ARTICLE   OPEN

Check for updates

Benchmarking quantum logic operations relative to thresholds for fault tolerance

Akel Hashim [1,2,3 ✉], Stefan Seritan [4,5], Timothy Proctor [4,5], Kenneth Rudinger [4,5], Noah Goss [1], Ravi K. Naik [1,3], John Mark Kreikebaum [1,6,7], David I. Santiago [3] and Irfan Siddiqi [1,3,6]

Contemporary methods for benchmarking noisy quantum processors typically measure average error rates or process infidelities. However, thresholds for fault-tolerant quantum error correction are given in terms of worst-case error rates—defined via the diamond norm—which can differ from average error rates by orders of magnitude. One method for resolving this discrepancy is to randomize the physical implementation of quantum gates, using techniques like randomized compiling (RC). In this work, we use gate set tomography to perform precision characterization of a set of two-qubit logic gates to study RC on a superconducting quantum processor. We find that, under RC, gate errors are accurately described by a stochastic Pauli noise model without coherent errors, and that spatially correlated coherent errors and non-Markovian errors are strongly suppressed. We further show that the average and worst-case error rates are equal for randomly compiled gates, and measure a maximum worst-case error of 0.0197(3) for our gate set. Our results show that randomized benchmarks are a viable route to both verifying that a quantum processor's error rates are below a fault-tolerance threshold, and to bounding the failure rates of near-term algorithms, if—and only if—gates are implemented via randomization methods which tailor noise.

INTRODUCTION

Quantum bits (qubits) in the noisy intermediate-scale quantum (NISQ)[1] era are short-lived and susceptible to a variety of errors and noise due to imperfect control signals and imperfect isolation from the surrounding environment. Therefore, utilizing quantum computers to solve classically intractable problems (e.g., integer factoring[2]) will likely require quantum error correction (QEC)[3–7]. QEC can protect logical qubits from errors, but it is only guaranteed to work if the error rate of each physical qubit is below some fault tolerance (FT) threshold[8–13]. Analytic lower bounds on FT thresholds for various QEC codes have been derived, ranging from ∼10^{-6} for generic local noise[13] to ∼10^{-5}–10^{-3} for stochastic and depolarizing noise[14–18]. More optimistic estimates obtained via numerical simulation are orders of magnitude larger than the lower bounds, ranging from ∼10^{-3}–10^{-1}[19–27], but often assume stochastic (e.g., Pauli, dephasing, or depolarizing) noise models. While recent claims of quantum gates approaching or surpassing FT thresholds boast impressive gate fidelities[28–30], there is a discrepancy between these claims and the formulation of FT thresholds, which are specified in terms of worst-case error rates.

Various error metrics and measures exist for quantifying the "error rate" of a quantum gate. Randomized benchmarks[31–35] typically define error rates in terms of the average gate fidelity, or, equivalently, the process fidelity

$$F = \langle\psi|(\mathbb{I} \otimes \mathcal{E})(\rho)|\psi\rangle, \tag{1}$$

where $\rho = |\psi\rangle\langle\psi|$ is a maximally entangled state, $\mathcal{E}$ is the error channel associated with some quantum gate, and $\mathbb{I}$ the identity operation. A gate's process *infidelity* $e_F(\mathcal{E}) = 1 - F$ (or *average*

error rate) quantifies the probability that the gate induces an error on a random input state, or, equivalently, the average failure rate of random circuits that contain one instance of this gate but that are otherwise perfect. However, FT thresholds are typically defined via each gate's worst-case error rate (also called the diamond norm)[36],

$$\epsilon_\diamond(\mathcal{E}) = \frac{1}{2}\|\mathcal{E} - \mathbb{I}\|_\diamond = \frac{1}{2}\sup_\rho \|[\mathbb{I} \otimes (\mathcal{E} - \mathbb{I})](\rho)\|_1, \tag{2}$$

where the supremum is taken over all pure states and $\|X\|_1 = \mathrm{Tr}\sqrt{X^\dagger X}$ is the trace norm. Operationally, $\epsilon_\diamond(\mathcal{E})$ represents the worst-case performance of a quantum gate in any circuit, whereas $e_F(\mathcal{E})$ represents the average-case performance for a single instance of the gate. While the diamond norm is a pessimistic estimate of the error rate of a quantum gate, it provides much more rigorous performance guarantees in the context of QEC. This is because the diamond norm upper bounds the accumulation of error in any quantum circuit, since the distance between the ideal and actual output probability distributions (measured via total variation distance) for any circuit is bounded above by the sum of the worst-case error rates of all its gates[36].

While $e_F(\mathcal{E})$ can be measured directly via randomized benchmarks, there exists no known scalable method for measuring $\epsilon_\diamond(\mathcal{E})$. Tomographic methods, such as gate set tomography (GST)[37–41], can be used to estimate $\epsilon_\diamond(\mathcal{E})$[40], but they are exponentially expensive in the number of qubits. If only $e_F(\mathcal{E})$ is known, $\epsilon_\diamond(\mathcal{E})$ can be bounded using[42–45]

$$e_F(\mathcal{E}) \le \epsilon_\diamond(\mathcal{E}) \le \sqrt{e_F(\mathcal{E})}d, \tag{3}$$

[1]Quantum Nanoelectronics Laboratory, Department of Physics, University of California at Berkeley, Berkeley, CA 94720, USA. [2]Graduate Group in Applied Science and Technology, University of California at Berkeley, Berkeley, CA 94720, USA. [3]Computational Research Division, Lawrence Berkeley National Lab, Berkeley, CA 94720, USA. [4]Quantum Performance Laboratory, Sandia National Laboratories, Albuquerque, NM 87185, USA. [5]Quantum Performance Laboratory, Sandia National Laboratories, Livermore, CA 94550, USA. [6]Materials Sciences Division, Lawrence Berkeley National Lab, Berkeley, CA 94720, USA. [7]Present address: Google Quantum AI, Mountain View, CA, USA. ✉email: ahashim@berkeley.edu

where $d = 2^n$ (for n qubits). The lower bound of $\epsilon_\diamond(\mathcal{E})$ is saturated when $\mathcal{E}$ is a stochastic Pauli channel, and the upper bound, which is quadratically larger in e_F and scales with the dimension d, is saturated by a unitary channel. While modern experimental platforms routinely report single- and two-qubit infidelities on the order of $e_{F,1Q} \lesssim 10^{-4}$ and $e_{F,2Q} \lesssim 10^{-2}$[30,46–51], respectively, even if coherent errors account for a tiny fraction of the infidelity, they can dominate the diamond norm, in which case worst-case error rates can be as large as $\epsilon_{\diamond,1Q} \sim \sqrt{e_{F,1Q}} \lesssim 10^{-2}$ and $\epsilon_{\diamond,2Q} \sim \sqrt{e_{F,2Q}} \lesssim 10^{-1}$[43,44,52]. Therefore, $e_F(\mathcal{E})$ and $\epsilon_\diamond(\mathcal{E})$ can differ by orders of magnitude in the presence of coherent errors. This means that randomized benchmarks are generally inadequate for testing whether gate error rates are below FT thresholds[43,45]. Refs. [28–30] report single- and two-qubit error rates below the FT threshold for the surface code, but base their claims on average error rates from randomized benchmarking (RB) or GST, not diamond norms. Refs. [53,54] make similar claims about state-preparation and measurement (SPAM) errors, but only report average preparation and assignment fidelities, not their worst-case infidelities. Notably, Ref. [30] includes estimates of the diamond norm that are an order of magnitude larger than their reported process infidelities.

While it is not generally possible to directly compare $e_F(\mathcal{E})$ to a FT threshold for QEC, if it can be guaranteed that $e_F(\mathcal{E}) \approx \epsilon_\diamond(\mathcal{E})$ then randomized benchmarks can be used to efficiently verify that gate error rates are below a FT threshold. One method for ensuring that an error budget is dominated by stochastic noise is randomized compiling (RC)[52,55], which converts all gate errors into stochastic Pauli channels via Pauli twirling. This ensures that the direct measurement of $e_F(\mathcal{E})$ (e.g., via cycle benchmarking[56]) accurately captures the worst-case error rate, which enables the comparison to FT thresholds as well as bounding the overall failure rate of any quantum circuit or application.

In this work, we use GST to study RC performed on two qubits (labeled Q5 and Q6; see Methods) on a superconducting transmon processor (`AQT@LBNL Trailblazer8-v5.c2`). GST enables measurements of both the process infidelity and diamond norm

for all gates in our gate set, allowing us to study the impact of RC on gate errors. We find that RC eliminates signatures of coherent errors, enabling one to accurately describe the gates' errors by stochastic Pauli noise. We further show that RC suppresses spatially correlated coherent errors and non-Markovian errors. Finally, we show that the diamond norm converges to the process infidelity under RC, saturating the lower bound of Eq. (3), providing strong experimental evidence that our quantum logic operations are approaching or below a threshold for fault tolerance. By combining RC with GST, our results provide a novel framework for verifying that FT-required assumptions are satisfied, demonstrating that FT thresholds can be accurately measured using randomized benchmarks as long as quantum circuits are implemented using RC or related randomization methods[57–61].

RESULTS

Gate set tomography

Gate set tomography is a robust method for tomographically reconstructing the errors and noise impacting all gate operations within a gate set[41]. Like traditional quantum process tomography (QPT)[62], GST fully characterizes the process matrix of a quantum gate; however, unlike QPT, it does so in a self-consistent manner that simultaneously characterizes the errors in all the gates and in the state-preparation and measurement (SPAM). Using the open-source Python library `pyGSTi`[63,64], one can obtain a best-fit model for the gate set, consisting of a process matrix for each gate, using maximum-likelihood estimation[40,41]. In this work, our gate set consists of all possible combinations of I, $X_{\pi/2}$, and $Y_{\pi/2}$ single-qubit gates applied simultaneously to both qubits, as well as the controlled-Z (CZ) gate: $\mathbb{G} = \{G_1 \otimes G_2 : G_1, G_2 \in \{I, X_{\pi/2}, Y_{\pi/2}\}\} \cup \{CZ\}$. We utilized GST up to depth $L = 128$ layers to benchmark the performance of all gates in the gate set.

