Fully Scalable Randomized Benchmarking Without Motion Reversal

Jordan Hines,[1,2,3,*] Daniel Hothem,[2,3] Robin Blume-Kohout,[2,3] Birgitta Whaley,[4] and Timothy Proctor[2,3,†]

[1] *Department of Physics, University of California, Berkeley, California 94720, USA*
[2] *Quantum Performance Laboratory, Sandia National Laboratories, Albuquerque, New Mexico 87185, USA*
[3] *Quantum Performance Laboratory, Sandia National Laboratories, Livermore, California 94550, USA*
[4] *Department of Chemistry, University of California, Berkeley, California 94720, USA*

(Received 9 October 2023; accepted 8 July 2024; published 15 August 2024)

We introduce *binary randomized benchmarking (BiRB)*, a protocol that streamlines traditional RB by using circuits consisting almost entirely of independent identically distributed (IID) layers of gates. BiRB reliably and efficiently extracts the average error rate of a Clifford gate set by sending tensor-product eigenstates of random Pauli operators through random circuits with IID layers. Unlike existing RB methods, BiRB does not use motion reversal circuits—i.e., circuits that implement the identity (or a Pauli) operator—which simplifies both the method and the theory proving its reliability. Furthermore, this simplicity enables scaling BiRB to many more qubits than the most widely used RB methods.

DOI: 10.1103/PRXQuantum.5.030334

I. INTRODUCTION

Randomized benchmarking (RB) [1–21] is a family of protocols that assess the average performance of a quantum processor's gates by running random circuits. RB experiments are ubiquitous, yet the most widely used RB protocols have important limitations that are caused by the kind of random circuits they use. Most RB protocols use *motion reversal* circuits that, if run without errors, implement the identity (or a Pauli) operator [1,4–6,9,10,16,18] [Fig. 1(a)]. This makes errors easily visible: each RB circuit, when run perfectly, always outputs a particular bit string, so the observation of any other bit string implies that an error occurred. However, random motion-reversal circuits must end with an *inversion subroutine* that undoes the preceding layers. The inversion subroutine causes challenges for RB theory [1,6,7,21–25] as well as practical problems. In most existing RB techniques—including standard Clifford group RB (CRB) [6] and its streamlined variant direct RB (DRB) [18]—the size of the inversion subroutine grows quickly with the number of qubits [26–28] [see Fig. 1(a)], severely limiting their applicability outside of the few-qubit setting [1,3,18].

In this work, we demonstrate that motion-reversal circuits are not required for reliable RB by introducing *binary randomized benchmarking (BiRB)*. BiRB is an efficient and scalable protocol for estimating the average error rate of a Clifford gate set. BiRB's circuits [Fig. 1(b)] consist of d IID layers of gates and two layers of single-qubit gates, for state and measurement preparation, and the measurement results are processed to obtain a binary-outcome Pauli measurement result. BiRB works because the average fidelity of highly scrambling random circuits decays exponentially in depth [18,21,23] and, for Clifford circuits, this fidelity can be efficiently estimated using random local state preparations and measurements [29,30]. Our method's local state preparation and measurement enables benchmarking of many more qubits than most existing RB techniques—including CRB and DRB—as shown in Fig. 2). Furthermore, we show that BiRB is more accurate than mirror RB (MRB) [1,21], which is the only other scalable RB protocol for Clifford gate sets.

BiRB connects RB and cross-entropy benchmarking (XEB) [31–34], another form of randomized benchmark. In contrast to RB, XEB uses random circuits consisting solely of IID (composite) layers of gates—these layers are typically sampled from a universal gate set, but a scalable form of XEB using Clifford gate sets has also been introduced [34]. While XEB circuits have no overhead from subroutines, in practice XEB decay curves exhibit non-exponential behavior at low depths for some Markovian error models, and therefore measuring a reliable error rate requires circuits with at least $O(n)$ depth [35]. The exact circuit depths required for exponential decay depend on

*Contact author: jordanh@berkeley.edu
†Contact author: tjproct@sandia.gov

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

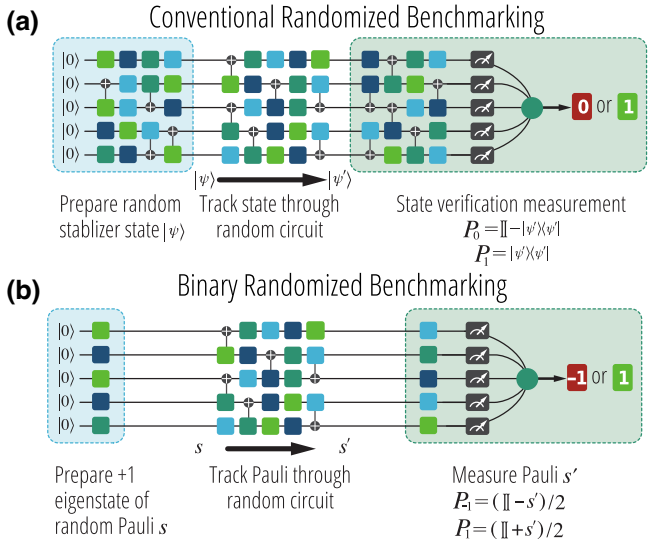

FIG. 1. RB without motion reversal. (a) Standard RB methods use motion-reversal circuits, which make errors easily visible but add complexity and limit scalability. (b) BiRB reliably estimates average gate error rates without motion reversal by tracking a single stabilizer of a random product state through a random circuit. (c) Results from CRB, DRB, and BiRB on `ibm_hanoi` show that BiRB is more scalable than both CRB and DRB. DRB estimates the same error rate as BiRB (r_Ω), and we find that their error rates are consistent, providing evidence for the reliability of BiRB.

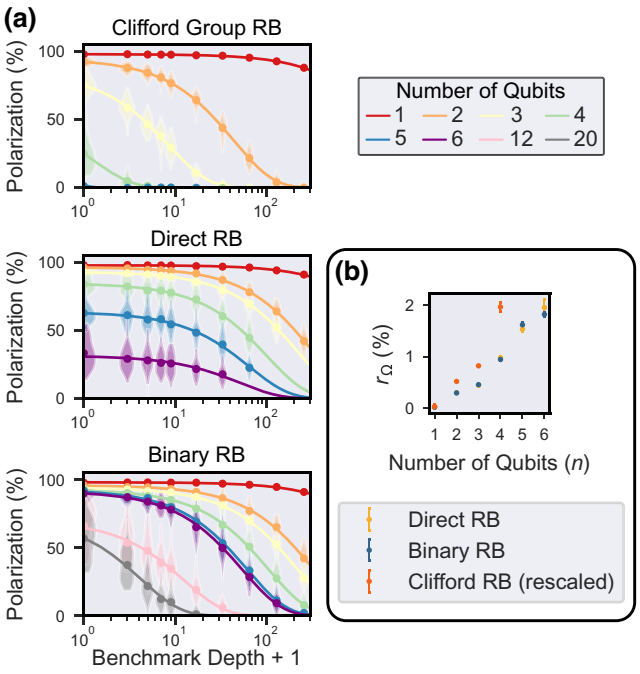

FIG. 2. BiRB and conventional RB on IBM Q. Results from CRB, DRB, and BiRB on `ibm_hanoi` show that BiRB is more scalable than both CRB and DRB. DRB estimates the same error rate as BiRB (r_Ω), and we find that their error rates are consistent, providing evidence for the reliability of BiRB.

the connectivity and gate set and must be estimated numerically for each distribution of layers benchmarked, adding additional complication to performing XEB. This issue arises in part because XEB estimates the fidelity of random circuits using the (linear) cross entropy, which is not an accurate fidelity estimator for general Markovian noise models [33,36]. BiRB shows how to add minimal overhead to circuits of IID layers to obtain a provably reliable RB protocol.

The remainder of this paper is structured as follows. In Sec. II we introduce our notation and review the existing results on which our method relies. In Sec. III we introduce the BiRB protocol. In Sec. IV we present a theory of BiRB that shows that our method is reliable: it accurately estimates the average error rate of an n-qubit circuit layer under assumptions commonly used in RB theory (e.g., Markovian errors). In Sec. V we demonstrate the reliability of our method with numerical simulations of BiRB on gate sets that experience both stochastic Pauli errors and (coherent) Hamiltonian errors. In Sec. VI, we demonstrate BiRB on IBM Q processors and validate it against the results of DRB and MRB. We then conclude in Sec. VII.

II. PRELIMINARIES

A. Definitions

In this section, we introduce our notation. An n-qubit *layer* L is an instruction to perform a particular unitary operation on those n qubits, typically specified in terms of one- and two-qubit gates. We use $U(L) \in \mathrm{SU}(2^n)$ to denote the unitary corresponding to L. The layers we use are randomly sampled, and we often treat a layer L as a layer-valued random variable. We use $\Omega : \mathbb{L} \to [0,1]$ to denote a probability distribution over the set of layers $\mathbb{L}$. We use L^{-1} to denote an instruction to perform the unitary $U(L)^{-1}$. An n-qubit, depth-d *circuit* is a sequence of n-qubit layers $C = L_d L_{d-1} \cdots L_2 L_1$, where we use the convention that the circuit is read right to left.

For a layer (or circuit) L, we use $\mathcal{U}(L)$ to denote the superoperator representation of its perfect implementation, i.e., $\mathcal{U}(L)[\rho] = U(L)\rho U^\dagger(L)$. We use $\phi(L)$ to denote the superoperator for an imperfect implementation of L, and we assume $\phi(L)$ is a completely positive trace preserving (CPTP) map. A layer L's *error map* is defined by $\mathcal{E}_L = \phi(L)\mathcal{U}^\dagger(L)$. The *entanglement fidelity* (also called the *process fidelity*) of $\phi(L)$ to $\mathcal{U}(L)$ is defined by

$$F\big(\phi(L),\mathcal{U}(L)\big) = F(\mathcal{E}_L) = \langle\varphi|\big(\mathbb{I}\otimes\mathcal{E}_L\big)[|\varphi\rangle\langle\varphi|]|\varphi\rangle \quad (1)$$

$$= \frac{1}{4^n}\mathrm{Tr}(\mathcal{U}(L)^\dagger\phi(L)) \quad (2)$$

$$= \mathop{\mathbb{E}}_{s\in\mathbb{P}_n} \mathrm{Tr}(s\mathcal{E}_L[s]), \quad (3)$$

where φ is any maximally entangled state of $2n$ qubits [37], and $\mathbb{P}_n$ is the set of all n-qubit Pauli operations with ± 1

global sign. Throughout, we use the term "(in)fidelity" to refer to the entanglement (in)fidelity. The *polarization* is a rescaling of fidelity given by

$$\gamma(\phi(L),\mathcal{U}(L)) = \gamma(\mathcal{E}_L) = \frac{4^n}{4^n-1}F(\mathcal{E}_L) - \frac{1}{4^n-1} \quad (4)$$

$$= \mathop{\mathbb{E}}_{s\in\mathbb{P}_n^*} \mathrm{Tr}(s\mathcal{E}_L[s]), \quad (5)$$

where $\mathbb{P}_n^* = \mathbb{P}_n \setminus \{\pm I_n\}$, and I_n denotes the n-qubit identity operator. We say that a state $|\psi\rangle$ is *stabilized* by a Pauli operator P if $P|\psi\rangle = |\psi\rangle$. An n-qubit *stabilizer state* $|\psi\rangle$ is a state that is stabilized by exactly 2^n Pauli operators. Equivalently, a stabilizer state is a state that can be prepared from $|0\rangle^{\otimes n}$ using only Clifford gates [26]. The *stabilizer group* of a stabilizer state $|\psi\rangle$ is $S_\psi = \{P \in \mathbb{P}_n \mid P|\psi\rangle = |\psi\rangle\}$. We use S_ψ^* to denote all nonidentity elements of the stabilizer group, i.e., $S_\psi^* = S_\psi \setminus \{I_n\}$.

B. Ω-distributed random circuits

BiRB uses Ω-*distributed random circuits* [1,18,21,23], which we now review. Ω-distributed random circuits consist of n-qubit layers of gates sampled from a distribution $\Omega(\mathbb{L})$ over a layer set $\mathbb{L}$. In this work, we restrict $\mathbb{L}$ to contain only Clifford gates. These circuit layers can be chosen to consist of a processor's native gates, or simple combinations thereof, thus eliminating the need for complicated compilation.

Ω-distributed random circuits are also used in DRB and MRB [1,18,21,23]. DRB and MRB are reliable if Ω satisfies certain conditions, and these same conditions are required for BiRB to be reliable. We require that the circuits generated by layers sampled from Ω are highly scrambling, meaning that for all Pauli operators $P, P' \neq I_n$, there exists constants $k \ll 1/\varepsilon$ and $\delta \ll 1$ such that

$$\frac{1}{4^n}\mathop{\mathbb{E}}_{L_1,\ldots,L_k} Tr(\mathcal{P}'\mathcal{U}(L_k\cdots L_1)\mathcal{P}\mathcal{U}(L_k\cdots L_1)^{-1}) \leq \delta + \frac{1}{4^n}. \quad (6)$$

Here, $\mathcal{P}[\rho] = P\rho P$ and $\mathcal{P}'[\rho] = P'\rho P'$ are Pauli superoperators, $L_1,\ldots,L_k$ are Ω-distributed random layers, and ε is the expected infidelity of an Ω-distributed random layer [21,23]. Informally, this condition means that an error is locally randomized (i.e., its basis is randomized over the X, Y, and Z bases) and delocalized across multiple qubits before a second error is likely to have occurred. While we require that $k \ll 1/\varepsilon$ for our theory, this condition on k can be relaxed when $n \gg 1$, because errors that occur on spatially separated qubits in close succession cannot cancel at all (see Refs. [21,23] for details).

