v . When the algorithm arrives at a leaf ℓ , the algorithm stops, with ℓ denoting the outcome of the experiments. Note that because of the tree structure, a leaf node ℓ uniquely determines the path leading from the root to it. By considering the rooted tree representation and allowing the algorithm to depend on each node in the tree adaptively, we cover all possible learning algorithms that can adaptively choose the experiment that it runs based on previous measurement outcomes.

For each node v on tree $\mathcal{T}$, we denote $p^{(\mathcal{T})}(\ell)$ and $p'^{(\mathcal{T})}(\ell)$ as the probabilities of arriving at the leaf node ℓ in the experiments when the unknown Hamiltonians are H and H' respectively, and the algorithm begins from the root of $\mathcal{T}$. We denote the resulting distribution over leaf nodes by $p^{(\mathcal{T})}$ and $p'^{(\mathcal{T})}$ respectively. The total evolution time T can be written as:

$$T = \sum_{v \in \text{p}(r, \ell)} T_v, \quad (87)$$

where $\text{p}(r, \ell)$ is the path from the root r to a leaf ℓ of $\mathcal{T}$. We will provide an upper bound of the TV distance between the probability distributions $p^{(\mathcal{T})}$ and $p'^{(\mathcal{T})}$ with a given T .

Lemma 6 (TV for multiple experiments): Consider a rooted tree representation $\mathcal{T}$ for a learning algorithm with total evolution time T and measurement noise $\gamma \in (0, 0.5)$. We have

$$\text{TV}(p^{(\mathcal{T})}, p'^{(\mathcal{T})}) \leq 1 - \gamma^{\|H - H'\| \cdot T}, \quad (88)$$

which is an upper bound for the total variation of the outcomes under multiple experiments.

Proof: The proof is almost identical to the one for [5, Lemma 31], which is by induction on the tree $\mathcal{T}$. The only change is to use Lemma 5 to upper bound the TV distance that can be generated in a single noisy experiment. □

The above theorem tells us how much the output probability distribution can be changed when the Hamiltonian is perturbed. It therefore quantifies how well we can distinguish a pair of Hamiltonians given limited total evolution time. With this tool we are ready to prove the lower bound result in Theorem 11.

The main tool we are going to use next is Assouad's lemma [70]. We consider a set of probability distributions $\mathcal{P}$ each of which correspond to a parameter $\theta(P) \in \mathcal{D}$, where the space $\mathcal{D}$ can be equipped with certain pseudo-distances. We restate the lemma here:

Lemma 7: (Assouad's Lemma [70, Lemma 2]) Let $M \geq 1$ be an integer and let $\mathcal{F}_M = \{P_\tau : \tau \in \{-1, 1\}^M\}$ contain 2^M probability measures. Write $\tau \sim \tau'$ if τ and τ' differ in only one coordinate, and write $\tau \sim_j \tau'$ when that coordinate is the j th. Suppose that there are M pseudo-distances on $\mathcal{D}$ such that for any $x, y \in \mathcal{D}$

$$d(x, y) = \sum_{j=1}^M d_j(x, y) \quad (89)$$

and further that, if $\tau \sim_j \tau'$,

$$d_j(\theta(P_\tau), \theta(P_{\tau'})) \geq \alpha_M. \quad (90)$$

Then

$$\begin{aligned} &\max_{P_\tau \in \mathcal{F}_M} E_\tau d(\hat{\theta}, \theta(P_\tau)) \\ &\geq M \cdot \frac{\alpha_M}{2} \min\{\|P_\tau \wedge P_{\tau'}\| : \tau \sim \tau'\}, \end{aligned} \quad (91)$$

where $\|\cdot \wedge \cdot\| = 1 - \text{TV}(\cdot, \cdot)$ is the affinity of two probability distributions.

We will use this lemma to prove Theorem 11.

Proof of Theorem 11: In the Hamiltonian learning scenario, each τ represents a coefficient vector $\mu = (\mu_1, \dots, \mu_M)$, where $\mu_i = h\tau_i \in \{-h, h\}$ and $h \in (0, 1)$. P_τ is the probability distribution p^τ over leaf nodes of $\mathcal{T}$ corresponding to adaptive noisy

experiments with the Hamiltonian represented by coefficient vector τ . The parameter $\theta(P_\tau) = \mu$. The distance on $\mathfrak{D}$ is set to be the ℓ^1 -norm distance between the corresponding coefficient vectors. For adaptive experiments with Hamiltonian coefficients μ and μ' , with output distributions $P_\tau = p^\tau$ and $P'_\tau = p'^\tau$, for $p \in [1, 2]$ we define

$$\begin{aligned} d^{(p)}(\theta(P_\tau), \theta(P'_\tau)) &= d^{(p)}(\mu, \mu') = \sum_{j=1}^M \frac{|\mu_j - \mu'_j|}{M^{1-1/p}} \\ &= \frac{\|\mu - \mu'\|_1}{M^{1-1/p}} \leq \|\mu - \mu'\|_p, \end{aligned} \quad (92)$$

where the inequality follows from Hölder's inequality. Using the notation $\sim_j$ introduced in Lemma 7, if $\tau \sim_j \tau'$, we have $\alpha_M^{(p)} = d^{(p)}(\mu, \mu') = d_j^{(p)}(\mu, \mu') = 2h/M^{1-1/p}$, where $\mu = h\tau$ and $\mu' = h\tau'$. For the affinity $\|P_\tau \wedge P'_\tau\|$, by Lemma 6 we have

$$\|P_\tau \wedge P'_\tau\| = 1 - \text{TV}(P_\tau, P'_\tau) \geq \gamma^{\|H-H'\|T} = \gamma^{2hT}.$$

With the above parameters, can directly apply Assouad's Lemma (Lemma 7) to obtain

$$\max_{P_\tau \in \mathcal{F}_M} \mathbb{E} \|\mu - \hat{\mu}\|_p \geq \max_{P_\tau \in \mathcal{F}_M} \mathbb{E}(d^{(p)}(\mu, \mu')) \geq M^{1/p} h \gamma^{2hT}, \quad (93)$$

where $\hat{\mu}$ is the estimate produced by any sequence of adaptive noisy experiments with total evolution time T . Note that this bound holds for any $h \in (0, 1)$, and we can therefore choose $h = 1/(2T \log(1/\gamma))$, and we will have

$$\max_{P_\tau \in \mathcal{F}_M} \mathbb{E} \|\mu - \hat{\mu}\|_p \geq \frac{M^{1/p}}{2Te \log(1/\gamma)}. \quad (94)$$

Therefore if we want to obtain coefficient estimates for Hamiltonians of the form $H = \sum_{a=1}^M \mu_a P_a$ that is accurate up to error ϵ_1 in ℓ^1 -distance, we need T as given in (80). $\square$

Remark 5: The ratio between the upper bound given by our algorithm and the lower bound shown above is $\tilde{\Theta}(M^{1/2})$ for all $p \in [1, 2]$.

VII. OPERATIONAL INTERPRETATION OF ERROR METRICS

In the previous sections we have discussed the method to learn the coefficients of a Hamiltonian H and the corresponding total evolution time lower bound. The ultimate goal of Hamiltonian learning is to use the learned Hamiltonian to predict the properties of the quantum system. In this section we will consider this task, and thereby provide an operational interpretation of the error metrics.

We consider the task of predicting observation expectation values at a given time t . Suppose that the exact Hamiltonian is $H = \sum_p \mu_p P$ and the learned Hamiltonian is $H' = \sum_p \mu'_p P$. Here, because we fix the global phase by enforcing $\mu_I = 0$, these two Hamiltonians are both traceless. Starting from a known quantum state ρ_0 , the exact state $\rho(t)$ and the state $\rho'(t)$ we predict will be

$$\rho(t) = e^{-iHt} \rho_0 e^{iHt}, \quad \rho'(t) = e^{-iH't} \rho_0 e^{iH't}. \quad (95)$$

If we focus on a specific observation O , then the expectation values are $\text{Tr}[O\rho(t)]$ and $\text{Tr}[O\rho'(t)]$. We will then discuss what the ℓ^1 - and ℓ^2 -error bounds imply for the error $|\text{Tr}[O\rho'(t)] - \text{Tr}[O\rho(t)]|$, which reflects how well we can make

predictions based on the learned Hamiltonian. Such considerations of the observable, or alternatively the initial state, have featured prominently in works on quantum simulation and related error analysis [73], [86], [87], [88], [89], [90], [91], [92].

Here we will first state our conclusions: the ℓ^1 -error bound provides a guarantee that we can accurately predict any observable O whose spectral norm $\|O\|$ is not too large, for any initial state (see (103)); the ℓ^2 -error bound guarantees that we can accurately predict any observable O whose Frobenius norm $\|O\|_2$ is not too large (these are typically low-rank operators) for an initial state drawn from a Haar distribution or a 2-design (see (111)).

