PHYSICS

Attention to quantum complexity

Hyejin Kim[1]†, Yiqing Zhou[1]†, Yichen Xu[1]†, Kaarthik Varma[1], Amir H. Karamlou[2], Ilan T. Rosen[3], Jesse C. Hoke[4,5], Chao Wan[6], Jin Peng Zhou[6], William D. Oliver[2,3,7], Yuri D. Lensky[1,4], Kilian Q. Weinberger[6,8], Eun-Ah Kim[1,4,9]*

Copyright © 2025 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY).

The imminent era of error-corrected quantum computing demands robust methods to characterize quantum state complexity from limited, noisy measurements. We introduce the Quantum Attention Network (QuAN), a classical artificial intelligence (AI) framework leveraging attention mechanisms tailored for learning quantum complexity. Inspired by large language models, QuAN treats measurement snapshots as tokens while respecting permutation invariance. Combined with our parameter-efficient miniset self-attention block, this enables QuAN to access high-order moments of bit-string distributions and preferentially attend to less noisy snapshots. We test QuAN across three quantum simulation settings: driven hard-core Bose-Hubbard model, random quantum circuits, and toric code under coherent and incoherent noise. QuAN directly learns entanglement and state complexity growth from experimental computational basis measurements, including complexity growth in random circuits from noisy data. In regimes inaccessible to existing theory, QuAN unveils the complete phase diagram for noisy toric code data as a function of both noise types, highlighting AI's transformative potential for assisting quantum hardware.

INTRODUCTION

Artificial intelligence (AI) and quantum information science are among the most active areas in cutting-edge science and technology, addressing the computational complexity frontier. Although these two domains have evolved separately in the past, recent breakthroughs in both fields create a unique opportunity to use AI to learn quantum complexity. The most paradigm-shifting aspect of the latest large language models, such as ChatGPT, is their generality: Generally trained big models can reason in many different complex settings using natural languages. As quantum hardware platforms enter an era with error correction within reach (*1*–*3*), a general-purpose method for deciphering quantum states with unprecedented levels of complexity and entanglement is critically needed. We ask a compelling question: Can the core mechanism of the large language model's success, the attention mechanism (*4*, *5*), drive a general-purpose machine for learning quantum complexity? The answer to this question will hinge upon whether an intelligent use of the attention mechanism can hit the core aspects of the quantum complexity using only a polynomial in system size number of measurements from noisy devices.

Quantum complexity, once an abstract information-theoretical concept, has become a lynchpin at the intersection of multiple subfields: the study of black holes and geometric theory of wormholes (*6*–*8*), quantum information (*9*–*11*), topological states (*12*–*14*), and thermalization studies (*15*–*20*). Relative complexity between two states represented by density matrices ρ_α and ρ_β aims to quantify the difficulty of producing one from the other. For pure states $|\alpha\rangle$ and $|\beta\rangle$, Brown and Susskind (*7*) defined it in terms of circuit complexity, i.e., the minimal number of gates required to implement a unitary U connecting the states ($|\beta\rangle = U|\alpha\rangle$). Although helpful in exploring quantum chaos, directly applying the definition by computing an efficient circuit is often impractical. The challenge is further aggravated by the fact that experimental systems are invariably in mixed states. Theoretical extensions like purification complexity (*21*) and complexity entropy (*22*) demand additional quantum computations, complicating practical use. In practice, our knowledge of states ρ_α and ρ_β is limited to the collection of their measurement bit-string outcomes, $\{B_j\}_\alpha$ and $\{B_j\}_\beta$, and the operations done to the states just before measurements (see Fig. 1A). From a data-centric point of view, the task of learning the relative complexity between ρ_α and ρ_β from available data is a binary classification task of distinguishing

$$\{B_j\}_\alpha \quad \text{vs} \quad \{B_j\}_\beta \tag{1}$$

where $j = 1, \ldots, M$, with the sample size M.

Expectation values of observables, such as order parameters, are the most common and classic choice for comparing two states ρ_α and ρ_β. Even when the observable is not known a priori, finding characteristic spatial motifs using machine learning has been powerful when the two states can be distinguished by quantities linear in density matrices (*23*–*29*). At the opposite extreme of ambition for learning to characterize a state ρ is to model the full probability distribution of bit-strings associated with the given state ρ in the space of 2^{N_q} possibilities

$$p(\{b_i\}|\rho), \quad i = 1, \ldots, 2^{N_q} \tag{2}$$

where N_q is the number of qubits, from the measurement data $\{B_j\}$. While the advantage of generative modeling has been explored (*30*–*35*), such modeling is ultimately restricted to relatively small systems. A natural middle ground would be to learn key features of the probability distribution in Eq. 2 through moments of the distribution. This approach can be motivated both from a purely statistical perspective and from the perspective of the analysis of local chaotic dynamics' approach to purely random evolution (*15*, *36*, *37*).

[1]Department of Physics, Cornell University, Ithaca, NY 14853, USA. [2]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [4]Google Research, Mountain View, CA 94043, USA. [5]Department of Physics, Stanford University, Stanford, CA 94305, USA. [6]Department of Computer Science, Cornell University, Ithaca, NY 14853, USA. [7]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [8]ASAPP, New York City, NY 10007, USA. [9]Department of Physics, Ewha Womans University, Seoul, South Korea.
*Corresponding author. Email: eun-ah.kim@cornell.edu
†These authors contributed equally to this work.

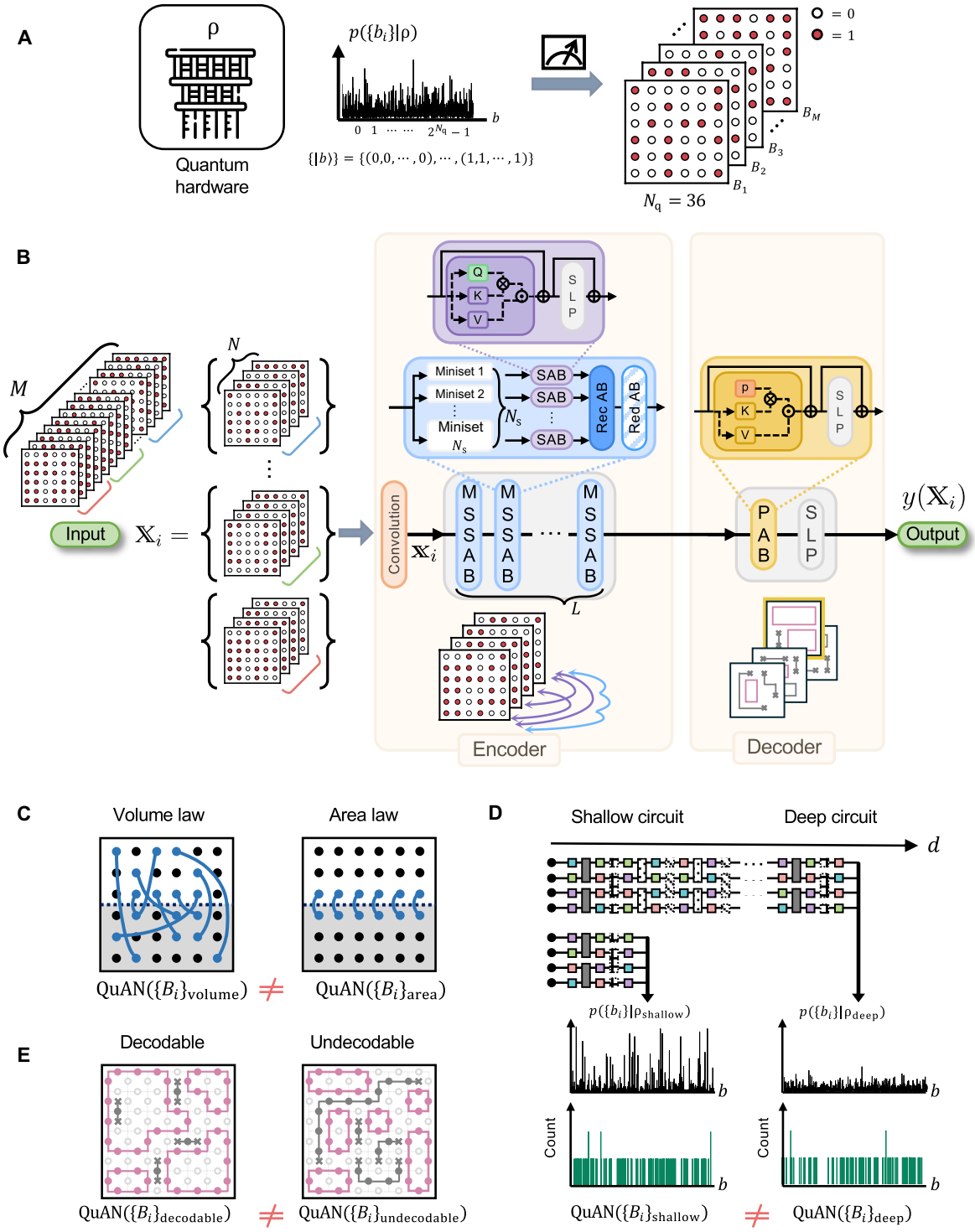


Fig. 1. Learning relative complexity between states ρ_α and ρ_β from bit-string collections. (**A**) Measurements of a quantum state ρ samples bit-strings $\{B_i\}$ from bit-string probability distribution $p(\{b_i\}|\rho)$ over the 2^{N_q}-dimensional Hilbert space. (**B**) Schematic architecture of the QuAN. The Z-basis snapshot collection of size M is partitioned into sets $\{\mathbb{X}_i\}$ of size N. In the encoder stage, after convolution registers positions of qubits, the set goes through L layers of the MSSAB. Inside the MSSAB, the input is further partitioned into N_s minisets to be parallel processed through SABs, recurrent attention block (RecAB), and reducing attention block (RedAB). The decoder stage compresses the output from the encoder, allowing for attending to different components in a permutation-invariant manner, using a PAB and single-layer perception (SLP). The output label is $y = 1$ for the state ρ_α and $y = 0$ for the state ρ_β. See Supplementary Materials section A for more details. (**C** to **E**) Examples of ρ_α and ρ_β for learning relative complexity using the binary classification output of the QuAN. (C) Volume-law entangled state versus area-law entangled state. The entanglement between two subsystems (white and gray) is indicated through blue links. (D) Random circuit state at depth d versus that at some deep reference depth. (E) Decodable versus undecodable states of an error-correcting code under noise. The incoherent noise depicted in gray suppresses large loops.