In this work, we represent gate errors using the Pauli transfer matrix (PTM) representation of superoperators (see Fig. 1a and Methods), denoted Λ. We fit our data to the following parameterized error models:
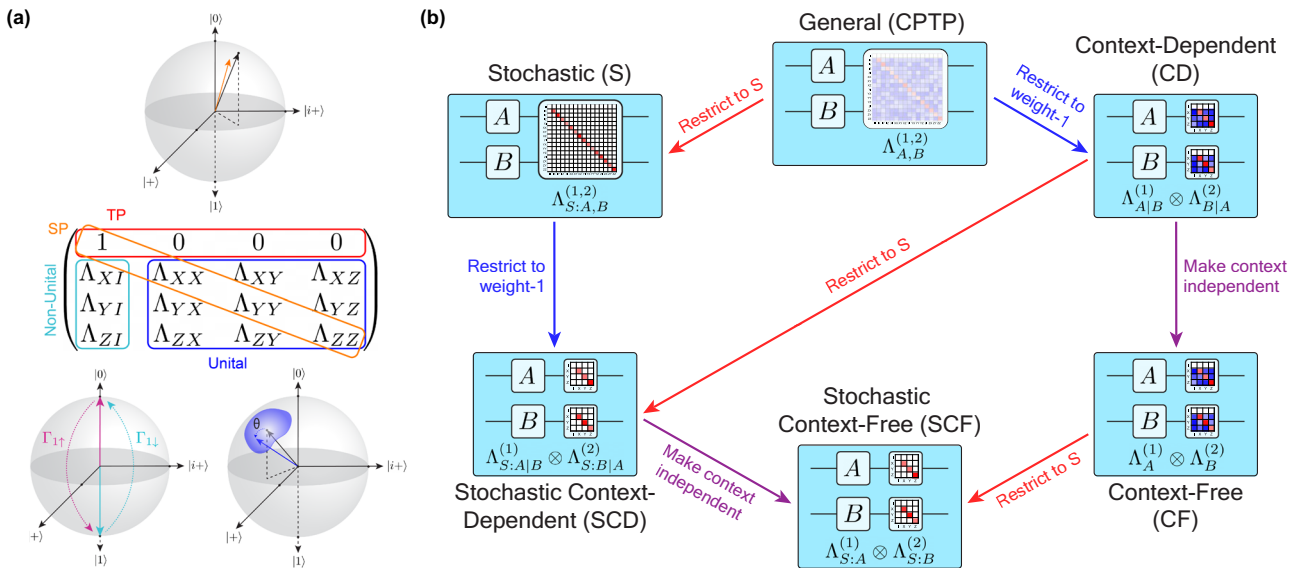

Fig. 1 Modeling gate errors in the Pauli transfer matrix (PTM) superoperator representation. (a) The PTM is divided into four blocks: the top row (red) represents trace-preservation (TP). The lower right-hand block (blue) captures unital processes, such as unitary errors. The leftmost column (cyan) captures non-unital processes, such as T_1 decay. And the diagonal of the PTM (orange) represents state preservation (SP); $\Lambda_{PP} = 1$ (< 1) if the Pauli channel P is (not) preserved under an error process such as stochastic Pauli noise. **(b)** Hierarchy of nested GST models. For single-qubit gates A and B acting on qubits 1 and 2, respectively, the general CPTP model contains all weight-1 and weight-2 error generators. The S model restricts the errors to be stochastic (i.e., the error generator is diagonal in the PTM) [red arrows]. The CD (SCD) model restricts the error generators to be weight-1 (stochastic) errors [blue arrows], but allows contextual dependence. The CF (SCF) model restricts the error generators to be context-independent weight-1 (stochastic) errors [purple arrows].

1. General completely positive and trace-preserving (CPTP) model: each gate's errors are modeled by a general CPTP two-qubit PTM $\Lambda_{A,B}^{(1,2)}$, where A (B) denotes the gate acting on qubit 1 (2).
2. General Pauli stochastic (S) model: the errors in the general model are restricted to be diagonal in the Pauli basis $\left(\Lambda_{S:A,B}^{(1,2)}\right)$.
3. Context-dependent (CD) model: the errors on each single-qubit gate is described by a single-qubit PTM that can depend on the gate acting on the other qubit $\left(\Lambda_{A|B}^{(1)} \otimes \Lambda_{B|A}^{(2)}\right)$.
4. Stochastic context-dependent (SCD) model: the errors in the CD model are restricted to be stochastic Pauli errors $\left(\Lambda_{S:A|B}^{(1)} \otimes \Lambda_{S:B|A}^{(2)}\right)$.
5. Context-free (CF) model: the errors on each single-qubit gate is described by a single-qubit PTM that is unconditional on the gate acting on the other qubit $\left(\Lambda_A^{(1)} \otimes \Lambda_B^{(2)}\right)$.
6. Stochastic context-free (SCF) model: the errors in the CF model are restricted to be stochastic Pauli errors $\left(\Lambda_{S:A}^{(1)} \otimes \Lambda_{S:B}^{(2)}\right)$.

The hierarchy of all of the nested GST models can be seen in Fig. 1b.

In general, models with greater complexity are able to capture more complex dynamics; however, they also require fitting a larger number of free parameters. The general CPTP model makes no assumption of locality and can completely capture all two-qubit interactions, including quantum crosstalk errors. The CD model assumes that errors on single-qubit gates must be local (weight-1), since the PTMs decompose as a tensor product of operations, but allows the errors to be classically correlated (i.e., the error impacting qubit 1 can be correlated with the gate being applied to qubit 2, but it cannot depend on the state of qubit 2). This can model errors due to classical crosstalk, but not entangling quantum crosstalk. The CF model assumes that the errors on

single-qubit gates are local and independent of the gate being applied to the other qubit. The corresponding S-type models make the same assumptions, but restrict the errors to be stochastic Pauli noise. All models make an assumption of Markovianity.

Randomized compiling

Randomized compiling is an efficient method for converting arbitrary Markovian errors into stochastic Pauli noise for each cycle (or layer) in a quantum circuit via Pauli twirling. This can benefit circuit performance in two different regimes (see Methods):

1. Single-randomization limit: a single randomization under RC interrupts the coherent accumulation of unitary errors between cycles of gates, similar to dynamical decoupling sequences[65].
2. Many-randomization limit: averaging over many (N) randomizations under RC tailors errors into stochastic Pauli channels, completely eliminating off-diagonal terms in the error process in the limit that $N \longrightarrow \infty$.

In this work, we apply RC to GST circuits using $N = 1$, $N = 10$, and $N = 100$ randomizations to study the transition of RC from the single- to many-randomization limit, and compare the results to GST performed without randomization (denoted by $N = 0$).

To study the types of errors present in a gate's process matrix G estimated using GST, we write a noisy quantum gate as

$$G = \Lambda G_0 = e^{\mathcal{L}} G_0, \qquad (4)$$

where G_0 is the ideal quantum gate, Λ the gate error channel, and $\mathcal{L}$ the gate error generator[66]. $\mathcal{L}$ is analogous to the Linbladian superoperator that generates all gate errors (coherent, stochastic, and non-unital).

Figure 2 shows the error generator $\mathcal{L}$ of the CZ gate for $N = 0, 1, 10, 100$ randomizations. We find that RC transforms the error generator from dense to sparse as we increase N, twirling $\mathcal{L}$ into a stochastic Pauli channel (these channels correspond to diagonal PTMs). Going from $N = 0$ (Fig. 2a) to $N = 1$ (Fig. 2b)
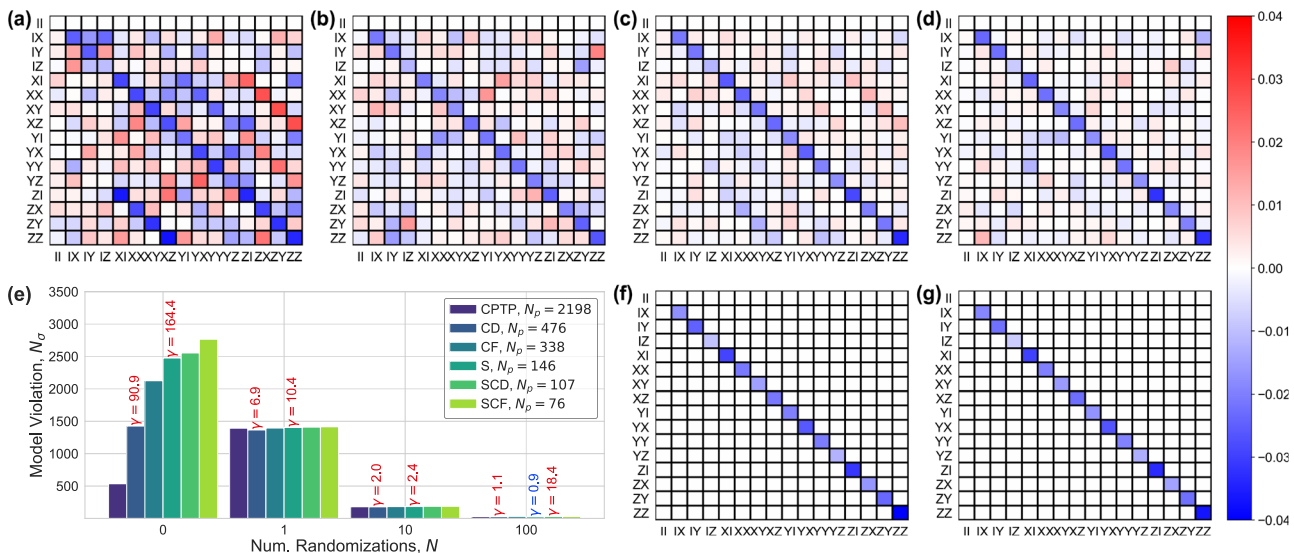

Fig. 2 Improving model accuracy via noise tailoring. The error generator $\mathcal{L}$ in the PTM representation for the CPTP model of the CZ gate is plotted for (a) $N = 0$, (b) $N = 1$, (c) $N = 10$, and (d) $N = 100$ randomizations under RC. If an error channel (i.e., PTM cell) in $\mathcal{L}$ is zero, then this component of the estimated gate matches the corresponding component of the ideal target gate. (e) The model violation N_σ for each GST model, plotted as a function of N. The evidence ratio γ is labeled above each nested model, and the number of free parameters N_p of each model is listed in the legend. For each N, we choose the model with the least number of parameters that satisfies $\gamma \lesssim 1$, finding that CPTP is the best fit for $N = 0$, $N = 1$, and $N = 10$, and S for $N = 100$; blue (red) text indicates that the model is accepted (rejected). The CZ gate error generator $\mathcal{L}$ is plotted for the S model for (f) $N = 10$ and (g) $N = 100$, showing that the stochastic models capture the dominant error for both. Even though it is ultimately rejected by the evidence ratio for $N = 10$, the S model could be reasonably selected for its simplicity. As the S model is selected for $N = 100$, all off-diagonal elements in (d) are statistically consistent with zero.

randomizations significantly reduces the magnitude of $\mathcal{L}$'s off-diagonal elements, which signifies the presence of coherent errors, but $\mathcal{L}$ for $N = 1$ still has significant magnitude in its off-diagonal elements. As N increases from 1 to 10, and from 10 to 100, the magnitude of the off-diagonal terms is greatly reduced. This is referred to as the "noise tailoring" property of RC.

Model Selection

GST enables the comparison of nested error models in a self-consistent manner[67,68]. To compare two nested models, we utilize the *evidence ratio* γ[67],

$$\gamma = \frac{\lambda_S - \lambda_L}{N_{p,L} - N_{p,S}}, \tag{5}$$

where λ is the model's log-likelihood ratio (see Methods), and L (S) denotes the larger (smaller) model defined in terms of the number of free parameters N_p describing the model. If $\gamma \lesssim 1$, then we automatically choose the smaller model, as it describes the data at least as well as the larger model without extra (potentially unused) parameters. If $\gamma > 1$, there is evidence for rejecting the smaller model, but even if $1 \lesssim \gamma \lesssim 10$ the smaller model may still be preferable due to its simplicity[51]. In this work, we compute the evidence ratio for each pair of nested models down the two separate branches of the hierarchy in Fig. 1b: if $\gamma \lesssim 1$ we select the smaller model and continue down the hierarchy until $\gamma > 1$, at which point we select the larger of the two models and stop. If two independent models satisfy this criteria, we select the model with the fewest free parameters, as this model represents the best fit estimate of our data without over-fitting.

