Title

Quasiprobabilistic Readout Correction of Midcircuit Measurements for Adaptive Feedback via Measurement Randomized Compiling

Authors

Hashim, Akel

Carignan-Dugas, Arnaud

Chen, Larry

et al.

Quasiprobabilistic Readout Correction of Midcircuit Measurements for Adaptive Feedback via Measurement Randomized Compiling

Akel Hashim[1,2,*,†] Arnaud Carignan-Dugas[3,†] Larry Chen[1] Christian Jünger[1,2]
Neelay Fruitwala[4] Yilun Xu[4] Gang Huang[4] Joel J. Wallman[3] and Irfan Siddiqi[1,2,5]

[1]*Quantum Nanoelectronics Laboratory, Department of Physics, University of California at Berkeley, Berkeley, California 94720, USA*

[2]*Applied Math and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

[3]*Keysight Technologies Canada, Kanata, Ontario K2K 2W5, Canada*

[4]*Accelerator Technology and Applied Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

[5]*Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

Quantum measurements are a fundamental component of quantum computing. However, on present-day quantum computers, measurements can be more error prone than quantum gates and are susceptible to nonunital errors as well as nonlocal correlations due to measurement crosstalk. While readout errors can be mitigated in postprocessing, this is inefficient in the number of qubits due to a combinatorially large number of possible states that need to be characterized. In this work, we show that measurement errors can be tailored into a simple stochastic error model using randomized compiling, enabling the efficient mitigation of readout errors via quasiprobability distributions reconstructed from the measurement of a single preparation state in an exponentially large confusion matrix. We demonstrate the scalability and power of this approach by correcting readout errors without matrix inversion on a large number of different preparation states applied to a register of eight superconducting transmon qubits. Moreover, we show that this method can be extended to midcircuit measurements used for active feedback via quasiprobabilistic error cancellation, and we demonstrate the correction of measurement errors on an ancilla qubit used to detect and actively correct bit-flip errors on an entangled memory qubit. Our approach enables the correction of readout errors on large numbers of qubits and offers a strategy for correcting readout errors in adaptive circuits in which the results of midcircuit measurements are used to perform conditional operations on nonlocal qubits in real time.

I. INTRODUCTION

Measurement plays a foundational role in quantum mechanics. It is the means by which we learn properties of quantum systems and it is fundamentally linked with the collapse of quantum wave functions. Measurement is also essential to quantum computing. In gate-based quantum computing, measurement is needed to translate quantum bits (qubits) to classical bits at the end of a computation, it

is the central component in teleportation-based protocols [1,2] and measurement-based quantum computing [3,4], it can be utilized to generate long-range entanglement in constant depth via adaptive quantum circuits [5,6], and it is necessary for syndrome extraction in quantum error correction [7–12]. However, measurements are inherently noisy and the nature of errors can depend not only on the quantum state prior to measurement but also contextually on the state of other qubits. Moreover, measurements are often slower and more error prone than the unitary gates used to prepare quantum states, which places limits on the speed and fidelity with which they can be used to perform real-time corrections in the middle of quantum circuits.

A common assumption in quantum computing is that readout errors are purely probabilistic—i.e., for a given projective measurement of some finite duration, a qubit has a defined probability of experiencing a bit flip during

---

*Contact author: ahashim@berkeley.edu

†These authors contributed equally to this work.

readout. However, this assumption is often violated in systems with multiplexed readout, in which measurement crosstalk can cause context-dependent coherent and correlated readout errors [13–16]. Moreover, the probability of a bit flip for a qubit in an excited state can vastly differ from the probability of a bit flip while sitting in the ground state due to nonunital processes such as energy relaxation and T_1 decay [17], leading to context-dependent errors that depend on the state of a qubit prior to readout. However, by twirling a process over a unitary 1-design, one can effectively *design* stochastic channels [18]. One such strategy for designing stochastic channels is randomized compiling (RC) [19,20], which is a robust and efficient method for tailoring arbitrary Markovian errors into Pauli channels in gate-based quantum computing. While RC was originally designed for tailoring gate noise, it can be adapted to tailor measurement noise [21] and it has been previously

shown to reduce worst-case error rates in state preparation and measurement (SPAM) [22,23]. In this work, we experimentally deploy RC for quantum measurements on an eight-qubit superconducting quantum processor [see Fig. 1(a)]. We show that, under measurement RC (MRC), quantum measurement noise can be accurately described by a stochastic error model in which the probability of a bit flip for any given qubit is independent of the preparation state or the state of other qubits on the quantum processor, thus enforcing common preexisting assumptions about the stochasticity of measurement errors.

Because measurements translate quantum bits to classical bits, errors in terminating measurements can be mitigated classically via postprocessing. One approach requires preparing and measuring all possible combinations of input basis states for n qubits, from which one can construct a *confusion matrix* of the measured results.
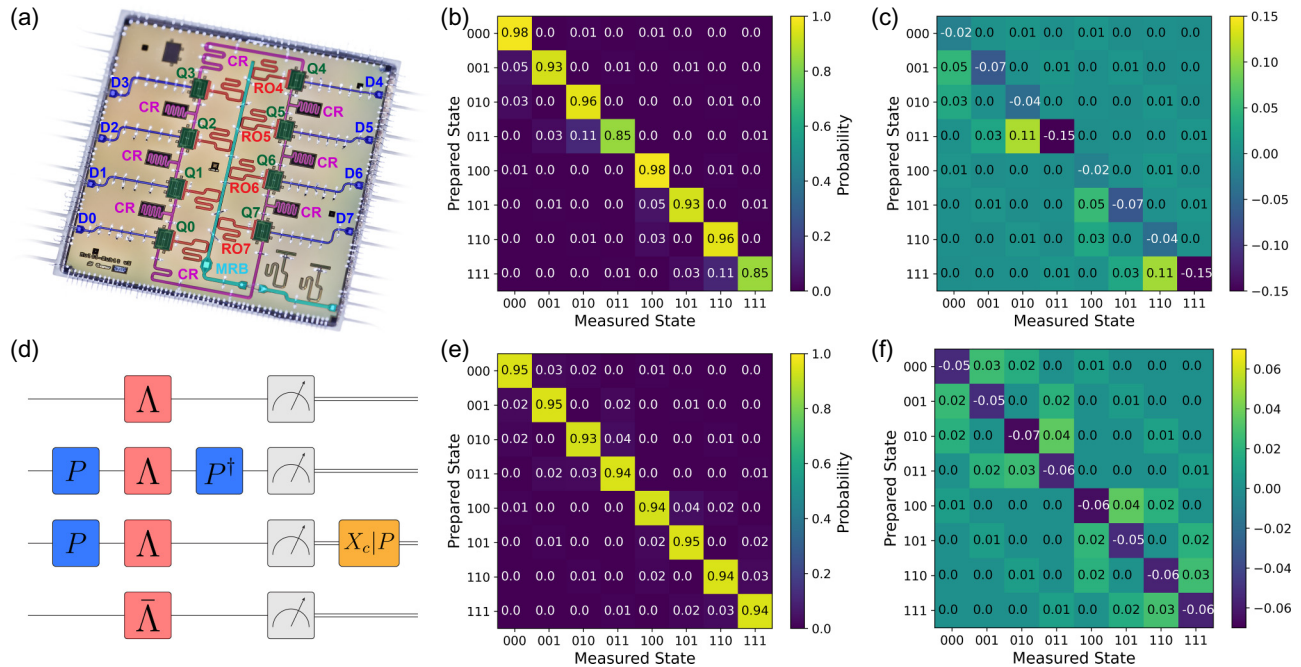