Here, we introduce the Quantum Attention Network (QuAN) shown in Fig. 1B as a general-purpose AI for a classical computer to learn relative quantum complexity between two states ρ_α and ρ_β through binary classification between the measurement bit-strings $\{B_j\}_\alpha$ and $\{B_j\}_\beta$. The QuAN capitalizes on the fact that the self-attention mechanism (*4*) learns the varying significance of the correlation between words at arbitrary distances within a sentence. However, while words in a sentence form a sequence, the order of bit-strings generated by identical experiments has no meaning. Through multiple layers of self-attention blocks (SABs) attending between bit-strings in a manner that is manifestly permutation invariant, the QuAN approximately learns high moments of the distribution $p(\{b_i\}|\rho)$ from a finite number of samples (see Supplementary Materials section A for more details). Furthermore, by attending to snapshots less affected by noise, the QuAN extracts target features of pure state in the presence of a finite amount of noise. In the rest of the paper, we present the design principles of the QuAN guided by the learning goals and demonstrate the QuAN's efficacy as a general-purpose AI by deploying QuAN to learn relative complexity between two states ρ_α and ρ_β that differ in ways that are fundamental to three frontiers: entanglement transition, chaotic dynamics, and error correction, as sketched in Fig. 1 (C to E).

Figure 1B shows the QuAN architecture and the data flow through QuAN. QuAN takes a set of snapshots $\mathbb{X}_i$ as input and outputs the machine's confidence $y(\mathbb{X}_i) \in [0, 1]$ on the input belonging to class α between the binary choice of α and β. To attend between snapshots, we first partition the complete data consisting of M snapshots into sets consisting of N snapshots. Typically, we train with 70 to 75% of data and validate with the rest (see Supplementary Materials sections C1, D1, and E1 for more details). Each set $\mathbb{X}_i$ then goes through three stages in sequence: convolution, encoder, and decoder stages (see Fig. 1B). The convolution stage incorporates local spatial features and maps the binary-valued $\mathbb{X}_i$ to vectors $\mathbb{x}_i$ with continuous entries with better algebraic properties for sampling moments. In language models, the encoding stage transforms the input into an informative, learned representation. For QuAN's encoding to target moments of $\mathbb{x}_i$, we introduce miniset self-attention blocks (MSSABs). Unique to the QuAN architecture, the MSSAB is adapted from a permutation-invariant version of the transformer (*38*) to access high moments of the distribution in a parameter-efficient manner. It accesses up to $2N_s^2$ order correlations for N_S minisets within one layer.

The decoder stage consists of the pooling attention block (PAB) and the single-layer perceptron that compresses all the information into the confidence output (see Supplementary Materials section A4). The PAB layer learns to attend more to the snapshots with features characteristic of the target state. When the confidence is $y(\mathbb{X}_i) < 0.5$, the QuAN assesses the data to belong to state α; otherwise, it is assessing the data to belong to state β. We train the QuAN by minimizing binary cross-entropy loss between the ground truth and QuAN's prediction through Adam optimization. The multifaceted use of attention mechanisms empowers the QuAN to be a versatile general-purpose AI capable of learning quantum complexity in various datasets.

RESULTS

First, we consider the relative complexity between states with different entanglement scaling: a volume-law scaling state and an area-law scaling state (see Fig. 1C). The distinction is substantial because classical computers cannot efficiently represent a general volume-law scaling state. Moreover, the change in entanglement scaling signals measurement-induced phase transitions (*39*–*44*). Nevertheless, extraction of the entanglement scaling is often challenging because it requires randomized multibasis measurements or state tomography for subsystems with varying sizes (*45*). Our key insight is that the QuAN can learn the change in the entanglement scaling by attending between snapshots within the set (see Fig. 2A). The self-attention score for the set $\mathbb{X}_i$ in each SAB is given by

$$\langle Q\mathbb{x}_i \,|\, K\mathbb{x}_i \rangle = \left(Q\mathbb{x}_i\right)\left(K\mathbb{x}_i\right)^T \tag{3}$$

up to normalization, where $\mathbb{x}_i$ is convolved from the snapshot set $\mathbb{X}_i$ as shown in Fig. 1B; Q and K are two trainable transformation matrices, often referred to as query (Q) and key (K) (see Supplementary Materials section A3a for more details). When two Z-basis snapshots are related by simultaneous bit-flips at a pair of qubits (j, k), the attention score reflects the correlation $\langle X_j X_k \rangle$, which is upper-bounded by their mutual information (see Fig. 2A and also Supplementary Materials section C4) (*46*). Hence, the QuAN can access Pauli X correlations from Z-basis measurement through intersnapshot attention. In an area-law state, an X correlation is concentrated between nearby qubits. On the other hand, in a volume-law state, the entanglement of each qubit shared with the entire system dilutes the X correlation. By accessing the X correlation through intersnapshot attentions, the QuAN may witness the entanglement transition from just Z-basis snapshots.

To verify QuAN's potential for witnessing the entanglement transition from the Z–basis measurements, we turn to coherent-like states prepared to be a superposition of hard-core Bose Hubbard model energy eigenstates reached through a driven Hamiltonian

$$H/\hbar = \sum_{\langle j,k \rangle} J_{jk} \hat{\sigma}_j^+ \hat{\sigma}_k^- + \frac{\delta}{2} \sum_j \hat{\sigma}_j^z + \Omega \sum_j \left(\alpha_j \hat{\sigma}_j^- + \text{h. c.}\right) \tag{4}$$

implemented using a superconducting, transmon-based quantum simulator (see Fig. 2B). Here, $\hat{\sigma}_j^+ \left(\hat{\sigma}_j^-\right)$ represents the raising (lowering) operator on qubit at site j, and $\hat{\sigma}_j^z$ represents the Pauli Z operator at site j; J_{jk} is the particle exchange interaction strength between sites j and k of average value J, δ is the detuning between the drive and qubit frequency, and Ω is the drive strength (see Materials and Methods). Upon tuning δ/J, the system prepares a coherent-like superposition of states at the center of the spectrum for small δ/J and that of states at the edge of the spectrum for large δ/J (*45*). Bipartite entanglement entropy calculated from subsystem tomography using subsystem measurements in an informationally complete basis set found the volume-law scaling at low values of $|\delta|/J$ and the area-law scaling at large values of $|\delta|/J$ (see Fig. 2C). Here, we use the experimental snapshots of the entire system in the particle number basis, which maps to Z-basis measurements in the hard-core limit over a range of δ/J.

To investigate the QuAN's capability and the role of "attention" in witnessing the entanglement transition, we compared the performance of three different architectures with varying degrees of attention. The simplest architecture is the set multilayer perceptron (SMLP) without the SABs and PABs (see Fig. 2D). The SMLP is a generalization of the usual multilayer perception (*47*) designed to

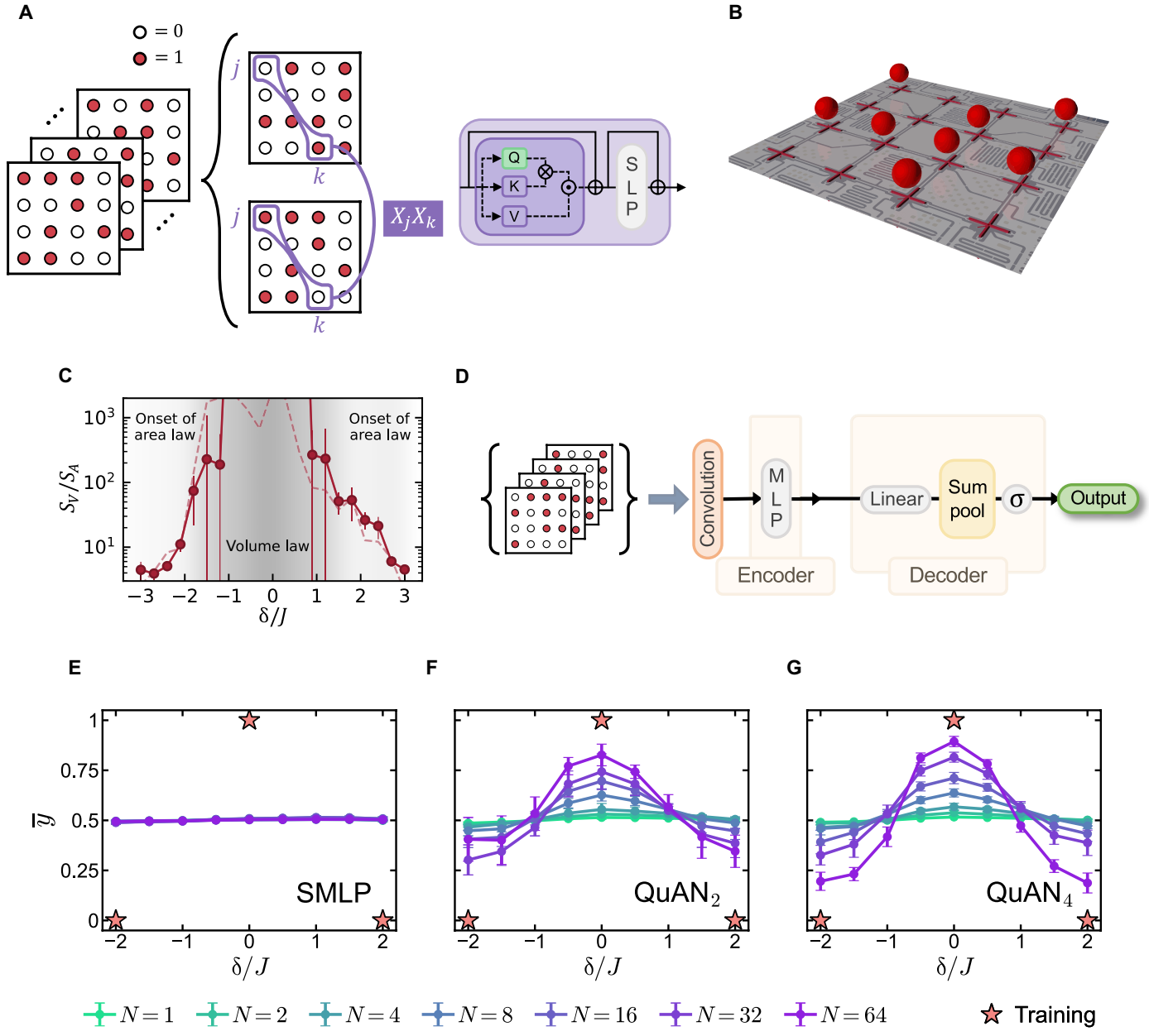


Fig. 2. Relative complexity between volume-law and area-law scaling states. (**A**) Intersnapshot correlation reveals X-X correlation of the quantum state. The purple box shows the schematic of the SAB capturing the intersnapshot correlation. (**B**) Schematic diagram of the 16-transmon-qubit chip used for quantum emulation of the driven hard-core Boson-Hubbard model. (**C**) Entanglement transition based on the scaling of bipartite entanglement entropy $S = S_A A + S_V V$, where A and V represent the area and volume of the subsystem, respectively. Adapted from Karamlou *et al.* (*45*) (https://creativecommons.org/licenses/by/4.0/). (**D**) Schematic of a contrast architecture: The SMLP respects the permutation symmetry. (**E** to **G**) Average confidence $\bar{y}$ as a function of detuning strength δ for different architectures using different set sizes N. The star symbol marks the training points. The average and errors are obtained from 10 independent model training. For machine learning details, see Supplementary Materials section C2. (E) The SMLP fails to train. (F) $QuAN_2$ ($N_s = 1, L = 1$). (G) $QuAN_4$ with two layers of self-attention ($N_s = 1, L = 2$).

take a set of snapshots as input and learn the positional information through convolution. We set the miniset size for the other two architectures to $N_s = 1$, which reduces the MSSAB to a single SAB. $QuAN_2$ and $QuAN_4$ each access up to the second and fourth moments through SAB layers (see Supplementary Materials section A3d).