C. The RB error rate

BiRB's output is an error rate r_Ω that quantifies the error in random n-qubit layers sampled from Ω. BiRB's r_Ω closely approximates an independent, physically motivated error rate ϵ_Ω—which is closely related to the average layer infidelity—introduced in Refs. [21,38] and reviewed here. ϵ_Ω is defined by the *rate of decay* of the fidelity of Ω-distributed random circuits. The expected fidelity of depth-d Ω-distributed random circuits C_d is given by

$$\bar{F}_d = \mathop{\mathbb{E}}_{C_d} F\big(\phi(C_d),\mathcal{U}(C_d)\big). \quad (7)$$

The scrambling requirements on Ω (see Sec. III A) ensure that $\bar{F}_d$ [Eq. (7)] decays exponentially, i.e., $\bar{F}_d \approx Ap_{\mathrm{RC}}^d + B$ for constants A, B, and p_{RC}. The average error rate of layers sampled from Ω is then defined as [21,38]

$$\epsilon_\Omega = \frac{4^n-1}{4^n}(1-p_{\mathrm{RC}}). \quad (8)$$

This rescaling of p_{RC} is used because p_{RC} corresponds to the effective polarization of a random layer in an Ω-distributed random circuit—i.e., the polarization in a depolarizing channel that would give the same fidelity decay—so ϵ_Ω is the effective average infidelity of a layer sampled from Ω. When stochastic Pauli errors are the dominant source of error, ϵ_Ω is approximately equal to the average layer infidelity,

$$\varepsilon_\Omega = 1 - \mathop{\mathbb{E}}_{L\in\mathbb{L}} F(\phi(L),\mathcal{U}(L)), \quad (9)$$

but this is not true more generally because gate infidelity is not "gauge-invariant"—see Refs. [21,22,24,38] for details.

D. Direct fidelity estimation

Our protocol can be interpreted as an application of *direct fidelity estimation* (DFE) [29,30] to varied-depth random Clifford circuits, so we now review DFE for the special case of Clifford circuits. Consider a Clifford circuit C and an imperfect implementation of that circuit $\phi(C) = \mathcal{U}(C)\mathcal{E}_C$, where $\mathcal{E}_C$ denotes the overall error map of the circuit. Using Eq. (5), the polarization of $\mathcal{E}_C$ can be written as

$$\gamma(\mathcal{E}_C) = \mathop{\mathbb{E}}_{s\in\mathbb{P}_n^*} \mathrm{Tr}(s\mathcal{E}_C[s]) \quad (10)$$

$$= \mathop{\mathbb{E}}_{s\in\mathbb{P}_n^*} \mathrm{Tr}(s\mathcal{U}(C)^\dagger\phi(C)[s]) \quad (11)$$

$$= \mathop{\mathbb{E}}_{s\in\mathbb{P}_n^*} \mathrm{Tr}(s'\phi(C)[s]), \quad (12)$$

where $s' = U(C)sU(C)^\dagger$ is a Pauli operator that can be efficiently computed classically [26], because C is a Clifford circuit. Equation (12) implies that polarization of $\mathcal{E}_C$ can be efficiently estimated as follows: (1) sample Pauli operators uniformly from $\mathbb{P}_n^*$, (2) for each sampled Pauli operator s, apply $\phi(C)$ to s and measure the evolved Pauli operator

s', and (3) average the measurement results. It is not physically possible to directly apply $\phi(C)$ to a Pauli operator s (Pauli operators are not valid quantum states), but DFE simulates doing so by applying $\phi(C)$ to randomly sampled eigenstates of s. BiRB also uses this approach, but, unlike DFE, BiRB is robust to state preparation and measurement (SPAM) error. BiRB separates SPAM error from gate error by applying DFE to variable-depth circuits and extracting gate error from the rate of decay of the polarization—as in cycle benchmarking [39] and Pauli noise learning techniques [40–42].

III. THE BINARY RB PROTOCOL

We now introduce BiRB circuits (Sec. III A) and the BiRB protocol (Sec. III B).

A. Binary RB circuits

We now state the procedure for constructing BiRB circuits [Fig. 3(b)]. Each BiRB circuit first generates an eigenstate of a random Pauli operator s, then applies a depth d random circuit, and then ends with a measurement of the evolved Pauli operator s'. A width n, benchmark depth d, Ω-distributed BiRB circuit is a circuit $C = L_{d+1}L_d \cdots L_1 L_0$ that begins with preparing $|0\rangle^{\otimes n}$ and ends with a computational basis measurement, and has layers sampled as follows:

(1) Sample a uniformly random n-qubit Pauli operator $s \in \mathbb{P}_n^*$ and a uniformly random state $|\psi(s)\rangle$ from the set of tensor-product stabilizer states stabilized by s. L_0 is a layer of single-qubit gates that prepares $|\psi(s)\rangle$.
(2) $L_1, L_2, \ldots, L_d$ are layers sampled from Ω. These layers form the *core circuit*, which is a depth-d Ω-distributed random circuit.

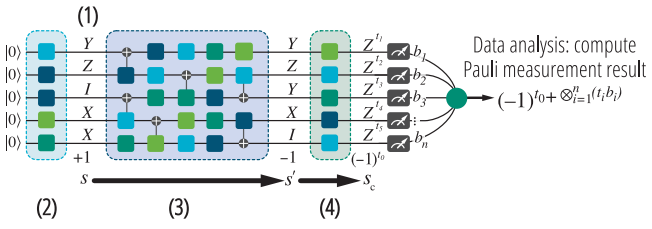

FIG. 3. Each BiRB circuit is constructed by (1) generating a random nonidentity Pauli operator s, then constructing a circuit consisting of (2) a layer of single-qubit gates generating a $+1$ eigenstate of a random Pauli operator s, (3) d layers of Clifford gates randomly sampled from some distribution Ω, and (4) a final layer of gates that transforms the evolved Pauli operator (s') into a tensor product of Z and I Pauli operators (s_C, represented by bit string t). The result of a computational basis measurement (bit string b) is used to compute a 1 ("success") or -1 ("fail") result by comparing it to the bit string t.

(3) L_{d+1} is a layer of single-qubit gates that transforms

$$s' = U(L_d) \cdots U(L_1) s U(L_1)^{-1} \cdots U(L_d)^{-1} \quad (13)$$

into a tensor product of Z and I operators.

The circuit has an associated "target" Pauli operator

$$s_C = U(L_{d+1}) s' U(L_{d+1})^{-1}. \quad (14)$$

If implemented without errors, the bit string b output by C will correspond to a $+1$ eigenstate of s_C, i.e., $s_C|b\rangle = |b\rangle$.

Step (1) can equivalently be formulated as (i) sampling a random unsigned Pauli P, (ii) picking a random tensor-product stabilizer state that is an eigenstate of P. Sampling from both $+1$ and -1 eigenstates ensures accurate fidelity estimation when there are nonunital errors in the circuits (see Sec. IV).

There is not a unique choice for either the initial layer (L_0) or the final layer (L_{d+1}) of gates in BiRB circuits. These layers may be chosen deterministically or at random from the set of all possible layers of single-qubit Clifford gates satisfying the criteria above. In our simulations and experiments, we choose to randomize L_0, but this is not required for our theory. There will always be a possible final layer satisfying the requirements in step (3). We can construct such a layer as follows: let $s' = \bigotimes_{i=1}^n s'_i$, where s'_i denotes the single-qubit Pauli operator acting on qubit i. On qubit i, apply H if $s'_i = X$, apply $HS^\dagger$ if $s'_i = Y$, and apply I if $s'_i = I$ or Z.

B. Binary RB protocol

The BiRB protocol is similar to other RB protocols: run BiRB circuits, compute a figure of merit for the circuits of each benchmark depth, then fit an exponential decay. A BiRB experiment is defined by a layer set $\mathbb{L}$, a sampling distribution Ω, and the usual RB sampling parameters (a set of benchmark depths d, the number of circuits K sampled per depth, and the number of times N each circuit is run). Our protocol is the following:

(1) For a range of integers $d \geq 0$, sample K Ω-distributed BiRB circuits with benchmark depth d, and run each circuit $N \geq 1$ times.
(2) For each circuit C, estimate the expected value $\langle s_C \rangle$ of the target Pauli observable s_C from the computational basis measurement results. Then, compute the average over all circuits of benchmark depth d,

$$\bar{f}_d = \frac{1}{K} \sum_{C_d} \langle s_C \rangle. \quad (15)$$

(3) Fit $\bar{f}_d$ to an exponential,

$$\bar{f}_d = Ap^d, \tag{16}$$

where A and p are fit parameters. Then compute [43]

$$r_\Omega = (4^n - 1)(1 - p)/4^n. \tag{17}$$

In Appendix B, we show that the number of circuits per depth (and therefore the total amount of data) required by our protocol to estimate r_Ω to within a fixed relative uncertainty is independent of n. For a fixed total number of shots per depth KN, it is statistically optimal to maximize the number of random circuits K and set $N = 1$. However, typically $N > 1$ for practical reasons—e.g., because of the time cost of generating and loading many distinct circuits onto the processor. See Refs. [44–46] for more detailed statistical analyses of RB protocols.

Note that if $\mathbb{L}$ is chosen to be the set of all n-qubit Clifford gates and Ω is the uniform distribution, then we obtain a version of standard RB (i.e., RB of the Clifford group) without an inversion gate. See Appendix A for further discussion of this variant of BiRB, whose reliability can be proven using the unitary 2-design twirling theory that underpins the theory of standard RB [6,7].

IV. THEORY OF BINARY RB

We now show that the error rate measured by BiRB [r_Ω, Eq. (17)] is a close approximation to the average layer error rate ϵ_Ω [Eq. (8)]. In Sec. IV A we show that BiRB estimates the expected fidelity of depth-d Ω-distributed circuits. In Sec. IV B, we show that this quantity decays exponentially in d, which allows us to conclude that the BiRB error rate is approximately the average layer error rate, i.e., $r_\Omega \approx \epsilon_\Omega$.

A. Relating measurement results to circuit polarizations

We start by showing that $\bar{f}_d$ [Eq. (15)] is approximately equal to the expected polarization [Eq. (4)] of an error map consisting of the composition of (1) the error map of a depth-d Ω-distributed random circuit, and (2) an error map absorbing all SPAM error. We then argue that the contribution of SPAM errors is approximately depth independent and can be factored out, so that $\bar{f}_d$ equals the polarization [Eq. (4)] of the error map of a random, depth-d Ω-distributed random circuit, multiplied by a d-independent prefactor.

We consider a BiRB circuit C with benchmark depth d and gate-dependent error channels on $L_1, L_2, \ldots, L_d$, i.e., $\phi(L) = \mathcal{E}_L\mathcal{U}(L)$. We model the error on L_0 and state preparation as a gate-independent global depolarizing channel $\mathcal{E}_0$ directly after L_0. We model the error on L_{d+1} and readout as a single gate- and measurement-independent global depolarizing channel $\mathcal{E}_{d+1}$ occurring directly before L_{d+1}. Therefore, the superoperator representing the imperfect implementation of the circuit C is given by

$$\phi(C) = \mathcal{U}(L_{d+1})\mathcal{E}_{d+1}\mathcal{E}_{L_d}\mathcal{U}(L_d)\cdots\mathcal{E}_{L_1}\mathcal{U}(L_1)\mathcal{E}_0\mathcal{U}(L_0). \tag{18}$$

We first rewrite the error in the circuit in terms of the core circuit's error map. We have

$$\phi(C) = \mathcal{U}(L_{d+1}L_d\cdots L_1)\mathcal{E}_{\text{tot}}\mathcal{U}(L_0), \tag{19}$$

where $\mathcal{U}(L_{d+1}L_d\cdots L_1) = \mathcal{U}(L_{d+1})\mathcal{U}(L_d)\cdots\mathcal{U}(L_1)$, and $\mathcal{E}_{\text{tot}} = \mathcal{E}_{d+1}\mathcal{E}_{L_1,\ldots,L_d}\mathcal{E}_0$ for

$$\mathcal{E}_{L_1,\ldots L_d} = \mathcal{U}(L_1)^{-1}\cdots\mathcal{U}(L_d)^{-1}\mathcal{E}_{L_d}\mathcal{U}(L_d)\cdots\mathcal{E}_{L_1}\mathcal{U}(L_1). \tag{20}$$