Additionally, we compute the model violation N_σ of each model, which captures the number of standard deviations that λ is from the expected mean (see Methods). If $N_\sigma \lesssim 1$, then the GST model faithfully captures all of the errors in the device. However, on actual NISQ hardware, typically $N_\sigma > 1$ (and sometimes even $N_\sigma \gg 1$) is observed[40,60,68], indicating the presence of significant model violation. Because GST can model all Markovian gate errors, large N_σ for the CPTP model indicates that there is strong evidence in the data for non-Markovian errors—which can be true even if these non-Markovian errors are small in magnitude (see the Non-Markovian Errors section for further discussion).

To choose a best-fit model for each dataset, we compute γ and N_σ for each model, shown in Fig. 2e. We find that the general CPTP model fits best for $N = 0$, $N = 1$, and $N = 10$, and that the S model fits as well as the general CPTP model for $N = 100$. For $N = 10$, none of the nested models satisfy the evidence ratio criteria, but the S or CD model could also be reasonably selected, as $\gamma = 2.0$ and 2.4, respectively (there is therefore only weak evidence for coherent errors). For $N = 100$, only the S model satisfies the evidence ratio criteria, and we therefore prefer this model over the general CPTP model. For the data presented in the rest of this work, we fix the model for each dataset using the best-fit models outlined above.

The large model violation for $N = 0$ ($N_\sigma \approx 535$) is strong evidence that there are non-Markovian errors, which cannot be modeled by a PTM. $N = 10$ ($N_\sigma \approx 180$) and $N = 100$ ($N_\sigma \approx 25$) have significantly less model violation, indicating less statistical evidence of non-Markovian errors. For a single randomization ($N = 1$) there is even larger model violation ($N_\sigma \approx 1394$) than with no randomization ($N = 0$). This effect arises because when $N = 1$, each circuit is replaced with a single randomization of that circuit, and no averaging occurs. Therefore, the implementation of a logical gate (in $\mathbb{G}$) will not correspond to the average action of multiple PTMs (which decoheres errors) corresponding to the different Pauli-equivalent implementations of the gate, but instead the action of a particular PTM corresponding to one randomly selected Pauli-equivalent implementation of the gate. Importantly, the PTM describing the action of a logic gate will change depending on the

randomization selected (and therefore it varies between circuits and uses of the logical gate within a circuit). Therefore, even when each physical gate is describable by a PTM (i.e., all errors are Markovian), when $N = 1$ each logical gate is not generally describable by a single PTM. This violates the assumptions of GST—which finds the best-fit PTM for each gate—and so a GST model fit to the $N = 1$ data can exhibit significant model violation. (See Methods for further discussion and examples.)

Impact of randomized compiling on error budgets

To explore how effective RC is at converting all errors into stochastic Pauli noise, we calculate the total amount of stochastic and coherent errors for each gate. To do so, we divide each gate's error generator $\mathcal{L}$ into stochastic and Hamiltonian components, and compute the *total error*,

$$\epsilon_{\text{tot}} = \epsilon_{\text{agg}} + \theta_{\text{agg}}, \tag{6}$$

where $\epsilon_{\text{agg}} = \sum_i \epsilon_i$ is the sum of the rates of all stochastic error generators, and $\theta_{\text{agg}} = \sqrt{\sum_i \theta_i^2}$ is the quadrature sum of all Hamiltonian error generators[51]. The total error is closely related to the diamond norm. For single qubit error generators $\mathcal{L}$, the total error upper-bounds the diamond norm error, $\epsilon_\diamond(e^{\mathcal{L}}) \leq \epsilon_{\text{tot}}(e^{\mathcal{L}})$. For two or more qubits, the bound is much weaker[51].

In Fig. 3 we plot the fraction of the total error due to stochastic and coherent (Hamiltonian) errors. We find that the error budget of the simultaneous single-qubit gates is dominated by coherent errors for $N = 0$, and that coherent errors account for approximately two-thirds of the total error for the CZ gate. For $N = 1$, coherent errors still dominate the single-qubit gates, but the contribution from coherent errors and stochastic noise are both approaching 50% for the *CZ gate*. By $N = 10$, stochastic noise makes up the largest contribution to the total error for the CZ gate, but we observe only a modest change for the single-qubit gates. However, by $N = 100$ the error budget for all gates is entirely due to stochastic noise. We note that for the $N = 100$ data, our model selection chose the S model, which enforces the constraint that the error budget is entirely due to stochastic Pauli noise (see the diagram in Fig. 1b). This is because there is no statistically significant evidence in the data for any coherent errors (if more data were taken, evidence for some residual coherent errors might be found). These data demonstrate the effectiveness of RC in eliminating the impact of coherent errors. However, this does not mean that coherent errors are physically not present; rather, each individual randomization under RC is impacted by coherent errors in a different manner, in such a way that the aggregate effect is to randomize the impact of coherent errors.

Alternative methods for quantifying the magnitude of coherent errors include purity RB (PRB)[69] and cross-entropy benchmarking (XEB)[70]. However, methods for measuring coherent errors that use random circuits have two disadvantages. First, these methods measure averages over gates, so they cannot separate out each gate's error rate into stochastic and coherent contributions (as done herein). Second, although the unitarity (estimated by PRB) and infidelity (estimated by RB) can be used to upper bound the diamond norm error[69], this method is inefficient (see Ref. [40] for details).

Process infidelity vs. diamond norm

Error rates measured via randomized benchmarks cannot in general be directly compared to FT thresholds. However, although the diamond norm [Eq. (2)] enables rigorous comparison to FT thresholds, it is difficult to measure in a scalable manner. If the error model of a quantum processor is dominated by stochastic noise, the process infidelity and diamond norm will be equal, in which case randomized benchmarks can be used to efficiently demonstrate that gate errors are below a FT threshold.
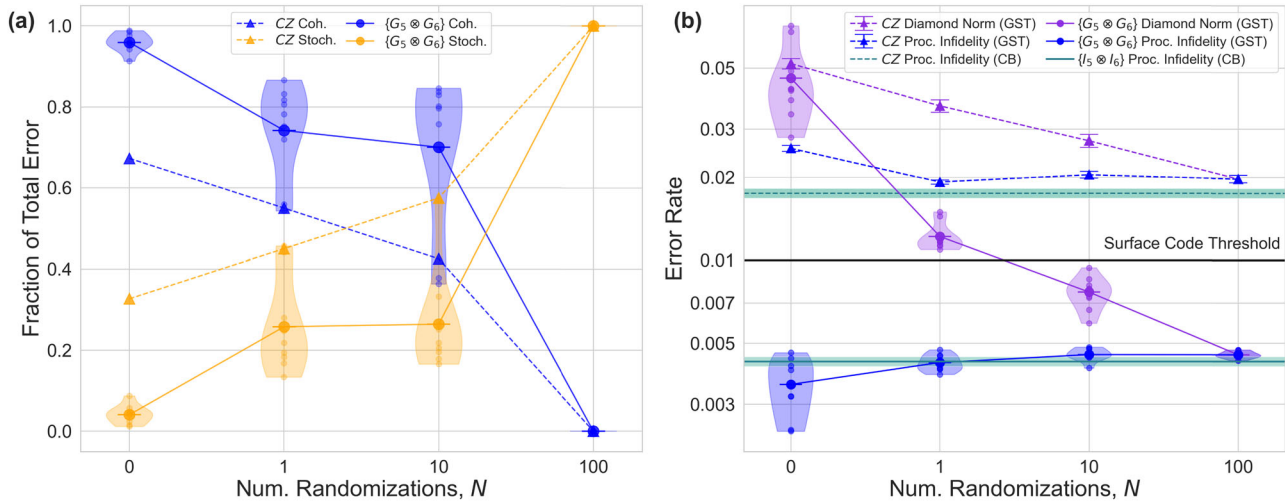
Fig. 3 Tailoring noise for improved estimates of the diamond norm. (**a**) Fraction of the total error due to coherent errors (blue) and stochastic noise (orange) as a function of the number of randomizations N under RC. As N increases, stochastic noise makes up a larger fraction of the total error, which is dominated by coherent errors for small N. For (**a**) and (**b**), triangular markers denote the CZ gate, and circular markers denote the single-qubit gates $\{G_5 \otimes G_6\}$ (small transparent markers depict the individual gates, large markers depict the averages, and violin plots outline the distribution). (**b**) Saturation of the lower bound of the diamond norm under RC. We plot the GST process infidelity e_F (blue) and diamond norm $\epsilon_\diamond$ (purple) as a function of the number of randomizations N. We observe that $\epsilon_\diamond > e_F$ for all gates for $N = 0, 1, 10$, but that the two are equal for $N = 100$. For visual clarity, we omit the idle cycle for $N = 0$, as $e_{F,II} = 0.0001$ and $\epsilon_{\diamond,II} = 0.001$ fall well below their respective averages. We compare the GST estimates with the process infidelity measured via cycle benchmarking (CB) for the CZ gate (dashed green line) and idle cycle $\{I_5 \otimes I_6\}$ (solid green line). The GST error bars and CB transparent bands indicate the 95% confidence intervals. Additionally, we compare the benchmarked error rates with the 1% FT threshold for the surface code (black line) and find that the single qubit gates are well below the threshold value, whereas the CZ gate is approaching, but does not surpass the threshold.

In Fig. 3, we plot the process infidelity and the diamond norm as a function of N for all gates in our gate set. We find that $\epsilon_\diamond(\mathcal{E})$ converges to $e_F(\mathcal{E})$ as N increases. This is strong experimental evidence that $\epsilon_\diamond(\mathcal{E}) = e_F(\mathcal{E})$ in the many-randomization limit for $N = 100$ [saturating the lower bound of Eq. (3)]. We also compare these results with process infidelities measured independently via CB and find good agreement between the two. These results demonstrate that randomized benchmarks are sufficient for benchmarking FT thresholds if—and only if—a quantum application is impacted only by stochastic noise, which can be guaranteed if implemented using methods which tailor noise. We note that similar results were previously reported using Pauli frame randomization[71,72] for single-qubit gates[60].

The largest diamond norm error over a gate set is arguably the most relevant quantity to compare to a FT threshold. The surface code[73] is a popular QEC code due to its high FT threshold, which is estimated to be between ~0.75−3%[22,74−78], with 1% being the threshold that is typically quoted in the literature[28,30]. In Fig. 3, we find that the diamond norm of simultaneous single-qubit gates is below the surface code threshold for $N = 10$ and $N = 100$, and that the error rate of our CZ gate is approaching—but does not surpass—the surface code threshold.

Correlated errors under Pauli twirling

A major requirement for reliable fault-tolerant QEC is the absence of correlated errors, which can occur temporally[79] or spatially[80]. Many-qubit correlated errors cannot be corrected by QEC (each QEC scheme has a maximal weight of error that it can correct), causing logical failures. Therefore, the rate of correlated errors must be low to achieve reliable fault-tolerant quantum computation. To characterize the extent to which spatially correlated errors are present in our system with and without Pauli twirling, we extract the weight-1 and weight-2 errors of each gate in our gate set. Figure 4 shows the weight-1 and weight-2 coherent (stochastic) errors for the idle cycle $\{I_5 \otimes I_6\}$ for the CPTP (S) model for $N = 0$ ($N = 100$) randomizations. We focus on coherent (stochastic) errors for $N = 0$ ($N = 100$) as they dominate the total

the error budget; see Fig. 3a. We observe significant weight-2 coherent errors for $N = 0$ (Fig. 4a), which corresponds to unintended entanglement (i.e., quantum crosstalk), such as static ZZ coupling[81−83]. In contrast, we observe that weight-1 Pauli errors dominate for $N = 100$, and that weight-2 errors are largely suppressed in comparison.