FIG. 1.   Randomized compiling for measurements. (a) An eight-qubit superconducting transmon processor. Qubits are labeled in green, individual drive lines are labeled in blue, individual readout resonators (RO) are labeled in red, and the multiplexed readout bus (MRB) is labeled in cyan. The qubits are coupled to nearest neighbors in a ring geometry via coupling resonators (CR, purple). (b) The full confusion matrix measured for three qubits (Q3, Q4, and Q5). Strong state-dependent errors are observed. For example, when $|011\rangle$ or $|111\rangle$ is prepared, $|010\rangle$ and $|110\rangle$ are measured approximately 11% of the time, respectively. (c) The confusion matrix in (b) minus the identity matrix. Context-dependent errors—such as errors that depend on the state of a qubit prior to measurement—appear as an asymmetry in the off-diagonal terms of the confusion matrix. (d) Top row: we can model the error in a measurement by a process matrix Λ preceding the measurement. Second row: in theory, it is possible to twirl this process matrix via Pauli twirling, $\Lambda \mapsto P^\dagger \Lambda P$. Third row: however, because this is a process matrix for a (nonunitary) measurement, the inversion operators must be implemented as classical bit flips X_c conditioned on which Pauli P was sampled before the measurement. Bottom row: by averaging this measurement many times over the full Pauli group, we obtain a twirled error process $\bar{\Lambda} = \frac{1}{4}\sum_{P\in\{I,X,Y,Z\}} P^\dagger \Lambda P$, in which measurement errors have been reduced to a stochastic bit-flip channel. (e) The full three-qubit confusion matrix measured using the scheme presented in (d) for qubits Q3, Q4, and Q5. We observe that the diagonal entries of the confusion matrix are all approximately equal in magnitude, showing that we have eliminated state-dependent readout errors. (f) The confusion matrix in (e) minus the identity matrix. All error probabilities in the off-diagonal elements are (approximately) symmetric along the diagonal. This indicates that under MRC, the probability of a bit flip for a given qubit is the same for all states.

Applying the inverse of the confusion matrix on the resulting outcome distribution often mitigates measurement errors. This strategy is limited to a subset of measurement errors, since the confusion matrix does not capture the effect of coherent measurement errors on quantum superpositions. Another limitation of this approach is its poor scalability: the size of this matrix grows exponentially in the number of qubits, making both the characterization and inversion steps intractable for large qubit numbers. As a result, experimentalists often resort to performing *local* readout correction [24], in which an individual confusion matrix is measured and inverted for each qubit. While this can correct individual readout errors, it cannot correct correlated bit flips. Alternative strategies for improving qubit readout include encoding qubits in a repetition code prior to readout [25,26], which comes at the cost of additional ancilla qubits, or correcting readout errors based on the results of detector tomography [15,16]. By tailoring noise in measurements into a stochastic bit-flip channel, we show that it is possible to correct readout errors for any input state without matrix inversion, ancilla qubits, or full tomography. To do so, it is sufficient to characterize readout errors on a single input state (e.g., $\left|0^{\otimes n}\right\rangle$ for n qubits) under MRC, from which a quasiprobability distribution can be constructed. Readout correction is then performed by inverting the quasiprobability distribution on the measured bit-string results. We compare full and local readout correction to our quasiprobabilistic protocol for a large number of structured and random input states on eight qubits and show that our protocol improves the results in over 90% of the circuits. Moreover, we show that this scheme extends to midcircuit measurements (MCMs) and we demonstrate the mitigation of readout errors used to perform real-time feedback to correct for bit-flip errors on an entangled qubit.

II. RANDOMIZED COMPILING FOR MEASUREMENTS

Generalized measurements of quantum states are described by positive operator-valued measures (POVMs), which are a set of positive semidefinite Hermitian matrices $\{E_i\}$ in d-dimensional Hilbert space $\mathcal{H}_d$ that obey the completeness relation:

$$\sum_i E_i = \mathbb{I}. \tag{1}$$

The probability of measuring an outcome i given a state ρ is governed by Born's rule,

$$p(i|\rho) = \mathrm{Tr}[E_i \rho]. \tag{2}$$

For a given system containing n qubits, the POVM set corresponding to computational-basis measurements contains 2^n elements, $\{E_i\}_{i=1}^{2^n}$, with each element indexed by

an n-qubit bit string i. For example, for a single qubit, the POVM set is $\{E_0, E_1\}$, for two qubits, the POVM set is $\{E_{00}, E_{01}, E_{10}, E_{11}\}$, etc. By preparing a system of n qubits in all 2^n possible combinations of basis states, represented by the set of input states $\{\rho_j\}$, and measuring the resulting POVMs $\{E_i\}$ for each basis state, one can construct a $2^n \times 2^n$ *confusion matrix* $\mathcal{M} = \langle\langle\{E_i\}|\{\rho_j\}\rangle\rangle$, the elements

$$\mathcal{M}_{ij} = \mathrm{Tr}[E_i \rho_j] \tag{3}$$

of which represent the probability $p(i|j)$ of measuring the outcome E_i given an input state ρ_j, where the double-bra (-ket) notation $\langle\langle\cdot|\ (|\cdot\rangle\rangle)$ denotes the *vectorization* of the POVM E_i (initial state ρ_j) into a $1 \times d^2$ row vector ($d^2 \times 1$ column vector). Classically, the confusion matrix is sufficient to predict the probability distribution of an outcome given any input. However, in quantum computing, while the $d \times d$ confusion matrix is an experimentally well-defined object [see Figs. 1(b)–1(f)], it generally only provides one part of the picture. Indeed, while Eq. (3) correctly describes the probabilities of observing the outcome i given the computational state j, it generally does not correctly prescribe the probability distribution expected for a quantum state involving quantum superpositions [21]. Fortunately, there is a way to compile quantum circuits such that, statistically, confusion matrices fully prescribe measurement errors as in the classical case. Such a method, which we call measurement randomized compiling (MRC), has been introduced in Ref. [21] and is described further below.

Let us consider the scenario in which the measurement error is such that given an n-qubit confusion matrix $\mathcal{M}$ and an ideal probability distribution p, the effect of measurement noise on the ideal outcomes produces a noisy probability distribution $q = \mathcal{M}p$. In such a case, correcting the effect of measurement noise on a probability distribution reduces to inverting $\mathcal{M}$ given a measured distribution q:

$$p = \mathcal{M}^{-1}q. \tag{4}$$

If $\mathcal{M}$ is known and if it correctly models measurement errors, then in theory one can correct the effect of measurement errors affecting the outcome of any quantum circuit. However, because $\mathcal{M}$ scales exponentially in the number of qubits n, in practice it is not feasible to construct a full n-qubit confusion matrix, nor is it always necessary if one can make reasonable assumptions about the locality and nature of correlated measurement noise. An alternative strategy is to assume that readout errors are uncorrelated and that measurement noise can be modeled as a tensor product of confusion matrices. In this case, it is sufficient to reconstruct the individual confusion matrix for each qubit,

such that $\mathcal{M}$ is given as

$$\mathcal{M} = \prod_{i=1}^{n} \otimes \mathcal{M}_i, \tag{5}$$

where $\mathcal{M}_i$ is the confusion matrix for the ith qubit. Now, the inversion process [Eq. (4)] only corrects readout errors on each qubit individually but cannot account for any correlated readout errors.