The task for the three architectures is to witness the change in the entanglement scaling upon an increase in δ/J by training the models to distinguish the $\delta/J = 0$ state from the $\delta/J = 2$ state using snapshots with the same number ($n = 8$; see Supplementary Materials section C1) of bosons. All three architectures were trained and tested using $M = 69{,}632$ experimental snapshots from $\delta/J = 0$ and the same number of snapshots from $\delta/J = \pm 2$ with the binary label: $y = 1$ for data from $\delta/J = 0$ with volume-law entanglement and $y = 0$ for data from $\delta/J = \pm 2$ with area-law entanglement.

Figure 2 (E to G) compares the performance of the three architectures on the basis of 10 independent training for each architecture. When an architecture learns to witness the entanglement transition, average confidence $\overline{y} = \langle y(\mathbb{X}_i) \rangle_i$ should span between $\overline{y} = 1$ at $\delta/J = 0$ and $\overline{y} = 0$ at $\delta/J = 2$ (see Supplementary Materials section C2). Figure 2E shows that the SMLP fails to learn the entanglement transition, given the average confidence remaining flat at $\overline{y} = 0.5$ independent of set size N (see Fig. 2E). By contrast, QuAN_2, which only uses second moments of bit-strings, already learns the entanglement transition once a set structure is used. Moreover, its learning improves with increasing set size within the bounds of the total sample size (see Fig. 2F and see Supplementary Materials section C3). Accessing up to the fourth moment, QuAN_4 shows sharper transition with reduced error bars across different training. The above comparison showcases the QuAN's ability to witness changes in entanglement scaling using Z-basis measurements alone through intersnapshot attention.

Next, we move on to learning the relative state complexity between shallow and deep random circuit states: ρ_{shallow} versus ρ_{deep}. The random circuits have become a theoretical paradigm for studying diverse sets of phenomena in quantum dynamics from information processing in black holes (*48*) to quantum chaos and thermalization (*15*, *49*–*51*). Despite being composed of local gates (see Fig. 3A), the ensemble of random circuits accurately approximates the ensemble of global random unitary transformations up to the first k-moments, where k grows with the depth of the circuit (*36*). Furthermore, the circuit complexity, which refers to the minimal number of gates required to evolve a fiduciary reference state to a state implemented by a given random circuit, is known to grow linearly with k (*15*). However, how to find such k for a particular random circuit from a finite amount of data is not known. Instead, we adopt our data-centric approach, which involves learning the relative complexity between the state reached at depth d and a reference deep circuit at $d = 20$ from the QuAN's accuracy at distinguishing data from the two states. We aim to learn the evolution of relative state complexity of states reached by random circuits as a function of the circuit depth by using the QuAN for multiple pairwise classification tasks that contrast data from variable depth d with data from $d = 20$. Training the QuAN to perform this classification probes its ability to learn higher-order features of the bit-string data, like aspects of the third and higher moments of the bit-string distribution (*15*), which reflect the relative complexity of the state at depth d compared to the state at depth $d = 20$. We anticipate that such relative complexity will shrink with increasing depth d. To coalesce the QuAN's learning across different training runs for each depth, we study the classification accuracy defined as the percentage of correctly classified inputs (by correct classification, meaning the machine's confidence in the correct labels exceeding 0.5).

While previous experiments primarily focused on deep circuits pushing against the limits of classical simulation, we systematically explore the increase in complexity as the depth increases. We consider both experimental data from quantum hardware (see Materials and Methods for experimental details) and data from the corresponding pure state implemented on classical hardware. At each circuit depth d, we sample $N_c = 50$ random circuit instances. For each circuit instance, we collect $M_s = M/N_c = 500{,}000$ bit-strings ($M_s = 2{,}000{,}000$ for $N_q = 36$). To prevent the QuAN from overfitting a specific circuit instance, we train the QuAN using data from 70% of all the circuit instances and reserve the remaining 30% circuits for testing (see Supplementary Materials section D1). We batch bit-strings from the same circuit instance into sets of N bit-strings, shown in Fig. 3B. See Supplementary Materials section D2 for more details on training and testing.

We benchmark the QuAN's learning of relative complexity at each circuit depth d against the depth dependence of the linear cross-entropy benchmark (XEB). The linear XEB, $\mathcal{F}_{\text{XEB}}$, for a collection of bit-strings $\{B_j\}$ is defined by

$$\mathcal{F}_{\text{XEB}}\left[\{B_j\}\right] = 2^{N_q} \left\langle p_U\left(B_j\right) \right\rangle_j - 1 \tag{5}$$

where $p_U(B_j)$ is the probability for the given bit-string B_j to be sampled from an ideal pure state $U|0\rangle^{\otimes N_q}$ for the random unitary circuit U of depth d. We first study the depth-dependent relative complexity of the ideal pure-state data. When $\{B_j\}$ is sampled from the same ideal pure state, $\mathcal{F}_{\text{XEB}}\left[\{B_j\}\right]$ measures the second moment of the bit-string distribution associated with the pure state. As shown in Fig. 3C, the XEB for such bit-string saturates the infinite depth asymptotic value of $\mathcal{F}_{\text{XEB}} = 1$ to polynomial precision near depth $d = 8$ for all system sizes considered, becoming blind to circuit dynamics past $d = 8$. The relative complexity learned by the QuAN quantified through classification accuracy shrinks upon increasing depth as expected, approaching the vanishing relative complexity at 50% accuracy of the classification task. Nevertheless, QuAN_{50} that accesses up to the 50th moment (see Supplementary Materials section A3) learns the relative complexity of $d = 8$ for all system sizes (see Fig. 3D). Furthermore, the classification accuracy at $d = 8$ shows promising subexponential scaling with the increase in the system size (see Supplementary Materials section D4 for more details). Given that the QuAN learns the relative complexity, we present a visualization of the relative XEBs considering the differences between XEBs from two circuit depths in Supplementary Materials section D3 as a benchmarking baseline.

Comparing various architectures' learning of relative complexity at depth $d = 8$ shown in Fig. 3E reveals how difficult this task is for most architectures. Despite all having the same number of hyperparameters, architectures other than the QuAN failed to learn the relative complexity of $d = 8$ and 20. Notably, the failure of three architectures that took individual bit-strings without forming a set structure [the MLP, convolutional neural network (CNN), and standard transformer (Transf.); see Supplementary Materials section B] establishes the importance of the set structure for learning relative complexity. The failure of the SMLP and PAB shows that using a set structure is not enough without self-attention. With just a single layer of MSSAB, the QuAN learns the relative complexity of depth $d = 8$ (QuAN_2 in Fig. 3E). Allowed to learn up to the 50th moment, QuAN_{50}'s accuracy shoots up to demonstrate the benefit of MSSAB accessing higher moments. We refer curious readers to Supplementary Materials section D6 for a detailed analysis addressing how the maximal accessible order of moments affects the QuAN's learning. In the rest of the paper, we focus on the performance of QuAN_{50} with a single MSSAB block ($L = 1$) containing $N_s = 5$ minisets.

The advantage of the QuAN's relative complexity learning is further pronounced when the bit-strings, $\{B_j\}$, are experimentally sampled. For such data, the XEB $\mathcal{F}_{\text{XEB}}\left[\{B_j\}\right]$ estimates the correspondence between experimental snapshots and the snapshot distribution of the ideal circuit up to the second order. As shown in Fig. 3F, $\mathcal{F}_{\text{XEB}}\left[\{B_j\}\right]$ for the experimental data smoothly evolves to 0, reflecting