Additionally, we compare the Pauli error rates for $N = 100$ (Fig. 4b) to independently estimated Pauli error rates (Fig. 4c) measured using cycle benchmarking (CB)[56] and reconstructed using cycle error reconstruction (CER)[55,84,85]. We find good agreement between the reconstructed Pauli noise in the two error maps, showing similar magnitudes of correlated errors, and demonstrating that the dominant errors in our system are weight-1. These results show that that Pauli twirling can suppress spatially correlated coherent errors due to entangling crosstalk and static ZZ coupling between superconducting qubits. (See Methods for a similar CER analysis of correlated errors in cycles acting on four qubits.)

In general, one can expect RC to provide a quadratic suppression of correlated coherent errors—RC converts a coherent error that contributes $\mathcal{O}(\theta)$ to the diamond norm (and the total error) into a stochastic error that contributes $\mathcal{O}(\theta^2)$ to the diamond norm. Therefore, while correlated errors might not be entirely suppressed by Pauli twirling, this reduction in the magnitude of correlated errors may reduce the cost of QEC[86]. While the current work focuses on spatial correlations, temporal correlations can also be detrimental to worst-case error rates[87]. However, we expect that a single randomization under RC can similarly suppress temporally correlated errors, although we leave an exploration of this topic to future work.

Non-Markovian errors

In the context of quantum computing, in particular for characterization and benchmarking, an error process is typically considered to be non-Markovian if it cannot be modeled by a process matrix. More precisely, non-Markovian errors are present if each n-qubit cycle (or layer) of gates cannot be modeled by a fixed, context-

independent n-qubit process matrix. Various common sources of non-Markovianity in NISQ systems include leakage out of the computation basis states[88–93], unwanted entangling interactions (e.g., static ZZ coupling) with qubits outside of the studied system, drift in qubit parameters[94] (e.g., stochastic fluctuations in transition frequencies), unwanted coupling to environmental systems with memory beyond the timescale of a cycle (e.g., two-level fluctuators and nonequilibrium quasiparticles[80,95–97]), and qubit heating[98]; see Fig. 5a. Studying and suppressing non-Markovian errors is important for at least two reasons: they interfere with the quantification of Markovian errors, and their impacts on QEC are less well-understood.

GST is designed to reconstruct all possible Markovian errors on any cycle of quantum gates. So, when GST cannot fit the data, this implies that there are non-Markovian errors. While model violation N_σ is useful for providing evidence of the existence of non-

Markovian errors, to quantify the magnitude of such errors, we add a *wildcard error model*[99] to each of our GST models. These wildcard error models assign a wildcard error rate $w_G \in [0, 1]$ to each gate G, and a wildcard error rate to the SPAM $w_{\text{SPAM}} \in [0, 1]$, which quantifies how much additional error on each operation is missing from the model (i.e., how much is required to make the model consistent with the data; see Methods for more details). By comparing w_G to $\epsilon_\diamond$ for each gate, we are able to quantify whether Markovian or unmodeled non-Markovian errors dominate the error model[68,99]. If $w_G \ll \epsilon_\diamond$ for all gates, then Markovian errors dominate and non-Markovian errors are negligible. In this case, the model captures the majority of the errors in the gate (or, more precisely, all those errors that were revealed in this experiment), despite the fact that there is evidence that the model is incomplete. On the other hand, if $w_G \geq \epsilon_\diamond$, the non-Markovian
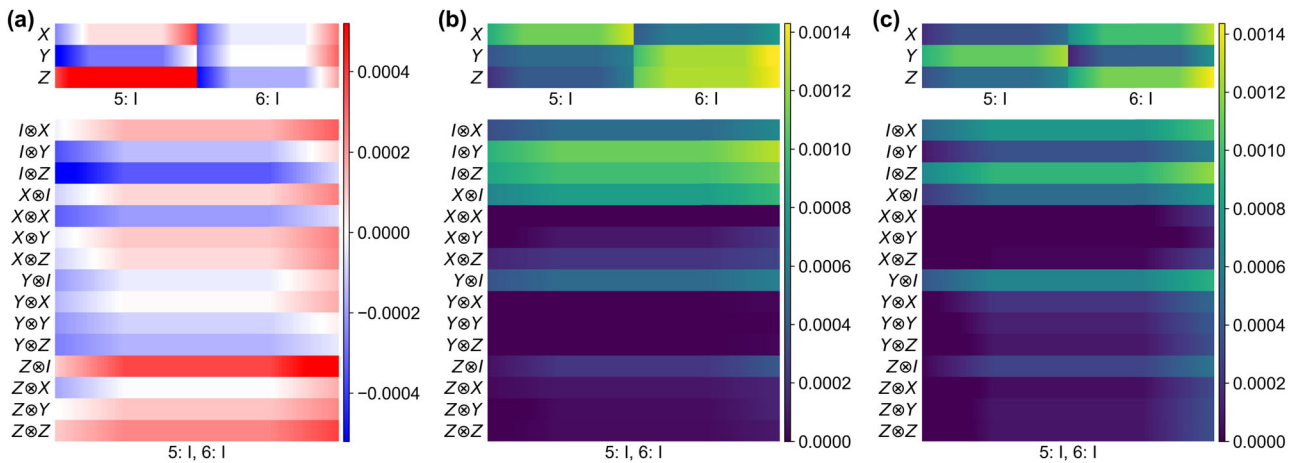

Fig. 4 Suppressing correlated errors. Heat maps of the weight-1 and weight-2 errors acting on Q5 and Q6 during the idle cycle $\{I_5 \otimes I_6\}$ reconstructed using (**a**) the Hamiltonian projection of the GST error generator $\mathcal{L}$ for $N = 0$, (**b**) error rates calculated from the stochastic projection of $\mathcal{L}$ for $N = 100$, and (**c**) cycle error reconstructing via CB measurements. The x-axis labels the target gate; the y-axis labels the Hamiltonian error for (**a**), and the Pauli Kraus error for (**b**) and (**c**); the cell color denotes the over-rotation angle for (**a**), and error rate for (**b**) and (**c**); the cell gradient defines the 95% confidence interval; the first (second) row of subplots shows marginalized weight-1 (correlated weight-1 and weight-2) errors. While weight-2 errors are dominant for $N = 0$, (**b**) and (**c**) show that Pauli twirling suppresses weight-2 errors to negligible levels.
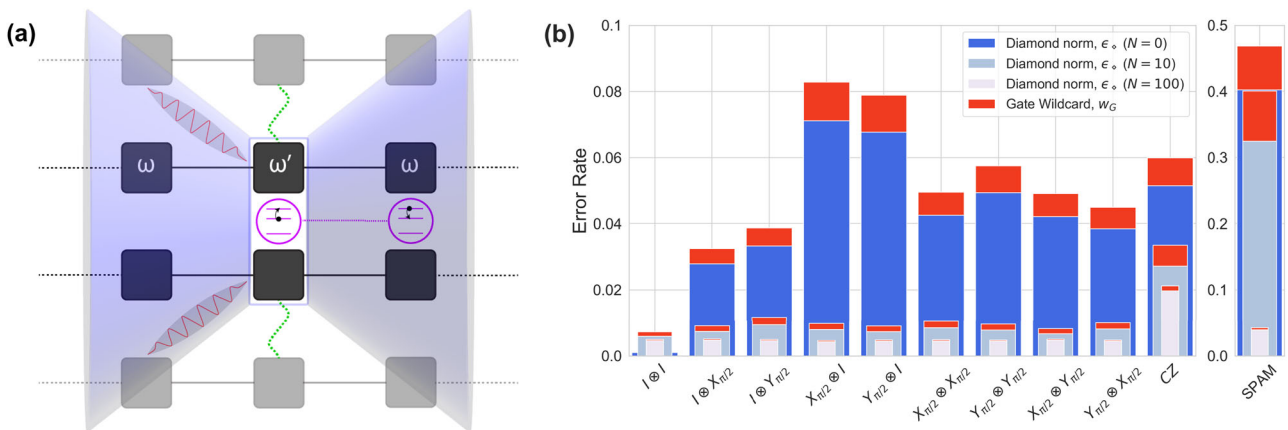

Fig. 5 Suppressing non-Markovian errors. (**a**) Non-Markovian errors in gate-based quantum computing. Markovian errors for a given system of qubits (black) are defined to occur within the timescale of a given cycle of gates (blue rectangle). Sources of non-Markovianity include leakage (seapage) out of (into) the computational basis states (purple), unintended entanglement (green) with external qubits (gray) or two-level fluctuators in the environment, drift in the system properties (e.g., fluctuations in the qubit frequency ω, with $\omega' = \omega + \delta\omega$), and classical EM signals from outside of the defined system (red) that reach the system qubits within their pseudo light cone (blue). (**b**) Unmodeled error versus the diamond norm. The per-gate wildcard w_G is plotted on top of the diamond norm for each gate in $\mathbb{G}$ (plus SPAM) as a function of N. As N increases, w_G becomes negligible, indicating that non-Markovian errors are suppressed under RC.

errors dominate the Markovian errors, and the so the GST estimate is unreliable.

Figure 5 b shows the wildcard error and the diamond norm error for all of the gates in our gate set and SPAM, for $N = 0, 10, 100$ (we omit the $N = 1$ data due to its systematic inconsistency with a Markovian error model; see Model Selection and Methods). Without RC ($N = 0$), we observe diamond norm errors as large as $\epsilon_\diamond \approx 0.07$, with up to 0.01 additional non-Markovian error. For $N = 10$ randomizations, the wildcard errors are still significant, but they are much smaller in magnitude than $N = 0$ *and* they are a small fraction of the diamond norm error rates. By $N = 100$ randomizations, the wildcard errors are negligible—in absolute terms and as a fraction of $\epsilon_\diamond$—contributing at most 0.0012 additional error per gate. This indicates that the S model accurately captures almost all of the $N = 100$ data (note that in this case the wildcard error quantifies the combined contribution of both non-Markovian errors and all non-S Markovian errors). Because $w_G \ll \epsilon_\diamond \forall G \in \mathbb{G}$ for all N, we consider all of our models trustworthy, even for $N = 0$. We additionally note that RC significantly improves both the wildcard error and worst-case error rate for SPAM, which is the sum of the trace distance for the state-preparation and the diamond norm for the positive operator-valued measure. However, even in the $N = 100$ case, the SPAM error rate remains around $\sim 4\%$, well above the 1% surface code threshold; these errors are equally as important as gate errors in the context of fault-tolerant QEC, and will need to be dramatically improved for achievable QEC in the future.

The $N = 100$ data was gathered over a period of over 40 hours, and no re-calibration of gates was performed during the experiment. Therefore, the negligible amount of unmodeled error speaks to the robustness of RC to the inevitable drift in gate and qubit parameters during this time period. While RC was not designed to specifically target non-Markovian errors, its apparent robustness to both non-Markovian errors and correlated errors is promising for future large-scale fault-tolerant applications.