While it is often assumed that readout errors are probabilistic and locally independent, in which case measuring individual confusion matrices for each qubit would be sufficient to correct all readout errors, in practice this is not the case. For example, in Fig. 1(b), we plot the full confusion matrix for three qubits [Q3, Q4, and Q5; see Fig. 1(a)]. We observe that for most preparation states, the combined readout fidelity is between 93% and 98%. However, for $|011\rangle$ and $|111\rangle$, we observe poor readout fidelities; these probabilities are inconsistent with an assumption of independent bit-flip rates. Such errors could be in part due to the fact that the readout frequencies of Q3, Q4, and Q5 are close in frequency (see Appendix A) resulting from fabrication inaccuracies, leading to readout crosstalk [13,15], which can result in context-dependent readout errors in which the error on a qubit depends on the state of *another* qubit. Moreover, even for the preparation states with higher readout fidelities, we generally observe that excited states have worse readout fidelity than ground states. This is due to nonunital errors such as T_1 decay, which results in state-dependent readout errors, placing fundamental limits on excited-state fidelities for a given readout time [17].

Fortunately, there is a way to drastically simplify the situation, by statistically tailoring the measurement error into a probability distribution over bit flips that remains independent of the measured state. The idea is to twirl the readout noise using compiling methods such as MRC [21]. To understand the principles behind MRC, let us express a noisy measurement $\langle\langle \tilde{E}_i|$ as an ideal measurement $\langle\langle E_i|$ preceded by a process matrix Λ that captures all measurement errors: $\langle\langle \tilde{E}_i| = \langle\langle E_i|\Lambda$ [see Fig. 1(d)]. The goal of MRC is to twirl Λ into diagonal Pauli channels, i.e., $\Lambda \mapsto 4^{-n} \sum_{P\in\mathbb{P}_n} P^\dagger \Lambda P$, where $\mathbb{P}_n = \{I,X,Y,Z\}^{\otimes n}$ is the n-qubit Pauli group. However, in reality, readout errors occur concurrently with measurement; therefore, we cannot simply conjugate Λ by Pauli gates. Rather, to twirl Λ, we compile random Paulis into the final cycle of single-qubit gates before measurement and perform classical bit flips on the measured results conditional on the inserted Pauli for each qubit. For example, if I or Z is inserted, these will not change the results of measurements in the computational basis; however, if X or Y is sampled, these will flip the qubit state prior to measurement, necessitating classical bit flips after measurement. By repeating this process many (K) times and recording the combined distribution of all results, we obtain an effective Pauli-twirled process matrix

$$\bar{\Lambda}_K = \frac{1}{K} \sum_{\substack{i=1 \\ P_i \in_R \mathbb{P}_n}}^{K} P_i^\dagger \Lambda P_i, \tag{6}$$

where R denotes that P_i is chosen at random from the n-qubit Pauli group $\mathbb{P}_n$ each time. $\bar{\Lambda}_K$ in Eq. (6) is a sample average and it converges quickly to the true average $\bar{\Lambda}_\infty =: 4^{-n} \sum_{P\in\mathbb{P}_n} P^\dagger \Lambda P$, as shown in the theoretical analysis of RC [19,27,28]. More importantly, the convergence of the sample average to the true average is almost independent of the system size (just as the required sample size of a poll is almost independent of the population size); this property is what makes RC applicable to any system size (i.e., number of qubits or Hilbert-space dimension) [29]. To demonstrate that MRC scales to more qubits in practice, we repeat the same analysis on all eight qubits on our quantum processor (see Appendix E), showing indeed that MRC tailors readout noise equally well on eight qubits as on three.

The appeal of RC is that the true average error $\bar{\Lambda}_\infty$ can provably be expressed as a probabilistic mixture of Pauli gates (also known as a Pauli stochastic channel) [19,27,28]:

$$\bar{\Lambda}_\infty[\rho] = \sum_{P\in\mathbb{P}_n} p(P)P^\dagger \rho P, \tag{7}$$

where $p(P)$ is the probability of the Pauli error $P \in \mathbb{P}_n$. In the case of measurement, the Z have no effect on computational-basis states, and Y has the same effect as X. Therefore, we get a classical stochastic error channel of the form

$$\bar{\Lambda}_K[\rho] \approx \bar{\Lambda}_\infty[\rho] = \sum_{x\in\mathbb{Z}_2^{\otimes n}} p_x X^x \rho X^x, \tag{8}$$

where $x \in \mathbb{Z}_2^{\otimes n}$ is the set of classical n-bit strings, $\{p_x\}_{x\in\mathbb{Z}_2^{\otimes n}}$ is a probability distribution over bit flips X^x, and where X^x is short for $X^{x_1}X^{x_2}\cdots X^{x_n}$.

In Fig. 1(e), we plot the full confusion matrix for qubits Q3, Q4, and Q5 reconstructed using MRC with $K = 100$ randomizations [22]. We observe that the diagonal readout fidelities $p(i|i)$ are all approximately equal and the off-diagonal probabilities $p(i|j) \ \forall \ j \neq i$ are approximately symmetric along the diagonal, suggesting that we have eliminated state- and context-dependent readout errors due to T_1 decay and readout crosstalk. This provides experimental evidence that under MRC, we can describe readout errors as a purely stochastic process in which the probability of a bit flip for any given qubit is independent of the preparation state (see also Appendix E).

It should be noted that a similar method to MRC has been introduced in Refs. [30,31] by inserting bit flips prior to measurement. However, bit-flip averaging does not provide a complete twirl of the readout noise, as phase randomization is also necessary in order to describe readout errors as purely stochastic. For example, suppose that a qubit is in the $|i+\rangle$ state prior to measurement; here, a coherent-X error during measurement [15] will result in an incorrect output distribution. However, by randomly inserting Pauli-Z gates prior to measurement, the impact of the coherent-X error will be averaged away on the ensemble level [32]. Finally, it is worth noting that because readout errors are state independent under MRC, the effect of readout errors on Pauli expectation values can be efficiently corrected by rescaling by the average readout fidelity [33].

III. QUASIPROBABILISTIC READOUT CORRECTION

As observed in Sec. II, applying RC to quantum measurements effectively tailors the measurement error channel Λ into a classical stochastic error channel. This has a few ramifications. First, the effective error channel $\bar{\Lambda}_K$ can be fully described by its corresponding probability distribution, and each probability p_x can be estimated up to $1/\sqrt{N_{\text{shots}}}$ simply by looking at the output distribution resulting from sending a single computational-basis input to the randomly compiled measurement channel. In other words, $\bar{\Lambda}_K$ can be approximately described with $\mathcal{O}(N_{\text{shots}})$ floating-point numbers and each number has a precision of $1/\sqrt{N_{\text{shots}}}$ [34]. Second, the effective measurement error $\bar{\Lambda}_K$ can be inverted by applying a linear operation on the noisy output distribution. The exact inversion can quickly become unscalable to describe but since the probabilities appearing in $\bar{\Lambda}_K$ are already estimated with $1/\sqrt{N_{\text{shots}}}$ precision, an approximation should suffice. Fortunately, there exist standard quasiprobabilistic correction techniques that provide different orders of approximation of the inverse of $\bar{\Lambda}_K$ [35–39]. The first-order approximation is described using $\mathcal{O}(N_{\text{shots}})$ floating-point numbers and in general the ith-order approximation is described using $\mathcal{O}(N_{\text{shots}}^i)$ floating-point numbers. We proceed with describing the quasiprobabilistic readout-correction (QPRC) protocol below.