Now we show that $\bar{f}_d$ is the expected polarization of $\mathcal{E}_{\text{tot}}$. $\bar{f}_d$ is the expectation value of circuit C's target Pauli operator s_C [Eq. (14)] averaged over all benchmark depth-d BiRB circuits, i.e.,

$$\bar{f}_d = \mathop{\mathbb{E}}_{L_1,\ldots,L_d}\mathop{\mathbb{E}}_{s\in\mathbb{P}_n^*}\mathop{\mathbb{E}}_{|\psi(s)\rangle} \text{Tr}(s_C\mathcal{U}(L_{d+1}\cdots L_1) \times \mathcal{E}_{\text{tot}}[|\psi(s)\rangle\langle\psi(s)|]), \tag{21}$$

where $|\psi(s)\rangle = U(L_0)|0\rangle^{\otimes n}$ is a uniformly random state from the set of all tensor-product states stabilized by s. Substituting in $s_c = \mathcal{U}(L_{d+1}\cdots L_1)[s]$, Eq. (21) becomes

$$\bar{f}_d = \mathop{\mathbb{E}}_{L_1,\ldots,L_d}\mathop{\mathbb{E}}_{s\in\mathbb{P}_n^*}\mathop{\mathbb{E}}_{|\psi(s)\rangle} \text{Tr}(s\mathcal{E}_{\text{tot}}[|\psi(s)\rangle\langle\psi(s)|]). \tag{22}$$

We now average over $|\psi(s)\rangle$. To do so, we first expand the initial state $|\psi(s)\rangle$ in terms of its stabilizer group:

$$\bar{f}_d = \frac{1}{2^n}\mathop{\mathbb{E}}_{L_1,\ldots,L_d}\mathop{\mathbb{E}}_{s\in\mathbb{P}_n^*}\mathop{\mathbb{E}}_{|\psi(s)\rangle} \text{Tr}\left(s\mathcal{E}_{\text{tot}}\left[\sum_{s'\in S_{|\psi(s)\rangle}} s'\right]\right) \tag{23}$$

$$= \frac{1}{2^n} \mathop{\mathbb{E}}_{L_1,\ldots,L_d} \mathop{\mathbb{E}}_{s\in\mathbb{P}_n^*} \mathrm{Tr}\left(s\mathcal{E}_{\mathrm{tot}} \left[I_n + s + \mathop{\mathbb{E}}_{|\psi(s)\rangle} \sum_{\substack{s'\in S_{|\psi(s)\rangle} \\ s'\neq I_n, s}} s' \right] \right) \tag{24}$$

$$= \frac{1}{2^n} \mathop{\mathbb{E}}_{L_1,\ldots,L_d} \mathop{\mathbb{E}}_{s\in\mathbb{P}_n^*} \left(\mathrm{Tr}(s\mathcal{E}_{\mathrm{tot}}[s]) + \mathrm{Tr}\left(s \mathop{\mathbb{E}}_{|\psi(s)\rangle} \sum_{\substack{s'\in S_{|\psi(s)\rangle} \\ s'\neq I_n, s}} \mathcal{E}_{\mathrm{tot}}[s'] \right) \right). \tag{25}$$

To get from Eq. (24) to Eq. (25), we use the fact that $\mathbb{E}_{s\in\mathbb{P}_n^*}\mathrm{Tr}(s\mathcal{E}_{\mathrm{tot}}[I_n]) = 0$ because we are averaging over signed nonidentity Pauli operators. The symmetry properties of the set of all local $+1$ eigenstates of s guarantee that the last term of Eq. (25) vanishes (see Appendix C), so that Eq. (25) becomes

$$\bar{f_d} = \frac{1}{2^n} \mathop{\mathbb{E}}_{L_1,\ldots,L_d} \mathop{\mathbb{E}}_{s\in\mathbb{P}_n^*} \mathrm{Tr}(s\mathcal{E}_{\mathrm{tot}}[s]) \tag{26}$$

$$= \mathop{\mathbb{E}}_{L_1,\ldots,L_d} \gamma(\mathcal{E}_{\mathrm{tot}}). \tag{27}$$

Equation (27) says that $\bar{f_d}$, which is measured in our protocol, is the expected polarization of $\mathcal{E}_{\mathrm{tot}}$. This error map is the composition of (1) the error map of an Ω-distributed random circuit and (2) the error maps of the state preparation and measurement layers. Because $\mathcal{E}_0$ and $\mathcal{E}_{d+1}$ are (by assumption) global depolarizing channels, we have

$$\bar{f_d} = \gamma(\mathcal{E}_0)\gamma(\mathcal{E}_{d+1}) \mathop{\mathbb{E}}_{L_1,\ldots,L_d} \gamma(\mathcal{E}_{L_1,\ldots,L_d}). \tag{28}$$

If $\mathcal{E}_0$ and $\mathcal{E}_{d+1}$ are stochastic Pauli channels (but not necessarily global depolarizing channels), or if $\mathcal{E}_{L_1,\ldots,L_d}$ is a stochastic Pauli channel, then Eq. (28) holds approximately. Specifically,

$$\bar{f_d} = \gamma(\mathcal{E}_0\mathcal{E}_{d+1}) \mathop{\mathbb{E}}_{L_1,\ldots,L_d} \gamma(\mathcal{E}_{L_1,\ldots,L_d}) + O(\varepsilon_{\mathrm{SPAM}}\varepsilon_{L_1,\ldots,L_d}), \tag{29}$$

where $\varepsilon_{\mathrm{SPAM}}$ is the infidelity of $\mathcal{E}_0\mathcal{E}_{d+1}$ and $\varepsilon_{L_1,\ldots,L_d}$ is the infidelity of $\mathcal{E}_{L_1,\ldots,L_d}$ [47]. The size of the $O(\varepsilon_{\mathrm{SPAM}}\varepsilon_{L_1,\ldots,L_d})$ term is determined by the amount of error cancellation between $\mathcal{E}_0\mathcal{E}_{d+1}$ and $\mathcal{E}_{L_1,\ldots,L_d}$ in expectation [1]. At low depths d, this correction term is small because $\varepsilon_{L_1,\ldots,L_d}$ is small, and at depths $d \gtrsim k$ [where k is the small constant in Eq. (6)], this term is small because the scrambling condition for Ω-distributed random layers implies that errors in that circuit are randomized and spread over many qubits. Equation (29) relies on the assumption of stochastic Pauli errors, and randomized compilation theory [48] implies that this can be enforced by (1) choosing Ω so that the distribution of $U(L)$ is invariant under left and right multiplication by Pauli operators, and (2) randomizing L_{d+1} and L_0. However, in practice, we find that these conditions on BiRB's circuits are not required, because Ω-distributed circuits rapidly scramble errors. This makes error cancellation negligible after constant depth k [23], implying that Eq. (28) hold to a good approximation for all kinds of small Markovian errors.

B. Deriving the exponential decay model

Our theory so far shows that $\bar{f_d}$ [Eq. (15)] is equal to the polarization of depth-d Ω-distributed random circuits multplied by a depth-independent prefactor. Recent work [1,21,23] has shown that the polarization of Ω-distributed random circuits decays exponentially—from which it follows that $r_\Omega \approx \epsilon_\Omega$—given the scrambling condition [Eq. (6)] that we require of Ω and $\mathbb{L}$. This is because Eq. (6) implies that errors within Ω-distributed random circuits cancel with negligible probability, which implies that the polarization of the BiRB core circuit is closely approximated by the product of the polarizations of its constituent layers [the error in this approximation is $O(d\varepsilon(\delta + k\varepsilon))$, which is negligible for small δ, where δ is as defined in Eq. (6) [21]]. Because the polarizations of Ω-distributed layers approximately multiply, $\bar{f_d}$ decays exponentially, i.e.,

$$\bar{f_d} \approx Ap^d \tag{30}$$

for some A and p, and $r_\Omega \approx \epsilon_\Omega$.

Here, we give an alternative, complementary proof that $\bar{f_d}$ decays exponentially, which uses the "$\mathcal{L}$ superchannel" framework from Refs. [23,24] and is similar to the most accurate theories for standard RB [22,24,25]. We start by expressing $\bar{f_d}$ [Eq. (26)] in terms of d applications of a linear operator acting on superoperators (i.e., a "superchannel"), given by

$$\mathcal{L}(\mathcal{M}) = \mathop{\mathbb{E}}_{L\in\mathbb{L}} \mathcal{U}(L)^{-1}\mathcal{M}\mathcal{E}_L\mathcal{U}(L). \tag{31}$$

When $\mathcal{E}_L = \mathbb{I}$ for all $L \in \mathbb{L}$, $\mathcal{L}$ has two unit eigenvalues (λ_0, λ_1) and all other eigenvalues (λ_i, $i > 1$) have absolute value strictly less than 1 [23]. The following theory requires that the gate errors are sufficiently small that this gap between the unit and nonunit modulus eigenvalues is preserved [49]. For this theory, we do not require that $\mathcal{E}_{d+1}$ and $\mathcal{E}_0$ are global depolarizing channels. Equation (26) can be expressed in terms of $\mathcal{L}$ as

$$\bar{f_d} = \frac{1}{2^n} \mathop{\mathbb{E}}_{s\in\mathbb{P}_n^*} \mathrm{Tr}\left(s\mathcal{L}^d(\mathcal{E}_{d+1})\left[\mathcal{E}_0[s]\right]\right). \tag{32}$$

In our theory so far, including our definition of $\mathcal{L}$ [Eq. (20)], we have used a particular representation of the imperfect gate set—the imperfect gates are given by $\{\mathcal{E}_L\mathcal{U}(L) \mid L \in \mathbb{L}\}$. However, we can express $\bar{f_d}$ in terms of a different representation of these gates with identical predictions, by performing a gauge transformation [50], i.e., we represent the gates as $\{\mathcal{M}\mathcal{E}_L\mathcal{U}(L)\mathcal{M}^{-1} \mid L \in \mathbb{L}\}$, where $\mathcal{M}$ is an invertible matrix. Below, we re-express the gate set in a particular gauge defined in terms of $\mathcal{L}$. Let $\mathcal{W} = \mathcal{E}_1 + \mathcal{E}_\lambda$, where $\mathcal{E}_1$ and $\mathcal{E}_\lambda$ are eigenoperators of $\mathcal{L}$ with eigenvalues 1 and λ, respectively (as defined in Ref. [23], Proposition 3), and where λ is the second largest eigenvalue of $\mathcal{L}$. Using the gauge-transformed gate set $\{\mathcal{W}\mathcal{E}_L\mathcal{U}(L)\mathcal{W}^{-1} \mid L \in \mathbb{L}\}$, Eq. (32) becomes

$$\bar{f_d} = \frac{1}{2^n}\mathbb{E}_{s\in\mathbb{P}_n^*}\mathrm{Tr}\left(s\tilde{\mathcal{L}}^d(\tilde{\mathcal{E}}_{d+1})[\tilde{\mathcal{E}}_0[s]]\right). \tag{33}$$

where

$$\tilde{\mathcal{L}}[C] = \mathcal{L}[C\mathcal{W}]\mathcal{W}^{-1}, \tag{34}$$

and

$$\tilde{\mathcal{E}}_{d+1} = \mathcal{E}_{d+1}\mathcal{W}^{-1} \tag{35}$$

$$\tilde{\mathcal{E}}_0 = \mathcal{W}\mathcal{E}_0. \tag{36}$$

If we assume that $\tilde{\mathcal{E}}_{d+1} = \tilde{\mathcal{D}}_{\mathrm{meas}}$, where $\tilde{\mathcal{D}}_{\mathrm{meas}}$ is a global depolarizing channel (which commutes with all unitary superoperators), it follows from Eq. (33) that

$$\bar{f_d} = \gamma\left(\tilde{\mathcal{L}}^d[\tilde{\mathcal{D}}_{\mathrm{meas}}]\tilde{\mathcal{E}}_0\right) \tag{37}$$

$$= \gamma\left(\tilde{\mathcal{D}}_{\mathrm{meas}}\tilde{\mathcal{L}}^d[\mathbb{I}]\tilde{\mathcal{E}}_0\right). \tag{38}$$

Reference [23] (Proposition 3) shows that $\mathcal{L}(\mathcal{W}) = \mathcal{D}_\lambda\mathcal{W}$, where $\mathcal{D}_\lambda$ is a global depolarizing channel with polarization λ. Therefore, $\tilde{\mathcal{L}}^d[\mathbb{I}] = \mathcal{D}_{\lambda^d}$, which implies that

$$\bar{f_d} = \gamma(\tilde{\mathcal{D}}_{\mathrm{meas}}\mathcal{D}_\lambda^d\tilde{\mathcal{E}}_0) \tag{39}$$

$$= \gamma(\tilde{\mathcal{D}}_{\mathrm{meas}}\tilde{\mathcal{E}}_0\mathcal{D}_\lambda^d) \tag{40}$$

$$= \lambda^d\gamma(\tilde{\mathcal{D}}_{\mathrm{meas}}\tilde{\mathcal{E}}_0). \tag{41}$$

Therefore, $\bar{f_d}$ decays exponentially in depth, at a rate determined by λ (the second largest eigenvalue of $\mathcal{L}$). Furthermore, Proposition 4 of Ref. [23] implies that λ is the average polarization of Ω-distributed layers computed in a particular gauge that is defined by $\mathcal{L}$.

V. SIMULATIONS

In this section, we present simulations of BiRB that show that it reliably estimates the average layer error rate ϵ_Ω.

A. BiRB with stochastic and Hamiltonian errors

To demonstrate that BiRB accurately estimates ϵ_Ω under broad conditions, we ran simulations of BiRB with varied error models containing stochastic Pauli and Hamiltonian errors. We simulated BiRB on $n = 1, 2$, and 4 qubits with all-to-all connectivity using the layer set consisting of all possible n-qubit layers constructed from parallel applications of $X_{\pi/2}$, $Y_{\pi/2}$, and CNOT gates. These layers were sampled so that the expected density of CNOT gates in a layer is $\xi = \frac{1}{4}$, and each of the two single-qubit gates appears with equal probability.