DISCUSSION

The field of quantum characterization, verification, and validation (QCVV) was developed in part to benchmark our progress toward fault-tolerant QEC. While full-scale fault-tolerant QEC is still many years away, contemporary quantum gates are rapidly approaching the necessary thresholds for many QEC codes, such as the surface code. Therefore, it is important to be able to accurately benchmark our progress toward this goal. While average error rates measured via randomized benchmarks are useful for tracking progress, they fall short of capturing the information required to determine whether all gate errors fall below a given FT threshold, for which the diamond norm is the relevant metric[43,45]. In this work, we demonstrate that FT thresholds can in fact be captured by randomized benchmarks, but only if the error-corrected application is performed using a randomization method which tailors noise. In fact, utilizing artificial randomness to suppress coherent dynamics is not new, and has been previously discussed for disorder-assisted error correction methods[100–102]. While our results were measured in the many-shot, many-randomization limit, they are still relevant for error correction, which operates in the single-shot, single-randomization limit, because results measured in latter limit sample from the same distribution that is estimated in the former limit; in other words, results measured in the single-shot, single-randomization limit are indistinguishable from those sampled from a distribution impacted by only stochastic noise.

The current approach also helps resolve a discrepancy regarding logical error rates in the Pauli twirling approximation (PTA). While some have argued that logical noise is well approximated by Pauli errors for large-distance codes[103], others have shown that the PTA is a dishonest approximation of the

impact of coherent errors[104–108] and breaks down for large numbers of qubits[109], and that despite the projective nature of QEC, coherent errors at the physical level can lead to coherent errors at the logical level[110–113], which can increase logical error rates[114]. Our results show that Pauli twirling can efficiently be implemented at the physical level, which may even impact the kinds of errors that manifest at the logical level, potentially negating any need to *approximate* the accuracy of Pauli twirling at the logical level. Moreover, while coherent errors may continue persist even at the logical level, recent studies suggest that the diamond norm of gates at the logical level is smaller than the diamond norm of gates at the physical level[115], which bodes favorably for the future of QEC. An exploration of this topic would be a natural extension of this work, and would provide insight into the necessity and/or potential benefits of using randomization methods for QEC.

While rapid improvements in two-qubit gates on many hardware platforms engender optimism for FT, caution must be taken in the claims inferred from gate fidelities, as results which include estimates of the diamond norm suggest that many contemporary two-qubit gates fall short of FT thresholds[30,51,116–118]. Furthermore, while isolated single- and two-qubit gates may be approaching the necessary requirements for fault-tolerance, any gate(s) performed in parallel with other qubits are likely to be impacted by crosstalk-induced coherent errors, potentially causing the diamond norm to scale with $\sqrt{e_F}$. The figure of merit for determining whether low logical error rates can be achieved via QEC is the error rate of a cycle containing all active qubits in a register, not simply the error rate of isolated gates within the cycle[86]. Thus, whether or not randomization methods such as RC will be effective at the scale of the large number of qubits needed for QEC is an open question. However, in theory the efficiency of noise tailoring via Pauli twirling does not depending on system size[119]. Therefore, future work could explore the efficacy of RC at saturating the lower bound of the diamond norm for larger cycles of simultaneous gate operations.

METHODS
Experiment

All experimental gate set tomography (GST) results were measured in a manner which normalizes shot statistics between circuits implemented with and without randomized compiling (RC). Bare GST sequences were measured $K = 1000$ times. For GST sequences implemented with N randomizations under RC, we fix $K = 1000$ and measure each randomization K/N times. By computing the union over the N different logically equivalent randomizations of each GST sequence, one obtains an equivalent statistical distribution for a single GST circuit measured K times.

Qubit parameters

Table 1 lists the relevant qubit parameters for the two transmon qubits used in this work. Qubit frequencies and anharmonicities are measured using Ramsey spectroscopy. Relaxation (T_1) and coherence (T_2^* and T_2^{echo}) times are extracted by fitting exponential

Table 1. Single-qubit parameters.

	Q5	Q6
Qubit freq. (GHz)	5.331004	5.490952
Anharm. (MHz)	-275	-271.35
T_1 (μs)	62(5)	52(4)
T_2^* (μs)	37(6)	36(6)
T_2^{echo} (μs)	73(7)	68(7)

decay curves to the excited state lifetime and Ramsey spectroscopy measurements (without and with an echo pulse), respectively. The reader is referred to Ref. [55] for details of the sample `AQT@LBNL Trailblazer8-v5.c2` and the experimental setup of the device, as well as Refs. [50,61,68,120] for the previous characterization of the qubits used in this experiment.

Representations of quantum processes

Various representations of quantum processes exist[39]. A common representation is the operator-sum, or Kraus, representation which maps any density matrix ρ into

$$\mathcal{E}(\rho) = \sum_i K_i \rho K_i^\dagger, \tag{7}$$

where the set $\{K_i\}$ is a general class of operators known as Kraus operators. For a Pauli channel,

$$\mathcal{E}(\rho) = \sum_{P \in \mathbb{P}^{\otimes n}} c_P P \rho P^\dagger \tag{8}$$

where the Kraus operators are $K = \sqrt{c_P}P$, with c_P the probability that P is applied (which is an error, except in the case of the identity Pauli), and $\mathbb{P}^{\otimes n} = \{I, X, Y, Z\}^{\otimes n}$ is the set of 4^n generalized Pauli operators.

Another useful representation of a quantum operator is the Pauli transfer matrix (PTM) representation, which is used throughout this work. The PTM representation can be defined by expanding any density matrix ρ in the Pauli basis,

$$\rho = \sum_{P \in \mathbb{P}^{\otimes n}} a_P P, \tag{9}$$

where a_P are the expansion coefficients. By vectorizing the expansion coefficients, we obtain a vectorized density matrix $|\rho\rangle\rangle = (a_{I^{\otimes n}} \quad \ldots \quad a_{Z^{\otimes n}})^T$. A linear completely positive and trace-preserving (CPTP) quantum map $\rho \mapsto \rho' = \Lambda(\rho)$ in the PTM representation is completely defined by a $4^n \times 4^n$ matrix Λ, with values $\Lambda_{ij} = \mathrm{Tr}[P_i \mathcal{E}(P_j)]/d$ that can derived directly from the Kraus representation [Eq. (8)] for a Hilbert space of dimension $d = 2^n$. All entries in the PTM are real and are bounded by $\Lambda_{ij} \in [-1, 1]$. PTMs have the useful property that the composite map (i.e. the PTM of a quantum circuit) can be constructed by taking the matrix products of the individual maps (i.e. the PTMs of the individual gates).

For a single qubit,

$$|\rho\rangle\rangle = \begin{pmatrix} a_I \\ a_X \\ a_Y \\ a_Z \end{pmatrix}. \tag{10}$$

The quantum map $\rho' = \Lambda(\rho)$ can be expressed in vector form, $|\rho'\rangle\rangle = \Lambda|\rho\rangle\rangle$, or more explicitly,

$$\begin{pmatrix} a_I' \\ a_X' \\ a_Y' \\ a_Z' \end{pmatrix} = \begin{pmatrix} \Lambda_{II} & \Lambda_{IX} & \Lambda_{IY} & \Lambda_{IZ} \\ \Lambda_{XI} & \Lambda_{XX} & \Lambda_{XY} & \Lambda_{XZ} \\ \Lambda_{YI} & \Lambda_{YX} & \Lambda_{YY} & \Lambda_{YZ} \\ \Lambda_{ZI} & \Lambda_{ZX} & \Lambda_{ZY} & \Lambda_{ZZ} \end{pmatrix} \begin{pmatrix} a_I \\ a_X \\ a_Y \\ a_Z \end{pmatrix}. \tag{11}$$

As shown in Fig. 1a, some important properties of a process can be easily extracted from the components of its PTM. The PTM can be divided into four blocks: the upper left-hand corner represents trace-preservation, with $\Lambda_{II} = 1$ if a process is trace-preserving (TP). This constraint can be succinctly summarized by stating that a process is TP if $\Lambda_{0j} = \delta_{0j}$ (i.e. the first row of the PTM is [1, 0, ..., 0]). The lower right-hand block is the unital block, which captures processes such as stochastic Pauli noise and unitary errors. A unital process is one that maps the identity operation to the identity operation $\Lambda(\mathbb{I}) = \mathbb{I}$ (it cannot purify a mixed state). The row above the unital block captures state-dependent leakage, represented by $\Lambda_{IP} \neq 0$ for $P \in \{X, Y, Z\}$; leakage is therefore not TP. The column to the left of the unital block

is the non-unital block, which captures processes such as spontaneous emission (i.e., T_1 decay) or amplitude damping. This constraint can be summarized by stating that a process is unital if $\Lambda_{i0} = \delta_{i0}$ (i.e. the first column of the PTM is $[1, 0, \ldots, 0]^T$). Finally, the diagonal of Λ represents state-preservation, with $\Lambda_{PP} = 1$ (< 1) for processes which (do not) preserve the Pauli channel $P \forall \{I, X, Y, Z\}$.

The process fidelity of a map in the PTM representation is the weighted average of the diagonal components,

$$F_\Lambda = \frac{1}{4^n} \sum_{P \in \mathbb{P}^{\otimes n}} \Lambda_{PP}. \tag{12}$$

Therefore, the process infidelity is given by $e_F = 1 - F_\Lambda$.

Gate set tomography

In GST, the quality of the model is quantified by computing the log-likelihood ratio λ of the likelihood $\mathcal{L}$ of the GST model with the likelihood $\mathcal{L}_{\max}$ of the "maximal model",

$$\lambda = -2 \ln\left(\frac{\mathcal{L}}{\mathcal{L}_{\max}}\right). \tag{13}$$

The maximal model is the one in which each independent measurement outcome in the data set is assigned a distinct probability equal to the observed frequencies. Wilks' theorem[121] states that if $\mathcal{L}$ and $\mathcal{L}_{\max}$ are both valid models, then the log-likelihood ratio is a χ_k^2 random variable, where $k = N_{\max} - N_p$ is the difference in the number of free parameters between the maximal and GST models. If λ is not consistent with χ_k^2 distribution (i.e. it does not lie within the interval $[k - \sqrt{2k}, k + \sqrt{2k}]$, with mean k and standard deviation $\sqrt{k}$), then this indicates that the GST data are inconsistent with the GST model.

We quantify the model violation, or "goodness of fit," by the number of standard deviations that λ is from the expected mean k,

$$N_\sigma = \frac{\lambda - k}{\sqrt{2k}}. \tag{14}$$

If $N_\sigma \leq 1$, then the GST model can be considered trustworthy and faithfully captures the behavior of the device. GST makes an assumption of Markovianity (i.e. any Markovian process can—by definition—be captured in a generalized model based on process matrices), therefore large N_σ indicates the presence of non-Markovian errors. However, N_σ does not quantify the magnitude of these non-Markovian errors. Because N_σ will grow linearly with the number of shots (and will typically increase with the depth of the GST circuits), a large N_σ can be observed even if the underlying non-Markovian errors are small in magnitude. Therefore, large N_σ simply indicates that some non-Markovianity is present with high statistical certainty.