Because measurement errors under MRC can be described by a stochastic bit-flip channel that is independent of the input state, it is sufficient to characterize the probability of bit flips on n qubits using a single preparation state. For simplicity, we choose to characterize measurement errors on $\left|0^{\otimes n}\right\rangle$ using the MRC protocol. For example, in Fig. 2(a), we depict a single cycle of Paulis applied to the all-zero state on eight qubits; if X or Y is applied before measurement, then a classical bit flip X_c is applied in postprocessing. This process should be

repeated many (K) times to construct a twirled measurement channel [Eq. (8)]. In Fig. 2(b), we plot the results of the characterization procedure in blue using $K = 100$ randomizations. We observe that the all-zero state is measured over 88% of the time, with the remaining approximately 12% distributed over various single-qubit bit-flip channels.

To model readout noise under MRC, let us denote linear combinations of bit strings as $\sum_x c_x x$, where $c_x \in \mathbb{R}$ are scalar coefficients and $x \in \mathbb{Z}_n$ are bit strings. Given an error-probability distribution $p = \sum_x p_x x$ and an ideal outcome distribution $a = \sum_y a_y y$, we can express the resulting noisy outcome distribution $b = \sum_z b_z z$ as

$$b = p \oplus a,$$

$$= \left(\sum_x p_x x\right) \oplus \left(\sum_y a_y y\right), \quad (9)$$

where $x \oplus y$ is the *bitwise* modulo 2 sum of the x and y bit strings. In other words, the probability of observing the outcome z given a noisy measurement is

$$b_z = \sum_{x \oplus y = z} p_x a_y. \quad (10)$$

Now, to invert the effect of readout errors on the distribution b, we perform probabilistic error cancellation (PEC) [36,37] for measurement. That is, we construct a quasiprobability distribution q that is an approximate inverse of p. If we denote the all-zero bit string as $\underline{0}$ (we underline bit strings to distinguish them from scalar coefficients), the goal is to construct q such that $q \oplus p \approx \underline{0}$ (note that $\underline{0}$ is the identity with respect to the $\oplus$ operator). Indeed, suppose that $q \oplus p = \underline{0}$; then it follows from Eq. (9) that applying q to b yields $q \oplus b = q \oplus p \oplus a = \underline{0} \oplus a = a$. A first-order construction of q can be expressed as follows:

$$q = \frac{1}{2p_{\underline{0}} - 1}\left(p_{\underline{0}}\underline{0} - \sum_{x \neq \underline{0}} p_x x\right). \quad (11)$$

Indeed, applying q to p yields

$$p' = q \oplus p = \frac{1}{2p_{\underline{0}} - 1}\left(p_{\underline{0}}\underline{0} - \sum_{x \neq \underline{0}} p_x x\right)$$

$$\oplus \left(p_{\underline{0}}\underline{0} + \sum_{x \neq \underline{0}} p_x x\right),$$

$$= \frac{p_{\underline{0}}^2}{2p_{\underline{0}} - 1}\underline{0} - \frac{1}{2p_{\underline{0}} - 1}\left(\sum_{x \neq \underline{0}} p_x x\right)^2, \quad (12)$$

which is approximately equal to the identity up to second order. To put it simply, the error amplitude goes from
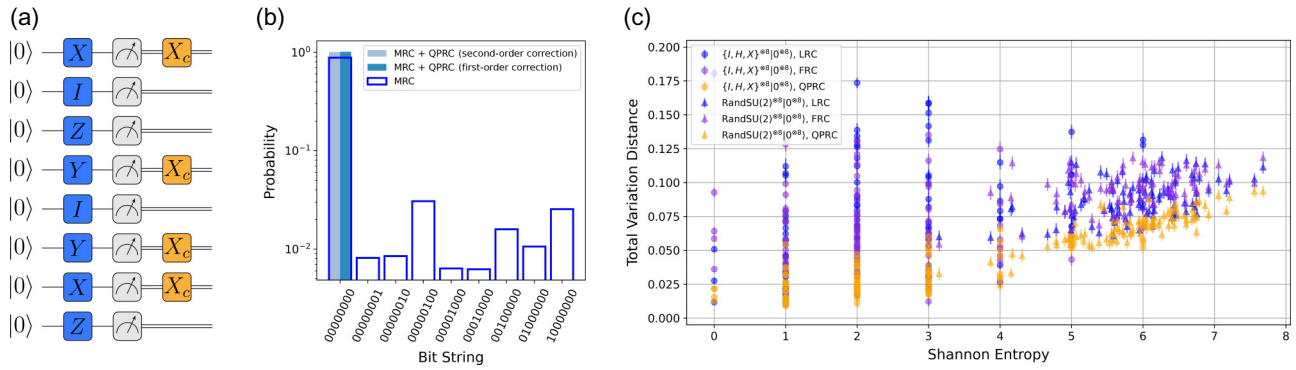
FIG. 2.   Quasiprobabilistic readout correction. (a) The readout characterization. The probability of bit-flip errors during readout can be characterized by preparing a single n-qubit input state (e.g., $|00000000\rangle$ for eight qubits) and measuring the resultant states under RC. To do so, a randomly sampled n-qubit Pauli operator should be inserted before measurement (e.g., $X \otimes I \otimes Z \otimes Y \otimes I \otimes Y \otimes X \otimes Z$); after measurement, a classical bit flip X_c should be applied to any qubit in which an X or Y gate was applied before measurement. This process should be repeated many times for many different randomly sampled Pauli operators. (b) The results for the protocol in (a) applied to eight qubits using $K = 100$ different randomizations are plotted with a blue outline. We observe single-qubit bit-flip errors on many qubits. For example, while $|00000000\rangle$ is measured over 88% of the time, we observe that Q5 has an approximately 3% chance of experiencing a bit-flip error during readout. This distribution can be used to perform quasiprobabilistic readout correction on any eight-qubit circuit in which the readout is performed using MRC. The first- and second-order corrections performed on the distribution itself are plotted in blue and light blue, respectively. The first-order correction has a probability greater than 1.0 for $|00000000\rangle$ (and small negative counts for other bit strings); however, the second-order correction reconstructs a distribution in which only the all-zero state remains. (Only the bit strings with significant counts are displayed for clarity.) (c) Readout-corrected single-qubit circuits. Structured circuits were generated by applying gates randomly sampled from $\{I, H, X\}$ to each qubit (circular data points) and random circuits were generated by applying a random $\mathsf{SU}(2)$ gate independently to each qubit (triangular data points). Each circuit was performed with and without MRC, denoted by the orange and blue or purple data points, respectively. For circuits without MRC, we apply local readout correction (LRC; blue) using confusion matrices measured for each qubit, or full-readout correction (FRC), using a full eight-qubit confusion matrix. For the circuits with MRC, we apply the quasiprobabilistic readout-correction (QPRC) protocol using the distribution measured in (b). We plot the total variation distance (TVD) of the experimental results with the ideal results as a function of the Shannon entropy S of the ideal result. $S = 0$ corresponds to a singular distribution and $S = 8$ is the uniform distribution for eight qubits; in general, the larger the entropy, the more uniform is the distribution. We observe that the TVD values of the MRC + QPRC results are lower than the results obtained with either LRC or FRC in over 90% of the circuits; thus, MRC + QPRC broadly outperforms both LRC and FRC. (The error bars for the TVD are on the order of the size of the markers.)