A. Operational Interpretation of the ℓ^1 -Norm

We first consider the ℓ^1 -norm distance between the coefficient vectors

$$\|\mu - \mu'\|_1 = \sum_p |\mu_p - \mu'_p|. \quad (96)$$

By Schrodinger's equation, we have the ordinary differential equation for ρ' :

$$\frac{d}{dt} \rho'(t) = -i[H', \rho'(t)] = -i[H, \rho'(t)] - i[H' - H, \rho'(t)]. \quad (97)$$

By Duhamel's principle, we have

$$\rho'(t) = \rho(t) - i \int_0^t ds e^{-iH(t-s)} [H' - H, \rho'(s)] e^{iH(t-s)}. \quad (98)$$

Therefore, the trace distance between $\rho(t)$ and $\rho'(t)$ can be bounded as:

$$\|\rho'(t) - \rho(t)\|_1 \leq 2t \|H' - H\|_\infty. \quad (99)$$

Here $\|\cdot\|_p$ denotes the Schatten p -norm. Specifically, $p = 1$ corresponds to the trace norm, and $p = \infty$ corresponds to the spectral norm. The spectral norm of the difference between H and H' can be bounded by

$$\|H - H'\|_\infty \leq \|\mu - \mu'\|_1. \quad (100)$$

Thus,

$$\|\rho(t) - \rho'(t)\|_1 \leq 2t \|\mu - \mu'\|_1. \quad (101)$$

The trace norm distance between $\rho(t)$ and $\rho'(t)$ is upper bounded by the ℓ^1 -norm distance between the two coefficient vectors, scaling linearly with evolution time.

Consider the expectation value of measuring $\rho(t)$ and $\rho'(t)$ with observable O ,

$$\begin{aligned} |\langle O \rangle_{\rho(t)} - \langle O \rangle_{\rho'(t)}| &= |\text{Tr}[O(\rho(t) - \rho'(t))]| \\ &\leq \|O\|_\infty \|\rho(t) - \rho'(t)\|_1. \end{aligned} \quad (102)$$

The second inequality comes from the Hölder's inequality for Schatten norms. Using the bound in (101), the difference between the expectation values can be bounded by

$$|\langle O \rangle_{\rho(t)} - \langle O \rangle_{\rho'(t)}| \leq 2t \|O\|_\infty \|\mu - \mu'\|_1. \quad (103)$$

Note that the above is true for all ρ_0 and O , and therefore the ℓ^1 -error bound provides us with a worst case guarantee on how well we can predict expectation values of observables.

B. Operational Interpretation of the ℓ^2 -Norm

We now consider the ℓ^2 -norm distance between the coefficient vectors

$$\|\mu - \mu'\|_2 = \sqrt{\sum_a (\mu_a - \mu'_a)^2}. \quad (104)$$

Parseval's identity provides that the ℓ^2 -norm of the distance between the coefficient vectors equals the Frobenious norm of $H - H'$:

$$\frac{1}{\sqrt{2^n}} \|H - H'\|_2 = \sqrt{\sum_i (\mu_i - \mu'_i)^2}. \quad (105)$$

Following [93], we consider the time evolution channel of H and H'

$$\mathcal{U}_t(\rho) = e^{-iHt} \rho e^{iHt}, \quad \mathcal{U}'_t(\rho) = e^{-iH't} \rho e^{iH't} \quad (106)$$

and the normalized Choi states of $\mathcal{U}_t$ and $\mathcal{U}'_t$ as

$$\mathcal{C}(\mathcal{U}_t) = (\mathcal{U}_t \otimes \mathcal{I})(\Omega) \quad \mathcal{C}(\mathcal{U}'_t) = (\mathcal{U}'_t \otimes \mathcal{I})(\Omega), \quad (107)$$

where $\Omega = |\Omega\rangle\langle\Omega|$ is the maximally entangled state, and $\mathcal{I}$ is the identity channel. We introduce the distance metric $\mathcal{D}(\mathcal{U}_t, \mathcal{U}'_t) = \frac{1}{\sqrt{2}} \|\mathcal{C}(\mathcal{U}_t) - \mathcal{C}(\mathcal{U}'_t)\|_2$.

Lemma 8: The following inequality holds

$$D(\mathcal{U}_t, \mathcal{U}'_t) \leq \frac{t}{\sqrt{2^n}} \|H - H'\|_2.$$

Proof: First note that

$$(e^{-iHt} \otimes I) |\Omega\rangle = (I \otimes e^{-iH'^T t}) |\Omega\rangle.$$

Then we have

$$\begin{aligned} & [(H - H') \otimes I, (e^{-iHt} \otimes I) \Omega (e^{iHt} \otimes I)] \\ &= [(H - H') \otimes I, (I \otimes e^{-iH'^T t}) \Omega (I \otimes e^{iH'^T t})] \\ &= (I \otimes e^{-iH'^T t}) [(H - H') \otimes I, \Omega] (I \otimes e^{iH'^T t}), \end{aligned}$$

where we have used the fact that $I \otimes e^{-iH'^T t}$ commute with $(H - H') \otimes I$. Therefore

$$\begin{aligned} & \|[(H - H') \otimes I, (e^{-iHt} \otimes I) \Omega (e^{iHt} \otimes I)]\|_2^2 \\ &= \|[(H - H') \otimes I, \Omega]\|_2^2 \\ &= \frac{2}{2^n} \|H - H'\|_2^2. \end{aligned}$$

Then because

$$\begin{aligned} & \mathcal{U}'_t(\Omega) - \mathcal{U}_t(\Omega) \\ &= (e^{-iH't} \otimes I) \Omega (e^{iH't} \otimes I) - (e^{-iHt} \otimes I) \Omega (e^{iHt} \otimes I) \\ &= -i \int_0^t e^{-iH's} [(H - H') \otimes I, (e^{-iHs} \otimes I) \Omega (e^{iHs} \otimes I)] e^{iH's} ds, \end{aligned}$$

we have

$$\begin{aligned} & D(\mathcal{U}_t, \mathcal{U}'_t) \\ &\leq \frac{1}{\sqrt{2}} \int_0^t \|[(H - H') \otimes I, (e^{-iHs} \otimes I) \Omega (e^{iHs} \otimes I)]\|_2 ds \\ &= \frac{1}{\sqrt{2^n}} \|H - H'\|_2 t. \end{aligned}$$

If the initial state is a pure state randomly sampled from the Haar distribution, $\rho_0 = |\psi_0\rangle\langle\psi_0|$, following [94], we have:

$$\begin{aligned} & \mathbb{E}_{|\psi_0\rangle \sim \text{Haar}_n} [\|U_t |\psi_0\rangle\langle\psi_0| U_t^\dagger - U'_t |\psi_0\rangle\langle\psi_0| U'^{\dagger}_t\|_2^2] \\ &= \frac{2^{n+1}}{2^n + 1} \mathcal{D}(\mathcal{U}_t, \mathcal{U}'_t)^2. \end{aligned} \quad (108)$$

Thus,

$$\begin{aligned} & \sqrt{\mathbb{E}_{|\psi_0\rangle \sim \text{Haar}_n} [\|\rho(t) - \rho'(t)\|_2^2]} \\ &\leq \sqrt{\frac{2}{2^n + 1}} t \|H - H'\|_2 \leq \sqrt{2} t \|\mu - \mu'\|_2. \end{aligned} \quad (109)$$

As in (102), we consider the difference between the expectation values of observable O when measuring $\rho(t)$ and $\rho'(t)$, with the same pure initial state $\rho_0 = |\psi_0\rangle\langle\psi_0|$. By using Hölder's inequality for Schatten norms, we have

$$\begin{aligned} & |\langle O \rangle_{\rho(t)} - \langle O \rangle_{\rho'(t)}| = |\text{Tr}[O(\rho(t) - \rho'(t))]| \\ &\leq \|O(\rho(t) - \rho'(t))\|_1 \leq \|O\|_2 \|\rho(t) - \rho'(t)\|_2. \end{aligned} \quad (110)$$

Taking the expectation value among the Haar random distributed initial state $|\psi_0\rangle$, we can obtain

$$\begin{aligned} & \mathbb{E}_{|\psi_0\rangle \sim \text{Haar}_n} [|\text{Tr}[O\rho(t)] - \text{Tr}[O\rho'(t)]|^2] \\ &\leq \|O\|_2^2 \mathbb{E}_{|\psi_0\rangle \sim \text{Haar}_n} [\|\rho(t) - \rho'(t)\|_2^2] \leq 2t^2 \|O\|_2^2 \|\mu - \mu'\|_2^2. \end{aligned} \quad (111)$$

The operators O for which $\|O\|_2$ is small include those that are low-rank or approximately low-rank. In the above, we have only used the second moment of $\rho_0 = |\psi_0\rangle\langle\psi_0|$, and therefore we only need $|\psi_0\rangle$ to be drawn from a 2-design rather than the exact Haar distribution. From the above discussion we have seen that an ℓ^2 -error bound guarantees the average accuracy of predicting the expectation value of a low-rank observable for initial states drawn from a 2-design.