While N_σ provides statistical evidence of non-Markovianity, a wildcard error model[99] can capture the magnitude of these errors. The wildcard error rate $w_G \in [0, 1]$ quantifies the unmodeled error per logic gate operation. A wildcard error can also be assigned for a circuit C containing gates by summing over the wildcard error rates for all gates $G \in C$: $w_C = \sum_{G \in C} w_G$. The wildcard model is chosen to be minimal, such that assigning w_G to a gate G is just sufficient to make the model's predictions consistent with the observed data. This is enforced by requiring that the total variation distance (TVD)

$$D_{\mathrm{TV}}(p, q) = \frac{1}{2}||p - q||_1 \tag{15}$$

between the observed probability distribution p_C and the wildcard-augmented probability distribution q_C be bounded by the total wildcard error for circuit C,

$$D_{\mathrm{TV}}(p_C, q_C) \leq w_C. \tag{16}$$

The wildcard-augmented model is therefore not unique, as q_C can be chosen from any distribution that satisfies Eq. (16). Because the

wildcard error quantifies the magnitude of unmodeled error per gate, and because unmodeled errors are often attributed to non-Markovianity in the system, the per-gate wildcard error budget is a good estimate of the magnitude of non-Markovian errors impacting the gate. The TVD is a useful metric for quantifying the magnitude of unmodeled errors because it captures the rate at which measurement outcomes are incorrectly predicted by a model. Because the TVD is upper-bounded by the diamond norm[36], one can compare the wildcard error w_G to the diamond error $\epsilon_\diamond$ for any gate G to inform whether unmodeled errors in the GST estimate are dominant or negligible, and thus whether the GST model can be trusted. By extension, comparing w_G to $\epsilon_\diamond$ quantifies whether Markovian or non-Markovian errors dominate the total error.

Randomization compiling: single-randomization limit

To understand how RC suppresses the coherent accumulation of unitary errors between cycles of gates, consider the simple example of applying many $R_x(2\pi)$ rotations to a qubit in the ground state, but each time the qubit over-rotates by a small angle θ. The resulting state of the qubit after M rotations is

$$|\psi\rangle = \prod^M e^{-i\theta\sigma_x}|0\rangle = \cos(M\theta)|0\rangle - i\sin(M\theta)|1\rangle. \quad (17)$$

The fidelity of this state with respect to $|0\rangle$ is $\mathcal{F} = |\langle 0|\psi\rangle|^2 = \cos^2(M\theta) \approx 1 - (M\theta)^2$, thus the infidelity $r = 1 - \mathcal{F} \approx (M\theta)^2$. Therefore, the infidelity scales quadratically in both the over-rotation angle θ and the number of rotations M. Under RC, the trajectory from the initial state to the final state is randomized, thus ensuring that coherent errors will not grow quadratically between gates. While exact quadratic growth is highly unlikely for longer-depth multi-qubit circuits due to the complex dynamics of crosstalk, and because coherent errors can act in any direction (not just along the axis of rotation), coherent errors can still accumulate in an adversarial fashion and grow faster than average error rates, especially in structured quantum circuits[122].

Randomization compiling: many-randomization limit

Regardless of the rate at which coherent errors accumulate between cycles of gates, they can still impact each computational gate G in a circuit. We can model this process as an ideal gate G_0 followed by an unwanted unitary operator $G = U(\hat{\mathbf{n}}, \theta)G_0$, where $U(\hat{\mathbf{n}}, \theta) = e^{-i\theta\hat{\mathbf{n}}\cdot\boldsymbol{\sigma}/2}$, $\hat{\mathbf{n}}$ is the axis of rotation, $\boldsymbol{\sigma}$ the Pauli vector, and θ is the rotation angle relative to the intended target state. For simplicity, consider a unitary error about the x-axis for a single qubit,

$$U(x, \theta) = \exp\left(-i\frac{\theta}{2}\sigma_x\right)$$
$$= \begin{pmatrix} \cos(\theta/2) & -i\sin(\theta/2) \\ i\sin(\theta/2) & \cos(\theta/2) \end{pmatrix}. \quad (18)$$

In the PTM representation (see Fig. 1a), this coherent error takes the following form,

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos(\theta) & -\sin(\theta) \\ 0 & 0 & \sin(\theta) & \cos(\theta) \end{pmatrix}. \quad (19)$$

For small θ, the diagonal components of Λ scale as $\cos(\theta) \approx 1 - \frac{1}{2}\theta^2$, and the off-diagonal terms scale as $\sin(\theta) \approx \theta$. While the infidelity of the diagonal terms is $e_F \approx \theta^2$, we see that the off-diagonal terms are quadratically larger, with $\sqrt{e_F} \approx \theta$. While not all error metrics are sensitive to the off-diagonal terms in an error process (e.g. fidelity-based measures), norm-based error metrics such as the diamond norm generally are sensitive to such terms, setting the $\sim \sqrt{e_F}$ upper bound of Eq. (3).

Under RC in the many-randomization limit, all off-diagonal terms in the error process are completely suppressed in the limit that $N \longrightarrow \infty$. To understand how this occurs, consider Pauli twirling Λ, i.e. $P\Lambda P^\dagger$ for any $P \in \{I, X, Y, Z\}$, where P represents the Pauli superoperator. Under Pauli conjugation, the signs of the off-diagonal terms remain the same (are reversed) if P commutes (does not commute) with Λ. Thus, the off-diagonal terms change sign with a 50% probability upon conjugation with a randomly selected Pauli. When averaging over N randomizations, the magnitude of the off-diagonal terms scale as $\theta/\sqrt{N}$, reminiscent of a random walk, and thus vanish as $N \longrightarrow \infty$ or if by luck the correct Paulis were sampled which average to zero. While the "noise tailoring" property of RC rests on assumption that the noise impacting the easy gates is gate-independent, Wallman et al.[52] prove that RC is robust to small gate-dependent errors, which are inevitable in modern-day experiments.

Method for performing randomized compiling on GST sequences

In order to preserve the circuit depth of GST sequences under RC, randomly sampled single-qubit Paulis and their correction gates are inserted between every layer. For circuits only containing single-qubit gates, the random Paulis are compiled into the previous layer and the correction gates are compiled into the subsequent layer. For circuits containing two-qubit gates, the correction gates are commuted through the two-qubit gate before being compiled into the subsequent layer.

To highlight this method, consider a circuit $\mathcal{C}$ containing N layers L of single-qubit gates:

$$\mathcal{C} = L_N L_{N-1} \dots L_3 L_2 L_1. \quad (20)$$

Under RC, a single randomized circuit takes the following form

$$\mathcal{C} = P_N L_N P^\dagger_{N-1} P_{N-1} L_{N-1} \dots L_3 P^\dagger_2 P_2 L_2 P^\dagger_1 P_1 L_1, \quad (21)$$

where $P^\dagger_N$ is omitted in the circuit but taken into account in the final ideal measurement results. The compiled circuit is

$$\mathcal{C} = \tilde{L}_N \tilde{L}_{N-1} \dots \tilde{L}_3 \tilde{L}_2 \tilde{L}_1, \quad (22)$$

where the kth layer $\tilde{L}_k = P_k L_k P^\dagger_{k-1}$ (except for the first layer, which does not contain a correction gate). For circuits containing two-qubit gates in layer $k - 1$, the kth layer becomes $\tilde{L}_k = P_k L_k P^c_{k-2}$, where $P^c_{k-2} = L_{k-1} P^\dagger_{k-2} L^\dagger_{k-1}$. This method therefore randomizes all layers of single-qubit gates, while also maintaining the original circuit depth, and was developed within the `pyGSTi` framework specifically for the purpose of randomizing GST and related benchmarking circuits.

As highlighted in the main text, we observe larger model violation for a single-randomization under RC than for no randomizations. This is expected behavior for $N = 1$ even when the physical gates are Markovian, which we illustrate with the following example: consider two single-qubit circuits, $\mathcal{C}_1 = G_I G_I$ and $\mathcal{C}_2 = G_I G_I G_I G_I$, and a physical gate set $\{G_I, G_X, G_Y, G_Z\}$, with $G_X = X_\pi$, etc. Consider a simple error model where each gate is followed by a small coherent X_θ rotation error. When we implement $\mathcal{C}_1$ with a single randomization under RC, we actually implement a circuit that should perform the identity rotation, but the gates have been randomized according to the method outlined above. Let's assume our randomized circuit $\tilde{\mathcal{C}}_1$ is the same as the original circuit, $\tilde{\mathcal{C}}_1 = \mathcal{C}_1 = G_I G_I$. If we initialize our qubit in the ground state, we will observe a small coherent over-rotation error by 2θ, which will result in some θ-dependent probability of not measuring 0. Similarly, when we run $\mathcal{C}_2$ we will actually actually implement one of the many length-4 combinations of Pauli gates that produces the identity rotation. For example, suppose we sampled $\tilde{\mathcal{C}}_2 = G_Z G_Z G_Z G_Z$; this combination of gates would echo away the X rotation error, and we would measure 0 with probability 1. Therefore, altogether we
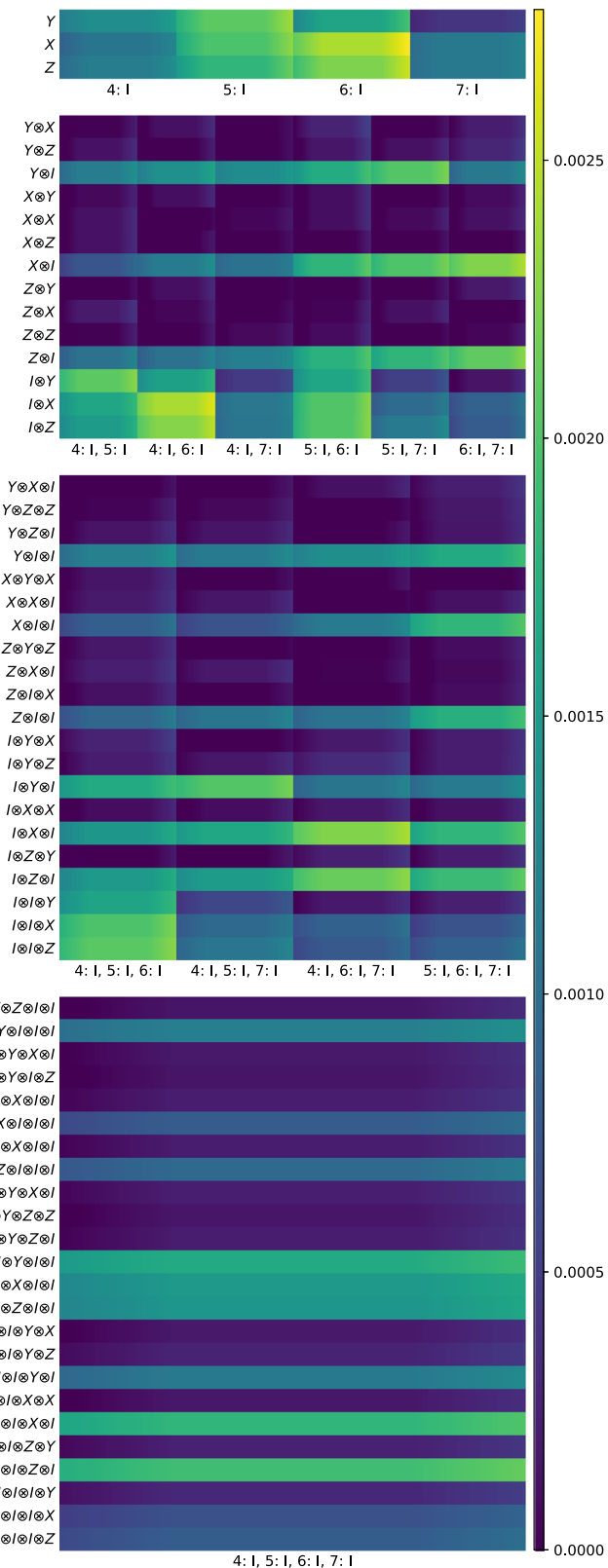
Fig. 6 Cycle error reconstruction of the identity cycle on four qubits $\{I_4 \otimes I_5 \otimes I_6 \otimes I_7\}$. Weight-1 errors dominate the error map; all weight-$k \geq 2$ errors are negligible in comparison. Rows in which all errors fall below 12% of the maximum value have been omitted for visual clarity.

observe $\Pr(0|\mathcal{C}_1) < 1$ and $\Pr(0|\mathcal{C}_2) = 1$. This is inconsistent with every possible process matrix for G_I (or at least every process matrix that is close to the target identity matrix), because repeating an identity gate amplifies all of its error parameters, but does not echo away errors. Finally, note that this argument is predicated on the assumption that we measure each circuit many times; if we measure each circuit only once, the results will not be inconsistent, but also will not be very informative.