$1 - p_{\underline{0}}$ to $(1 - p_{\underline{0}})^2/(2p_{\underline{0}} - 1)$. However, the inverse operation can be improved further. Consider the family of quasiprobability distributions

$$q^{(k)} = \frac{p_{\underline{0}}^{2k-1}}{p_{\underline{0}}^{2k} - (1 - p_{\underline{0}})^{2k}} \left(\underline{0} + \sum_{j=1}^{2k-1} \left(\frac{-1}{p_{\underline{0}}} \right)^j \left(\sum_{x \neq \underline{0}} p_x x \right)^j \right), \tag{13}$$

for $k \in \mathbb{N}_+$. Note that $q^{(1)} = q$ from Eq. (11). Applying $q^{(k)}$ to p yields

$$q^{(k)} \oplus p = \frac{p_{\underline{0}}^{2k}}{p_{\underline{0}}^{2k} - (1 - p_{\underline{0}})^{2k}} \underline{0}$$

$$- \frac{1}{p_{\underline{0}}^{2k} - (1 - p_{\underline{0}})^{2k}} \left(\sum_{x \neq \underline{0}} p_x x \right)^{2k}. \tag{14}$$

Thus, in the generalized case, the error amplitude goes from $1 - p_{\underline{0}}$ to $1/[1 - (p_{\underline{0}}/(1 - p_{\underline{0}}))^{2k}]$. This strategy is

expected to be effective as long as $p_{\underline{0}} > 1/2$, which is the turning point for which the coefficient in front of $\underline{0}$ in Eq. (14) does not converge to 1 by increasing k. In principle, the constraint $p_{\underline{0}} > 1/2$ limits the scaling of this method. However, in Appendix D, we generalize the above quasiprobability inversion method, allowing it to scale for large systems in which the total error probability $1 - p_{\underline{0}}$ reaches well above $1/2$. Of course, like any mitigation method, there is a limit to its scalability [40]. However, given the current observations, since it does not involve any error propagation through a circuit, it is expected to apply well to outcome distributions *marginalized* over tens of physical qubits (given the same error rates).

In practice, to correct readout errors on any noisy experimental probability distribution b that has been measured using MRC, we sum over all corrected results for which the counts for each experimental result x have been redistributed according to $q^{(k)}$:

$$b^{(k)} = \sum_x \left(b_x x \oplus q^{(k)} \right), \tag{15}$$

where the readout-corrected distribution $b^{(k)}$ is the union over all of the redistributed counts $b_x x \oplus q^{(k)}$. To demonstrate that our procedure corrects readout errors, we perform a first- (i.e., using $q^{(1)} = q$) and second-order (i.e., using $q^{(2)}$) correction on the characterized probability distribution in Fig. 2(b) that is used to construct the quasiprobability distribution. We observe that the first-order correction redistributes most of the results to $\underline{0}$ but that $p_{\underline{0}}^{(1)}$ is slightly greater than 1.0. Because the corrected distribution is itself a quasiprobability distribution that has been normalized to preserve the total probability, $\underline{0}$ has a quasiprobability $p_{\underline{0}}^{(1)}$ greater than 1.0 to account for the negative quasiprobabilities in the other states (not shown). It is reasonable for the first-order correction to have negative probabilities, because the quasiprobability distribution is only an *approximate* inverse distribution; thus, small residual biases can remain after the first-order correction. However, after performing a second-order correction on the characterized distribution, we find that $p_{\underline{0}}^{(2)} = 1.0$, as we would expect if all readout errors were corrected. This process highlights the fact that readout correction (or, more generally, error-mitigation strategies) can introduce nonphysical outcomes into the results of experiments. For example, if one wants to preserve the total probability of a process, then the small residual negative values that remain after the quasiprobabilistic readout correction should be preserved, which equates to enforcing trace-preservation (TP). However, negative probabilities violate complete positivity (CP) and these values could be reasonably set to zero depending on the nature of the final computation. Therefore, one cannot always enforce both CP and TP on the outcomes of error corrected results and the choice of which to preserve is up to the experimenter.

To demonstrate the efficacy of our protocol on a wide variety of input states, we perform a second-order correction [i.e., using $q^{(2)}$] on 200 different eight-qubit input states prepared using a single cycle of gates, shown in Fig. 2(c). For half of the circuits, we sample gates from $\{I, H, X\}$ at random for each qubit, and for the other half of the circuits we sample random $\mathsf{SU(2)}$ gates for each qubit independently. To compute the accuracy of the readout-corrected results, we compute the total variation distance (TVD) between the experimental distribution b and the ideal distribution a,

$$D_{\mathrm{TV}}(b, a) = \frac{1}{2} \sum_x |b_x - a_x|, \tag{16}$$

plotted as a function of the Shannon entropy of the ideal results,

$$S = -\sum_x a_x \log_2(a_x). \tag{17}$$

A lower value for D_{TV} means that the results are closer to the ideal distribution. For the Shannon entropy, $S = 0$ corresponds to a probability distribution that is peaked around a single value, and for eight qubits $S = 8$ is the uniform distribution; in general, the higher the entropy, the closer the output is to a uniform distribution. We compare the results of our QPRC protocol to results obtained using full-readout correction (FRC) using matrix inversion of the full confusion matrix, which is not scalable, and local readout correction (LRC), which is scalable. We find that while FRC and LRC perform approximately equivalently over all circuits (i.e., their average TVDs are equivalent), our protocol produces better results (i.e., has a lower TVD) in over 90% of the circuits compared to both FRC and LRC. Additionally, we observe a positive linear correlation between the TVD of the corrected results and the entropy of the ideal results, with better performance at lower entropy. This can be explained by the fact that for higher entropy, the approximate correction has to be applied to more outputs, meaning that the systematic error in the approximated inverse is applied more often.

It should be noted that QPRC makes no distinction between the different sources of physical errors that lead to readout errors (e.g., T_1 decay, misclassification due to low measurement signal to noise, etc.). Therefore, it can correct different readout errors equally well, as under MRC they all appear as stochastic bit flips. However, like other methods for performing readout correction, periodic recharacterization of the readout errors under MRC is necessary for accurate readout correction via QPRC. While MRC is robust to drift and, e.g., fluctuations in qubit T_1 times (which would cause excited-state readout fidelities to also fluctuate in time), QPRC requires an accurate characterization of the error-probability distribution p in order to construct the inverse quasiprobability distribution q. Thus, it is recommended that one recharacterize the bit-flip error rates under MRC periodically, depending on how often one expects the system to drift or T_1 times to fluctuate.

IV. READOUT CORRECTION FOR MIDCIRCUIT MEASUREMENTS

The QPRC protocol presented in Sec. III provides a clear strategy for correcting readout errors afflicting terminating measurements. However, it is less clear how to correct readout errors in midcircuit measurements (MCMs), which are subject to complex error processes that are not always present in terminating measurements [41], and the results of which can be used to adapt circuits in real time via classical feedback [5]. While the result of a single measurement used for decision branching in feed-forward schemes cannot be corrected in real time, the results of MCMs can still be corrected quasiprobabilistically in the paradigm where we end up with an ensemble distribution at the very end of a circuit. To do so requires characterizing the

probability of a bit flip for a given MCM and quasiprobabilistically cancelling this error via random insertion of artificial Pauli-X errors. We describe this procedure below.