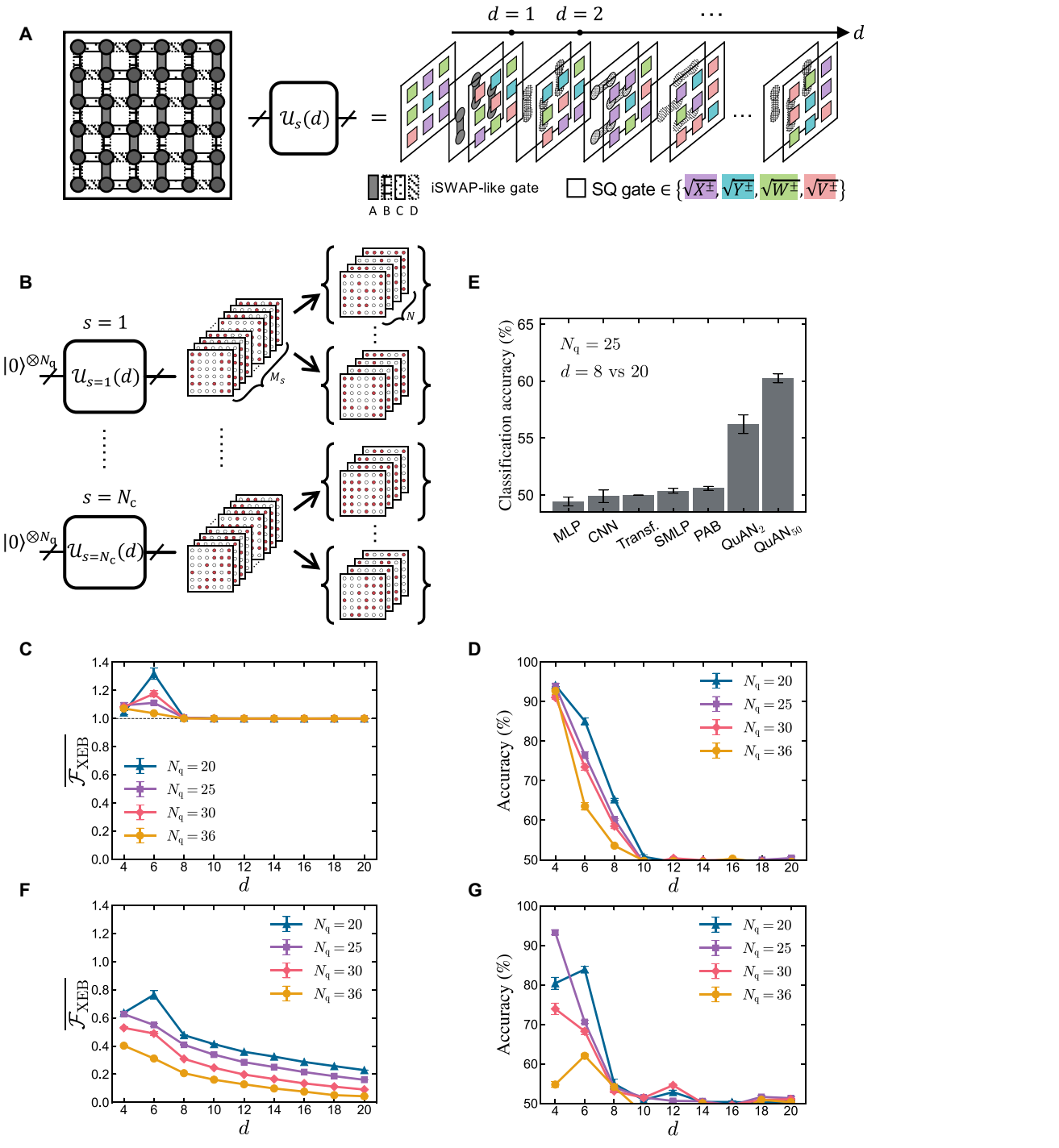


Fig. 3. Relative complexity between the random circuit state at depth *d* and the reference state at depth $d = 20$. (**A**) Schematic illustration of the 6-by-6 subarray of qubits from Google's "Sycamore" quantum processor. A random circuit of depth d alternates entangling iSWAP-like gates (gray) and single-qubit (SQ) gates randomly chosen from the set $\left\{\sqrt{X^{\pm1}}, \sqrt{Y^{\pm1}}, \sqrt{W^{\pm1}}, \sqrt{V^{\pm1}}\right\}$, with $W = (X+Y)/\sqrt{2}$ and $V = (X-Y)/\sqrt{2}$. The two-qubit gates are applied in a repeating series of ABCDCDAB patterns. (**B**) Data structure. For each depth d, we sample $N_c = 50$ circuits. For each circuit instance s, we sample M_s bit-strings and partition them into sets of size N, resulting in a total of $N_c \times M_s / N$ sets for each circuit depth d. (**C**) XEB (Eq. 5) for bit-strings from noiseless simulations as a function of circuit depth d with varying system sizes N_q. The markers show the averaged XEB over $N_c = 50$ different circuit instances and the error bars for the standard errors. (**D**) Pure-state trained QuAN$_{50}$'s classification accuracy for pure-state data. We train eight independent models at each circuit depth d and show the averaged accuracy (marker) and the standard error (error bar). QuAN$_{50}$ successfully learns the relative complexity of $d = 8$. (**E**) Comparison of the performances of QuAN$_2$, QuAN$_{50}$, and other architectures in learning the relative complexity of depth $d = 8$ on an $N_q = 25$ qubit system. The models maintain approximately the same total number of trainable parameters to make a controlled comparison between different architectures. (**F**) Averaged XEB for experimentally collected bit-strings. The plot shows the averaged XEB over 50 circuit instances (markers) and the standard error (error bars). The XEB smoothly decays as a function of depth d. (**G**) Learning relative complexity from experimental data using QuAN$_{50}$ trained on noiseless data.

the increase in noise with increasing depth, which pushes the system closer to the maximally mixed state rather than the intended pure state. Dominated by the noise, $\mathcal{F}_{\mathrm{XEB}}$ on experimental data does not reveal the increase in complexity driven by the intended unitary evolution. The QuAN also confirms the notable influence of increasing noise when the depth dependence of relative complexity is probed by training and testing the QuAN entirely on experimental data (see Supplementary Materials section D4). The key question is whether learning the remnants of the pure state dynamics from noisy experimental data is possible.

One way to learn the aspect of relative complexity that corresponds to pure state dynamics from the experimental data is to quantify the relative complexity of experimental bit-strings using the QuAN trained on noiseless pure state data. Unexpectedly, the depth dependence of the pure-state trained QuAN's classification accuracy of experimental data in Fig. 3G shows a trend that closely follows corresponding depth dependence in its inspection of the noiseless simulated data shown in Fig. 3D. Specifically, they both exhibit a transition at depth $d = 10$. Hence, the QuAN was able to reveal traces of pure-state complexity evolution from noisy experimental data.

Overcoming noise in relative complexity learning is particularly important for noisy states from error-correcting codes. As the quantum hardware approaches the breakeven point, the community is increasingly focused on learning the topological order, a phase of matter that supports quantum error correction, in a mixed state (*12*, *52*–*56*). The earlier literature on the complexity of topological states focused on the relative complexity of topological states compared to a simple product state $|0\rangle^{\otimes N_q}$. It is known that topological states' relative complexity with respect to $|0\rangle^{\otimes N_q}$ grows linearly with the system size. However, much less is known about the relative complexity between a topological state and an undecodably noisy state.

One leading candidate for fault-tolerant quantum memory is the toric code or $\mathbb{Z}_2$ topological order. In its ideal ground state $|\mathrm{TC}\rangle$, a closed Z-loop operator around a loop γ, $Z_{\mathrm{closed}}(\gamma) \equiv \prod_{i\in\gamma} Z_i$, has a length-independent expectation value

$$\langle \mathrm{TC}|\, Z_{\mathrm{closed}}(\gamma)\,|\mathrm{TC}\rangle = 1 \tag{6}$$

Infinitesimal noise introduces tension to the above loop operator expectation value, resulting in its exponential decay with the loop perimeter (*57*), as illustrated in Fig. 1E for incoherent noise. Hence, any amount of noise destroys the topological order from the loop tension perspective. Nevertheless, mapping the limiting cases to statistical mechanics models reveals hidden phase transitions. Specifically, with just the coherent noise g_X modeled through

$$|\Psi(g_X)\rangle\!\rangle = \frac{1}{\sqrt{\mathcal{N}}}\exp\left(g_X \sum_i X_i\right)|TC\rangle\!\rangle \tag{7}$$

where $\mathcal{N}$ is a normalization factor; the error threshold of $g_X \approx 0.22$ was established in (*58*) via mapping the problem to a classical two-dimensional (2D) Ising model. Alternatively, with just an incoherent noise p_{flip} modeled through the bit-flip error channel

$$\mathcal{E}_i(\rho) = (1-p_{\mathrm{flip}})\rho + p_{\mathrm{flip}} X_i \rho X_i \tag{8}$$

for each qubit i; the error threshold of $p_{\mathrm{flip}} \approx 0.11$ was established in (*52*) via mapping the error model to the random bond Ising model (*59*). However, the phase diagram interpolating between the two axes has yet to be achieved. Motivated by the QuAN's successes, we use the QuAN to solve this open problem.

To study the effect of coherent and incoherent noise in a controlled way, we use classically simulated toric code ground state modified by coherent noise strength g_X, available in (*60*) as a part of Cong *et al.* (*61*). To the Z-basis bit-string data sampled from this state, we implement the error channel (Eq. 8) through random bit-flips (see Materials and Methods). We then transform the resulting bit-strings into measurements of the smallest loops, building on the insight in (*28*). Now, the collection of these plaquette values goes into the QuAN as input. To arrive at the QuAN that interpolates between the two axes of the noise phase space (g_X, p_{flip}) in its learning of the relative complexity between decodable and undecodable states, we train the QuAN with nearly coherent data over the range of g_X value and deeply incoherent data over the same range of g_X value (see Fig. 4, C and D, and Supplementary Materials section E1). We then classify the data from the rest of the phase space.

Figure 4 (C and D) shows that with a sufficiently large set size, QuAN confidence marks a sharp distinction between decodable and undecodable states, saturating the error threshold for incoherent noise $p_{\mathrm{flip}} \approx 0.11$. The cut along the $g_X = 0$ axis (Fig. 4E) shows that the set structure and the attention mechanisms in the QuAN are essential. Upon increasing the set size, the transition is sharpening toward $p_{\mathrm{flip}} \approx 0.11$ along the $g_X = 0$ axis. With the set size of $N = 64$, the QuAN observes a sharp transition close to the theoretically predicted coherent noise threshold of $g_X \approx 0.22$ along the $p_{\mathrm{flip}} = 0$ axis (see Fig. 4F) (*62*). Furthermore, predictions of both phase boundaries are robust against shifts in the training parameter region (see Supplementary Materials section E3a). This is unexpected given that we did not explicitly train the QuAN to contrast $g_X = 0$ versus large $g_X \neq 0$ and warrants further theoretical investigation into the potential existence of a unifying statistical mechanics model.

The model ablation studies (see Fig. 4, G and H, and Supplementary Materials section E3c) revealed the critical role of the pooling attention decoder as an automatic importance sampler with excellent sample complexity. Specifically, we tried removing the self-attention stage entirely, leaving only the pooling attention stage; we refer to the resulting architecture as the PAB. When comparing the average confidence of the PAB architecture shown in Fig. 4 (G and H) with that of QuAN_2, we observe that the PAB effectively saturates the known thresholds along both axes for the set size $N = 64$. This allows in-depth interpretation of the machine's learning because the pooling attention score of the PAB can be traced to individual snapshots $\mathbb{X}_i$ (see Supplementary Materials section E4).

By comparing the snapshots with high and low pooling attention scores, we can gain insight into the features of the data recognized as that of a topological phase. For this, we inspect the distribution of the pooling attention score across all the snapshots with $p_{\mathrm{flip}} = 0.05$ and $g_X = 0$ shown in Fig. 4I. Selecting the snapshots with top 15% and bottom 15% of the distribution, we analyze the subset average value of the Z-loop operator Z_{closed} for a closed loop γ as a function of the length of the perimeter (see Fig. 4J). This expectation value can be readily calculated from each subset of Z-basis snapshots as an average. The contrast in the length dependence of the loop expectation value $\langle Z_{\mathrm{closed}}\rangle$ between the high and low attention groups