Cycle error reconstruction

In Fig. 6, we plot the results of CER applied to the identity cycle $\{I_4 \otimes I_5 \otimes I_6 \otimes I_7\}$ across all four qubits on the quantum processor. We measure all $4^4 = 256$ Pauli channels, but omit values which fall below 12% of the maximum value for visual clarity. We observe no significant contribution from weight-$k \geq 2$ in the plot, demonstrating that Pauli twirling can break larger-scale correlations across an entire quantum processor. Furthermore, Refs. [55,120] present similar results for multi-qubit cycles containing two-qubit gates, and the conclusions were the same: the most dominant errors under Pauli twirling are single-body errors (i.e. weight-1 errors for single-qubit gates, weight-1 or weight-2 errors for two-qubit gates, etc.). Moreover, because two-body errors are the sum of all errors that act non-trivially on the corresponding two bodies, irrespective of their action on other qubits, the fact that two-body errors are negligible indicates that higher-body errors are also negligible.

DATA AVAILABILITY

All data are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

All GST sequences were generated and analyzed using `pyGSTi`[63,64], an open-source Python package. All CB and CER sequences were generated and analyzed using `True-Q`[85], a proprietary software package.

REFERENCES

1. Preskill, J. Quantum computing in the nisq era and beyond. *Quantum* **2**, 79 (2018).
2. Shor, P. W. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Rev.* **41**, 303–332 (1999).
3. Shor, P. W. Scheme for reducing decoherence in quantum computer memory. *Phys. Rev. A* **52**, R2493 (1995).
4. Gottesman, D. Class of quantum error-correcting codes saturating the quantum hamming bound. *Phys. Rev. A* **54**, 1862 (1996).
5. Steane, A. Multiple-particle interference and quantum error correction. *Proc. Roy. Soc. A* **452**, 2551–2577 (1996).
6. Steane, A. Introduction to quantum error correction. *Philos. Trans. Royal Soc. A* **356**, 1739–1758 (1998).
7. Calderbank, A. R., Rains, E. M., Shor, P. & Sloane, N. J. Quantum error correction via codes over gf (4). *IEEE Trans. Inf. Theory* **44**, 1369–1387 (1998).
8. Shor, P. W. Fault-tolerant quantum computation. In: *Proceedings of 37th Conference on Foundations of Computer Science* 56–65 (IEEE, 1996).
9. Knill, E. & Laflamme, R. Concatenated quantum codes. Preprint at https://arxiv.org/abs/quant-ph/9608012 (1996).
10. Knill, E., Laflamme, R. & Zurek, W. H. Resilient quantum computation. *Science* **279**, 342–345 (1998).
11. Preskill, J. in *Introduction to Quantum Computation and Information* 213–269 (World Scientific, 1998).
12. Kitaev, A. Y. Fault-tolerant quantum computation by anyons. *Ann. Phys.* **303**, 2–30 (2003).
13. Aharonov, D. & Ben-Or, M. Fault-tolerant quantum computation with constant error rate. *SIAM J. Sci. Comput.* **38** (2008).

14. Aliferis, P., Gottesman, D. & Preskill, J. Quantum Accuracy Threshold for Concatenated Distance-3 Codes. *Quantum Info. Comput.* **6**, 97–165 (2006).

15. Aliferis, P. & Cross, A. W. Subsystem fault tolerance with the bacon-shor code. *Phys. Rev. Lett.* **98**, 220502 (2007).

16. Aliferis, P., Gottesman, D. & Preskill, J. Accuracy Threshold for Postselected Quantum Computation. *Quantum Info. Comput.* **8**, 181–244 (2008).

17. Aliferis, P. & Preskill, J. Fibonacci scheme for fault-tolerant quantum computation. *Phys. Rev. A* **79**, 012332 (2009).

18. Chamberland, C., Jochym-O'Connor, T. & Laflamme, R. Thresholds for universal concatenated quantum codes. *Phys. Rev. Lett.* **117**, 010501 (2016).

19. Knill, E. Quantum computing with realistically noisy devices. *Nature* **434**, 39–44 (2005).

20. Aliferis, P. et al. Fault-tolerant computing with biased-noise superconducting qubits: a case study. *New J. Phys.* **11**, 013061 (2009).

21. Duclos-Cianci, G. & Poulin, D. Fast decoders for topological quantum codes. *Phys. Rev. Lett.* **104**, 050504 (2010).

22. Wang, D. S., Fowler, A. G. & Hollenberg, L. C. Surface code quantum computing with error rates over 1%. *Phys. Rev. A* **83**, 020302 (2011).

23. Bombin, H., Andrist, R. S., Ohzeki, M., Katzgraber, H. G. & Martin-Delgado, M. A. Strong resilience of topological codes to depolarization. *Phys. Rev. X* **2**, 021004 (2012).

24. Wootton, J. R. & Loss, D. High threshold error correction for the surface code. *Phys. Rev. Lett.* **109**, 160503 (2012).

25. Stephens, A. M. Fault-tolerant thresholds for quantum error correction with the surface code. *Phys. Rev. A* **89**, 022321 (2014).

26. Auger, J. M., Anwar, H., Gimeno-Segovia, M., Stace, T. M. & Browne, D. E. Fault-tolerance thresholds for the surface code with fabrication errors. *Phys. Rev. A* **96**, 042316 (2017).

27. Tuckett, D. K., Bartlett, S. D., Flammia, S. T. & Brown, B. J. Fault-tolerant thresholds for the surface code in excess of 5% under biased noise. *Phys. Rev. Lett.* **124**, 130501 (2020).

28. Barends, R. et al. Superconducting quantum circuits at the surface code threshold for fault tolerance. *Nature* **508**, 500–503 (2014).

29. Rong, X. et al. Experimental fault-tolerant universal quantum gates with solid-state spins under ambient conditions. *Nat. Commun.* **6**, 1–7 (2015).

30. Xue, X. et al. Quantum logic with spin qubits crossing the surface code threshold. *Nature* **601**, 343–347 (2022).

31. Emerson, J., Alicki, R. & Życzkowski, K. Scalable noise estimation with random unitary operators. *J. opt., B Quantum semiclass. opt.* **7**, S347 (2005).

32. Knill, E. et al. Randomized benchmarking of quantum gates. *Phys. Rev. A* **77**, 012307 (2008).

33. Dankert, C., Cleve, R., Emerson, J. & Livine, E. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Phys. Rev. A* **80**, 012304 (2009).

34. Magesan, E., Gambetta, J. M. & Emerson, J. Scalable and robust randomized benchmarking of quantum processes. *Phys. Rev. Lett.* **106**, 180504 (2011).

35. Magesan, E. et al. Efficient measurement of quantum gate error by interleaved randomized benchmarking. *Phys. Rev. Lett.* **109**, 080505 (2012).

36. Kitaev, A. Y. Quantum computations: algorithms and error correction. *Usp. Mat. Nauk* **52**, 53–112 (1997).

37. Merkel, S. T. et al. Self-consistent quantum process tomography. *Phys. Rev. A* **87**, 062119 (2013).

38. Blume-Kohout, R. et al. Robust, self-consistent, closed-form tomography of quantum logic gates on a trapped ion qubit. Preprint at https://arxiv.org/abs/1310.4492 (2013).

39. Greenbaum, D. Introduction to quantum gate set tomography. Preprint at https://arxiv.org/abs/1509.02921 (2015).

40. Blume-Kohout, R. et al. Demonstration of qubit operations below a rigorous fault tolerance threshold with gate set tomography. *Nat. Commun.* **8**, 1–13 (2017).

41. Nielsen, E. et al. Gate set tomography. *Quantum* **5**, 557 (2021).

42. Wallman, J. J. & Flammia, S. T. Randomized benchmarking with confidence. *New J. Phys.* **16**, 103032 (2014).

43. Sanders, Y. R., Wallman, J. J. & Sanders, B. C. Bounding quantum gate error rate based on reported average fidelity. *New J. Phys.* **18**, 012002 (2015).

44. Wallman, J. J. Bounding experimental quantum error rates relative to fault-tolerant thresholds. Preprint at https://arxiv.org/abs/1511.00727 (2015).

45. Kueng, R., Long, D. M., Doherty, A. C. & Flammia, S. T. Comparing experiments to the fault-tolerance threshold. *Phys. Rev. Lett.* **117**, 170502 (2016).

46. Sheldon, S. et al. Characterizing errors on qubit operations via iterative randomized benchmarking. *Phys. Rev. A* **93**, 012301 (2016).

47. Arute, F. et al. Quantum supremacy using a programmable superconducting processor. *Nature* **574**, 505–510 (2019).

48. Wang, Y. et al. High-fidelity two-qubit gates using a microelectromechanical-system-based beam steering system for individual qubit addressing. *Phys. Rev. Lett.* **125**, 150505 (2020).

49. Pino, J. M. et al. Demonstration of the trapped-ion quantum ccd computer architecture. *Nature* **592**, 209–213 (2021).

50. Mitchell, B. K. et al. Hardware-efficient microwave-activated tunable coupling between superconducting qubits. *Phys. Rev. Lett.* **127**, 200502 (2021).

51. Mądzik, M. T. et al. Precision tomography of a three-qubit electron-nuclear quantum processor in silicon. *Nature* **601**, 348–353 (2022).

52. Wallman, J. J. & Emerson, J. Noise tailoring for scalable quantum computation via randomized compiling. *Phys. Rev. A* **94**, 052325 (2016).

53. Ryan-Anderson, C. et al. Realization of real-time fault-tolerant quantum error correction. *Phys. Rev. X* **11**, 041058 (2021).

54. Chen, L. et al. Transmon qubit readout fidelity at the threshold for quantum error correction without a quantum-limited amplifier. *npj Quantum Inf.* **9**, 26 (2023).

55. Hashim, A. et al. Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor. *Phys. Rev. X* **11**, 041039 (2021).

56. Erhard, A. et al. Characterizing large-scale quantum computers via cycle benchmarking. *Nat. Commun.* **10**, 1–7 (2019).

57. Hastings, M. B. Turning gate synthesis errors into incoherent errors. Preprint at https://arxiv.org/abs/1612.01011 (2016).

58. Campbell, E. Shorter gate sequences for quantum computing by mixing unitaries. *Phys. Rev. A* **95**, 042306 (2017).

59. Cai, Z., Xu, X. & Benjamin, S. C. Mitigating coherent noise using pauli conjugation. *npj Quantum Inf.* **6**, 1–9 (2020).