VIII. DISCUSSION

In this work we developed a protocol that efficiently learns a k -body Hamiltonian consisting of M Pauli terms with nearly Heisenberg-limited scaling. Compared to previous works such as [5], [40], and [45], our main contribution is that the new algorithm does not rely on any kind of geometric locality: the Pauli terms can involve qubits anywhere in the system, and we do not require the combined interaction strength on a qubit to be bounded, which is needed in [45]. Moreover, our protocol only requires single-qubit interleaving operations as opposed to the multi-qubit adaptive operations used in [40] and [45], and is robust to a constant amount of SPAM error. These features are all essential to practical implementation on current hardware.

Our results naturally lead to several open problems to be considered in future works. It can be readily seen that between the total evolution time lower bound for learning with the ℓ^p -metric in Theorem 11 and the upper bound in Theorem 9, there is a gap of order $\sqrt{M}$, ignoring logarithmic factors, for all $p \in [1, 2]$. In fact, the lower bound is tight for the ℓ^1 -metric, up to logarithmic factors, for Hamiltonians with geometrically local Pauli terms, as the learning protocol in [5] provides a matching upper bound in this scenario. However, it remains unclear whether, in the general non-geometrically local setting,

the current gap comes from looseness in the upper bound, the lower bound, or both. On the one hand, the lower bound does not take into account that the support of the M terms is unknown, nor possible structural relations among the terms such as commutativity, and therefore may not fully capture the difficulty of the problem. On the other hand, the gap may not simply be an artifact of compressed-sensing bounds: for example, if the M unknown terms are all commuting Z -strings, then recovery can be carried out by direct inversion of the Hadamard transform rather than by compressed sensing, yet the error analysis still indicates that estimating the coefficients to ϵ precision in ℓ^2 -norm requires total evolution time of order M/ϵ , matching our upper bound. The gap between the current upper bound and lower bound then points to two interesting directions: can we further strengthen our algorithm to achieve better dependence on M , or can we strengthen the lower bound by accounting for the non-geometrically local nature of the Hamiltonian? If the latter can be achieved, we will be able to show rigorously that the lack of geometric locality fundamentally changes the difficulty of the task.

In this work we focused on qubit Hamiltonians, and it is natural to consider how the present approach extends to bosonic and fermionic Hamiltonians. The bosonic Hamiltonians, because of the infinite local degrees of freedom, may require us to change our experimental design and post-processing, and so do the fermionic Hamiltonians because of the more complicated commutation relations between terms.

Our learning protocol requires the preparation of GHZ states to achieve the desired M and n scaling. While such states can in principle be prepared in constant depth using measurements and feedback, maintaining high-fidelity large GHZ states remains experimentally challenging in practice. If restricted to only using product states as done in [5], can we still achieve the current scaling, or can one derive a lower bound that rules it out? These are all questions that should be explored in future works.

APPENDIX A

THE HAMILTONIAN RESHAPING ERROR BOUND

In this section we prove Theorem 4, which we restate here:

Theorem 12: Let β be a length- n string of x, y, z , and let $\mathcal{K}_\beta$ be as defined in (16). Let U be the random unitary defined in (17). Let $V = e^{-iH_{\text{eff}}t}$ for H_{eff} given in (21), and $t = r\tau$. We define the quantum channels $\mathcal{U}$ and $\mathcal{V}$ be

$$\mathcal{U}(\rho) = \mathbb{E}[U\rho U^\dagger], \quad \mathcal{V}(\rho) = V\rho V^\dagger.$$

Then

$$\|\mathcal{U} - \mathcal{V}\|_\diamond \leq \frac{4M^2t^2}{r}.$$

Proof: Recall from (17) that

$$U = Q_r e^{-iH\tau} Q_r \cdots Q_1 e^{-iH\tau} Q_1.$$

We define $U_j = Q_j e^{-iH\tau} Q_j$, $\mathcal{U}_j(\rho) = \mathbb{E}_{Q_j} U_j \rho U_j^\dagger$, where $\mathcal{U}_j$ is the quantum channel.

Because each Q_j is chosen independently, we have

$$\mathcal{U} = \mathcal{U}_r \mathcal{U}_{r-1} \cdots \mathcal{U}_1.$$

We then define $\mathcal{V}_\tau(\rho) = e^{-iH_{\text{eff}}\tau} \rho e^{iH_{\text{eff}}\tau}$, which then gives us $\mathcal{V} = \mathcal{V}_\tau^r$. We will then focus on obtaining a bound for $\|\mathcal{U}_j - \mathcal{V}_\tau\|_\diamond$. Consider a quantum state ρ on the current system and an auxiliary system denoted by α , and an observable O on the combined quantum system such that $\|O\| \leq 1$.

We have the equation shown at the bottom of the next page, where we are using Taylor's theorem with the remainder in the Lagrange form. This tells us that see the equation shown at the bottom of the next page. Because the two terms on the left-hand side, when we take the expectation value with respect to Q_j , become the equations shown at the bottom of the next page, we have the equation shown at the bottom of the next page.

Because this is true for all O such that $\|O\| \leq 1$,

$$\|(\mathcal{I}_\alpha \otimes \mathcal{U}_j)(\rho) - (\rho - i\tau[I_\alpha \otimes H_{\text{eff}}, \rho])\|_1 \leq 2\|H\|^2 \tau^2. \quad (112)$$

Similarly we can also show that

$$\|(\mathcal{I}_\alpha \otimes \mathcal{V}_\tau)(\rho) - (\rho - i\tau[I_\alpha \otimes H_{\text{eff}}, \rho])\|_1 \leq 2\|H_{\text{eff}}\|^2 \tau^2. \quad (113)$$

Combining (112) and (113), and by the triangle inequality, we have

$$\|(\mathcal{I}_\alpha \otimes \mathcal{U}_j)(\rho) - (\mathcal{I}_\alpha \otimes \mathcal{V}_\tau)(\rho)\|_1 \leq 2\|H\|^2 \tau^2 + 2\|H_{\text{eff}}\|^2 \tau^2.$$

Because the above is true for all auxiliary system α and for all quantum states on the combined quantum system, we have

$$\|\mathcal{U}_j - \mathcal{V}_\tau\|_\diamond \leq 2\|H\|^2 \tau^2 + 2\|H_{\text{eff}}\|^2 \tau^2. \quad (114)$$

Next, we observe that

$$\mathcal{U} - \mathcal{V} = \sum_{k=1}^r \mathcal{U}_r \cdots \mathcal{U}_{k+1} (\mathcal{U}_k - \mathcal{V}_\tau) \mathcal{V}_\tau^{k-1}.$$

Therefore

$$\|\mathcal{U} - \mathcal{V}\|_\diamond \leq \sum_{k=1}^r \|\mathcal{U}_k - \mathcal{V}_\tau\|_\diamond \leq r(2\|H\|^2 \tau^2 + 2\|H_{\text{eff}}\|^2 \tau^2).$$

The bound in the Theorem follows by observing that $\|H\|, \|H_{\text{eff}}\| \leq M$, and $\tau = t/r$. $\square$

APPENDIX B

TECHNICAL LEMMAS FOR COMPRESSED SENSING

We start by introducing the orthogonal systems and the bounded orthogonal systems (BOS). Then we will move on to show that the weight- k Hadamard matrix introduced in Definition 1 is a BOS. After that, we will introduce the restricted isometry property, which is essential for the successful reconstruction of sparse vectors with compressed sensing. With that, we will show that the weight- k Hadamard matrix has the desired restricted isometry property, which in turn leads to the error bound for reconstructing the sparse vector.

Definition 8 (Orthonormal Systems): Let $\mathcal{D} \subset \mathbb{R}^N$ be endowed with a probability measure ν . We call $\Phi = \{\phi_1, \dots, \phi_N\}$ an orthonormal system of complex-valued functions on $\mathcal{D}$ if, for $j, k \in [N]$,

$$\int_{\mathcal{D}} \phi_j(\mathbf{t}) \overline{\phi_k(\mathbf{t})} d\nu(\mathbf{t}) = \delta_{j,k} = \begin{cases} 0 & \text{if } j \neq k, \\ 1 & \text{if } j = k, \end{cases} \quad (115)$$

Upon orthonormal systems, we can define bounded orthonormal systems.