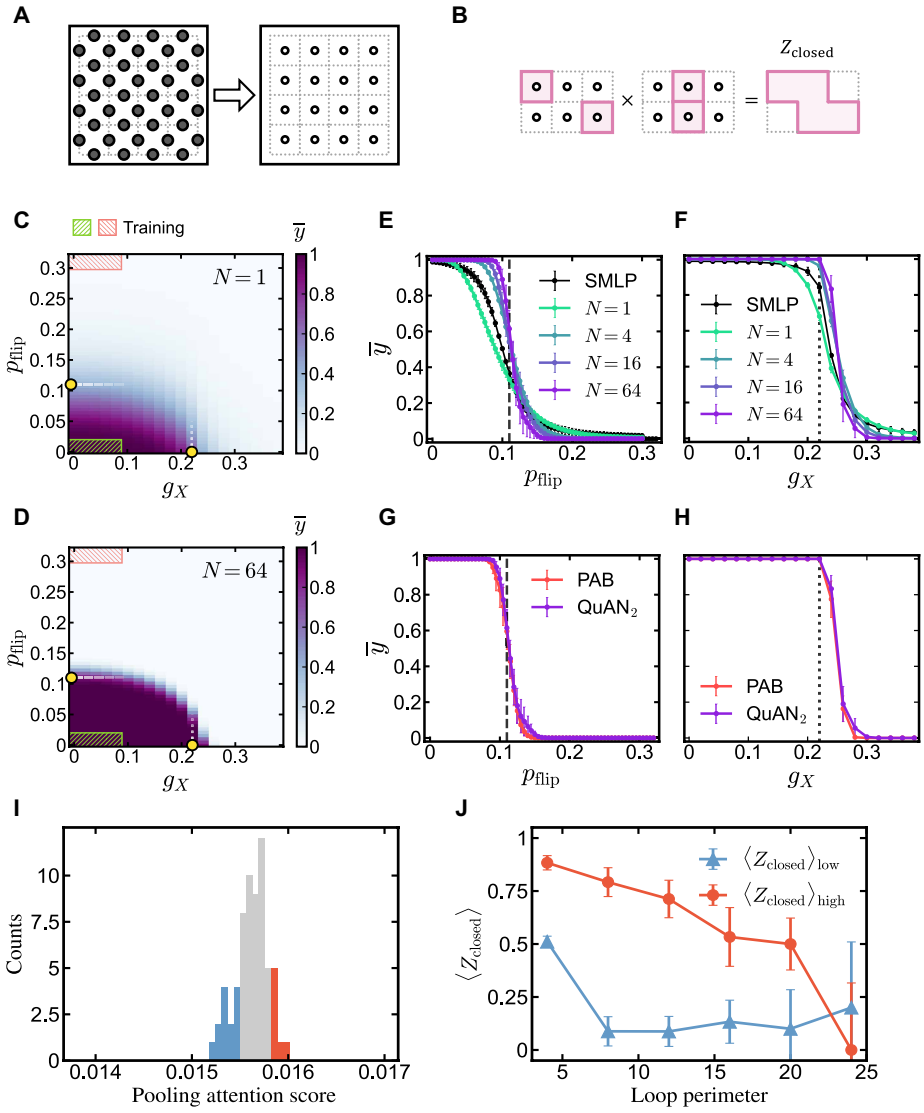


Fig. 4. Learning the relative complexity of decodable and undecodable states of the toric code. (**A**) Transformation from the *Z*-basis measurements to the smallest-loop, plaquette variables. (**B**) The QuAN can build larger closed loops through multiplication. (**C** and **D**) Decodability phase diagram of the toric code state under coherent and incoherent noise for two different set sizes: $N = 1$ in (C) and $N = 64$ in (D). The regions in the phase space that support the training data are marked with hatch marks. The average confidence $\bar{y}$ averages over 10 independent model training. The known thresholds are marked along the $g_X = 0$ axis at $p_c \approx 0.11$ and along the $p_{\text{flip}} = 0$ at $g_c \approx 0.22$. (**E**) Average confidence $\bar{y}$ by QuAN$_2$ for different set sizes N and by the SMLP with $N = 64$ along the axis $g_X = 0$. The error bar shows the standard error for $\bar{y}$ over 10 independent model training. (**F**) Average confidence $\bar{y}$ by QuAN$_2$ with varying set sizes N and by the SMLP with $N = 64$ along the axis $p_{\text{flip}} = 0$. (**G**) Average confidence $\bar{y}$ by QuAN$_2$ and the PAB with $N = 64$ along the axis $g_X = 0$, where the PAB is defined as the model without self-attention and has only pooling attention. (**H**) Average confidence $\bar{y}$ by QuAN$_2$ and the PAB with $N = 64$ along the axis $p_{\text{flip}} = 0$. (**I**) Pooling attention score histogram from the topological state with $(g_X, p_{\text{flip}}) = (0,0.05)$. (**J**) Loop expectation value $\langle Z_{\text{closed}}\rangle$ as a function of the loop perimeter for high- and low-attention-score snapshots in the topological state with $(g_X, p_{\text{flip}}) = (0,0.05)$. The error bars represent the standard error of $\langle Z_{\text{closed}}\rangle$ over different loop configurations in corresponding snapshots.

is notable. The snapshots in the high-attention-score group show large $\langle Z_{\text{closed}}\rangle$ with weak perimeter length dependence until the length hits the system size. On the other hand, for the snapshots in the low-attention-score group, $\langle Z_{\text{closed}}\rangle$ decays immediately after the smallest loop perimeter. Hence, it appears that the QuAN learned to selectively attend to snapshots with vanishing loop tension. The QuAN's importance sampling is a data-efficient alternative to seeking a dressed loop operator (*57*, *61*) with a length-independent expectation value or information-theoretical measures (*12*, *54*, *58*).

DISCUSSION

To summarize, we introduced the QuAN, a versatile general-purpose architecture for a classical computer that adopts the attention mechanism for learning relative quantum complexity between two states. We choose to restrict the QuAN's access to *Z*-basis measurements as such measurements are most broadly accessible to quantum hardware. While this means that off-diagonal elements of the density matrix cannot be learned, our analysis shows that much can be learned from such data when the salient aspects of the bit-string distributions are effectively learned using the QuAN. Specifically,

the bit-string distribution for a relatively more complex state compared to a product state will be spread out rather than concentrated. The QuAN can learn such a change in the distribution shape through the moments of the distribution. Also, a volume-law entangled state is more anticoncentrated. Hence, although the QuAN cannot directly learn entanglement from Z-basis measurements, it still witnesses the consequences of the entanglement transition. The QuAN is built on three principles: (i) treat the snapshots as a set with permutation invariance, (ii) attend between snapshots to access higher moments of bit-string distribution, and (iii) attend over snapshots to the importance sample. The QuAN treats each snapshot as a "token" and leverages the capability of stacked L-layers of N_s miniset self-attention to sample $\left(2N_s^2\right)^L$-th moments of snapshots.

We put the QuAN to work on three challenging sets of Z-basis data to showcase the power of the QuAN and gain insights. With the driven hard-core Bose-Hubbard (HCBH) model data, we found that the entanglement transition between the volume-law and area-law scaling regimes can be witnessed entirely with the Z-basis measurements. With random circuit sampling data, we revealed the evolution of complexity with increasing depth from noisy experimental data that reflect noise-free evolution. Last, with the mixed state data of the toric code with coherent and incoherent noise, we obtained the decodability phase diagram that saturates known error thresholds. The QuAN's findings set challenges for the theoretical understanding. Simultaneously, the QuAN's ability to learn relative quantum complexity through the adaptive use of attention mechanisms holds promise for quantum error correction, the key data-centric problem for the application of quantum hardware. While an AI's learning is typically hard to interpret, we revealed the critical role of self-attention and pulling attention through ablation studies. Promising future directions for the application of QuAN include the development of a decoder for quantum error correction codes and learning measurement-induced phase transitions.

MATERIALS AND METHODS

Data acquisition

The data used for learning the entanglement transition in the driven HCBH model is acquired from a 4-by-4 array of superconducting transmon qubits as described in (*45*). In this system, the on-site interaction, determined by the anharmonicity of the transmon qubits, is much stronger than the exchange interaction. Therefore, the system can be described by an HCBH Hamiltonian

$$\hat{H}_{\mathrm{HCBH}}/\hbar = \sum_{\langle j,k\rangle} J_{jk}\hat{\sigma}_j^{+}\hat{\sigma}_k^{-} - \sum_j \frac{\epsilon_j}{2}\hat{\sigma}_j^{z} \tag{9}$$

where $\hat{\sigma}_j^{+}$ ($\hat{\sigma}_j^{-}$) is the raising (lowering) operator for a two-level system at site j, and $\hat{\sigma}_j^{z}$ is the Pauli-Z operator. The first term describes the particle exchange interaction between neighboring lattice sites with strength J_{jk}, with an average strength of $J/2\pi = 5.9 \pm 0.4\,\mathrm{MHz}$. The second term represents the site energies, tuned by the transmon transition frequencies with an accuracy of $300\,\mathrm{kHz}$ ($\approx 5\times10^{-2}J$) in the device (*63*). This system features site-resolved, multiplexed single-shot dispersive qubit readout and enables simultaneous tomographic measurements of the qubit states when combined with single-qubit gates. To prepare superposition states across the energy spectrum of the lattice, the interacting qubits are simultaneously driven via a common control line. The Hamiltonian of the driven lattice is thus the one in Eq. 4. The drive strength Ω can be tuned by varying the amplitude of the applied drive pulse. The common drive couples to each qubit with a complex coefficient α_k [see Karamlou *et al.* (*45*) for details]. By changing the drive detuning δ, the distribution of the superposition states across the HCBH energy spectrum can be controlled: With detuning = 0, the superposition state will be concentrated near the center of the energy band, whereas as the magnitude of δ increases, the superposition state approaches the edge of the energy band. At each detuning strength δ, a snapshot is a collection of 4-by-4 projective measurement outcomes of $\hat{\sigma}_j^{z}$ for each site j. For more details on the data preprocessing of the driven HCBH model, see Supplementary Materials section C1.

The random quantum circuit experiment was conducted on a Google Sycamore processor composed of 70 frequency-tunable transmon qubits with tunable couplers. The quantum processor used has a similar design to Arute *et al.* (*64*) and was carried out on the same processor used in previous works, where typical coherence times, readout errors, and single- and two-qubit gate errors on this particular chip can be found in (*65*, *66*). The two-qubit gates used for this experiment are iSWAP-like gates with an iSWAP angle $\theta \approx 0.5\pi$ and conditional phase angle $\phi \approx 0.1\pi$ (*66*). We collected data on rectangular subarrays of $N_q = 20, 25, 30$, and 36 qubits (see fig. S11) with variable circuit depths $d = 4, 6, 8, 10, 12, 14, 16, 18$, and 20. For every N_q and d, we collected data on 50 different random circuit instances. Each instance contains a different sequence of single-qubit gates randomly chosen from gate set $\left\{\sqrt{X^{\pm1}}, \sqrt{Y^{\pm1}}, \sqrt{W^{\pm1}}, \sqrt{V^{\pm1}}\right\}$, with $W = (X+Y)/\sqrt{2}$ and $V = (X-Y)/\sqrt{2}$. For each of the $N_c = 50$ random circuit instances, we performed $M_s = 500{,}000$ ($M_s = 2{,}000{,}000$) Z-basis measurements for $N_q = 20{,}25{,}30$ ($N_q = 36$). Thus, $N_c \times M_s$ bit-strings were collected for each (N_q, d) pair.