We simulated BiRB with three types of error models for these gates: (1) Pauli stochastic errors, (2) Hamiltonian errors, and (3) Pauli stochastic and Hamiltonian errors. To generate each error model, we assign each gate random error rates specified using elementary error generators [51]. For each k-qubit gate ($k = 1, 2$), we specify a post-gate error of the form $e^{\mathcal{G}}$ for each of $\{X_{\pi/2}, Y_{\pi/2}, \mathrm{CNOT}\}$, where

$$\mathcal{G} = \sum_{i=1}^{4^k-1} s_i\mathcal{S}_i + \sum_{i=1}^{4^k-1} h_i\mathcal{H}_i. \tag{42}$$

Here, $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{4^k-1}$ denote the k-qubit stochastic Pauli error generators, and $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{4^k-1}$ denote the k-qubit Hamiltonian error generators. For each error model, we sample s_i and h_i at random (see Appendix D for details) to produce a range of expected layer error rates. These models contain no crosstalk errors (but our theory encompasses error models with crosstalk errors) and no state preparation or measurement error.

Figure 4 shows the results of these simulations. Figures 4(a)–4(c) compares the true average layer error rate per qubit,

$$\epsilon_{\Omega,\mathrm{perQ}} = 1 - (1 - \epsilon_\Omega)^{\frac{1}{n}} \approx \frac{\epsilon_\Omega}{n} \tag{43}$$

to the estimate of the BiRB error rate per qubit

$$r_{\Omega,\mathrm{perQ}} = 1 - (1 - r_\Omega)^{\frac{1}{n}} \approx \frac{r_\Omega}{n} \tag{44}$$

in each simulation, separated into the three families of error models. Error bars (1σ) are shown, computed using a standard bootstrap (there are error bars on ϵ_Ω as well as on r_Ω

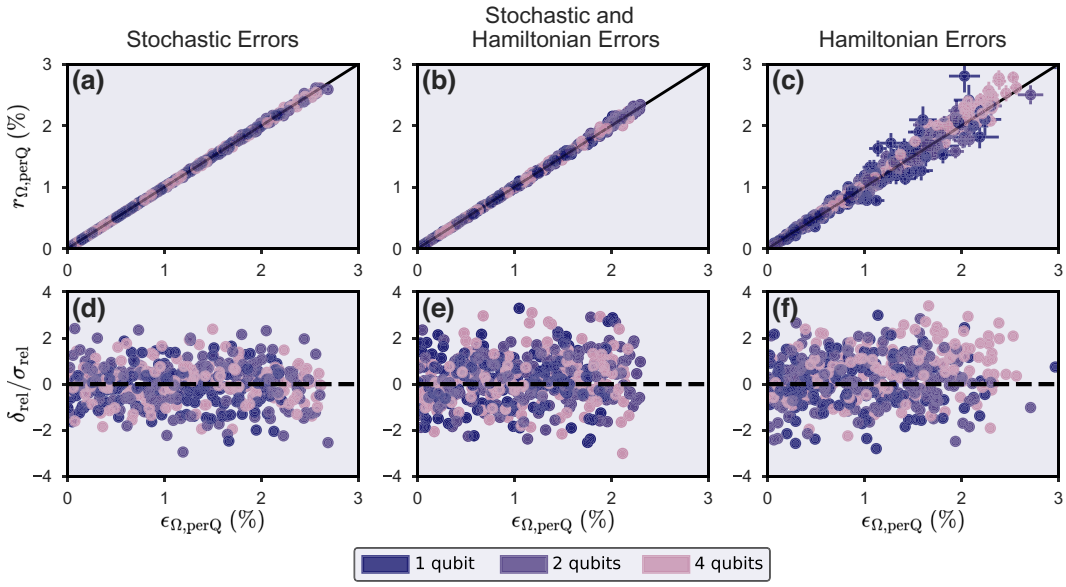

FIG. 4. Simulations of BiRB for gates with stochastic and Hamiltonian errors. We simulated BiRB on one, two, and four qubits with randomly sampled error models. These error models consist of randomly sampled (a),(d) stochastic Pauli errors, (c),(f) Hamiltonian errors, (b),(e) stochastic and Hamiltonian errors. (a)–(c) We compare the estimated BiRB error rate r_Ω to ϵ_Ω. Error bars are 1σ and are calculated using a standard bootstrap (there are error bars on ϵ_Ω, as well as r_Ω, as ϵ_Ω is estimated via sampling). (d)–(f) The relative error $\delta_{\mathrm{rel}} = (r_\Omega - \epsilon_\Omega)/\epsilon_\Omega$, divided by its standard deviation (σ_{rel}), for each randomly sampled error model. For all error models, we find that r_Ω is approximately equal to ϵ_Ω, and all discrepancies between r_Ω and ϵ_Ω are consistent with finite sample fluctuations.

because ϵ_Ω is computed by random sampling). We observe that for each error model, r_Ω approximately equals ϵ_Ω, as predicted by our theory of BiRB.

The statistical uncertainty in r_Ω (and ϵ_Ω) is typically much larger in simulations of BiRB experiments on gates with purely Hamiltonian errors, due to higher variance in the performance of circuits of the same depth for this kind of error (as is the case with other RB methods). To quantify any systematic differences between r_Ω and ϵ_Ω, in Figs. 4(d)–4(f) we show the relative error $\delta_{\mathrm{rel}} = (r_\Omega - \epsilon_\Omega)/\epsilon_\Omega$ divided by its uncertainty σ_{rel}, which is computed from 1σ uncertainties for r_Ω and ϵ_Ω. We see that r_Ω is typically within 2σ of ϵ_Ω for all three classes of error model. The distribution of δ_{rel} is similar across all error models, suggesting that BiRB is similarly reliable for all three types of error model. Furthermore, we observe that r_Ω does not systematically under- or overestimate ϵ_Ω. This contrasts with the only other method for scalable RB of Clifford gates: MRB. Simulations and theory for MRB both show that MRB systematically underestimates ϵ_Ω [1,21]. Therefore, our results suggest that BiRB is more accurate than MRB (although note that, unlike BiRB, MRB can scalably benchmark non-Clifford gates).

B. Binary RB with measurement error

The simulations presented above (Sec. V A) did not include SPAM errors, but SPAM errors are often large in current quantum processors. Like other RB protocols, BiRB is designed to be robust to SPAM errors—the effect of SPAM errors is absorbed into a depth-independent prefactor in the exponential fit (see Sec. IV). Here, we present simulations that demonstrate the robustness of BiRB in the presence of SPAM errors.

We simulated BiRB on one, two, and four qubits with single-qubit bit flip and amplitude-damping measurement errors. These BiRB simulations used the same layer set and sampling distribution as the simulations presented in Sec. V A. For these simulations, we simulated BiRB with error models in which the gates have both stochastic Pauli and Hamiltonian errors with rates sampled so that ϵ_Ω is approximately the same for every error model (see Appendix D for details). From each set of gate error rates, we construct five error models, each of which has a different type of measurement error. These five error models are as follows: (1) no error on the measurements, (2) bit-flip errors on the measurements for all n qubits, (3) bit-flip errors on the measurements for only a single qubit, (4) amplitude-damping errors on the measurements for all n qubits, and (5) amplitude-damping errors on the measurement for only a single qubit. The measurement error rates on each qubit are chosen so that the expected measurement error rate is a constant p, which we varied over a range of values (see Appendix D for details).

Figure 5 shows the results of our simulations of BiRB with measurement errors. We see that the BiRB error rate is not systematically affected by bit flip or

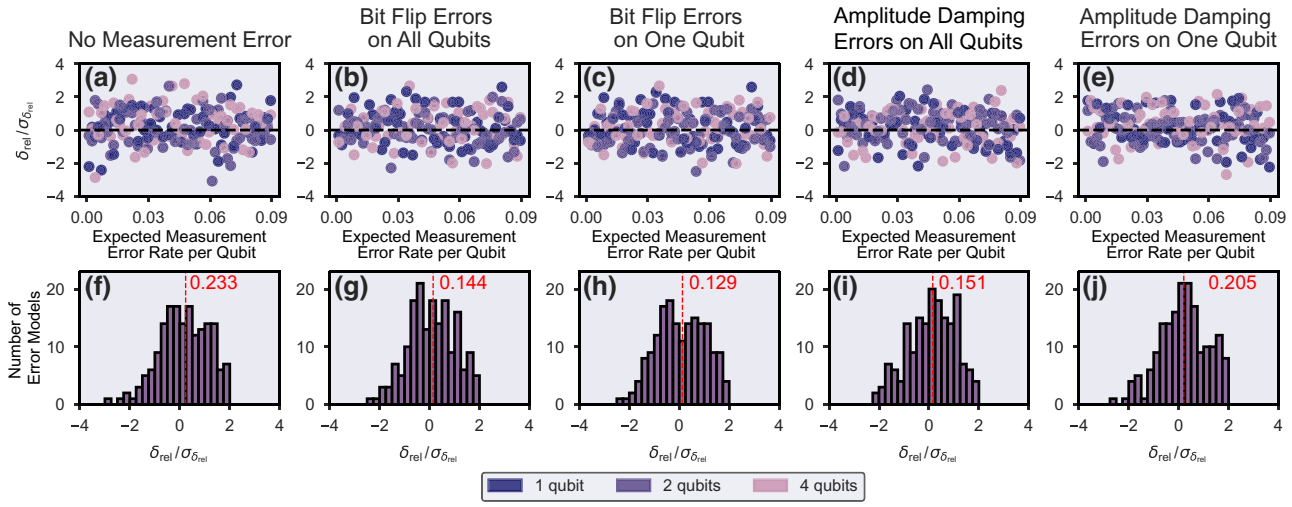

FIG. 5. Simulations of BiRB with measurement errors. We simulated BiRB with five types of error models: (a),(f) no error on the measurements, (b),(g) bit-flip errors on the measurements for all n qubits, (c),(h) bit-flip errors on the measurements for only a single qubit, (d),(i) amplitude damping errors on the measurements for all n qubits, and (e),(j) amplitude damping errors on the measurement for only a single qubit. (a)–(e) The relative error in r_Ω divided by its uncertainty ($\delta_{\text{rel}}/\sigma_{\delta_{\text{rel}}}$) versus the strength of the measurement error. (f)–(j) Histograms of $\delta_{\text{rel}}/\sigma_{\delta_{\text{rel}}}$ for each type of measurement error. We observe no evidence that r_Ω is affected by measurement error, which is consistent with our theory of BiRB and provides further evidence that BiRB is robust to measurement errors.

amplitude-damping error. Figures 5(a)–5(e) shows the relative error (δ_{rel}) in r_Ω, divided by its standard deviation ($\sigma_{\delta_{\text{rel}}}$), for all error models. We observe no systematic change in $\delta_{\text{rel}}/\sigma_{\delta_{\text{rel}}}$ as the strength of measurement error (p) is varied. Figures 5(f)–5(j) show the distribution of $\delta_{\text{rel}}/\sigma_{\delta_{\text{rel}}}$ for all error models with each type of measurement error. We see that the distributions are similar for all types of measurement error. These simulations show no evidence that r_Ω is affected by measurement error, which is consistent with our theory for BiRB.

VI. DEMONSTRATIONS ON IBM Q

In this section we present demonstrations of BiRB on 7- and 27-superconducting qubit IBM Q devices. We provide experimental evidence that BiRB works by comparing it to two other RB protocols—DRB and MRB—that are designed to measure the same error rate.

A. Validating binary RB in the few-qubit regime

We ran two experiments comparing BiRB and DRB. We chose to compare the results of BiRB and DRB because (i) DRB is designed to measure the same error rate as BiRB, (ii) BiRB is equivalent to DRB when $n = 1$, and (iii) DRB theory [18,23] shows that DRB is a highly accurate method for estimating the average error rate (ϵ_Ω). In these BiRB and DRB experiments, each layer in the core circuit consists of randomly sampled native CNOT gates (i.e., CNOT gates on connected qubits) and uniformly random single-qubit Clifford gates on all other qubits. We sampled the two-qubit gates using the "edgegrab" sampler from Ref. [2] with an expected two-qubit gate density of $\xi = \frac{1}{4}$. In our experiment on `ibm_perth`, we sampled $K = 30$ circuits at exponentially spaced benchmark depths. In our experiment on `ibm_hanoi`, we also ran CRB on up to $n = 5$ qubits, and we sampled $K = 60$ circuits at exponentially spaced benchmark depths. See Appendix E for further details.

Figures 2 and 6(a)–6(d) show the results of these demonstrations. In all of our BiRB experiments, we observe that the polarization decays exponentially. The DRB and BiRB error rates [Figs. 2(b) and 6(d)] are consistent with each other on all qubit subsets we tested [52], which is consistent with the theories for both DRB and BiRB. These results demonstrate that BiRB is a reliable method for measuring the average layer error rate. In Fig. 2(b) we also compare the BiRB error rate to an ad hoc heuristic estimate of the average layer error rate obtained by rescaling the results of CRB (see Appendix E for details). The rescaled CRB error rate is systematically higher than both the BiRB and DRB error rates. While rescaling CRB error rates to estimate native gate error rates is common practice [53–56], this is not typically accurate, as these results demonstrate.