Definition 9: (Bounded Orthonormal Systems, [9, Theorem 12.32]) We call $\Phi = \{\phi_1, \dots, \phi_N\}$ a bounded orthonormal system (BOS) with constant K if it satisfies Equation (115) and if

$$\|\phi_j\|_\infty := \sup_{\mathbf{t} \in \mathcal{D}} |\phi_j(\mathbf{t})| \leq K \quad \text{for all } j \in [N]. \quad (116)$$

The smallest value that the constant K can take is 1 since

$$1 = \int_{\mathcal{D}} |\phi_j(\mathbf{t})|^2 d\nu(\mathbf{t}) \leq \sup_{\mathbf{t} \in \mathcal{D}} |\phi_j(\mathbf{t})|^2 \int_{\mathcal{D}} d\nu(\mathbf{t}) \leq K^2. \quad (117)$$

When the equality $K = 1$ is reached, it is necessary to have $|\phi_j(\mathbf{t})| = 1$ for ν .

Consider the weight- k Hadamard matrix defined in Definition 1 with normalization factor $\frac{1}{\sqrt{2^n}}$. Denote the columns of this normalized weight- k Hadamard matrix as $c_1, c_2, \dots, c_D$. Since they are columns of the original Hadamard matrix, they are mutually orthogonal. Moreover, the BOS that consists of all columns of the original Hadamard matrix has the bounding constant $K = 1$. Then the bounding constant for its subsystem is at most $K = 1$. Since $K \geq 1$, we have the bounding constant K for the BOS of columns of weight- k Hadamard matrix is one.

We can consider functions in a BOS, of the form

$$f(\mathbf{t}) = \sum_{k=1}^N x_k \phi_k(\mathbf{t}), \quad \mathbf{t} \in \mathcal{D}. \quad (118)$$

Let $\mathbf{t}_1, \dots, \mathbf{t}_\Gamma \in \mathcal{D}$ be some sampling points and suppose we have given the sample values

$$y_\ell = f(\mathbf{t}_\ell) = \sum_{k=1}^N x_k \phi_k(\mathbf{t}_\ell), \quad \ell \in [\Gamma]. \quad (119)$$

We can introduce the sampling matrix $\mathbf{A} \in \mathbb{C}^{\Gamma \times N}$ with entries

$$A_{\ell,k} = \phi_k(\mathbf{t}_\ell), \quad \ell \in [\Gamma], k \in [N]. \quad (120)$$

the vector $\mathbf{y} = [y_1, \dots, y_\Gamma]^\top$ of sample values (measurements) can be written in the form

$$\mathbf{y} = \mathbf{A}\mathbf{x} \quad (121)$$

where $\mathbf{x} = [x_1, \dots, x_N]^\top$ is the vector of coefficients in Equation (118).

Sometimes, we are interested in discrete systems, where instead of having $\mathcal{D} \subset \mathbb{R}^N$, we have $\mathcal{D} = \mathbb{F}_2^n$. In this case, we can consider the Γ rows of the sampling matrix $\mathbf{A} \in \mathbb{C}^{\Gamma \times 2^n}$ to be sampled from the 2^n rows of the transform matrix describing this system $\mathbf{T} \in \mathbb{C}^{2^n \times 2^n}$ at random.

For a discrete BOS described by a transfer matrix $\mathbf{T} \in \mathbb{C}^{N \times N}$, we consider a random sampling matrix $\mathbf{A} \in \mathbb{C}^{\Gamma \times N}$, where $\mathbf{A}$ is a sub-matrix of $\mathbf{T}$ with its rows sampled from all rows of $\mathbf{T}$ uniformly at random. As in Equation (121), we define the vector $\mathbf{y} = [y_1, \dots, y_\Gamma]^\top$ as the sampling result, where

$$y_\ell = \sum_{k=1}^N T_{\ell,k} x_k + e_\ell, \quad \ell \in [\Gamma], \quad (122)$$

and $\mathbf{x} = [x_1, \dots, x_N]^\top$ is the vector of coefficients to be reconstructed. Here, a term e is added to represent noisy sampling. To write it in a more compact form

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}. \quad (123)$$

To reconstruct $\mathbf{x}$ with random sampling results in the presence of noise $\mathbf{e}$, robust sparse recovery methods can be used, which is introduced in Theorem 5. These methods require the sampling matrix $\mathbf{A}$ to have the restricted isometry property (RIP).

Definition 10: (The M th restricted isometry constant, [9, Definition 6.1]) The M th restricted isometry constant $\delta_M = \delta_M(\mathbf{A})$ of a matrix $\mathbf{A} \in \mathbb{C}^{\Gamma \times N}$ is the smallest $\delta > 0$ such that

$$(1 - \delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta)\|\mathbf{x}\|_2^2 \quad (124)$$

for all M -sparse vector $\mathbf{x} \in \mathbb{C}^N$. Equivalently, it is given by

$$\delta_M = \max_{\mathbf{M} \subseteq [N], \text{card}(\mathbf{M}) \leq M} \|\mathbf{A}_\mathbf{M}^* \mathbf{A}_\mathbf{M} - \mathbf{Id}\|_{2 \rightarrow 2} \quad (125)$$

where $\mathbf{A}_\mathbf{M}$ is the sub-matrix of $\mathbf{A}$ with only those columns with indices in set $\mathbf{M}$.

With the restricted isometry constant, we can then define the restricted isometry property.

$$\begin{aligned} \text{Tr}[O(I_\alpha \otimes U_j) \rho(I_\alpha \otimes U_j^\dagger)] &= \text{Tr}[O(I_\alpha \otimes (Q_j e^{-iH\tau} Q_j)) \rho(I_\alpha \otimes (Q_j e^{iH\tau} Q_j))] \\ &= \text{Tr}[O\rho] - i\tau \text{Tr}[O[I_\alpha \otimes (Q_j H Q_j), \rho]] - \frac{\tau^2}{2} \text{Tr}[O(I_\alpha \otimes Q_j)[I_\alpha \otimes H, [I_\alpha \otimes H, e^{-iH\tau}(I_\alpha \otimes Q_j) \rho(I_\alpha \otimes Q_j) e^{iH\tau}]](I_\alpha \otimes Q_j)], \end{aligned}$$

$$\left| \text{Tr}[O(I_\alpha \otimes U_j) \rho(I_\alpha \otimes U_j^\dagger)] - (\text{Tr}[O\rho] - i\tau \text{Tr}[O[I_\alpha \otimes (Q_j H Q_j), \rho]]) \right| \leq 2\|H\|^2 \tau^2$$

$$\begin{aligned} \mathbb{E}_{Q_j} \text{Tr}[O(I_\alpha \otimes U_j) \rho(I_\alpha \otimes U_j^\dagger)] &= \text{Tr}[O(\mathcal{I}_\alpha \otimes \mathcal{U}_j)(\rho)], \\ \mathbb{E}_{Q_j} [\text{Tr}[O\rho] - i\tau \text{Tr}[O[I_\alpha \otimes (Q_j H Q_j), \rho]]] &= \text{Tr}[O\rho] - i\tau \text{Tr}[O[I_\alpha \otimes H_{\text{eff}}, \rho]], \end{aligned}$$

$$\left| \text{Tr}[O(I_\alpha \otimes U_j)(\rho)] - (\text{Tr}[O\rho] - i\tau \text{Tr}[O[I_\alpha \otimes H_{\text{eff}}, \rho]]) \right| \leq 2\|H\|^2 \tau^2.$$

Definition 11 (Restricted isometry property): A matrix $\mathbf{A}$ is said to satisfy the restricted isometry property is $\delta_M = \delta_M(\mathbf{A})$ is small for reasonably large M .

To achieve sparse reconstruction from noisy samples, one way is to utilize the basic pursuit, i.e. doing ℓ^1 -minimization. To be more specific, with sparse vector $\mathbf{x}$, sampling matrix $\mathbf{A}$, sampling noise $\mathbf{e}$, and noisy sample $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$. With $\|\mathbf{e}\|_2 \leq \eta$, the basis pursuit reconstruct $\mathbf{x}$ from $\mathbf{y}$ and $\mathbf{A}$ by solving a ℓ^1 -minimization problem

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta. \quad (126)$$

Notice that ℓ^1 -minimization problem is a convex optimization problem, so the solution $\mathbf{x}^\#$ to this problem is unique and is guaranteed to be found. As shown in [9, Theorem 6.11], if $\mathbf{x}$ is a M -sparse vector, and $\mathbf{A}$ has the restricted isometry property with the $2M$ -th restricted isometry constant $\delta_{2M} \leq \frac{77 - \sqrt{1337}}{82}$, then the unique solution $\mathbf{x}^\#$ to the ℓ^1 -minimization problem in Eq (126) is a good approximation of $\mathbf{x}$

$$\|\mathbf{x} - \mathbf{x}^\#\|_p \leq CM^{1/p-1/2}\eta, \quad 1 \leq p \leq 2 \quad (127)$$

where C is a constant only depends on δ_{2M} . The proof of this is based on the interpretation of the restricted isometry constant in Eq (125). $\mathbf{A}_M^* \mathbf{A}_M$ is close to identity, and minimizing $\|\mathbf{z}\|_1$ will guarantee finding location of the support of $\mathbf{x}$.