In Figs. 3 (C and F), we show the linear XEB $\mathcal{F}_{\mathrm{XEB}}(N_q, d)$ as a function of circuit depth d for different system sizes N_q. The XEB is calculated as follows. Parallelized over eight NVIDIA A100 graphics processing units (GPUs), we can exactly simulate random quantum circuits up to $N_q = 36$ qubits. We use Cirq (*67*) to simulate the same random quantum circuit instances used in the experiment. For each circuit instance, we evolve an all-zero product state $|0\rangle^{\otimes N_q}$ with the circuit to get a state vector $|\psi_s(N_q, d)\rangle$, where $d \in [4, 6, \ldots, 20]$ represents the depth of the circuit, and $s \in [1, 2, \ldots, N_c]$ represents $N_c = 50$ different circuit instances. To simulate measuring a state in the Z-basis, we sample from the distribution given by $|\psi_s(N_q, d)|^2$. Similar to the sample size we have in experiments, we draw $M_s = 500{,}000$ samples for $N_q = 20, 25$, and 30 and $M_s = 2{,}000{,}000$ for $N_q = 36$. The linear XEB is defined as

$$\mathcal{F}_{\mathrm{XEB}}(N_q, d, s) = 2^{N_q}\left\langle p(B_j)\right\rangle_j - 1 = 2^{N_q}\sum_{b\in(0,1)^{\otimes N_q}} p(b)^2 - 1 \tag{10}$$

where $p(B_j) = |\langle \psi_s(N_q, d)\,|\,B_j\rangle|^2$. From a finite set of samples, we can get an estimate of $\mathcal{F}_{\mathrm{XEB}}$

$$\mathcal{F}_{\mathrm{XEB}}(N_q, d, s) \approx 2^{N_q}\left[\frac{1}{M}\sum_{j=1}^{M} p(B_j)\right] - 1 \tag{11}$$

We exactly calculate simulated $\mathcal{F}_{\mathrm{XEB}}(N_q, d, s)$ using Eq. 10 for each state $|\psi_s(N_q, d)\rangle$ and then average over $N_c = 50$ circuit instances with the same circuit depth to get $\overline{\mathcal{F}_{\mathrm{XEB}}}(N_q, d)$ in Fig. 3C. The error

bar in the plot is the standard error over multiple circuit instances. For experimental XEB (see the main text and Fig. 3E), we estimate $\mathcal{F}_{\mathrm{XEB}}(N_{\mathrm{q}}, d, s)$ using Eq. 11 for each state $|\psi_s(N_{\mathrm{q}}, d)\rangle$ and then average over $N_{\mathrm{c}} = 50$ circuit instances to get estimated $\overline{\mathcal{F}_{\mathrm{XEB}}}(N_{\mathrm{q}}, d)$. For more details on the data preprocessing of the random quantum circuit, see Supplementary Materials section D1.

The data of toric code state under coherent and incoherent errors are generated as follows. We start with the Z-basis bit-string measurement from the deformed toric code state with coherent noise, now available in an open-source database (*60*) as a part of Cong *et al.* (*61*). The database includes bit-strings obtained by simulating measuring the toric code state $|\mathrm{TC}\rangle$ deformed by coherent X, Z noise with varying strengths

$$|\psi(g_X, g_Z)\rangle = \frac{1}{\mathcal{N}} e^{-g_X \sum_i X_i - g_Z \sum_i Z_i} |\mathrm{TC}\rangle \tag{12}$$

The bit-strings available in the database are simulated and sampled using the projected entangled pair state on a 300-by-1000 vertex square lattice (*61*). We then introduce incoherent noise through bit-flip with probability p_{flip} (see the main text and Eq. 6). Thus, the resulting bit-strings are effectively sampled from mixed states $\rho(g_X, g_Z, p_{\mathrm{flip}})$. For more details on the data preprocessing of the noisy toric code state, see Supplementary Materials section E1.

QuAN architecture

The QuAN is a machine learning model designed to learn high-order statistical correlations between quantum measurement snapshots while respecting permutation invariance. It consists of three main components: a convolutional preprocessing module, an encoder based on the MSSAB, and a decoder using a PAB. The architecture is outlined in Fig. 1E, and see Supplementary Materials section A for more details.

Input and convolution

The input to the QuAN is a set $\mathbb{X}_i = \{B_{i,\alpha}\}_{\alpha=1}^{N}$ consisting of N binary-valued arrays $B_{i,\alpha}$, each representing a 2D bit-string of N_{q} qubits. Each element $B_{i,\alpha,\mu} \in \{0, 1\}$ is indexed by spatial coordinate μ. For clarity, we denote $B_{i,\alpha,\mu}$ as $B^{\alpha}_{i,\mu}$. The goal of the model is to predict a label $y(\mathbb{X}_i)$, trained via binary cross-entropy loss

$$\mathcal{L} = -\sum_i \hat{y}_i \log y(\mathbb{X}_i) \tag{13}$$

where $\hat{y}_i$ is the ground truth.

To prepare the inputs, we apply a 2D convolutional layer with kernel size 2 and stride 1 to each bit-string, followed by batch normalization and flattening. If the input bit-string is of shape $(N_{\mathrm{r}}, N_{\mathrm{c}})$ and n_{c} convolution filters are used, the output dimension per element becomes $d_x = n_{\mathrm{c}}(N_{\mathrm{r}} - 1)(N_{\mathrm{c}} - 1)$. The output of this stage is $\mathbb{x}^{\alpha}_{\mu}$, a matrix of shape (N, d_x).

MSSAB encoder

The encoder processes the preprocessed input using the MSSAB, which accesses high-order intersnapshot correlations efficiently.

Parallel SAB. The input set is randomly shuffled and partitioned into N_{s} minisets, each of size N / N_{s}. Each miniset is processed independently by a SAB with shared parameters. The input to the encoder is denoted as $\mathbb{x}^{\alpha}_{\mu}$, where α indexes the set element and μ represents the dimension of the feature space after the convolution layer. The SAB transforms the input set to a set of hidden state vectors

$$\mathbb{h}^{\alpha}_{\mu} = \sum_{\nu} Q_{\mu\nu} \mathbb{x}^{\alpha}_{\mu} + \sum_{\beta} A^{\alpha\beta} \sum_{\nu} V_{\mu\nu} \mathbb{x}^{\beta}_{\nu} \tag{14}$$

where Q, K, and V are learnable weights of shape (d_h, d_x). The attention score matrix $A^{\alpha\beta}$ is given by

$$A^{\alpha\beta} = \mathrm{Softmax}\left[\sum_{\rho\lambda\eta} \frac{1}{\sqrt{d_h}} Q_{\rho\lambda} \mathbb{x}^{\alpha}_{\lambda} K_{\rho\eta} \mathbb{x}^{\beta}_{\eta}\right] \tag{15}$$

The output undergoes residual connection, layer normalization, and feed-forward transformation

$$\begin{aligned} \mathbb{h}'^{\alpha}_{\mu} &= \mathrm{LayerNorm}\left(\mathbb{h}^{\alpha}_{\mu}\right) \\ \mathbb{y}^{\alpha}_{\mu} &= \mathrm{Sigmoid}\left(\mathrm{LayerNorm}\left(\mathbb{h}'_{\mu} + \mathrm{FF}_{\mu\nu}(\mathbb{h}'_{\nu})\right)\right) \end{aligned} \tag{16}$$

where $\mathrm{FF}(\mathbb{x}) = \mathrm{Re}LU(O\mathbb{x})$ or $Sigmoid(O\mathbb{x})$ for learned weight matrix $\mathbb{x}$.

Recurrent AB (RecAB). Each miniset output $\mathbb{y}_{(m)}$ undergoes attention with other $N_{\mathrm{s}} - 1$ minisets using multihead attention blocks (MABs)

$$\mathbb{h}^{\alpha}_{(t+1),\mu} = \sum_{\nu} \mathrm{Q}'_{\mu\nu} \mathbb{y}^{\alpha}_{[(\mathrm{m+t+1})\mathrm{modN_s}],\nu} \tag{17}$$

$$+ \sum_{\beta} \mathrm{Softmax}\left\{\left[\sum_{\rho\lambda\eta} \frac{1}{\sqrt{d_h}} Q'_{\rho\lambda} \mathbb{y}^{\alpha}_{[(\mathrm{m+t+1})\mathrm{modN_s}]} K'_{\rho\eta} \mathbb{h}^{\beta}_{(t),\eta}\right]\right\} \tag{18}$$

$$\times \sum_{\nu} V'_{\mu\nu} \mathbb{h}^{\beta}_{(t),\nu}\Bigg\} \tag{19}$$

where $\mathbb{h}_{(0)} \equiv \mathbb{y}_{(m)}$, and $\mathbb{h}_{(N_{\mathrm{S}}-1)} \equiv \mathbb{y}'_{(m)}$. This enables capturing up to $2N_{\mathrm{s}}$-th order moments in the input.

Reducing AB (RedAB). To merge N_{s} minisets into one while preserving permutation invariance, RedAB performs $N_{\mathrm{s}} - 1$ recurrent MAB operations in a randomized order

$$\mathbb{h}'_{(t+1)} = \mathrm{MAB}\left\{\mathbb{h}'_{(t)}, \mathbb{y}'_{[\sigma(t+1)]}\right\} \tag{20}$$

resulting in a final output $\mathbb{z}^{\alpha}_{\mu}$ of size $(N/N_{\mathrm{s}}, d_h)$. The MSSAB with one layer accesses up to $2N_{\mathrm{s}}^2$-th moments. For further details on moment order and computational complexity, see Supplementary Materials section A3d.

PAB decoder

The decoder compresses the encoder output $\mathbb{z}^{\alpha}_{\mu}$ into a scalar prediction $y(\mathbb{X})$. The PAB uses a seed vector S and computes attention scores s'^{β} over set elements

$$\mathbb{p}_{\mu} = S_{\mu} + \sum_{\beta} s'^{\beta} \sum_{\nu} V''_{\mu\nu} \mathbb{z}^{\beta}_{\nu} \tag{21}$$

where the pooling attention score is given by

$$s'^{\beta} = \mathrm{Softmax}\left[\sum_{\rho\lambda} \frac{1}{\sqrt{d_h}} S_{\rho} K''_{\rho\lambda} \mathbb{z}^{\beta}_{\lambda}\right] \tag{22}$$

After layer normalization, a final linear transformation produces the output

$$\mathbb{p}'_\mu = \mathrm{LayerNorm}\left(\mathbb{p}_\mu\right) \tag{23}$$

$$y(\mathbb{X}) = \mathrm{Sigmoid}\left(\sum_\mu W_\mu \mathrm{LayerNorm}\left(\mathbb{p}'_\mu + rFF_{\mu\lambda}\left(\mathbb{p}'_\nu\right)\right) + b \right) \tag{24}$$

where W, b, K'', V'', and S are learnable parameters. This design enables the QuAN to learn from measurement data while ensuring permutation invariance and leveraging high-order moment correlations effectively and efficiently.