Our results demonstrate that BiRB is more scalable than both DRB and CRB. Although DRB is more scalable than CRB [Fig. 2], the initial (i.e., $d = 0$) polarization of DRB circuits [Fig. 6(c)] still drops off rapidly with increasing n, which is due to the $O(n^2/\log n)$ gate overhead from the stabilizer state preparation and measurement subroutines [23]. The decrease in initial polarization with n for BiRB circuits is much smaller (note that we expect some

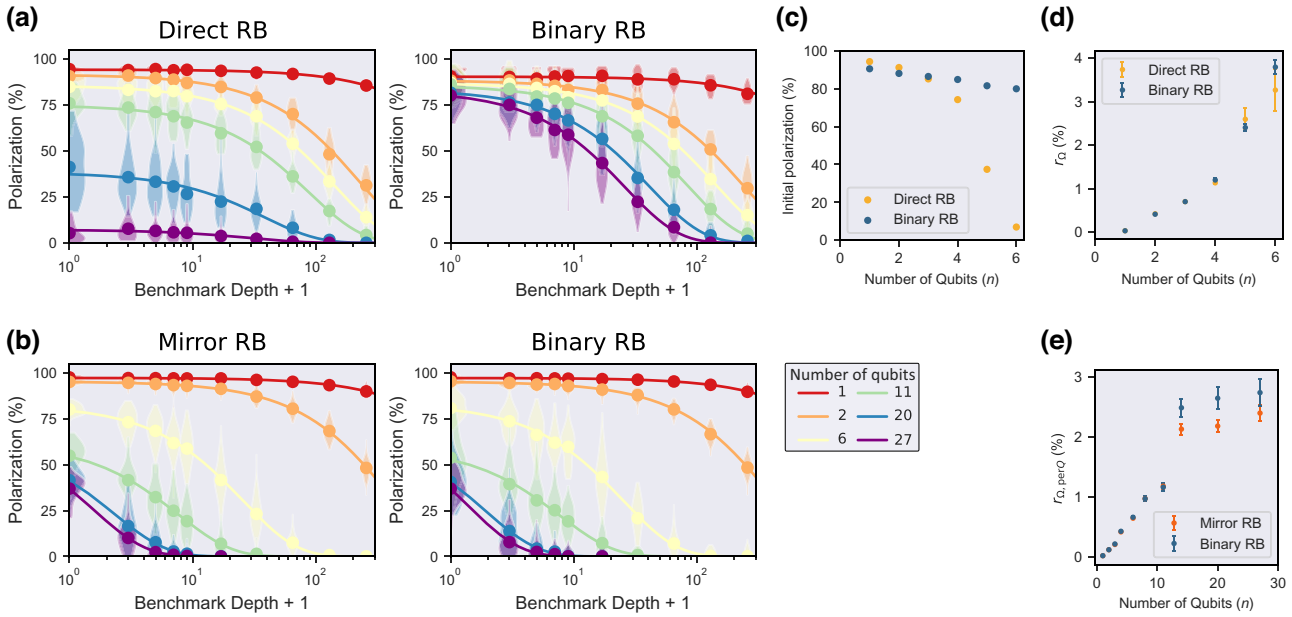

FIG. 6. BiRB on IBM Q processors. (a) The results of DRB and BiRB on 1–6 qubits on `ibm_perth`. (b) The results of MRB and BiRB on 1–27 qubits on `ibmq_kolkata`. (c) DRB's initial (i.e., $d = 0$) polarization decreases rapidly as a function of the number of qubits (n), due to the random stabilizer state preparation and measurement subroutines in DRB circuits whose size grows quickly with n [18]. In contrast, the initial polarization in BiRB decreases slowly with increasing n. (d) The BiRB and DRB error rates (r_Ω) are consistent on all qubit subsets. As DRB is a robust technique that is designed to measure the same error rate as BiRB, this provides evidence of BiRB's reliability. (e) The BiRB and MRB error rates are consistent up to $n = 11$ qubits, but the BiRB error rate is systematically higher than the MRB error rate for $n > 11$. This is consistent with the theories of BiRB and MRB: MRB theory [1,21] predicts that MRB's r_Ω slightly underestimates ϵ_Ω (the average layer error rate) whereas BiRB theory predicts that BiRB's r_Ω accurately estimates ϵ_Ω.

decrease in polarization with increased n due to increasing SPAM error).

B. Demonstrating the scalability of binary RB

To demonstrate that BiRB is reliable in the $n \gg 1$ regime, where DRB is infeasible, we ran BiRB and MRB on all 27 qubits of `ibmq_kolkata`. MRB is designed to measure the same error rate as BiRB (ϵ_Ω) and it is also scalable. However, the theory of MRB shows that it slightly but systematically underestimates ϵ_Ω due to correlations between the random layers used in MRB circuits [1,21]. MRB circuits consist of (1) a depth $d/2$ Ω-distributed random circuit, followed by (2) its layer-by-layer inverse, with Pauli frame randomization. MRB theory shows that if the error rates of a Ω-distributed layer and its inverse are uncorrelated, then MRB accurately estimates ϵ_Ω, but that if these error rates are correlated then MRB slightly underestimates ϵ_Ω. In real systems, these error rates are typically correlated.

In the BiRB and MRB circuits we ran, each randomly sampled layer in the core circuit has the form $L = L_1 L_2$, where L_1 consists of single-qubit gates on all qubits, and L_2 consists of parallel CNOT gates on pairs of connected qubits. We sampled the single-qubit gates in L_1 uniformly from the single-qubit Clifford gates, and we sampled the two-qubit gates in L_2 using the "edgegrab" sampler [2] with an expected two-qubit gate density of $\xi = \frac{1}{4}$. We ran circuits with exponentially spaced benchmark depths and sampled $K = 60$ circuits of each circuit shape.

Figure 6(b) shows the results of our BiRB and MRB experiments on six sets of qubits. Figure 6(e) compares the MRB and BiRB error rates for all sets of qubits we tested. The MRB and BiRB error rates are consistent on up to 11 qubits. For $n > 11$ qubits, the MRB error rate is systematically lower than the BiRB error rate. This result is consistent with the theory of MRB, which predicts that MRB's r_Ω systematically underestimates ϵ_Ω, and the theory of BiRB, which does not predict a systematic under- or overestimate of ϵ_Ω. MRB theory predicts that MRB's underestimate of ϵ_Ω is larger when the error rate of a layer and its inverse are highly correlated [1]. We therefore conjecture that the observed discrepancy between the BiRB and MRB error rates is caused by high variance in the layer error rates in many-qubit circuits, which could occur due to, e.g., large crosstalk error caused by some two-qubit gates. The largest difference we observe between MRB and BiRB is in the $n = 20$ qubit experiments, where the BiRB error rate is $r_\Omega \approx 41.5\%$ and the MRB error rate is $r_\Omega \approx 35.7\%$. This discrepancy is consistent with BiRB and MRB theory if the variance in the layers' error rates are sufficiently large. For example, a simple error model that

leads to the observed BiRB and MRB error rates is one where each layer experiences purely global depolarizing error and half of the layers have 85% polarization whereas the other half of the layers have 32% polarization.

VII. DISCUSSION

In this paper, we introduced BiRB, a highly streamlined RB protocol for Clifford gate sets. Unlike most RB protocols, BiRB does not use motion reversal circuits. Instead, BiRB works by tracking a single random Pauli operator through each random circuit—using ideas first developed for DFE [29,30] and later leveraged by Pauli noise learning methods [39,40,42]. This enables BiRB to scale to many more qubits than most RB methods. Many-qubit BiRB allows for benchmarking of large many-qubit layer sets when individually characterizing all those layers is infeasible, and it is able to accurately capture crosstalk. We have presented a theory for BiRB that proves that BiRB reliably estimates the average error rate of random layers under common assumptions used in RB theory (e.g., Markovian errors), and we have supported this theory with simulations and experimental demonstrations. Our results on IBM Q processors demonstrate BiRB error rates consistent with DRB error rates on up to six qubits, and they show that BiRB scales well beyond the limits of DRB and standard CRB.

BiRB enables RB on many more qubits than most existing RB protocols, but it also has advantages in the few-qubit setting. For example, simultaneous few-qubit RB experiments are widely used to quantify crosstalk errors [57], but simultaneous CRB and DRB on $n > 1$ qubits are complicated in practice by scheduling problems that arise due to the variable depths of compiled subroutines [53]. In contrast, simultaneous BiRB experiments are simple to run because the state preparation and measurement layers in BiRB circuits are each just a single layer of single-qubit gates. Finally, because BiRB does not rely on motion reversal, we anticipate that BiRB can be adapted to benchmark operations that are not intended to be unitary. In particular, in subsequent work we will show that BiRB can be adapted to benchmark gate sets containing midcircuit measurements—a computational primitive that is essential for quantum error correction.

Circuit sampling and data analysis code for BiRB is available in `pyGSTi` [58]. Data and code for the simulations and IBM *Q* demonstrations in this work are available upon reasonable request.

The DOE will provide public access to results of federally sponsored research in accordance with the DOE Public Access Plan.

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Quantum Testbed Pathfinder and the Laboratory Directed Research and Development program at Sandia National Laboratories. This research was funded, in part, by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA). Sandia National Laboratories is a multiprogram laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. We acknowledge the use of IBM Quantum services for this work. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the U.S. Department of Energy, or the U.S. Government, or IBM, or the IBM Quantum team.

This written work is authored by an employee of NTESS. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S. Government. The publisher acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for U.S. Government purposes.

The employee, not NTESS, owns the right, title and interest in and to the written work and is responsible for its contents.

APPENDIX A: CLIFFORD GROUP BINARY RB

In this Appendix, we discuss Clifford group BiRB—i.e., BiRB with uniformly random n-qubit Clifford layers. Furthermore, we use 2-design twirling theory to show that the expected polarization ($\bar{f}_d$) decays exponentially in depth for Clifford group BiRB.

1. Clifford group binary RB protocol

We start by introducing the Clifford group BiRB protocol. A depth-d Clifford group BiRB circuit is a circuit $C = C_d C_{d-1} \cdots C_1 C_0$ of $d+1$ random n-qubit Clifford gates $C_i \in \mathbb{C}_n$. The first random Clifford C_0 produces a uniformly random n-qubit stabilizer state $|\psi\rangle^{\otimes n} = C_0|0\rangle^{\otimes n}$. Picking a uniformly random nonidentity element of the stabilizer group of $|\psi\rangle$ is equivalent to picking a uniformly random element of $\mathbb{P}_n^*$, which allows us to do Clifford group BiRB without the initial tensor-product state preparation. At the end of the circuit, the evolved Pauli operator is

$$s'_C = U(C_d \cdots C_2 C_1) s U(C_1^{-1} \cdots C_{d-1}^{-1} C_d^{-1}). \tag{A1}$$

For simplicity, we will assume the ability to measure in all Pauli bases in the following discussion, so that we do

not need a final layer of gates to transform s' into a tensor product of Z and I Pauli operators. We note that the method stated here is equivalent to BiRB as stated in the main text, but with the first layer of gates L_0 recompiled into the first benchmarking layer (here, C_0). We choose to discuss this variant because it uses circuits consisting entirely of IID layers.

Our protocol is the following:

(1) For a range of integers $d \geq 0$, sample K circuits $C_d C_{d-1} \cdots C_1 C_0$. For each circuit, sample a random $s \neq I_n$ in the stabilizer group of $U(C_0)|0\rangle^{\otimes n}$.
(2) Run each circuit C $N \geq 1$ times and compute $\langle s'_C \rangle$. Then, compute the average over all circuits of benchmark depth d,

$$\bar{f_d} = \frac{1}{K} \sum_{C_d} \langle s'_C \rangle. \tag{A2}$$

(3) Fit $\bar{f_d}$ to an exponential, $\bar{f_d} = Ap^d$, where A and p are fit parameters. The RB error rate (scaled to correspond to average gate infidelity) is given by

$$r = (2^n - 1)(1 - p)/2^n. \tag{A3}$$

2. Extracting the RB error rate in Clifford group binary RB

We now show that $\bar{f_d}$ decays exponentially in benchmark depth using 2-design twirling. We assume arbitrary gate-independent Markovian error on each n-qubit Clifford—i.e., $\phi(C_i) = \mathcal{E}\mathcal{U}(C_i)$ for all C_i. An imperfect implementation of a benchmark depth-d Clifford group BiRB circuit is given by

$$\phi(C) = \phi(C_d) \cdots \phi(C_2)\phi(C_1)\phi(C_0) \tag{A4}$$

$$= \mathcal{E}\mathcal{U}(C_d)\mathcal{E}\mathcal{U}(C_{d-1}) \cdots \mathcal{U}(C_1)\mathcal{E}\mathcal{U}(C_0). \tag{A5}$$