With basis pursuit, as long as the sampling matrix $\mathbf{A}$ associated with a BOS has sufficiently good restricted isometry property, robust sparse reconstruction with high accuracy can be guaranteed. In the following, we will use in [9, Theorem 12.32] to show that the sampling matrix associated with the BOS with columns of the normalized weight- k Hadamard matrix can have a sufficiently small restricted isometry constant with high probability, as long as the number of samples Γ satisfies Eq (25) in Theorem 5. This, in turn, leads to the proof of the reconstruction error bound in Theorem 5.

Proof: (Proof of Theorem 5) Consider the weight- k Hadamard matrix $\mathbf{H}^{(k)}$, which is associated with a BOS system with constant $K = 1$. The sampling matrix $\mathbf{A} \in \mathbb{C}^{\Gamma \times D}$ is a matrix whose rows are independently randomly sampled from the rows of $\mathbf{H}^{(k)}$, where Γ is the number of samples and D is the number all possible n -qubit, k -body Pauli-Z strings. When $k = 1$ and Γ satisfies the conditions in Eq (25), [9, Theorem 12.32] shows that the sampling matrix $\mathbf{A}$ has the restricted isometry property with the restricted isometry constant $\delta_M \leq \eta_1 + \eta_1^2 + \eta_2$, where $\eta_1, \eta_2 \in (0, 1)$ are constants that can be chosen to be small. With the restricted isometry property of $\mathbf{A}$, [9, Theorem 6.11] shows that the error for robust reconstruction is bounded by the sparsity M and sampling noise η , which leads to the error bound for $\mathbf{x}^\#$ in Theorem 5. $\square$

APPENDIX C

TECHNICAL LEMMAS FOR THE LOWER BOUND

Lemma 9: Consider two quantum channels $\mathcal{U}(t)$ and $\mathcal{V}(t)$ defined through

$$\mathcal{U}(t)(\rho) = e^{-iH_1 t} \rho e^{iH_1 t}, \quad \mathcal{V}(t)(\rho) = e^{-iH_2 t} \rho e^{iH_2 t}.$$

Then

$$\|\mathcal{U}(t) - \mathcal{V}(t)\|_\diamond \leq 2\|H_1 - H_2\| \cdot t. \quad (128)$$

Proof: We define

$$U(t) = e^{-iH_1 t}, \quad V(t) = e^{-iH_2 t}. \quad (129)$$

Consider

$$\frac{d(e^{-iH_1 s} e^{-iH_2(t-s)})}{ds} = -ie^{-iH_1 s} (H_1 - H_2) e^{-iH_2(t-s)}, \quad (130)$$

we have

$$\int_0^t \frac{d(e^{-iH_1 s} e^{-iH_2(t-s)})}{ds} ds = e^{-iH_1 s} e^{-iH_2(t-s)} \Big|_0^t = e^{-iH_1 t} - e^{-iH_2 t}. \quad (131)$$

Hence,

$$e^{-iH_1 t} - e^{-iH_2 t} = -i \int_0^t e^{-iH_1 s} (H_1 - H_2) e^{-iH_2(t-s)} ds. \quad (132)$$

Since the diamond norm between $\mathcal{U}(t)$ and $\mathcal{V}(t)$ is bounded by the operator norm between the time evolution operators $e^{-iH_1 t}$, $e^{-iH_2 t}$, we have

$$\begin{aligned} \|\mathcal{U}(t) - \mathcal{V}(t)\|_\diamond &\leq 2\|e^{-iH_1 t} - e^{-iH_2 t}\| \\ &= 2\left\| -i \int_0^t e^{-iH_1 s} (H_1 - H_2) e^{-iH_2(t-s)} ds \right\| \\ &\leq 2 \int_0^t \|H_1 - H_2\| ds \\ &= 2\|H_1 - H_2\|t. \end{aligned} \quad (133)$$

$\square$

APPENDIX D

THE ROBUST FREQUENCY ESTIMATION PROTOCOL

In this section we introduce the robust frequency estimation protocol in [43], which largely follows the idea of the robust phase estimation protocol in [76]. We modify the protocol in [43] because our goal is to obtain an accurate estimate with large probability rather than having an optimal mean-squared error scaling. The key tool is the following lemma that allows us to incrementally refine the frequency estimate:

Lemma 10: Let $\theta \in [a, b]$. Let $Z(t)$ be a random variable such that

$$|Z(t) - e^{i\theta t}| < 1/2. \quad (134)$$

Then we can correctly distinguish between two overlapping cases $\theta \in [a, (a+2b)/3]$ and $\theta \in [(2a+b)/3, b]$ with one sample of $Z(\pi/(b-a))$.

Proof: These two situations can be distinguished by looking at the value of

$$f(\theta) = \sin\left(\frac{\pi}{b-a}\left(\theta - \frac{a+b}{2}\right)\right).$$

We know from (134) that

$$\left| \text{Im}\left(e^{-i\frac{(a+b)\pi}{2(b-a)}} Z(\pi/(b-a))\right) - f(\theta) \right| < 1/2.$$

where t in (134) is substituted by $\pi/(b-a)$ and a phase factor is added.

$$T = m \sum_{l=1}^L s_l = \frac{m\pi}{2A} \sum_{l=1}^L \left(\frac{3}{2}\right)^{l-1} \leq \frac{3\pi m}{2\epsilon} = \mathcal{O}((1/\epsilon)(\log(1/q) + \log \log(A/\epsilon))) \quad (139)$$

If

$$\operatorname{Im} \left(e^{-i \frac{(a+b)\pi}{2(b-a)}} Z(\pi/(b-a)) \right) \leq 0,$$

then $f(\theta) < 1/2$, which implies $\theta \in [a, (a+2b)/3]$. If

$$\operatorname{Im} \left(e^{-i \frac{(a+b)\pi}{2(b-a)}} Z(\pi/(b-a)) \right) > 0,$$

then $f(\theta) > -1/2$, which implies $\theta \in [(2a+b)/3, b]$. $\square$

Using this lemma we will prove Theorem 6, which we restate below:

Theorem 13: Let $\theta \in [-A, A]$. Let $X(t)$ and $Y(t)$ be independent random variables satisfying

$$\begin{aligned} |X(t) - \cos(\theta t)| &< 1/\sqrt{2}, \text{ with probability at least } 2/3, \\ |Y(t) - \sin(\theta t)| &< 1/\sqrt{2}, \text{ with probability at least } 2/3. \end{aligned} \quad (135)$$

Then with independent non-adaptive samples $X(t_1), X(t_2), \dots, X(t_K)$ and $Y(t_1), Y(t_2), \dots, Y(t_K)$, $t_j \geq 0$, for

$$K = \mathcal{O}(\log(A/\epsilon)(\log(1/q) + \log \log(A/\epsilon))), \quad (136)$$

$$\begin{aligned} T = \sum_{j=1}^K t_j &= \mathcal{O}((1/\epsilon)(\log(1/q) + \log \log(A/\epsilon))), \\ \max_j t_j &= \mathcal{O}(1/\epsilon), \end{aligned} \quad (137)$$

we can obtain a random variable $\hat{\theta}$ such that

$$\Pr[|\hat{\theta} - \theta| > \epsilon] \leq q. \quad (138)$$

Proof: We let $\lambda = \epsilon/3$. We build a random variable $S(t)$ satisfying (134), with which we will iteratively narrow down the interval $[a, b]$ containing θ until $|b - a| \leq 2\lambda$, at which point we choose $\hat{\theta} = (a + b)/2$. If $\theta \in [a, b]$ we will then ensure $|\hat{\theta} - \theta| \leq \lambda$. However, each iteration will involve some failure probability, which we will analyze later.