60. Ware, M. et al. Experimental pauli-frame randomization on a superconducting qubit. *Phys. Rev. A* **103**, 042604 (2021).

61. Hashim, A. et al. Optimized swap networks with equivalent circuit averaging for qaoa. *Phys. Rev. Res.* **4**, 033028 (2022).

62. Chuang, I. L. & Nielsen, M. A. Prescription for experimental determination of the dynamics of a quantum black box. *Journal of Modern Optics* **44**, 2455–2467 (1997).

63. Nielsen, E. et al. Python gst Implementation (pygsti) v. 0.9. Tech. Rep., Sandia National Lab.(SNL-NM), Albuquerque, NM (United States) (2019).

64. Nielsen, E. et al. Probing quantum processor performance with pygsti. *Quantum Sci. Technol.* **5**, 044002 (2020).

65. Viola, L., Knill, E. & Lloyd, S. Dynamical decoupling of open quantum systems. *Phys. Rev. Lett.* **82**, 2417 (1999).

66. Blume-Kohout, R. et al. A taxonomy of small markovian errors. *PRX Quantum* **3**, 020335 (2022).

67. Nielsen, E., Rudinger, K., Proctor, T., Young, K. & Blume-Kohout, R. Efficient flexible characterization of quantum processors with nested error models. *New J. Phys.* **23**, 093020 (2021).

68. Rudinger, K. et al. Experimental characterization of crosstalk errors with simultaneous gate set tomography. *PRX Quantum* **2**, 040338 (2021).

69. Wallman, J., Granade, C., Harper, R. & Flammia, S. T. Estimating the coherence of noise. *New J. Phys.* **17**, 113020 (2015).

70. Boixo, S. et al. Characterizing quantum supremacy in near-term devices. *Nat. Phys.* **14**, 595–600 (2018).

71. Knill, E. Fault-tolerant postselected quantum computation: threshold analysis. Preprint at https://arxiv.org/abs/quant-ph/0404104 (2004).

72. Kern, O., Alber, G. & Shepelyansky, D. L. Quantum error correction of coherent errors by randomization. *Eur. Phys. J. D* **32**, 153–156 (2005).

73. Fowler, A. G., Mariantoni, M., Martinis, J. M. & Cleland, A. N. Surface codes: Towards practical large-scale quantum computation. *Phys. Rev. A* **86**, 032324 (2012).

74. Wang, C., Harrington, J. & Preskill, J. Confinement-higgs transition in a disordered gauge theory and the accuracy threshold for quantum memory. *Ann. Phys.* **303**, 31–58 (2003).

75. Raussendorf, R., Harrington, J. & Goyal, K. A fault-tolerant one-way quantum computer. *Ann. Phys.* **321**, 2242–2270 (2006).

76. Raussendorf, R., Harrington, J. & Goyal, K. Topological fault-tolerance in cluster state quantum computation. *New J. Phys.* **9**, 199 (2007).

77. Raussendorf, R. & Harrington, J. Fault-tolerant quantum computation with high threshold in two dimensions. *Phys. Rev. Lett.* **98**, 190504 (2007).

78. Fowler, A. G., Whiteside, A. C. & Hollenberg, L. C. Towards practical classical processing for the surface code. *Phys. Rev. Lett.* **108**, 180501 (2012).

79. Lu, F. & Marinescu, D. C. Quantum error correction of time-correlated errors. *Quantum Inf. Process.* **6**, 273–293 (2007).

80. Wilen, C. D. et al. Correlated charge noise and relaxation errors in superconducting qubits. *Nature* **594**, 369–373 (2021).

81. Mundada, P., Zhang, G., Hazard, T. & Houck, A. Suppression of qubit crosstalk in a tunable coupling superconducting circuit. *Phys. Rev. App.* **12**, 054023 (2019).

82. Zhao, P. et al. High-contrast zz interaction using superconducting qubits with opposite-sign anharmonicity. *Phys. Rev. Lett.* **125**, 200503 (2020).

83. Ni, Z. et al. Scalable method for eliminating residual zz interaction between superconducting qubits. *Phys. Rev. Lett.* **129**, 040502 (2022).

84. Flammia, S. T. & Wallman, J. J. Efficient estimation of pauli channels. *ACM Trans. Quantum Comput.* **1**, 1–32 (2020).

85. Beale, S. J. et al. True-q (2020). https://doi.org/10.5281/zenodo.3945250.

86. Fowler, A. G. & Martinis, J. M. Quantifying the effects of local many-qubit errors and nonlocal two-qubit errors on the surface code. *Phys. Rev. A* **89**, 032316 (2014).

87. Ball, H., Stace, T. M., Flammia, S. T. & Biercuk, M. J. Effect of noise correlations on randomized benchmarking. *Phys. Rev. A* **93**, 022303 (2016).

88. Ghosh, J., Fowler, A. G., Martinis, J. M. & Geller, M. R. Understanding the effects of leakage in superconducting quantum-error-detection circuits. *Phys. Rev. A* **88**, 062329 (2013).

89. Wallman, J. J., Barnhill, M. & Emerson, J. Robust characterization of leakage errors. *New J. Phys.* **18**, 043021 (2016).

90. Chen, Z. et al. Measuring and suppressing quantum state leakage in a super-conducting qubit. *Phys. Rev. Lett.* **116**, 020501 (2016).

91. Wood, C. J. & Gambetta, J. M. Quantification and characterization of leakage errors. *Phys. Rev. A* **97**, 032306 (2018).

92. Hayes, D. et al. Eliminating leakage errors in hyperfine qubits. *Phys. Rev. Lett.* **124**, 170501 (2020).

93. Babu, A. P., Tuorila, J. & Ala-Nissila, T. State leakage during fast decay and control of a superconducting transmon qubit. *npj Quantum Inf.* **7**, 1–8 (2021).

94. Proctor, T. et al. Detecting and tracking drift in quantum information processors. *Nat. Commun.* **11**, 1–9 (2020).

95. Serniak, K. et al. Hot nonequilibrium quasiparticles in transmon qubits. *Phys. Rev. Lett.* **121**, 157701 (2018).

96. de Graaf, S. et al. Two-level systems in superconducting quantum devices due to trapped quasiparticles. *Sci. Adv.* **6**, eabc5055 (2020).

97. Berlin-Udi, M. et al. Changes in electric field noise due to thermal transformation of a surface ion trap. *Phys. Rev. B* **106**, 035409 (2022).

98. Webb, A. E. et al. Resilient entangling gates for trapped ions. *Phys. Rev. Lett.* **121**, 180501 (2018).

99. Blume-Kohout, R., Rudinger, K., Nielsen, E., Proctor, T. & Young, K. Wildcard error: Quantifying unmodeled errors in quantum processors. Preprint at https://arxiv.org/abs/2012.12231 (2020).

100. Wootton, J. R. & Pachos, J. K. Bringing order through disorder: Localization of errors in topological quantum memories. *Phys. Rev. Lett.* **107**, 030503 (2011).

101. Stark, C., Pollet, L., Imamoğlu, A. & Renner, R. Localization of toric code defects. *Phys. Rev. Lett.* **107**, 030504 (2011).

102. Bravyi, S. & König, R. Disorder-assisted error correction in majorana chains. *Commun. Math. Phys.* **316**, 641–692 (2012).

103. Bravyi, S., Englbrecht, M., König, R. & Peard, N. Correcting coherent errors with surface codes. *npj Quantum Inf.* **4**, 1–6 (2018).

104. Geller, M. R. & Zhou, Z. Efficient error models for fault-tolerant architectures and the Pauli twirling approximation. *Phys. Rev. A* **88**, 012314 (2013).

105. Tomita, Y. & Svore, K. M. Low-distance surface codes under realistic quantum noise. *Phys. Rev. A* **90**, 062320 (2014).

106. Katabarwa, A. & Geller, M. R. Logical error rate in the Pauli twirling approximation. *Sci. Rep.* **5**, 1–6 (2015).

107. Puzzuoli, D. et al. Tractable simulation of error correction with honest approximations to realistic fault models. *Phys. Rev. A* **89**, 022306 (2014).

108. Gutiérrez, M. & Brown, K. R. Comparison of a quantum error-correction threshold for exact and approximate errors. *Phys. Rev. A* **91**, 022335 (2015).

109. Katabarwa, A. A dynamical interpretation of the pauli twirling approximation and quantum error correction. Preprint at https://arxiv.org/abs/1701.03708 (2017).

110. Fern, J., Kempe, J., Simic, S. N. & Sastry, S. Generalized performance of concatenated quantum codes—a dynamical systems approach. *IEEE Trans. Autom. Control* **51**, 448–459 (2006).

111. Gutiérrez, M., Smith, C., Lulushi, L., Janardan, S. & Brown, K. R. Errors and pseudo-thresholds for incoherent and coherent noise. *Phys. Rev. A* **94**, 042338 (2016).

112. Darmawan, A. S. & Poulin, D. Tensor-network simulations of the surface code under realistic noise. *Phys. Rev. Lett.* **119**, 040502 (2017).

113. Greenbaum, D. & Dutton, Z. Modeling coherent errors in quantum error correction. *Quantum Science and Technology* **3**, 015007 (2017).

114. Hakkaku, S., Mitarai, K. & Fujii, K. Sampling-based quasiprobability simulation for fault-tolerant quantum error correction on the surface codes under coherent noise. *Phys. Rev. Res.* **3**, 043130 (2021).

115. Huang, E., Doherty, A. C. & Flammia, S. Performance of quantum error correction with coherent errors. *Phys. Rev. A* **99**, 022313 (2019).

116. Dehollain, J. P. et al. Optimization of a solid-state electron spin qubit using gate set tomography. *New J. Phys.* **18**, 103018 (2016).

117. Hughes, A. et al. Benchmarking a high-fidelity mixed-species entangling gate. *Phys. Rev. Lett.* **125**, 080504 (2020).

118. White, G. A., Hill, C. D. & Hollenberg, L. C. Performance optimization for drift-robust fidelity improvement of two-qubit gates. *Phys. Rev. App.* **15**, 014023 (2021).

119. Goss, N. et al. Extending the computational reach of a superconducting qutrit processor. Preprint at https://arxiv.org/abs/2305.16507 (2023).

120. Ferracin, S. et al. Efficiently improving the performance of noisy quantum computers. Preprint at https://arxiv.org/abs/2201.10672 (2022).

121. Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9**, 60–62 (1938).

122. Proctor, T., Rudinger, K., Young, K., Nielsen, E. & Blume-Kohout, R. Measuring the capabilities of quantum computers. *Nat. Phys.* **18**, 75–79 (2021).

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

A.H. conceptualized and performed the experiments. K.R., S.S., T.P., and N.G. assisted in the design of the experiments. A.H. analyzed the data with the help of K.R., S.S., and T.P. J.M.K. fabricated the sample. R.K.N., D.I.S., and I.S. supervised all experimental work. A.H. and T.P. wrote the manuscript with input from all coauthors. A.H. acknowledges fruitful discussions with Joel Wallman, Joseph Emerson, Robin Blume-Kohout, Samuele Ferracin, and Long Nguyen. T.P. acknowledges helpful discussions with Kevin Young and Robin Blume-Kohout.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Akel Hashim.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.