Here, we will use S_{C_0} to denote the stabilizer group of $U(C_0)|0\rangle^{\otimes n}$. The expected polarization of a benchmark depth d circuit is

$$\bar{f_d} = \mathop{\mathbb{E}}_{C_0,\ldots,C_d} \mathop{\mathbb{E}}_{s \in S^*_{C_0}} \mathrm{Tr}\left(s'_C \phi(C)[(|0\rangle\langle 0|)^{\otimes n}]\right), \tag{A6}$$

$$= \mathop{\mathbb{E}}_{C_0,\ldots,C_d} \mathop{\mathbb{E}}_{s \in S^*_{C_0}} \mathrm{Tr}\left(\mathcal{U}(C_d \cdots C_1)[s]\phi(C)[(|0\rangle\langle 0|)^{\otimes n}]\right) \tag{A7}$$

$$= \mathop{\mathbb{E}}_{C_0,\ldots,C_d} \mathop{\mathbb{E}}_{s \in S^*_{C_0}} \mathrm{Tr}\left(s\mathcal{U}(C_1^{-1} \cdots C_d^{-1})[\phi(C)[(|0\rangle\langle 0|)^{\otimes n}]]\right). \tag{A8}$$

Equation (A7) follows from applying the definition of s'_C [Eq. (A1)] to Eq. (A6), and Eq. (A8) follows from Eq. (A7) and the cyclic property of the trace. Furthermore, we have

$$\mathcal{U}(C_1^{-1} \cdots C_d^{-1})\phi(C) = \mathcal{U}(C_1^{-1}) \cdots \mathcal{U}(C_d^{-1})\mathcal{E}\mathcal{U}(C_d) \times \mathcal{E}\mathcal{U}(C_{d-1}) \cdots \mathcal{U}(C_1)\mathcal{E}\mathcal{U}(C_0). \tag{A9}$$

Therefore, averaging over $C_1, \ldots C_d$ in Eq. (A8) twirls the error channels $\mathcal{E}$ into global depolarizing error [59]. Equation (A8) becomes

$$\bar{f_d} = \mathop{\mathbb{E}}_{C_0} \mathop{\mathbb{E}}_{s \in S^*_{C_0}} \mathrm{Tr}\left(s\tilde{\mathcal{E}}^d[\phi(C_0)[(|0\rangle\langle 0|)^{\otimes n}]]\right), \tag{A10}$$

where $\tilde{\mathcal{E}} = \mathbb{E}_{C \in \mathbb{C}_n} \mathcal{U}(C)^{-1}\mathcal{E}\mathcal{U}(C)$. Since $\mathcal{E}$ is perfectly twirled by the n-qubit Clifford group, $\tilde{\mathcal{E}}$ is an n-qubit depolarizing error channel $\tilde{\mathcal{E}}[\rho] = \gamma\rho + (1 - \gamma)I_n/2^n$. Using this result in Eq. (A10), we find that

$$\bar{f_d} = \gamma^d \mathop{\mathbb{E}}_{C_0} \mathop{\mathbb{E}}_{s \in S^*_{C_0}} \mathrm{Tr}\left(s\phi(C_0)[(|0\rangle\langle 0|)^{\otimes n}]\right) + \frac{1}{2^n}(1 - \gamma^d) \mathop{\mathbb{E}}_{C_0} \mathop{\mathbb{E}}_{s \in S^*_{C_0}} \mathrm{Tr}\left(s\phi(C_0)[I_n]\right) \tag{A11}$$

$$= A\gamma^d, \tag{A12}$$

where

$$A = \mathop{\mathbb{E}}_{C_0} \mathop{\mathbb{E}}_{s \in S^*_{C_0}} \mathrm{Tr}\left(s\phi(C_0)[(|0\rangle\langle 0|)^{\otimes n}]\right). \tag{A13}$$

Therefore, $\bar{f_d}$ decays exponentially in circuit depth, at a rate determined by the fidelity of $\mathcal{E}$. This implies that the Clifford group BiRB error rate is the same as the (standard) CRB error rate.

APPENDIX B: BINARY RB STATISTICS

In this Appendix, we show that the number of circuits required for the BiRB protocol is independent of the number of qubits n. We work in the single-shot limit, so that each measurement result is an independent random variable $f_i \in [-1, 1]$. At each circuit depth, we compute the estimate $\hat{f}_d = (1/K)\sum_{i=1}^K f_i$ of the expected polarization of benchmark depth-d BiRB circuits ($\bar{f_d}$). Hoeffding's inequality says that

$$Pr\left[|\hat{f}_d - \bar{f_d}| \geq \delta\right] \leq 2\exp\left(-\frac{1}{2}\delta^2 K\right), \tag{B1}$$

where K is the number of circuits ran. We have $\bar{f_d} \approx A\overline{\gamma}^d$, where $\overline{\gamma} = \mathbb{E}_L \gamma(\mathcal{E}_L)$ is the expected layer polarization.

Replacing δ with a relative uncertainty $\delta = \alpha A\overline{\gamma}^d$, we have

$$Pr\left[|\hat{f}_d - \bar{f}_d| \geq \alpha\overline{\gamma}^d\right] \leq 2\exp\left(-\frac{1}{2}\alpha^2 A^2\overline{\gamma}^{2d}K\right). \quad \text{(B2)}$$

The number of circuits required to obtain an estimate of $\bar{f}_d$ to within relative uncertainty α with probability at least $1-\nu$ is therefore

$$K = \frac{2\log(2/\nu)}{\alpha^2 A^2\overline{\gamma}^{2d}}. \quad \text{(B3)}$$

Importantly, this does not scale with the number of qubits n.

In order to obtain an accurate estimate of the decay rate of $\bar{f}_d$, we need to estimate $\bar{f}_d$ for at least two depths d_0, d_1 with $d_1 - d_0 = O(\log(1/\bar{\gamma}))$. For simplicity, we take $d_0 = 0$ and $d_1 = \log(1/\bar{\gamma})$. We consider the simplified scenario of estimating $\bar{\gamma}$ using the ratio of these two polarization estimates,

$$\bar{\gamma} = \left(\frac{f_{d_1}}{f_{d_0}}\right)^{\frac{1}{d_1}}. \quad \text{(B4)}$$

In order to estimate $\bar{\gamma}$ to multiplicative accuracy β, we need to estimate f_{d_1} and f_{d_0} to multiplicative accuracy $d_1\beta/2$. At depth d_1, the number of shots required is

$$K(d_1) = \frac{8\log(2/\nu)}{d_1^2\beta^2 A^2\overline{\gamma}^{2d_1}}, \quad \text{(B5)}$$

and the number of shots required at depth 0 is

$$K(d_0) = \frac{8\log(2/\nu)}{d_1^2\beta^2 A^2}. \quad \text{(B6)}$$

APPENDIX C: STABILIZERS OF TENSOR-PRODUCT STATES

Here, we prove the result used in Sec. IV A to go from Eqs. (25) to (26): For any $s, p \in \mathbb{P}_n^*$ with $s \neq \pm p$ and $[s,p] = 0$, there is a bijection between tensor-product stabilizer states $|\psi(s)\rangle$ that are stabilized by p and tensor-product stabilizer states $|\psi(s)\rangle$ that are stabilized by $-p$.

To construct our bijection, we pick an ordering of the n qubits, and we express s and p as tensor products of single qubit Pauli operators—i.e., $s = \otimes_{i=0}^n s_i$ and $p = \otimes_{i=0}^n p_i$. For any $|\psi(s)\rangle = L_0|0\rangle^{\otimes n}$ satisfying $s'|\psi(s)\rangle = |\psi(s)\rangle$, we can create another tensor-product state $|\psi'(s)\rangle = L_0'|0\rangle^{\otimes n}$ satisfying $-p'|\psi'(s)\rangle = |\psi'(s)\rangle$ as follows: we start with $L_0' = L_0$ and modify some of the single-qubit gates in L_0'. Find the lowest index j such that $s_j = I$ and $p_j \neq I$, if it exists. There are two cases:

(1) Such a j exists: because $p_j \neq I$, the gate on qubit j produces a $+1$ eigenstate of some $q \in \mathbb{P}_n^*$. Replace this gate in L_0' with a gate that produces a $+1$ eigenstate of $-q$.

(2) No such j exists: if there is no j such that $s_j = I$ and $p_j \neq I$, then there must be a qubit j such that $s_j \neq I$ and $p_j = I$. The gate on qubit j produces a $+1$ eigenstate of some $q \in \mathbb{P}_n^*$. Replace this gate in L_0' with a gate that produces a $+1$ eigenstate of $-q$. In addition, because we know that $p \neq I$, there must be some other qubit j' such that $p_{j'} \neq I$, and it must be that $s_{j'} = p_{j'}$. Suppose the gate on qubit j' produces a $+1$ eigenstate of some $q' \in \mathbb{P}_n^*$. Replace this gate with a gate that produces a $+1$ eigenstate of $-q'$.

The new state $|\psi'(s)\rangle$ produced by the new layer of gate is stabilized by s and $-p$, and this mapping is bijective. From this result it follows that

$$\sum_{\substack{s' \in S_\psi \\ s' \neq I, s}} s' = 0, \quad \text{(C1)}$$

from which Eq. (26) follows.

APPENDIX D: SIMULATIONS OF BINARY RB

1. Binary RB with stochastic Pauli and Hamiltonian errors

We simulated BiRB on $n = 1, 2, 4$ qubits with all-to-all connectivity using layers constructed from the gate set $\{X_{\pi/2}, Y_{\pi/2}, \text{CNOT}\}$. The error models we use in our BiRB simulations are defined in terms of the stochastic and Hamiltonian elementary error generators defined in Ref. [51]. For each k-qubit gate ($k = 1, 2$), we specify a post-gate error of the form $e^{\mathcal{G}}$ for each of $\{X_{\pi/2}, Y_{\pi/2}, \text{CNOT}\}$, where

$$\mathcal{G} = \sum_{i=1}^{4^k-1} s_i\mathcal{S}_i + \sum_{i=1}^{4^k-1} h_i\mathcal{H}_i, \quad \text{(D1)}$$

where $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{4^k-1}$ denote the k-qubit stochastic Pauli error generators, and $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{4^k-1}$ denote the k-qubit stochastic Pauli error generators. For each error model, we sample s_i and h_i at random to produce a range of expected layer error rates. To generate error models, we start with an overall error parameter p that determines the expected gate error rates in the model. We generate models with $p \in [0, 0.01875]$ for 150 evenly spaced values for the single-qubit models and $p \in [0, 0.0750]$ for 150 evenly spaced values for the two- and four-qubit models. We use p to determine the expected rates of stochastic and Hamiltonian errors. In the stochastic Pauli error models, we set $h = 0$ and $s = 1.2p$. In the Hamiltonian error models, we set $s = 0$ and $h = \sqrt{8p}$ for $n = 4$ qubit models, and we set $s = 0$ and $= \sqrt{6p}$ for $n = 1, 2$ qubit models. In the stochastic Pauli and Hamiltonian error models, we

TABLE I. BiRB and DRB on IBM Perth. The RB error rates from every BiRB and DRB experiment we ran on `ibm_perth`.

Qubit subset	r_Ω (BiRB)	r_Ω (DRB)
Q0	0.027(1)	0.0279(5)
(Q0, Q1)	0.41(1)	0.42(1)
(Q0, Q1, Q2)	0.70(2)	0.70(1)
(Q0, Q1, Q2, Q3)	1.20(4)	1.13(3)
(Q0, Q1, Q2, Q3, Q5)	2.40(9)	2.6(2)
(Q0, Q1, Q2, Q3, Q5, Q6)	3.8(2)	3.3(4)
(Q0, Q1, Q2, Q3, Q4, Q5, Q6)	5.6(3)	fit failed

generate $s \in [0, p]$ at random and set $h = \sqrt{2p - s}$. These sampling parameters are chosen to produce models with a similar range of per-qubit error rates across all error model types and all values of n.

We include qubit-dependent Hamiltonian errors and stochastic Pauli errors on each gate, with Hamiltonian error rates sampled in the range $[0, \chi h]$, and stochastic Pauli error rates sampled in the range $[0, \chi s]$, where $\chi = 0.1$ if $n = 2, 4$ and $k = 1$ (i.e., single-qubit gate error rates are sampled so that their expected error rate is $\frac{1}{10}$ the error rate of two-qubit gates) and $\chi = 1$ otherwise. The stochastic and Hamiltonian errors are each split randomly across the $4^k - 1$ error generators.

For each error model, we run $K = 100$ BiRB circuits at each depth $d \in \{0\} \cup \{2^j \mid 0 \leq j \leq 8\}$. Each layer in the core circuit consists of randomly sampled CNOT gates and uniformly random gates from the set $\{X_{\pi/2}, Y_{\pi/2}, I\}$ on all other qubits. We sampled the two-qubit gates using the "edgegrab" sampler from Ref. [2] with an expected two-qubit gate density of $\xi = \frac{1}{2}$.

We also approximate the average layer error rate ϵ_Ω via sampling. We sample $K = 100$ Ω-distributed random circuits at each depth $d \in \{0\} \cup \{2^j \mid 0 \leq j \leq 8\}$ (using the same layer sampling as described above) and determine their polarization, then fit the resulting data to an exponential to obtain an estimate of ϵ_Ω.

2. Binary RB with measurement errors

We simulated BiRB on $n = 1, 2, 4$ qubits with single-qubit bit flip and amplitude damping measurement error. These BiRB circuits used layers constructed from the gates $\{X_{\pi/2}, Y_{\pi/2}, \text{CNOT}\}$. In these simulations, we simulated BiRB with error models in which the gates have both stochastic Pauli and Hamiltonian errors. We generated 30 models with Hamiltonian and stochastic errors. Each error model had randomly chosen error rates sampled so that the expected stochastic error rate was $p/2$ and the expected Hamiltonian error rate was $\sqrt{p/2}$, and we set $p = 0.015n$. In our $n = 2, 4$ qubit simulations, we sampled the errors on single-qubit gates so that their expected error rates were approximately $\frac{1}{10}$ of the expected two-qubit gate error rate.