To build the random variable $S(t)$, we first use m independent samples of $X(t)$ and then take median $X_{\text{median}}(t)$, which satisfies

$$|X_{\text{median}}(t) - \cos(\theta t)| \leq 1/\sqrt{2}$$

with probability at least $1 - \delta/2$, where by (135) and the Chernoff bound

$$\delta = c_1 e^{-c_2 m},$$

for some universal constant c_1, c_2 . Similarly we can obtain $Y_{\text{median}}(t)$ such that

$$|Y_{\text{median}}(t) - \sin(\theta t)| \leq 1/\sqrt{2}$$

with probability at least $1 - \delta/2$. With these medians we then define

$$S(t) = X_{\text{median}}(t) + iY_{\text{median}}(t).$$

This random variable satisfies

$$|S(t) - e^{i\theta t}| \leq 1/2$$

with probability at least $1 - \delta$ using the union bound. It therefore allows us to solve the discrimination task in Lemma 10 with probability at least $1 - \delta$.

Whether each iteration proceeds correctly or not, the algorithm terminates after

$$L = \lceil \log_{3/2}(A/\lambda) \rceil$$

iterations. In the l th iteration we sample $X(s_l)$ and $Y(s_l)$ where $s_l = (\pi/2A)(3/2)^{l-1}$. We use m samples of $X(s_l)$ and $Y(s_l)$ for computing the median in each iteration, and therefore the failure probability is at most $\delta = c_1 e^{-c_2 m}$. With probability at most $L\delta$ using the union bound, one of the iteration fails. In order to ensure the protocol succeeds with probability at least $1 - q$, it suffices to let $\delta \leq q/L$, and it therefore suffices to choose

$$m = \lceil c_3 \log(L/q) \rceil = \mathcal{O}(\log(1/q) + \log \log(A/\epsilon)).$$

The total number of samples (for either $X(t)$ or $Y(t)$) is therefore as described in (136). All the t in each sample added together is (139), shown at the top of the page, thus giving us (137). $\square$

ACKNOWLEDGMENT

The authors thank Anurag Anshu, Zofia Adamska, Ainesh Bakshi, Sary Bseiso, Sitan Chen, Matthew C. Ding, Zihan Hao, Hong-Ye Hu, Jun Ikeda, Liang Jiang, Akshar D. Ramkumar, Thomas Schuster, Jamie Sikora, Samson Wang, John Wright, Qi Ye, Haimeng Zhao, and Xincheng Zhang for helpful discussions. The Institute for Quantum Information and Matter is an NSF Physics Frontiers Center.

REFERENCES

- [1] J. Haah, R. Kothari, and E. Tang, "Optimal learning of quantum Hamiltonians from high-temperature Gibbs states," in *Proc. IEEE 63rd Annu. Symp. Found. Comput. Sci. (FOCS)*, Oct. 2022, pp. 135–146.
- [2] D. S. França, L. A. Markovich, V. V. Dobrovitski, A. H. Werner, and J. Borregaard, "Efficient and robust estimation of many-qubit Hamiltonians," *Nature Commun.*, vol. 15, no. 1, p. 311, Jan. 2024, doi: 10.1038/s41467-023-44012-5.
- [3] W. Yu, J. Sun, Z. Han, and X. Yuan, "Robust and efficient Hamiltonian learning," *Quantum*, vol. 7, p. 1045, Jun. 2023.
- [4] A. Gu, L. Cincio, and P. J. Coles, "Practical Hamiltonian learning with unitary dynamics and Gibbs states," *Nature Commun.*, vol. 15, no. 1, p. 312, Jan. 2024.
- [5] H.-Y. Huang, Y. Tong, D. Fang, and Y. Su, "Learning many-body Hamiltonians with Heisenberg-limited scaling," *Phys. Rev. Lett.*, vol. 130, no. 20, May 2023, Art. no. 200403.
- [6] M. C. Caro, "Learning quantum processes and Hamiltonians via the Pauli transfer matrix," *ACM Trans. Quantum Comput.*, vol. 5, no. 2, pp. 1–53, Jun. 2024, doi: 10.1145/3670418.
- [7] A. Anshu, S. Arunachalam, T. Kuwahara, and M. Soleimanifar, "Sample-efficient learning of interacting quantum systems," *Nature Phys.*, vol. 17, no. 8, pp. 931–935, Aug. 2021.
- [8] A. Bakshi, A. Liu, A. Moitra, and E. Tang, "Learning quantum Hamiltonians at any temperature in polynomial time," in *Proc. 56th Annu. ACM Symp. Theory Comput.*, Jun. 2024, pp. 1470–1477.
- [9] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Cham, Switzerland: Springer, 2013, doi: 10.1007/978-0-8176-4948-7.