When MCMs are used to perform conditional feedforward operations, the readout fidelity of each MCM will dictate the rate at which incorrect decision branching occurs, and thus the rate at which the incorrect conditional operation is performed, which will add up linearly as a function of the number of MCMs in the circuit. In a model in which readout errors are purely probabilistic, this rate can be measured *a priori* by characterizing the probability of a bit-flip error on the measured qubit(s). For example, suppose that a qubit prepared in the ground state has a probability p_1 of experiencing a bit flip during a MCM; then the probability with which a single instance of the MCM performs the correct conditional operation is $1 - p_1 = p_0$. According to the QPRC protocol presented in Sec. III, the results of an imperfect measurement can be corrected by assigning a negative weight to the incorrect outcomes and subtracting them from the ideal outcomes. To do so in circuits with MCMs, we probabilistically insert an artificial bit flip X_p prior to the measured qubit with probability $p = p_1$. Now, for a circuit measured N_s times, on average the correct conditional operation will have been applied $(1 - p)N_s$ times and the incorrect conditional operation will have been applied pN_s times. To mitigate the impact of the noisy MCM, we subtract the raw counts of the circuit measured with X_p from the raw counts of the circuit measured without X_p. For circuits with multiple rounds of MCMs, we assign a negative weight to each instance in which X_p appears in the circuit; thus, for

odd (even) occurrences, the results are subtracted (added) to the bare results. This process generally increases the shot noise, since the error-mitigated results only have $(1 - p)N_s - pN_s = (1 - 2p)N_s$ shots; one can choose to compensate for this at the cost of a larger overhead by increasing the total number of shots to $N_s' = N_s/(1 - 2p)$.

We demonstrate the correction of readout errors on MCMs by performing the above protocol on a circuit designed to protect the memory of a qubit in the $|1\rangle$ state, shown in Fig. 3(a). Real-time active feedback is performed using the open-source control hardware QubiC [42,43]. When MRC is utilized for MCMs, the conditional readout value of the measured qubit now depends on the Pauli that is sampled before readout, and thus the conditional operation on the memory qubit is a function of this Pauli, $f(P)$; the random sampling of P and the calculation of $f(P)$ are performed many times before run time but the conditional bit flip X_c is performed in real time, with a feedback latency of 150 ns. In Fig. 3(b), we plot the probability of measuring the memory qubit in $|0\rangle$ as a function of the number (N) of rounds of MCMs. We find that for the raw output, the probability is approximately 5% for $N = 1$, growing to approximately 23% for $N = 10$. When we perform the MCMs with MRC, this probability is reduced significantly, growing to only approximately 18% for $N = 10$ rounds of MCMs. This difference can be explained by the fact that, in the ideal scenario, the ancilla qubit should be in $|1\rangle$ before each MCM—indicating that the memory qubit has not experienced a bit flip—which normally has a lower readout fidelity than the ground state. However, with MRC, the readout fidelities are equal
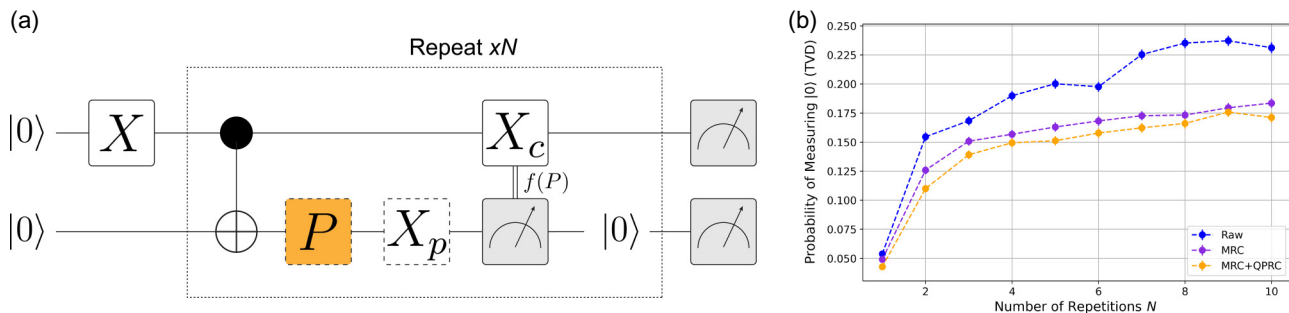

FIG. 3. Quasiprobabilistic readout correction of midcircuit measurements. (a) The schematic for active bit-flip protection. A memory qubit (top row) is prepared in the $|1\rangle$ state and a controlled-NOT (CNOT) gate is applied between the data qubit and an ancilla qubit (bottom row), which is subsequently measured. A conditional bit flip (X_c) is performed on the memory qubit depending on the results of the MCM of the ancilla qubit, after which the ancilla qubit is reset. This process is performed for N repetitions to protect the memory qubit from decaying to the ground state. Under MRC, the MCM is performed with RC by insertion of random Paulis (P, dashed orange box) before the MCM, in which case the conditional operation on the memory qubit is a function of the Pauli inserted before measurement, $f(P)$. To perform quasiprobabilistic readout correction on the MCM, a Pauli-X gate is probabilistically inserted before the measurement (X_p, dashed white box) and the final results of the MCM with X_p are subtracted from the results without X_p. (b) Results from performing the scheme presented in (a) for ten rounds of bit-flip protection. In the bare case, the probability of measuring the memory qubit in the $|0\rangle$ state grows from approximately 5% to approximately 23%. When the MCM is performed with MRC it only grows to approximately 18%. When the results measured with X_p are subtracted from the results measured without X_p (MRC+QPRC), the probability of measuring the memory qubit in $|0\rangle$ reduces by approximately 1% for all N compared to MRC alone. This is consistent with a measured bit-flip rate of 1.1% for the ancilla qubit. (The error bars are on the order of the size of the markers.)

$[p(0|0) = p(1|1)]$, so the error in the raw output will grow faster than the results with MRC. Additionally, inserting a random Pauli before the measurement can decouple the ancilla qubit from the memory qubit, ensuring that the error grows smoothly and monotonically. Finally, when we add QPRC to the MCMs performed with MRC, this reduces the probability of measuring the memory qubit in $|0\rangle$ by approximately 1% compared to just using MRC. The difference between MRC and MRC + QPRC is consistent with a bit-flip rate of 1.1% measured for the ancilla qubit prior to the experiment.

It should be noted that in Ref. [44], a related method is proposed for mitigating Pauli errors that occur during MCMs using a quasiprobabilistic error-cancellation scheme that depends on randomized compiling [39]. There are advantages and disadvantages to both techniques. A key distinction is that the protocol in Ref. [44] utilizes cycle benchmarking [45] to characterize the rates of Pauli errors, which has a much higher characterization overhead than our technique—whose characterization overhead is constant in the number of qubits n—and depends on the ability to measure exponential decays for the cycle or subcircuit of interest. However, the protocol in Ref. [44] can mitigate global errors that occur across an entire register of qubits, including correlated gate errors, whereas QPRC is only designed to mitigate readout errors. Future work could explore the scalability and trade-offs between these related methods as they relate to MCMs and adaptive circuits.