From each set of gate error rates, we construct five error models, each of which has different measurement errors. These five error models are as follows: (1) no error on the measurements, (2) bit-flip errors on the measurements for all n qubits, (3) bit-flip errors on the measurements for only a single qubit, (4) amplitude-damping errors on the measurements for all n qubits, and (5) amplitude-damping errors on the measurement for only a single qubit. We define our measurement error using the single-qubit elementary error generators S_X, S_Y, and $A_{X,Y}$ defined in Ref. [51], and an error strength parameter p_m. In our bit-flip error models, we add the error $\mathcal{E} = e^{p_m S_X}$ immediately before the measurement. In our amplitude damping error models, we add the error $\mathcal{E} = e^{p_m (S_X + S_Y + A_{X,Y})}$ immediately before measurement. In error models with measurement error on a single qubit, we generate error models with 60 evenly spaced values of $p_m \in [0.0001, 0.09]$. In error models with measurement error on all qubits we sample a uniform random $p_m \in [0, 2p/n]$ independently for each qubit, for 60 evenly spaced values of $p \in [0.0001, 0.09]$,

For each error model, we run $K = 100$ BiRB circuits at each depth $d \in \{0\} \cup \{2^j \mid 1 \leq j \leq 8\}$ using the same gate set and layer sampling distribution as in Appendix D 1. We approximate ϵ_Ω via sampling using the method described in Appendix D 1.

APPENDIX E: DEMONSTRATING BINARY RB ON IBM Q

1. Details of demonstration on IBM Hanoi

We ran BiRB and other RB protocols on `ibm_hanoi`, `ibm_perth`, and `ibmq_kolkata`. In this Appendix,

TABLE II. IBM Perth calibration data. Calibration data from `ibm_perth` from the time of our BiRB demonstrations.

Qubit	T_1 (us)	T_2 (us)	Frequency (GHz)	Anharmonicity (GHz)	Readout error	Pr(prep 1, measure 0)	Pr(prep 0, measure 1)	Readout length (ns)
Q0	122.58	84.49	5.16	−0.34	0.019	0.018	0.020	675.56
Q1	96.44	36.20	5.03	−0.34	0.019	0.022	0.016	675.56
Q2	298.92	61.66	4.86	−0.35	0.010	0.011	0.009	675.56
Q3	170.68	179.99	5.13	−0.34	0.012	0.016	0.008	675.56
Q4	77.10	109.16	5.16	−0.33	0.012	0.011	0.012	675.56
Q5	148.18	69.49	4.98	−0.35	0.014	0.015	0.013	675.56
Q6	167.33	245.76	5.16	−0.34	0.007	0.008	0.006	675.56

TABLE III. BiRB and MRB on IBMQ Kolkata. The RB error rates for every BiRB and MRB experiment we ran on `ibmq_kolkata`.

Qubit subset	r_Ω (BiRB)	r_Ω (MRB)
Q0	0.0230(4)	0.0228(3)
(Q0, Q1)	0.245(3)	0.246(3)
(Q0, Q1, Q2)	0.65(1)	0.63(1)
(Q0, Q1, Q2, Q3)	1.71(4)	1.67(4)
(Q0, Q1, Q2, Q3, Q4, Q5)	3.9(1)	3.8(1)
(Q0, Q1, Q2, Q3, Q4, Q5, Q6, Q7)	7.5(3)	7.5(3)
(Q0, Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10)	12.1(6)	12.2(4)
(Q0, Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10, Q11, Q12, Q13)	30(2)	26.0(9)
(Q0, Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10, Q11, Q12, Q13, Q14, Q15, Q16, Q17, Q18, Q19)	42(2)	36(1)
(Q0, Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10, Q11, Q12, Q13, Q14, Q15, Q16, Q17, Q18, Q19, Q20, Q21, Q22, Q23, Q24, Q25, Q26)	53(3)	48(2)

we provide the details of our RB experiments on `ibm_hanoi`. Details of our experiments on `ibm_perth` and `ibmq_kolkata` can be found in Sec. VI.

We ran DRB, BiRB, and CRB on `ibm_hanoi` [Fig. 2]. For our DRB and BiRB experiments, we benchmarked a gate set consisting of the 24 single-qubit Clifford gates

TABLE IV. IBMQ Kolkata calibration data. Calibration data from `ibmq_kolkata` from the time of our BiRB demonstrations.

Qubit	T_1 (us)	T_2 (us)	Frequency (GHz)	Anharmonicity (GHz)	Readout error	Pr(prep 1, measure 0)	Pr(prep 0, measure 1)	Readout length (ns)
Q0	126.85	24.35	5.20	−0.34	0.011	0.010	0.012	675.56
Q1	152.85	147.77	4.99	−0.35	0.014	0.013	0.014	675.56
Q2	87.82	37.07	5.11	−0.34	0.012	0.016	0.008	675.56
Q3	113.23	47.66	4.87	−0.35	0.043	0.061	0.026	675.56
Q4	120.73	104.49	5.22	−0.34	0.031	0.032	0.031	675.56
Q5	120.73	42.50	5.10	−0.34	0.027	0.026	0.027	675.56
Q6	118.32	65.44	5.20	−0.34	0.031	0.028	0.034	675.56
Q7	128.19	23.16	5.02	−0.35	0.065	0.026	0.104	675.56
Q8	171.87	193.75	4.96	−0.35	0.017	0.022	0.013	675.56
Q9	168.79	57.68	5.06	−0.34	0.155	0.165	0.144	675.56
Q10	141.42	67.37	5.18	−0.34	0.017	0.017	0.018	675.56
Q11	11.32	17.23	4.88	−0.37	0.042	0.052	0.031	675.56
Q12	140.45	54.01	4.96	−0.35	0.015	0.025	0.006	675.56
Q13	130.20	154.80	5.02	−0.35	0.011	0.012	0.011	675.56
Q14	165.62	120.93	5.12	−0.34	0.007	0.009	0.004	675.56
Q15	150.20	162.20	5.03	−0.34	0.008	0.009	0.006	675.56
Q16	88.63	63.20	5.22	−0.34	0.017	0.014	0.019	675.56
Q17	100.91	33.29	5.24	−0.34	0.006	0.007	0.004	675.56
Q18	120.14	57.40	5.09	−0.34	0.011	0.008	0.014	675.56
Q19	132.28	117.35	5.00	−0.34	0.011	0.013	0.009	675.56
Q20	135.13	155.12	5.19	−0.34	0.008	0.010	0.007	675.56
Q21	115.65	103.94	5.27	−0.34	0.005	0.005	0.005	675.56
Q22	149.22	42.50	5.12	−0.34	0.010	0.013	0.008	675.56
Q23	153.04	129.84	5.14	−0.34	0.006	0.007	0.006	675.56
Q24	136.85	30.46	5.00	−0.35	0.011	0.017	0.005	675.56
Q25	266.91	163.80	4.92	−0.35	0.007	0.011	0.004	675.56
Q26	138.55	100.89	5.12	−0.34	0.007	0.010	0.003	675.56

TABLE V. BiRB, DRB, and CRB on IBM Hanoi. The error rates for every RB experiment we ran on `ibm_hanoi`. We ran CRB on up to five qubits, we ran DRB on up to six qubits, and we ran BiRB on up to 20 qubits.

Qubit subset	r_Ω (BiRB)	r_Ω (DRB)	r (CRB)
Q0	0.0224(3)	0.0222(3)	0.0318(3)
(Q0, Q1)	0.296(5)	0.300(5)	2.13(3)
(Q0, Q1, Q2)	0.456(7)	0.445(8)	10.9(2)
(Q0, Q1, Q2, Q3)	0.94(2)	0.98(3)	55(2)
(Q0, Q1, Q2, Q3, Q4)	1.61(3)	1.53(4)	99(1)
(Q0, Q1, Q2, Q3, Q4, Q5)	1.82(4)	1.9(2)	
(“Q0”, “Q1”, “Q2”, “Q3”, “Q4”, “Q5”, “Q6”, “Q7”, “Q10”, “Q12”, “Q13”, “Q14”)	7.59(3)		
(“Q0”, “Q1”, “Q2”, “Q3”, “Q4”, “Q5”, “Q6”, “Q7”, “Q10”, “Q12”, “Q13”, “Q14”, “Q16”, “Q19”, “Q22”, “Q25”)	11.9(5)		
(“Q0”, “Q1”, “Q2”, “Q3”, “Q4”, “Q5”, “Q6”, “Q7”, “Q10”, “Q12”, “Q13”, “Q14”, “Q19”, “Q16”, “Q21”, “Q22”, “Q25”, “Q24”, “Q23”, “Q26”)	21.0(7)		

and CNOT. Each benchmarking layer consisted of random CNOT gates, respecting the device connectivity, and random single-qubit gates on all other qubits. The CNOT gates were sampled using edgegrab sampling with expected two-qubit gate density of $\xi = \frac{1}{4}$. We ran $K = 60$ circuits at exponentially spaced benchmark depths for each of DRB,

TABLE VI. IBM Hanoi calibration data. Calibration data from `ibm_hanoi` from the time of our BiRB demonstrations.

Qubit	T_1 (us)	T_2 (us)	Frequency (GHz)	Anharmonicity (GHz)	Readout error	Pr(prep 1, measure 0)	Pr(prep 0, measure 1)	Readout length (ns)
Q0	170.13	240.96	5.04	−0.34	0.010	0.012	0.007	817.78
Q1	119.38	125.05	5.16	−0.34	0.013	0.013	0.013	817.78
Q2	139.61	206.70	5.26	−0.34	0.014	0.018	0.010	817.78
Q3	120.10	32.35	5.10	−0.34	0.011	0.014	0.007	817.78
Q4	196.74	17.02	5.07	−0.34	0.006	0.006	0.007	817.78
Q5	148.10	186.60	5.21	−0.34	0.006	0.008	0.004	817.78
Q6	97.99	143.88	5.02	−0.34	0.024	0.027	0.021	817.78
Q7	177.40	255.79	4.92	−0.35	0.012	0.014	0.010	817.78
Q8	205.98	341.43	5.03	−0.34	0.012	0.012	0.011	817.78
Q9	96.37	208.45	4.87	−0.35	0.008	0.012	0.004	817.78
Q10	54.74	55.26	4.82	−0.35	0.020	0.021	0.020	817.78
Q11	150.80	259.36	5.16	−0.34	0.077	0.075	0.080	817.78
Q12	96.62	175.47	4.72	−0.35	0.173	0.215	0.130	817.78
Q13	241.95	274.17	4.96	−0.34	0.047	0.045	0.050	817.78
Q14	130.40	23.22	5.05	−0.34	0.009	0.011	0.007	817.78
Q15	80.32	35.15	4.92	-0.32	0.029	0.023	0.035	817.78
Q16	194.51	316.64	4.88	−0.35	0.009	0.010	0.008	817.78
Q17	152.12	66.53	5.22	−0.34	0.018	0.019	0.016	817.78
Q18	155.96	138.19	4.97	−0.35	0.012	0.017	0.006	817.78
Q19	201.17	238.68	5.00	−0.35	0.006	0.007	0.004	817.78
Q20	170.75	67.15	5.10	−0.34	0.006	0.005	0.006	817.78
Q21	128.89	31.88	4.84	−0.35	0.008	0.011	0.005	817.78
Q22	198.10	108.40	4.92	−0.35	0.011	0.012	0.010	817.78
Q23	173.22	256.72	4.92	−0.34	0.011	0.008	0.014	817.78
Q24	158.95	36.15	4.99	−0.34	0.007	0.008	0.005	817.78
Q25	158.51	47.02	4.81	−0.35	0.010	0.010	0.009	817.78
Q26	79.50	28.58	5.02	−0.34	0.008	0.010	0.006	817.78

BiRB, and CRB. We randomized the order of the circuit list and ran each circuit with 1000 shots.

The CRB error rate is an estimate of the average error rate of a (compiled) n-qubit Clifford gate. To directly compare the CRB error rate to the DRB and BiRB error rates, we use a heuristic to approximate the average error of a layer from the distribution Ω we used to sample the DRB and BiRB circuits. Our estimate for the n-qubit layer error rate is $r_{\Omega,\mathrm{est}} = (r_n/k_n)n\xi$, where r_n is the n-qubit CRB error rate and k_n is the average number of two-qubit gates per n-qubit Clifford.

2. RB error rates and calibration data

Here, we provide the RB error rates and device calibration data from all of our RB experiments on `ibm_perth` [Tables I and II], `ibmq_kolkata` [Tables III and IV], and `ibm_hanoi` [Tables V and VI].

[1] Timothy Proctor, Stefan Seritan, Kenneth Rudinger, Erik Nielsen, Robin Blume-Kohout, and Kevin Young, Scalable randomized benchmarking of quantum computers using mirror circuits, Phys. Rev. Lett. **129**, 150502 (2022).

[2] Timothy Proctor, Kenneth Rudinger, Kevin Young, Erik Nielsen, and Robin Blume-Kohout, Measuring the capabilities of quantum computers, Nat. Phys. **18**, 75 (2022).