- [10] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.
- [11] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, Jan. 2001.
- [12] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.
- [13] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Commun. Pure Appl. Math.*, vol. 61, no. 8, pp. 1025–1045, Aug. 2008.
- [14] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.
- [15] E. J. Candès and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 7235–7254, Nov. 2011.
- [16] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
- [17] A. Seif, M. Hafezi, and Y.-K. Liu, "Compressed sensing measurement of long-range correlated noise," 2021, *arXiv:2105.12589*.
- [18] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Phys. Rev. Lett.*, vol. 105, no. 15, Oct. 2010, Art. no. 150401.
- [19] A. Shabani et al., "Efficient measurement of quantum dynamics via compressive sensing," *Phys. Rev. Lett.*, vol. 106, no. 10, Mar. 2011, Art. no. 100401.
- [20] A. Shabani, M. Mohseni, S. Lloyd, R. L. Kosut, and H. Rabitz, "Estimation of many-body quantum Hamiltonians via compressive sensing," *Phys. Rev. A, Gen. Phys.*, vol. 84, no. 1, Jul. 2011, Art. no. 012107.
- [21] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert, "Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators," *New J. Phys.*, vol. 14, no. 9, Sep. 2012, Art. no. 095022.
- [22] A. Kalev, R. L. Kosut, and I. H. Deutsch, "Quantum tomography protocols with positivity are compressed sensing protocols," *Npj Quantum Inf.*, vol. 1, no. 1, p. 15018, Dec. 2015.
- [23] I. Roth et al., "Recovering quantum gates from few average gate fidelities," *Phys. Rev. Lett.*, vol. 121, no. 17, Oct. 2018, Art. no. 170502.
- [24] R. Harper, W. Yu, and S. T. Flammia, "Fast estimation of sparse quantum noise," *PRX Quantum*, vol. 2, no. 1, Feb. 2021, Art. no. 010322.
- [25] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, Apr. 1995.
- [26] T. J. Evans, R. Harper, and S. T. Flammia, "Scalable Bayesian Hamiltonian learning," 2019, *arXiv:1912.07636*.
- [27] H.-Y. Huang, R. Kueng, and J. Preskill, "Predicting many properties of a quantum system from very few measurements," *Nature Phys.*, vol. 16, no. 10, pp. 1050–1057, Oct. 2020, doi: [10.1038/s41567-020-0932-7](https://doi.org/10.1038/s41567-020-0932-7).
- [28] A. Elben et al., "The randomized measurement toolbox," *Nature Rev. Phys.*, vol. 5, no. 1, pp. 9–24, Dec. 2022, doi: [10.1038/s42254-022-00535-2](https://doi.org/10.1038/s42254-022-00535-2).
- [29] Z. Li, L. Zou, and T. H. Hsieh, "Hamiltonian tomography via quantum quench," *Phys. Rev. Lett.*, vol. 124, no. 16, Apr. 2020, Art. no. 160502.
- [30] L. Che et al., "Learning quantum Hamiltonians from single-qubit measurements," *Phys. Rev. Res.*, vol. 3, no. 2, Jun. 2021, Art. no. 023246.
- [31] D. Hangleiter, I. Roth, J. Fuksa, J. Eisert, and P. Roushan, "Robustly learning the Hamiltonian dynamics of a superconducting quantum processor," *Nature Commun.*, vol. 15, no. 1, p. 9595, Nov. 2024, doi: [10.1038/s41467-024-52629-3](https://doi.org/10.1038/s41467-024-52629-3).
- [32] A. Zubida, E. Yitzhaki, N. H. Lindner, and E. Bairey, "Optimal short-time measurements for Hamiltonian learning," 2021, *arXiv:2108.08824*.
- [33] E. Bairey, I. Arad, and N. H. Lindner, "Learning a local Hamiltonian from local measurements," *Phys. Rev. Lett.*, vol. 122, no. 2, Jan. 2019, Art. no. 020504.
- [34] E. Bairey, C. Guo, D. Poletti, N. H. Lindner, and I. Arad, "Learning the dynamics of open quantum systems from their steady states," *New J. Phys.*, vol. 22, no. 3, Mar. 2020, Art. no. 032001.
- [35] C. E. Granade, C. Ferrie, N. Wiebe, and D. G. Cory, "Robust online Hamiltonian learning," *New J. Phys.*, vol. 14, no. 10, Oct. 2012, Art. no. 103013.
- [36] F. Wilde, A. Kshetrimayum, I. Roth, D. Hangleiter, R. Sweke, and J. Eisert, "Scalably learning quantum many-body Hamiltonians from dynamical data," 2022, *arXiv:2209.14328*.
- [37] S. Krastanov, S. Zhou, S. T. Flammia, and L. Jiang, "Stochastic estimation of dynamical variables," *Quantum Sci. Technol.*, vol. 4, no. 3, Jul. 2019, Art. no. 035003.
- [38] T. Möbus, A. Bluhm, M. C. Caro, A. H. Werner, and C. Rouzé, "Learning and simulating bosonic systems via finite-energy locality," 2023, *arXiv:2307.15026*.
- [39] M. Holzäpfel, T. Baumgratz, M. Cramer, and M. B. Plenio, "Scalable reconstruction of unitary processes and Hamiltonians," *Phys. Rev. A, Gen. Phys.*, vol. 91, no. 4, Apr. 2015, Art. no. 042129.
- [40] A. Dutkiewicz, T. E. O'Brien, and T. Schuster, "The advantage of quantum control in many-body Hamiltonian learning," *Quantum*, vol. 8, p. 1537, Nov. 2024.
- [41] A. Mirani and P. Hayden, "Learning interacting fermionic Hamiltonians at the Heisenberg limit," *Phys. Rev. A, Gen. Phys.*, vol. 110, no. 6, Dec. 2024, Art. no. 062421.
- [42] H. Ni, H. Li, and L. Ying, "Quantum Hamiltonian learning for the Fermi-hubbard model," *Acta Applicandae Mathematicae*, vol. 191, no. 1, pp. 1–16, Jun. 2024.
- [43] H. Li, Y. Tong, T. Gefen, H. Ni, and L. Ying, "Heisenberg-limited Hamiltonian learning for interacting bosons," *Npj Quantum Inf.*, vol. 10, no. 1, p. 83, Sep. 2024.
- [44] S. Boixo and R. D. Somma, "Parameter estimation with mixed-state quantum computation," *Phys. Rev. A, Gen. Phys.*, vol. 77, no. 5, May 2008, Art. no. 052320.
- [45] A. Bakshi, A. Liu, A. Moitra, and E. Tang, "Structure learning of Hamiltonians from real-time evolution," in *Proc. IEEE 65th Annu. Symp. Found. Comput. Sci. (FOCS)*, Oct. 2024, pp. 1037–1050.
- [46] K. Wang and X. Li, "Simulation-assisted learning of open quantum systems," *Quantum*, vol. 8, p. 1407, Jul. 2024.
- [47] T. Odake, H. Kristjánsson, P. Taranto, and M. Murao, "Universal algorithm for transforming Hamiltonian eigenvalues," *Phys. Rev. Res.*, vol. 7, no. 1, Mar. 2025, Art. no. 013331.
- [48] B. L. Higgins, D. W. Berry, S. D. Bartlett, H. M. Wiseman, and G. J. Pryde, "Entanglement-free Heisenberg-limited phase estimation," *Nature*, vol. 450, no. 7168, pp. 393–396, Nov. 2007.
- [49] S. Sachdev and J. Ye, "Gapless spin-fluid ground state in a random quantum Heisenberg magnet," *Phys. Rev. Lett.*, vol. 70, no. 21, pp. 3339–3342, May 1993.
- [50] E. H. Lieb and D. W. Robinson, "The finite group velocity of quantum spin systems," *Commun. Math. Phys.*, vol. 28, no. 3, pp. 251–257, Sep. 1972. [Online]. Available: <https://link.springer.com/10.1007/BF01645779>
- [51] M. C. Tran, A. Y. Guo, C. L. Baldwin, A. Ehrenberg, A. V. Gorshkov, and A. Lucas, "Lieb-robinson light cone for power-law interactions," *Phys. Rev. Lett.*, vol. 127, no. 16, Oct. 2021, Art. no. 160401, doi: [10.1103/physrevlett.127.160401](https://doi.org/10.1103/physrevlett.127.160401).
- [52] M. B. Hastings and T. Koma, "Spectral gap and exponential decay of correlations," *Commun. Math. Phys.*, vol. 265, no. 3, pp. 781–804, Apr. 2006, doi: [10.1007/s00220-006-0030-4](https://doi.org/10.1007/s00220-006-0030-4).
- [53] D. V. Else, F. Machado, C. Nayak, and N. Y. Yao, "Improved lieb-robinson bound for many-body Hamiltonians with power-law interactions," *Phys. Rev. A, Gen. Phys.*, vol. 101, no. 2, Feb. 2020, Art. no. 022333, doi: [10.1103/physrev.101.022333](https://doi.org/10.1103/physrev.101.022333).
- [54] M. C. Tran et al., "Locality and digital quantum simulation of power-law interactions," *Phys. Rev. X*, vol. 9, no. 3, Jul. 2019, Art. no. 031006, doi: [10.1103/physrevx.9.031006](https://doi.org/10.1103/physrevx.9.031006).
- [55] C.-F. Chen and A. Lucas, "Finite speed of quantum scrambling with long range interactions," *Phys. Rev. Lett.*, vol. 123, no. 25, Dec. 2019, Art. no. 250605, doi: [10.1103/physrevlett.123.250605](https://doi.org/10.1103/physrevlett.123.250605).
- [56] T. Kuwahara and K. Saito, "Strictly linear light cones in long-range interacting systems of arbitrary dimensions," *Phys. Rev. X*, vol. 10, no. 3, Jul. 2020, Art. no. 031010, doi: [10.1103/physrevx.10.031010](https://doi.org/10.1103/physrevx.10.031010).
- [57] Z. Eldredge, Z.-X. Gong, J. T. Young, A. H. Moosavian, M. Foss-Feig, and A. V. Gorshkov, "Fast quantum state transfer and entanglement renormalization using long-range interactions," *Phys. Rev. Lett.*, vol. 119, no. 17, Oct. 2017, Art. no. 170503, doi: [10.1103/physrevlett.119.170503](https://doi.org/10.1103/physrevlett.119.170503).
- [58] M. C. Tran et al., "Hierarchy of linear light cones with long-range interactions," *Phys. Rev. X*, vol. 10, no. 3, Jul. 2020, Art. no. 031009, doi: [10.1103/physrevx.10.031009](https://doi.org/10.1103/physrevx.10.031009).
- [59] M. C. Tran, A. Y. Guo, A. Deshpande, A. Lucas, and A. V. Gorshkov, "Optimal state transfer and entanglement generation in power-law interacting systems," *Phys. Rev. X*, vol. 11, no. 3, Jul. 2021, Art. no. 031016, doi: [10.1103/physrevx.11.031016](https://doi.org/10.1103/physrevx.11.031016).
- [60] B. Nachtergaele, Y. Ogata, and R. Sims, "Propagation of correlations in quantum lattice systems," *J. Stat. Phys.*, vol. 124, no. 1, pp. 1–13, Jul. 2006, doi: [10.1007/s10955-006-9143-6](https://doi.org/10.1007/s10955-006-9143-6).
- [61] B. Nachtergaele and R. Sims, "Lieb-robinson bounds and the exponential clustering theorem," *Commun. Math. Phys.*, vol. 265, no. 1, pp. 119–130, Mar. 2006, doi: [10.1007/s00220-006-1556-1](https://doi.org/10.1007/s00220-006-1556-1).