Training and testing procedure

All models presented in this work are implemented using PyTorch and trained using the Adam optimizer to minimize binary cross-entropy loss. A key feature across all tasks is the use of set-structured input data, where each data point comprises multiple quantum measurement snapshots. For the task of learning the entanglement transition in the driven HCBH model, we train binary classifiers to distinguish between states with different entanglement scaling laws. Each model is trained independently for a given set size, and the one with the highest validation accuracy is used for testing. The testing region corresponds to unknown intermediate labels; thus, we analyze the phase transition by averaging the machine confidence $y\left(\mathbb{X}_i\right)$ over multiple testing sets, and 10 independent models are trained to see good convergence in model performance. We train each model using one A100 GPU for 500 epochs, with a total wall-clock time of ~150 s (see Supplementary Materials section C2 for detailed hyperparameter settings). For learning the complexity between random circuit states, we classify measurement outcomes by circuit depth using depth $d = 20$ as a reference and by varying the shallow depth. We use a fixed set size of $N = 10{,}000$ and enhance training with a learning rate scheduler and Xavier initialization. Eight independent models are trained for each depth pair $(d, 20)$ on simulated data, and performance is evaluated on both simulated and experimental data collected from Google's Sycamore processor. We train each model using one A100 GPU for 400 epochs, with a total wall-clock time of ~24 hours (see Supplementary Materials section D2 for detailed hyperparameter settings). For learning the decodable and undecodable states of the toric code, we train binary classifiers to distinguish between two states (which also correspond to the topological and trivial phases). Unlike other tasks, convolution layers are omitted to prioritize closed-loop feature learning using MLPs. The resulting phase diagram is derived from the averaged model confidence over testing sets at each parameter point $\left(g_X, p_{\mathrm{flip}}\right)$. We train each model using one A100 GPU for 200 epochs, with a total wall-clock time of ~500 s (see Supplementary Materials section E2 for detailed hyperparameter settings).

To ensure robust generalization and avoid overfitting, we adopt a "dataset shuffling period," where input sets are regenerated or shuffled every 10 epochs. For each model and configuration, we perform multiple independent training runs and report either the average highest validation accuracy or the average machine confidence and its standard error over the trained models. We maintain a consistent classification threshold of $y = 0.5$ for label prediction.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S24
Tables S1 to S17
References

REFERENCES AND NOTES

1. D. Bluvstein, S. J. Evered, A. A. Geim, S. H. Li, H. Zhou, T. Manovitz, S. Ebadi, M. Cain, M. Kalinowski, D. Hangleiter, J. P. Bonilla Ataides, N. Maskara, I. Cong, X. Gao, P. Sales Rodriguez, T. Karolyshyn, G. Semeghini, M. J. Gullans, M. Greiner, V. Vuletić, M. D. Lukin, Logical quantum processor based on reconfigurable atom arrays. *Nature* **626**, 58–65 (2024).
2. Google Quantum AI, Suppressing quantum errors by scaling a surface code logical qubit. *Nature* **614**, 676–681 (2023).
3. Z. Ni, S. Li, X. Deng, Y. Cai, L. Zhang, W. Wang, Z.-B. Yang, H. Yu, F. Yan, S. Liu, C.-L. Zou, L. Sun, S.-B. Zheng, Y. Xu, D. Yu, Beating the break-even point with a discrete-variable-encoded logical qubit. *Nature* **616**, 56–60 (2023).
4. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2017), p. 6000–6010; https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
5. A. Parikh, O. Täckström, D. Das, J. Uszkoreit, A decomposable attention model for natural language inference. arXiv:1606.01933 (2016).
6. L. Susskind, Computational complexity and black hole horizons. *Fortschr. Phys.* **64**, 24–43 (2016).
7. A. R. Brown, L. Susskind, Second law of quantum complexity. *Phys. Rev. D* **97**, 086015 (2018).
8. A. Bouland, B. Fefferman, U. Vazirani, Computational pseudorandomness, the wormhole growth paradox, and constraints on the AdS/CFT duality. arXiv:1910.14646 (2019).
9. S. Aaronson, Y. Atia, L. Susskind, On the hardness of detecting macroscopic superpositions. arXiv:2009.07450 (2020).
10. D. Gross, S. T. Flammia, J. Eisert, Most quantum states are too entangled to be useful as computational resources. *Phys. Rev. Lett.* **102**, 190501 (2009).
11. D. Girolami, F. Anzà, Quantifying the difference between many-body quantum states. *Phys. Rev. Lett.* **126**, 170502 (2021).
12. M. B. Hastings, Topological order at nonzero temperature. *Phys. Rev. Lett.* **107**, 210501 (2011).
13. M. Schwarz, K. Temme, F. Verstraete, D. Perez-Garcia, T. S. Cubitt, Preparing topological projected entangled pair states on a quantum computer. *Phys. Rev. A* **88**, 032321 (2013).
14. J. Miller, A. Miyake, Latent computational complexity of symmetry-protected topological order with fractional symmetry. *Phys. Rev. Lett.* **120**, 170503 (2018).
15. F. G. Brandão, W. Chemissany, N. Hunter-Jones, R. Kueng, J. Preskill, Models of quantum complexity growth. *PRX Quantum* **2**, 030316 (2021).
16. J. Haferkamp, P. Faist, N. B. T. Kothakonda, J. Eisert, N. Yunger Halpern, Linear growth of quantum circuit complexity. *Nat. Phys.* **18**, 528–532 (2022).
17. K. Kaneko, E. Iyoda, T. Sagawa, Characterizing complexity of many-body quantum dynamics by higher-order eigenstate thermalization. *Phys. Rev. A* **101**, 042126 (2020).
18. E. Bianchi, L. Hackl, M. Kieburg, M. Rigol, L. Vidmar, Volume-law entanglement entropy of typical pure quantum states. *PRX Quantum* **3**, 030201 (2022).
19. J. Wang, M. H. Lamann, J. Richter, R. Steinigeweg, A. Dymarsky, J. Gemmer, Eigenstate thermalization hypothesis and its deviations from random-matrix theory beyond the thermalization time. *Phys. Rev. Lett.* **128**, 180601 (2022).
20. B. D. S. L. Torres, E. Martín-Martínez, Toward a physically motivated notion of Gaussian complexity geometry. *Phys. Rev. D* **110**, 025006 (2024).
21. H. A. Camargo, L. Hackl, M. P. Heller, A. Jahn, T. Takayanagi, B. Windt, Entanglement and complexity of purification in (1+1)-dimensional free conformal field theories. *Phys. Rev. Res.* **3**, 013248 (2021).
22. N. Yunger Halpern, N. B. T. Kothakonda, J. Haferkamp, A. Munson, J. Eisert, P. Faist, Resource theory of quantum uncomplexity. *Phys. Rev. A* **106**, 062417 (2022).
23. C. Miles, R. Samajdar, S. Ebadi, T. T. Wang, H. Pichler, S. Sachdev, M. D. Lukin, M. Greiner, K. Q. Weinberger, E.-A. Kim, Machine learning discovery of new phases in programmable quantum simulator snapshots. *Phys. Rev. Res.* **5**, 013026 (2023).
24. H.-Y. Huang, R. Kueng, G. Torlai, V. V. Albert, J. Preskill, Provably efficient machine learning for quantum many-body problems. *Science* **377**, eabk3333 (2022).
25. C. Miles, A. Bohrdt, R. Wu, C. Chiu, M. Xu, G. Ji, M. Greiner, K. Q. Weinberger, E. Demler, E.-A. Kim, Correlator convolutional neural networks as an interpretable architecture for image-like quantum matter data. *Nat. Commun.* **12**, 3905 (2021).
26. J. Carrasquilla, Machine learning for quantum matter. *Adv. Phys. X.* **5**, 1797528 (2020).