[3] David C. McKay, Andrew W. Cross, Christopher J. Wood, and Jay M. Gambetta, Correlated randomized benchmarking, ArXiv:2003.02354.

[4] Joseph Emerson, Robert Alicki, and Karol Życzkowski, Scalable noise estimation with random unitary operators, J. Opt. B Quantum Semiclass. Opt. **7**, S347 (2005).

[5] Joseph Emerson, Marcus Silva, Osama Moussa, Colm Ryan, Martin Laforest, Jonathan Baugh, David G. Cory, and Raymond Laflamme, Symmetrized characterization of noisy quantum processes, Science **317**, 1893 (2007).

[6] Easwar Magesan, Jay M. Gambetta, and Joseph Emerson, Scalable and robust randomized benchmarking of quantum processes, Phys. Rev. Lett. **106**, 180504 (2011).

[7] Easwar Magesan, Jay M. Gambetta, and Joseph Emerson, Characterizing quantum gates via randomized benchmarking, Phys. Rev. A **85**, 042311 (2012).

[8] Emanuel Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Randomized benchmarking of quantum gates, Phys. Rev. A **77**, 012307 (2008).

[9] Arnaud Carignan-Dugas, Joel J. Wallman, and Joseph Emerson, Characterizing universal gate sets via dihedral benchmarking, Phys. Rev. A **92**, 060302 (2015).

[10] Andrew W. Cross, Easwar Magesan, Lev S. Bishop, John A. Smolin, and Jay M. Gambetta, Scalable randomised benchmarking of non-Clifford gates, npj Quantum Inf. **2**, 16012 (2016).

[11] Winton G. Brown and Bryan Eastin, Randomized benchmarking with restricted gate sets, Phys. Rev. A **97**, 062323 (2018).

[12] A. K. Hashagen, S. T. Flammia, D. Gross, and J. J. Wallman, Real randomized benchmarking, Quantum **2**, 85 (2018).

[13] Jonas Helsen, Xiao Xue, Lieven M. K. Vandersypen, and Stephanie Wehner, A new class of efficient randomized benchmarking protocols, npj Quantum Inf. **5**, 71 (2019).

[14] Jonas Helsen, Sepehr Nezami, Matthew Reagor, and Michael Walter, Matchgate benchmarking: Scalable benchmarking of a continuous family of many-qubit gates, Quantum **6**, 657 (2022).

[15] Jahan Claes, Eleanor Rieffel, and Zhihui Wang, Character randomized benchmarking for non-multiplicity-free groups with applications to subspace, leakage, and matchgate randomized benchmarking, PRX Quantum **2**, 010351 (2021).

[16] J. Helsen, I. Roth, E. Onorati, A. H. Werner, and J. Eisert, General framework for randomized benchmarking, PRX Quantum **3**, 020357 (2022).

[17] A. Morvan, V. V. Ramasesh, M. S. Blok, J. M. Kreikebaum, K. O'Brien, L. Chen, B. K. Mitchell, R. K. Naik, D. I. Santiago, and I. Siddiqi, Qutrit randomized benchmarking, Phys. Rev. Lett. **126**, 210504 (2021).

[18] Timothy J. Proctor, Arnaud Carignan-Dugas, Kenneth Rudinger, Erik Nielsen, Robin Blume-Kohout, and Kevin Young, Direct randomized benchmarking for multiqubit devices, Phys. Rev. Lett. **123**, 00 (2019).

[19] Andrew W. Cross, Lev S. Bishop, Sarah Sheldon, Paul D. Nation, and Jay M. Gambetta, Validating quantum computers using randomized model circuits, Phys. Rev. A **100**, 032328 (2019).

[20] Karl Mayer, Alex Hall, Thomas Gatterman, Si Khadir Halit, Kenny Lee, Justin Bohnet, Dan Gresh, Aaron Hankin, Kevin Gilmore, and John Gaebler, Theory of mirror benchmarking and demonstration on a quantum computer, ArXiv:2108.10431.

[21] Jordan Hines, Marie Lu, Ravi K. Naik, Akel Hashim, Jean-Loup Ville, Brad Mitchell, John Mark Kriekebaum, David I. Santiago, Stefan Seritan, Erik Nielsen, Robin Blume-Kohout, Kevin Young, Irfan Siddiqi, Birgitta Whaley, and Timothy Proctor, Demonstrating scalable randomized benchmarking of universal gate sets, Phys. Rev. X **13**, 041030 (2023).

[22] Joel J. Wallman, Randomized benchmarking with gate-dependent noise, Quantum **2**, 47 (2018).

[23] Anthony M. Polloreno, Arnaud Carignan-Dugas, Jordan Hines, Robin Blume-Kohout, Kevin Young, and Timothy Proctor, A theory of direct randomized benchmarking, ArXiv:2302.13853.

[24] Timothy Proctor, Kenneth Rudinger, Kevin Young, Mohan Sarovar, and Robin Blume-Kohout, What randomized benchmarking actually measures, Phys. Rev. Lett. **119**, 130502 (2017).

[25] Seth T. Merkel, Emily J. Pritchett, and Bryan H. Fong, Randomized benchmarking as convolution: Fourier analysis of gate dependent errors, Quantum **5**, 581 (2021).

[26] Scott Aaronson and Daniel Gottesman, Improved simulation of stabilizer circuits, Phys. Rev. A **70**, 052328 (2004).

[27] Ketan N. Patel, Igor L. Markov, and John P. Hayes, Efficient synthesis of linear reversible circuits, Quantum Inf. Comput. **8**, 282 (2008).

[28] Sergey Bravyi and Dmitri Maslov, Hadamard-free circuits expose the structure of the Clifford group, IEEE Trans. Inf. Theory **67**, 4546 (2021).

[29] Steven T. Flammia and Yi-Kai Liu, Direct fidelity estimation from few Pauli measurements, Phys. Rev. Lett. **106**, 230501 (2011).

[30] Marcus P. da Silva, Olivier Landon-Cardinal, and David Poulin, Practical characterization of quantum devices without tomography, Phys. Rev. Lett. **107**, 210404 (2011).

[31] Sergio Boixo, Sergei V. Isakov, Vadim N. Smelyanskiy, Ryan Babbush, Nan Ding, Zhang Jiang, Michael J. Bremner, John M. Martinis, and Hartmut Neven, Characterizing quantum supremacy in near-term devices, Nat. Phys. **14**, 595 (2018).

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

[33] Yunchao Liu, Matthew Otten, Roozbeh Bassirianjahromi, Liang Jiang, and Bill Fefferman, Benchmarking near-term quantum computers via random circuit sampling, ArXiv:2105.05232.

[34] Jianxin Chen, Dawei Ding, Cupjin Huang, and Linghang Kong, Linear cross entropy benchmarking with Clifford circuits, ArXiv:2206.08293.

[35] Markus Heinrich, Martin Kliesch, and Ingo Roth, Randomized benchmarking with random quantum circuits, ArXiv:2212.06181.

[36] Xun Gao, Marcin Kalinowski, Chi-Ning Chou, Mikhail D. Lukin, Boaz Barak, and Soonwon Choi, Limitations of linear cross-entropy as a measure for quantum advantage, PRX Quantum **5**, 010334 (2024).

[37] Michael A. Nielsen, A simple formula for the average gate fidelity of a quantum dynamical operation, Phys. Lett. A **303**, 249 (2002).

[38] Arnaud Carignan-Dugas, Kristine Boone, Joel J. Wallman, and Joseph Emerson, From randomized benchmarking experiments to gate-set circuit fidelity: How to interpret randomized benchmarking decay parameters, New J. Phys. **20**, 092001 (2018).

[39] Alexander Erhard, Joel James Wallman, Lukas Postler, Michael Meth, Roman Stricker, Esteban Adrian Martinez, Philipp Schindler, Thomas Monz, Joseph Emerson, and Rainer Blatt, Characterizing large-scale quantum computers via cycle benchmarking, Nat. Commun. **10**, 5347 (2019).

[40] Steven T. Flammia, in *17th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2022)*, edited by F. Le Gall and T. Morimae (Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2022), Vol. 232, pp. 4:1–4:10.

[41] Robin Harper, Steven T. Flammia, and Joel J. Wallman, Efficient learning of quantum noise, Nat. Phys. **16**, 1 (2020).

[42] Steven T. Flammia and Joel J. Wallman, Efficient estimation of Pauli channels, ACM Trans. Quantum Comput. **1**, 3 (2020).

[43] We choose this definition of r_Ω so that it corresponds to the average layer entanglement infidelity. It is also acceptable to set $r_\Omega = (2^n - 1)(1 - p)/2^n$, so that it corresponds to average gate infidelity.

[44] Robin Harper, Ian Hincks, Chris Ferrie, Steven T. Flammia, and Joel J. Wallman, Statistical analysis of randomized benchmarking, Phys. Rev. A **99**, 052350 (2019).

[45] Joel J. Wallman and Steven T. Flammia, Randomized benchmarking with confidence, New J. Phys. **16**, 103032 (2014).

[46] Alex Kwiatkowski, Laurent J. Stephenson, Hannah M. Knaack, Alejandra L. Collopy, Christina M. Bowers, Dietrich Leibfried, Daniel H. Slichter, Scott Glancy, and Emanuel Knill, Optimized experiment design and analysis for fully randomized benchmarking, ArXiv:2312.15836.

[47] If $\mathcal{E}_0\mathcal{E}_{d+1}$ is a stochastic Pauli channel, then $\gamma(\mathcal{E}_0\mathcal{E}_{d+1}\mathcal{E}_{L_1,\ldots,L_d}) = \mathbb{E}_{P\in\mathbb{P}_n}\gamma(\mathcal{P}\,\mathcal{E}_0\,\mathcal{E}_{d+1}\mathcal{P}^\dagger\mathcal{E}_{L_1,\ldots,L_d}) = \gamma(\mathcal{E}_0\,\mathcal{E}_{d+1}(\mathbb{E}_{P\in\mathbb{P}_n}\mathcal{P}^\dagger\mathcal{E}_{L_1,\ldots,L_d}\mathcal{P}))$. Equation (29) then follows because $\mathcal{E}_0\mathcal{E}_{d+1}$ and $\mathbb{E}_{P\in\mathbb{P}_n}\mathcal{P}^\dagger\mathcal{E}_{L_1,\ldots,L_d}\mathcal{P}$ are stochastic Pauli channels [1].

[48] Joel J. Wallman and Joseph Emerson, Noise tailoring for scalable quantum computation via randomized compiling, Phys. Rev. A **94**, 052325 (2016).

[49] The spectral gap is known to not close when n is small, because Eq. (6) implies that Ω-distributed random circuits rapidly converge to a 2-design. For $n \gg 1$, Ω-distributed random circuits will typically not rapidly converge to a 2-design. Other proofs of exponential decay based on arguing that the polarizations of layer multiply (assuming highly scrambling layers) do not require this fact, and they apply in the $n \gg 1$ regime.

[50] Erik Nielsen, John King Gamble, Kenneth Rudinger, Travis Scholten, Kevin Young, and Robin Blume-Kohout, Gate set tomography, Quantum **5**, 557 (2021).

[51] Robin Blume-Kohout, Marcus P. da Silva, Erik Nielsen, Timothy Proctor, Kenneth Rudinger, Mohan Sarovar, and Kevin Young, A taxonomy of small Markovian errors, PRX Quantum **3**, 020335 (2022).

[52] In these DRB experiments, the target circuit outcomes were not randomized, whereas BiRB has randomization of the outcome by design. This is why the distribution of individual circuit polarizations differs between our BiRB and DRB data in Fig. 6(a), which changes the effect of biased readout error on the distribution of circuit outcomes. However, because this is a depth-independent effect from measurement error, we do not expect the lack of randomization to affect the resulting DRB error rate.

[53] David C. McKay, Sarah Sheldon, John A. Smolin, Jerry M. Chow, and Jay M. Gambetta, Three qubit randomized benchmarking, Phys. Rev. Lett. **122**, 200502 (2019).

[54] J. T. Muhonen, A. Laucht, S. Simmons, J. P. Dehollain, R. Kalra, F. E. Hudson, S. Freer, Kohei M. Itoh, D. N. Jamieson, J. C. McCallum, *et al.*, Quantifying the quantum gate fidelity of single-atom spin qubits in silicon by randomized benchmarking, J. Phys. Condens. Matter **27**, 154205 (2015).

[55] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell, *et al.*, Superconducting quantum circuits at the surface code threshold for fault tolerance, Nature **508**, 500 (2014).

[56] J. Raftery, A. Vrajitoarea, G. Zhang, Z. Leng, S. J. Srinivasan, and A. A. Houck, Direct digital synthesis of microwave waveforms for quantum computing, ArXiv:1703.00942.

[57] Jay M. Gambetta, A. D. Córcoles, Seth T. Merkel, Blake R. Johnson, John A. Smolin, Jerry M. Chow, Colm A. Ryan, Chad Rigetti, S. Poletto, Thomas A. Ohki, *et al.*, Characterization of addressability by simultaneous randomized benchmarking, Phys. Rev. Lett. **109**, 240504 (2012).

[58] Erik Nielsen, Kenneth Rudinger, Timothy Proctor, Antonio Russo, Kevin Young, and Robin Blume-Kohout, Probing quantum processor performance with pyGSTi, Quantum Sci. Technol. **5**, 044002 (2020).

[59] Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, Phys. Rev. A **80**, 012304 (2009).