- [62] Z.-X. Gong, M. Foss-Feig, S. Michalakis, and A. V. Gorshkov, "Persistence of locality in systems with power-law interactions," *Phys. Rev. Lett.*, vol. 113, no. 3, Jul. 2014, Art. no. 030602, doi: [10.1103/physrevlett.113.030602](https://doi.org/10.1103/physrevlett.113.030602).
- [63] D.-M. Storch, M. V. den Worm, and M. Kastner, "Interplay of sound-cone and supersonic propagation in lattice models with power law interactions," *New J. Phys.*, vol. 17, no. 6, Jun. 2015, Art. no. 063021, doi: [10.1088/1367-2630/17/6/063021](https://doi.org/10.1088/1367-2630/17/6/063021).
- [64] B. Nachtergaele, H. Raz, B. Schlein, and R. Sims, "Lieb-robinson bounds for harmonic and anharmonic lattice systems," *Commun. Math. Phys.*, vol. 286, no. 3, pp. 1073–1098, Sep. 2008, doi: [10.1007/s00220-008-0630-2](https://doi.org/10.1007/s00220-008-0630-2).
- [65] I. Prémont-Schwarz, A. Hamma, I. Klich, and F. Markopoulou-Kalamara, "Lieb-robinson bounds for commutator-bounded operators," *Phys. Rev. A, Gen. Phys.*, vol. 81, no. 4, Apr. 2010, Art. no. 040102(R), doi: [10.1103/physreva.81.040102](https://doi.org/10.1103/physreva.81.040102).
- [66] I. Prémont-Schwarz and J. Hnybida, "Lieb-Robinson bounds on the speed of information propagation," *Phys. Rev. A, Gen. Phys.*, vol. 81, no. 6, Jun. 2010, Art. no. 062107. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevA.81.062107>
- [67] R. Scheibler, S. Haghshatshoar, and M. Vetterli, "A fast Hadamard transform for signals with sublinear sparsity in the transform domain," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 2115–2132, Apr. 2015, doi: [10.1109/TIT.2015.2404441](https://doi.org/10.1109/TIT.2015.2404441).
- [68] X. Li, J. K. Bradley, S. Pawar, and K. Ramchandran, "SPRIGHT: A fast and robust framework for sparse Walsh–Hadamard transform," 2015, *arXiv:1508.06336*.
- [69] H.-Y. Huang, S. T. Flammia, and J. Preskill, "Foundations for learning from noisy quantum experiments," 2022, *arXiv:2204.13691*.
- [70] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*. Cham, Switzerland: Springer, 1997, pp. 423–435.
- [71] E. Campbell, "Random compiler for fast Hamiltonian simulation," *Phys. Rev. Lett.*, vol. 123, no. 7, Aug. 2019, Art. no. 070503.
- [72] D. W. Berry, A. M. Childs, Y. Su, X. Wang, and N. Wiebe, "Time-dependent Hamiltonian simulation with L^1 -norm scaling," *Quantum*, vol. 4, p. 254, Oct. 2020.
- [73] C.-F. Chen, H.-Y. Huang, R. Kueng, and J. A. Tropp, "Concentration for random product formulas," *PRX Quantum*, vol. 2, no. 4, Oct. 2021, Art. no. 040305.
- [74] L. Viola and S. Lloyd, "Dynamical suppression of decoherence in two-state quantum systems," *Phys. Rev. A, Gen. Phys.*, vol. 58, no. 4, pp. 2733–2744, Oct. 1998.
- [75] L. Viola, E. Knill, and S. Lloyd, "Dynamical decoupling of open quantum systems," *Phys. Rev. Lett.*, vol. 82, no. 12, pp. 2417–2421, Mar. 1999.
- [76] S. Kimmel, G. H. Low, and T. J. Yoder, "Robust calibration of a universal single-qubit gate set via robust phase estimation," *Phys. Rev. A, Gen. Phys.*, vol. 92, no. 6, Dec. 2015, Art. no. 062315.
- [77] H. J. Briegel and R. Raussendorf, "Persistent entanglement in arrays of interacting particles," *Phys. Rev. Lett.*, vol. 86, no. 5, pp. 910–913, Jan. 2001.
- [78] T.-C. Lu, L. A. Lessa, I. H. Kim, and T. H. Hsieh, "Measurement as a shortcut to long-range entangled quantum matter," *PRX Quantum*, vol. 3, no. 4, Dec. 2022, Art. no. 040337.
- [79] Y. Ye and E. Tse, "An extension of Karmarkar's projective algorithm for convex quadratic programming," *Math. Program.*, vol. 44, nos. 1–3, pp. 157–179, May 1989.
- [80] E. Knill et al., "Randomized benchmarking of quantum gates," *Phys. Rev. A, Gen. Phys.*, vol. 77, no. 1, Jan. 2008, Art. no. 012307.
- [81] E. Magesan, J. M. Gambetta, and J. Emerson, "Scalable and robust randomized benchmarking of quantum processes," *Phys. Rev. Lett.*, vol. 106, no. 18, May 2011, Art. no. 180504.
- [82] A. Erhard et al., "Characterizing large-scale quantum computers via cycle benchmarking," *Nature Commun.*, vol. 10, no. 1, pp. 1–7, Nov. 2019.
- [83] R. Harper, S. T. Flammia, and J. J. Wallman, "Efficient learning of quantum noise," *Nature Phys.*, vol. 16, no. 12, pp. 1184–1188, Dec. 2020.
- [84] S. Chen, J. Cotler, H.-Y. Huang, and J. Li, "Exponential separations between learning with and without quantum memory," in *Proc. IEEE 62nd Annu. Symp. Found. Comput. Sci. (FOCS)*, Feb. 2022, pp. 574–585.
- [85] H.-Y. Huang et al., "Quantum advantage in learning from experiments," *Science*, vol. 376, no. 6598, pp. 1182–1186, Jun. 2022.
- [86] Q. Zhao, Y. Zhou, A. F. Shaw, T. Li, and A. M. Childs, "Hamiltonian simulation with random inputs," *Phys. Rev. Lett.*, vol. 129, no. 27, Dec. 2022, Art. no. 270502.
- [87] Y. Borns-Weil and D. Fang, "Uniform observable error bounds of trotter formulae for the semiclassical Schrödinger equation," *Multiscale Model. Simul.*, vol. 23, no. 1, pp. 255–277, Mar. 2025.
- [88] Y. Su, H.-Y. Huang, and E. T. Campbell, "Nearly tight trotterization of interacting electrons," *Quantum*, vol. 5, p. 495, Jul. 2021.
- [89] D. An, D. Fang, and L. Lin, "Time-dependent unbounded Hamiltonian simulation with vector norm scaling," *Quantum*, vol. 5, p. 459, May 2021.
- [90] D. Fang and A. T. Vilanova, "Observable error bounds of the time-splitting scheme for quantum-classical molecular dynamics," *SIAM J. Numer. Anal.*, vol. 61, no. 1, pp. 26–44, Feb. 2023.
- [91] Y. Tong, V. V. Albert, J. R. McClean, J. Preskill, and Y. Su, "Provably accurate simulation of gauge theories and bosonic systems," *Quantum*, vol. 6, p. 816, Sep. 2022.
- [92] Y. Cai, Y. Tong, and J. Preskill, "Stochastic error cancellation in analog quantum simulation," in *Proc. 19th Conf. Theory Quantum Comput., Commun. Cryptography*, 2024, pp. 2:1–2:15. [Online]. Available: <https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.TQC.2024.2>
- [93] A. Bluhm, M. C. Caro, and A. Oufkir, "Hamiltonian property testing," *Quantum*, vol. 10, p. 1979, Jan. 2026, doi: [10.22331/q-2026-01-21-1979](https://doi.org/10.22331/q-2026-01-21-1979).
- [94] Z. Bao and P. Yao, "On testing and learning quantum junta channels," in *Proc. 36th Annu. Conf. Learn. Theory*, 2023, pp. 1064–1094.

Muzhou Ma received the B.S. degree in electronic engineering from Tsinghua University in 2025. He is currently pursuing the Ph.D. degree in computer science with California Institute of Technology. His research interests include theoretical computer science and information theory.

Steven T. Flammia received the B.S. degree in physics and mathematics from The Pennsylvania State University in 2002 and the Ph.D. degree in physics from The University of New Mexico in 2007. He is currently a Professor of computer science with Virginia Tech and the Director of the Center for Quantum Architecture and Software Development, Virginia Tech Institute for Advanced Computing. Previously, he held a research and faculty positions with the AWS Center for Quantum Computing, The University of Sydney, the University of Washington, Caltech, and the Perimeter Institute. His research interests include quantum error correction, fault-tolerant quantum computing, quantum characterization and verification, and the learning of quantum states, dynamics, and noise.

John Preskill received the A.B. degree in physics from Princeton University in 1975 and the Ph.D. degree in physics from Harvard University in 1980. He is currently a Richard P. Feynman Professor of theoretical physics with California Institute of Technology (Caltech) and the Director of the Institute for Quantum Information and Matter. Before joining Caltech in 1983, he was a Junior Fellow of the Harvard Society of Fellows and an Associate Professor of physics with Harvard University. His research interests include quantum computation, quantum information theory, quantum error correction, quantum many-body physics, and quantum gravity. He is a member of the National Academy of Sciences and a fellow of American Physical Society.

Yu Tong received the B.S. degree in computational mathematics from Peking University in 2017 and the Ph.D. degree in applied mathematics from the University of California, Berkeley, in 2022. From 2022 to 2024, he was an Institute for Quantum Information and Matter Post-Doctoral Scholar with California Institute of Technology. He is currently an Assistant Professor of mathematics and electrical and computer engineering with Duke University and a member of the Duke Quantum Center. His research interests include quantum algorithms, quantum learning theory, and numerical and analytical methods for quantum many-body systems.