V. DISCUSSION AND OUTLOOK

Improving the fidelity of qubit readout is equally as important as improving gate fidelities. However, in recent years, much more focus has been placed on improving gate fidelities, leaving readout errors (or, more generally, SPAM errors) much larger than contemporary gate errors. To compensate for this, experimentalists typical correct readout errors by inverting a $2^n \times 2^n$ confusion matrix or, alternatively, inverting local 2×2 readout confusion matrices for each qubit independently. While the former method can correct n-qubit readout errors that occur on computational-basis states, it is not scalable; on the other hand, while the latter method is scalable, it cannot correct correlated readout errors.

In this work, we introduce a quasiprobabilistic method for correcting measurement noise that utilizes randomized compiling for enforcing a stochastic bit-flip model of readout errors. Our method requires a minimal characterization overhead that is constant in the number of qubits and is scalable in the limit that probabilistically small readout errors can be ignored. We demonstrate that our method outperforms both full and local readout correction on a large number of different possible input states for eight qubits. Moreover, we show that it can be extended to scenarios in

which MCMs are used in the single-shot limit for adaptive feedback, as long as the end goal is to collect ensemble statistics of the outputs.

While significant research and development is required to improve the readout fidelities of contemporary qubits on many hardware platforms, scalable matrix-inversion-free readout-correction methods such as QPRC are useful tools for correcting readout errors in the noisy intermediate-scale quantum (NISQ) era and beyond. Our method is fully compatible with MCMs and future work could demonstrate the utility of utilizing QPRC for correcting readout errors in adaptive circuits used for preparing nonlocal entangled states [6]. Furthermore, the machinery needed for adaptive circuits is the same as is needed for quantum error correction, so combining QPRC with quantum error correction would be an intriguing avenue for exploration.

Note added.—We recently became aware of a related but independently developed error-mitigation technique for midcircuit measurements that has appeared at the same time [46].

APPENDIX A: QUBIT AND READOUT CHARACTERIZATION

The quantum processing unit (QPU) used in this work consists of eight superconducting transmon qubits arranged in a ring geometry [Fig. 1(a)]. The frequency spectra of the GE and EF transitions of each qubit are plotted in Fig. 4(a). Some frequency crowding is observed at the lower end of the frequency spectrum. For example, the GE transition of Q2 is close to the EF transitions of Q0, Q4, and Q3. This can lead to microwave-line crosstalk between qubits, which can result in coherent leakage on the EF transitions when the GE transition of Q2 is driven. A similar effect can occur between the GE transition of Q5 and the EF transition of Q6, which is spectrally far from the rest of the qubits on the QPU due to fabrication inaccuracies. The qubit coherences are listed in Table I.

In Fig. 4(c), we plot the readout calibration for all eight qubits on this QPU, which supports qutrit-state discrimination. In qubit computations, qutrit readout can be used
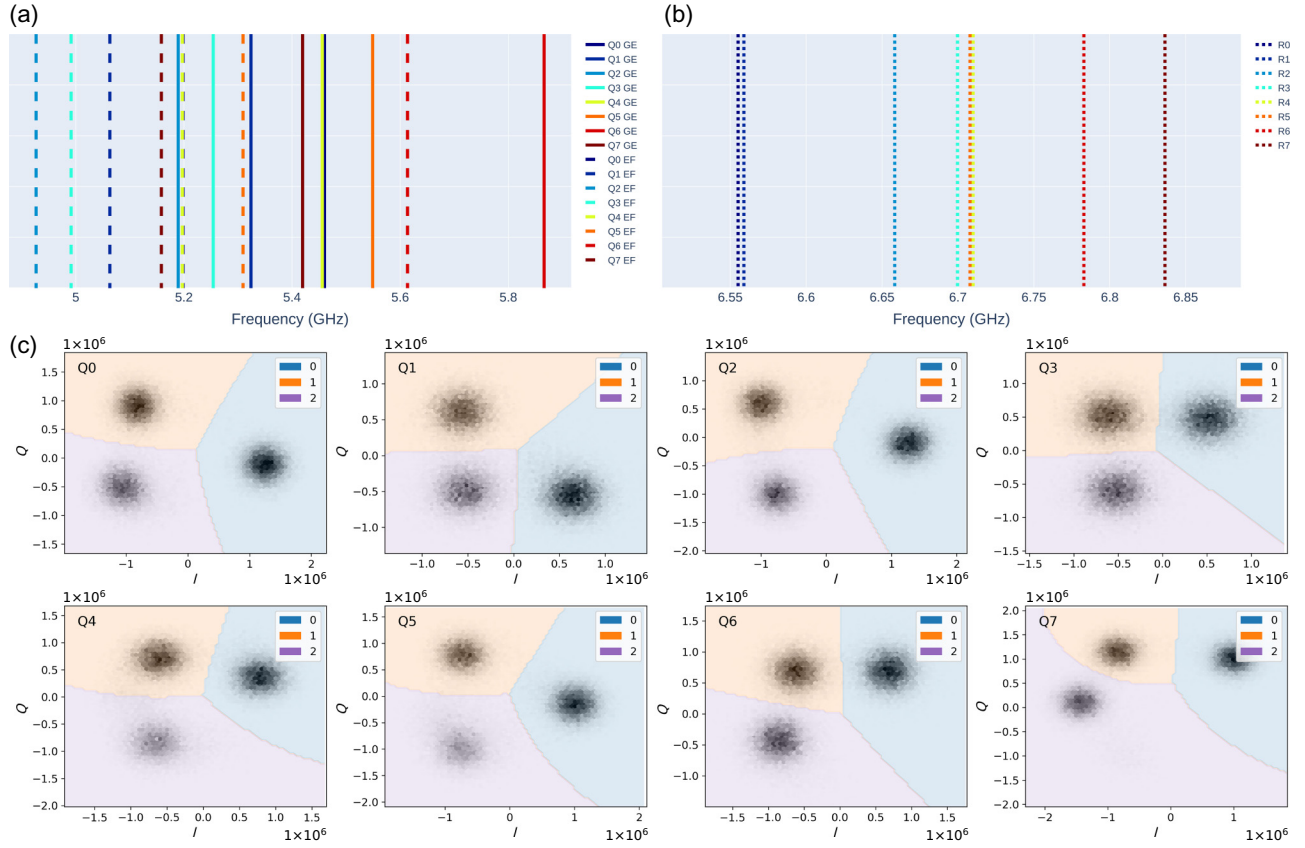
FIG. 4. Frequencies and readout characterization. (a) The frequency spectra of the GE (i.e., $|0\rangle \leftrightarrow |1\rangle$; solid lines) and EF (i.e., $|1\rangle \leftrightarrow |2\rangle$; dashed lines) transitions of the eight qubits on the quantum processor. (b) The frequency spectrum for the readout resonators coupled to the qubits. (c) Qutrit-state discrimination is supported for all qubits on the quantum processor.

to measure leakage rates. Alternatively, qutrit state discrimination can be used for excited-state promotion (ESP) [17,47] for improving qubit readout fidelities, whereby a $\pi_{1 \to 2}$ pulse is applied to each qubit before readout, after which all $|2\rangle$-state results are reclassified as $|1\rangle$ in postprocessing. ESP can protect qubits against energy relaxation during readout, which can include readout-induced decay. We utilize ESP to improve qubit readout and calibrate readout amplitudes to maximize readout fidelity with ESP turned on. The simultaneous readout fidelities are listed in Table II.