27. Y. Zhang, P. Ginsparg, E.-A. Kim, Interpreting machine learning of topological quantum phase transitions. *Phys. Rev. Res.* **2**, 023283 (2020).
28. Y. Zhang, R. G. Melko, E.-A. Kim, Machine learning $\mathbb{Z}_2$ quantum spin liquids with quasiparticle statistics. *Phys. Rev. B* **96**, 245119 (2017).
29. Y. Zhang, E.-A. Kim, Quantum loop topography for machine learning. *Phys. Rev. Lett.* **118**, 216401 (2017).
30. S. Ahmed, C. Sánchez Muñoz, F. Nori, A. F. Kockum, Quantum state tomography with conditional generative adversarial networks. *Phys. Rev. Lett.* **127**, 140502 (2021).
31. M. Y. Niu, A. M. Dai, L. Li, A. Odena, Z. Zhao, V. Smelyanskyi, H. Neven, S. Boixo, Learnability and complexity of quantum samples. arXiv:2010.11983 (2020).
32. P. Cha, P. Ginsparg, F. Wu, J. Carrasquilla, P. L. McMahon, E.-A. Kim, Attention-based quantum tomography. *Mach. Learn. Sci. Technol.* **3**, 01LT01 (2021).
33. J. Carrasquilla, G. Torlai, R. G. Melko, L. Aolita, Reconstructing quantum states with generative models. *Nat. Mach. Intell.* **1**, 155–161 (2019).
34. J. Carrasquilla, D. Luo, F. Pérez, A. Milsted, B. K. Clark, M. Volkovs, L. Aolita, Probabilistic simulation of quantum circuits using a deep-learning architecture. *Phys. Rev. A* **104**, 032610 (2021).
35. Y.-H. Zhang, M. Di Ventra, Transformer quantum state: A multipurpose model for quantum many-body problems. *Phys. Rev. B* **107**, 075147 (2023).
36. F. G. S. L. Brandão, A. W. Harrow, M. Horodecki, Local random quantum circuits are approximate polynomial-designs. *Commun. Math. Phys.* **346**, 397–434 (2016).
37. J. Haferkamp, Random quantum circuits are approximate unitary t-designs in depth $O\left(nt^{5+o(1)}\right)$. arXiv:2203.16571 (2022).
38. J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *Proceedings of the 36th International Conference on Machine Learning* (PMLR, 2019), pp. 3744–3753; https://proceedings.mlr.press/v97/lee19d.html.
39. D. N. Page, Average entropy of a subsystem. *Phys. Rev. Lett.* **71**, 1291–1294 (1993).
40. Y. Li, X. Chen, M. P. A. Fisher, Measurement-driven entanglement transition in hybrid quantum circuits. *Phys. Rev. B* **100**, 134306 (2019).
41. B. Skinner, J. Ruhman, A. Nahum, Measurement-induced phase transitions in the dynamics of entanglement. *Phys. Rev. X* **9**, 031009 (2019).
42. C.-M. Jian, Y.-Z. You, R. Vasseur, A. W. W. Ludwig, Measurement-induced criticality in random quantum circuits. *Phys. Rev. B* **101**, 104302 (2020).
43. M. J. Gullans, D. A. Huse, Dynamical purification phase transition induced by quantum measurements. *Phys. Rev. X* **10**, 041020 (2020).
44. S. Choi, Y. Bao, X.-L. Qi, E. Altman, Quantum error correction in scrambling dynamics and measurement-induced phase transition. *Phys. Rev. Lett.* **125**, 030505 (2020).
45. A. H. Karamlou, I. T. Rosen, S. E. Muschinske, C. N. Barrett, A. Di Paolo, L. Ding, P. M. Harrington, M. Hays, R. Das, D. K. Kim, B. M. Niedzielski, M. Schuldt, K. Serniak, M. E. Schwartz, J. L. Yoder, S. Gustavsson, Y. Yanay, J. A. Grover, W. D. Oliver, Probing entanglement in a 2D hard-core Bose-Hubbard lattice. *Nature* **629**, 561–566 (2024).
46. M. M. Wolf, F. Verstraete, M. B. Hastings, J. I. Cirac, Area laws in quantum systems: Mutual information and correlations. *Phys. Rev. Lett.* **100**, 070502 (2008).
47. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016); http://www.deeplearningbook.org.
48. P. Hayden, J. Preskill, Black holes as mirrors: Quantum information in random subsystems. *J. High Energy Phys.* **2007**, 120–120 (2007).
49. A. Nahum, J. Ruhman, S. Vijay, J. Haah, Quantum entanglement growth under random unitary dynamics. *Phys. Rev. X* **7**, 031016 (2017).
50. C. W. Von Keyserlingk, T. Rakovszky, F. Pollmann, S. L. Sondhi, Operator hydrodynamics, OTOCs, and entanglement growth in systems without conservation laws. *Phys. Rev. X* **8**, 021013 (2018).
51. A. Nahum, S. Vijay, J. Haah, Operator spreading in random unitary circuits. *Phys. Rev. X* **8**, 021014 (2018).
52. E. Dennis, A. Kitaev, A. Landahl, J. Preskill, Topological quantum memory. *J. Math. Phys.* **43**, 4452–4505 (2002).
53. Y. Bao, R. Fan, A. Vishwanath, E. Altman, Mixed-state topological order and the errorfield double formulation of decoherence-induced transitions. arXiv:2301.05687 (2023).
54. R. Fan, Y. Bao, E. Altman, A. Vishwanath, Diagnostics of mixed-state topological order and breakdown of quantum memory. *PRX Quantum* **5**, 020343 (2024).
55. Y.-H. Chen, T. Grover, Unconventional topological mixed-state transition and critical phase induced by self-dual coherent errors. *Phys. Rev. B* **110**, 125152 (2024).
56. R. Sohal, A. Prem, Noisy approach to intrinsically mixed-state topological order. *PRX Quantum* **6**, 010313 (2025).
57. M. B. Hastings, X.-G. Wen, Quasiadiabatic continuation of quantum states: The stability of topological ground-state degeneracy and emergent gauge invariance. *Phys. Rev. B* **72**, 045141 (2005).
58. C. Castelnovo, C. Chamon, Topological order and topological entropy in classical systems. *Phys. Rev. B* **76**, 174416 (2007).
59. A. Honecker, M. Picco, P. Pujol, Universality class of the Nishimori point in the 2D $\pm J$ random-bond Ising model. *Phys. Rev. Lett.* **87**, 047201 (2001).
60. N. Maskara, "Enhancing detection of topological order by local error correction" (2023); https://osf.io/k8up2/.
61. I. Cong, N. Maskara, M. C. Tran, H. Pichler, G. Semeghini, S. F. Yelin, S. Choi, M. D. Lukin, Enhancing detection of topological order by local error correction. *Nat. Commun.* **15**, 1527 (2024).
62. C. Castelnovo, C. Chamon, Quantum topological phase transition at the microscopic level. *Phys. Rev. B* **77**, 054433 (2008).
63. C. N. Barrett, A. H. Karamlou, S. E. Muschinske, I. T. Rosen, J. Braumüller, R. Das, D. K. Kim, B. M. Niedzielski, M. Schuldt, K. Serniak, M. E. Schwartz, J. L. Yoder, T. P. Orlando, S. Gustavsson, J. A. Grover, W. D. Oliver, Learning-based calibration of flux crosstalk in transmon qubit arrays. *Phys. Rev. Appl.* **20**, 024070 (2023).
64. F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, T. S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven, J. M. Martinis, Quantum supremacy using a programmable superconducting processor. *Nature* **574**, 505–510 (2019).
65. Google Quantum AI and Collaborators, Measurement-induced entanglement and teleportation on a noisy quantum processor. *Nature* **622**, 481–486 (2023).
66. A. Morvan, B. Villalonga, X. Mi, S. Mandrà, A. Bengtsson, P. V. Klimov, Z. Chen, S. Hong, C. Erickson, I. K. Drozdov, J. Chau, G. Laun, R. Movassagh, A. Asfaw, L. T. A. N. Brandão, R. Peralta, D. Abanin, R. Acharya, R. Allen, T. I. Andersen, K. Anderson, M. Ansmann, F. Arute, K. Arya, J. Atalaya, J. C. Bardin, A. Bilmes, G. Bortoli, A. Bourassa, J. Bovaird, L. Brill, M. Broughton, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, J. Campero, H.-S. Chang, B. Chiaro, D. Chik, C. Chou, J. Cogan, R. Collins, P. Conner, W. Courtney, A. L. Crook, B. Curtin, D. M. Debroy, A. D. T. Barba, S. Demura, A. D. Paolo, A. Dunsworth, L. Faoro, E. Farhi, R. Fatemi, V. S. Ferreira, L. F. Burgos, E. Forati, A. G. Fowler, B. Foxen, G. Garcia, É. Genois, W. Giang, C. Gidney, D. Gilboa, M. Giustina, R. Gosula, A. G. Dau, J. A. Gross, S. Habegger, M. C. Hamilton, M. Hansen, M. P. Harrigan, S. D. Harrington, P. Heu, M. R. Hoffmann, T. Huang, A. Huff, W. J. Huggins, L. B. Ioffe, S. V. Isakov, J. Iveland, E. Jeffrey, Z. Jiang, C. Jones, P. Juhas, D. Kafri, T. Khattar, M. Khezri, M. Kieferová, S. Kim, A. Kitaev, A. R. Klots, A. N. Korotkov, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, K.-M. Lau, L. Laws, J. Lee, K. W. Lee, Y. D. Lensky, B. J. Lester, A. T. Lill, W. Liu, W. P. Livingston, A. Locharla, F. D. Malone, O. Martin, S. Martin, J. R. McClean, M. McEwen, K. C. Miao, A. Mieszala, S. Montazeri, W. Mruczkiewicz, O. Naaman, M. Neeley, C. Neill, A. Nersisyan, M. Newman, J. H. Ng, A. Nguyen, M. Nguyen, M. Y. Niu, T. E. O'Brien, S. Omonije, A. Opremcak, A. Petukhov, R. Potter, L. P. Pryadko, C. Quintana, D. M. Rhodes, C. Rocque, E. Rosenberg, N. C. Rubin, N. Saei, D. Sank, K. Sankaragomathi, K. J. Satzinger, H. F. Schurkus, C. Schuster, M. J. Shearn, A. Shorter, N. Shutty, V. Shvarts, V. Sivak, J. Skruzny, W. C. Smith, R. D. Somma, G. Sterling, D. Strain, M. Szalay, D. Thor, A. Torres, G. Vidal, C. V. Heidweiller, T. White, B. W. K. Woo, C. Xing, Z. J. Yao, P. Yeh, J. Yoo, G. Young, A. Zalcman, Y. Zhang, N. Zhu, N. Zobrist, E. G. Rieffel, R. Biswas, R. Babbush, D. Bacon, J. Hilton, E. Lucero, H. Neven, A. Megrant, J. Kelly, P. Roushan, I. Aleiner, V. Smelyanskiy, K. Kechedzhi, Y. Chen, S. Boixo, Phase transitions in random circuit sampling. *Nature* **634**, 328–333 (2024).
67. Cirq Developers, Cirq (2023); https://zenodo.org/records/10247207.
68. M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, A. J. Smola, "Deep sets," in *Advances in Neural Information Processing Systems* (Curran Associates Inc., 2017), vol. 30; https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

Acknowledgments: We thank J. Carrasquilla, S. Gopalakrishnan, T. Grover, R. Huang, J. Preskill, N. Read, and P. Roushan for helpful discussions. We thank I. Cong, N. Maskara, and M. Lukin for sharing the toric code simulation data before publication and for discussions. **Funding:** H.K., Y.X., K.V., Y.D.L., C.W., J.P.Z., K.Q.W., and E.-A.K. acknowledge support from the NSF through OAC-2118310. Y.Z. acknowledges support from NSF Materials Research Science and Engineering Center (MRSEC) through DMR-1719875 and from Platform for the Accelerated Realization, Analysis, and Discovery of Interface Materials (PARADIM), supported by the NSF under Cooperative Agreement No. DMR-2039380. I.T.R. is supported by an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at the Massachusetts Institute of Technology administered by Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy and the Office of the Director of National Intelligence (ODNI).Y.D.L. and E.-A.K. acknowledge New Frontier Grant from Cornell University's College of Arts and Sciences. The work at MIT (to A.H.K., I.T.R., and W.D.O.) was supported in part by the US Department of Energy, Office of Science, National Quantum Information Science Research Centers, Quantum

Systems Accelerator (QSA), and in part by the Defense Advanced Research Projects Agency under the Quantum Benchmarking contract. The computation was done using a high-powered computing cluster established through the support of the Gordon and Betty Moore Foundation's EPiQS Initiative, grant GBMF10436 to E.-A.K. The superconducting processor used in the random circuits study was made by the Google Quantum AI team who fabricated the processor, built the cryogenic and control systems, optimized the processor performance, and provided the tools that enabled the execution of this experiment. **Author contributions:** H.K., Y.D.L., E.-A.K., C.W., J.P.Z., and K.Q.W. designed the QuAN architecture. H.K., C.W., and J.P.Z. wrote the code for the QuAN and carried out its training. K.Q.W. guided training. Y.Z. led the exploration of the entanglement transition in the driven HCBH model. H.K. led the study of increasing complexity with the depth in random quantum circuits. Y.X. led the analysis and interpretation of the mixed-state topological order phase diagram. A.H.K., I.T.R., and W.D.O. conducted the driven HCBH model experiment and provided feedback on the machine learning outcome of experimental data. K.V. produced the simulation data for the random quantum circuit and calculated XEB. J.C.H. conducted the random quantum circuit experiment and provided feedback on the machine learning outcome of experimental data. Y.D.L., Y.Z., C.W., K.Q.W., and E.-A.K. initiated the project concept and guided the work. H.K., Y.Z., Y.X., and E.-A.K. wrote the paper with input and modifications from all authors. E.-A.K. led the project. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The data for this study have been deposited in the Zenodo database, available at https://doi.org/10.5281/zenodo.16805719.

Submitted 19 October 2024
Accepted 11 September 2025
Published 10 October 2025
10.1126/sciadv.adu0059