Even with improved readout fidelities using ESP, qubits can experience readout crosstalk during measurement. In Fig. 4(b), we plot the frequency spectra of the readout resonators for all eight qubits. We observe that several readout resonators are close in frequency. For example, the readout

resonators for Q0 and Q1 are within approximately 4 MHz of each other and the readout resonators for Q3, Q4, and Q5 are all within approximately 11 MHz of each other. Readout crosstalk can lead to context-dependent readout errors, in which the error on one qubit depends on the state of another qubit. This effect is apparent in the results presented in Fig. 1, in which the $|011\rangle$ and $|111\rangle$ states have drastically worse readout fidelities than the other preparation states.

APPENDIX B: GATE BENCHMARKING

The single- and two-qubit gates used in this work are benchmarked using randomized benchmarking (RB) and cycle benchmarking (CB). Infidelities for single-qubit gates are listed in Table III. Infidelities for two-qubit gates

TABLE I. Qubit coherences: the qubit coherence times T_1, T_2^*, and T_{2E} are listed.

	Q0	Q1	Q2	Q3	Q4	Q5	Q6	Q7
T_1 (µs)	96.6(2.6)	130.0(2.7)	142.0(3.0)	140.0(6.3)	77.0(5.2)	30.4(0.95)	55.6(1.3)	22.5(0.32)
T_2^* (µs)	120.0(14.0)	41.0(7.2)	92.0(16.0)	61.0(6.1)	38.0(5.4)	8.5(1.3)	26.0(3.7)	39.0(1.7)
T_{2E} (µs)	120.0(8.3)	130.0(7.5)	140.0(12.0)	90.0(13.0)	110.0(11.0)	33.0(3.6)	90.0(14.0)	43.0(2.2)

TABLE II.   Readout fidelities. Simultaneous readout fidelities for all qubits with excited-state promotion.

	Q0	Q1	Q2	Q3	Q4	Q5	Q6	Q7
$P(0|0)$	0.995(1)	0.995(1)	0.995(1)	0.992(2)	0.990(2)	0.998(1)	0.997(1)	0.987(3)
$P(1|1)$	0.983(2)	0.962(7)	0.994(2)	0.986(2)	0.966(7)	0.969(4)	0.994(2)	0.986(2)

are listed in Table IV. It should be noted that all quoted infidelities are the *process infidelity* e_F, not the *average gate infidelity* r. These two are equal up to a simple dimensionality factor:

$$e_F = \frac{d+1}{d}r, \tag{B1}$$

where $d = 2^n$ for n qubits.

APPENDIX C: QUANTUM HARDWARE

In this work, we use the open-source control system QubiC [42,43] to perform these experiments. QubiC is a field-programmable-gate-array- (FPGA) based control system for superconducting qubits developed at Lawrence Berkeley National Laboratory. The QubiC system used for these experiments has been implemented on the Xilinx ZCU216 RFSoC (RF system-on-chip) evaluation board and uses custom gateware for real-time pulse sequencing and synthesis.

The QubiC gateware has a bank of distributed processor cores for performing pulse sequencing, parametrization, and conditional execution (i.e., control flow) [48]. The QubiC readout DSP (digital signal processing) chain includes on-FPGA demodulation and qubit state discrimination using a threshold mechanism. Currently, the discrimination is performed for MCMs using the in-phase (I) component of the integrated readout pulse. If $I > 0$, the discriminator returns a 0; if $I < 0$, the discriminator returns a 1. For this reason, all of the $|0\rangle$ states are calibrated to be on the right side of $I = 0$, and all of the $|1\rangle$ and $|2\rangle$ states are calibrated to be on the left side of $I = 0$ [see Fig. 4(c)]. These state-discriminated results can then be requested by any processor core (using a special instruction) and used as inputs to arbitrary control flow and/or branching decisions (e.g., a WHILE loop or IF_ELSE code block). The total feedback latency (not including readout time) is 150 ns. After these experiments were performed, a neural-network-based readout discriminator has been developed for MCMs, performed on QubiC, which is capable of distinguishing $|1\rangle$ from $|2\rangle$ [49].

APPENDIX D: SCALING THE QUASIPROBABILISTIC INVERSE FOR A HIGHER NUMBER OF QUBITS

In Sec. III, we have proposed to invert the effect of measurement errors by applying an approximate inverse operation described in Eq. (13). While this has worked well in our experiment, such an inversion technique may fail when the error probability grows above $1/2$. Without a workaround, this constraint would make it impossible for such a quasiprobabilistic inversion technique to scale, since the error probability grows exponentially in the number of qubits. Indeed, with independent local error rates of 2%, the total error probability reaches $1/2$ after 34 qubits. As such, in this appendix, we provide a generalization of our quasiprobabilistic inversion technique that ensures a proper scaling of our mitigation method for measurements.

To explain the generalization, let us consider a two-qubit toy example where the error-probability distribution takes the form:

$$p_{AB}(\epsilon) = \left(\frac{9}{16} - \epsilon\right)\underline{00}_{AB} + \frac{3}{16}\underline{01}_{AB} + \frac{3}{16}\underline{10}_{AB}$$
$$+ \left(\frac{1}{16} + \epsilon\right)\underline{11}_{AB}, \tag{D1}$$

where $\epsilon \in [0, 1/4]$ and where the underscore notation indicates a classical bit string in the probability distribution. Note that in the case in which $\epsilon = 0$, we fall back on marginalized independent bit-flip errors of 25%. As such, ϵ denotes a correlated bit-flip error probability on top of the independent bit flips. Note that for all values of $\epsilon > 1/16$, the total error probability exceeds $1/2$. It should be noted that the first-order inversion makes things worse when

$$\left|\frac{p_{\underline{0}}^2}{2p_{\underline{0}} - 1} - 1\right| < |p_{\underline{0}} - 1|, \tag{D2}$$

which occurs when $p_{\underline{0}} < 2/3$.

As we demonstrate here, even in the milder case in which $\epsilon = 0$, $p_{\underline{0}} < 2/3$, and the first-order inversion

TABLE III.   Single-qubit gate infidelities: the process infidelities for isolated and simultaneous single-qubit gates measured via RB for each qubit are listed.

	Q0	Q1	Q2	Q3	Q4	Q5	Q6	Q7
RB isolated (10^{-3})	1.5(1)	0.33(2)	0.54(3)	1.0(1)	3.2(1)	1.92(9)	2.4(3)	2.58(9)
RB simultaneous (10^{-3})	2.1(2)	3.1(2)	2.0(3)	1.9(2)	6.1(7)	5.7(3)	3.4(3)	7.6(9)

TABLE IV.   Two-qubit gate infidelities: the process infidelities for two-qubit RB are listed for each qubit pair used in this work. The native (CZ) gate fidelities are measured via CB.

	(Q0, Q1)	(Q1, Q2)	(Q2, Q3)	(Q3, Q4)	(Q4, Q5)	(Q5, Q6)	(Q6, Q7)	(Q7, Q0)
RB isolated (10^{-2})	5.0(5)	1.62(7)	1.64(8)	2.6(2)	3.8(2)	4.6(2)	6.7(6)	3.8(3)
CB (CZ) (10^{-2})	1.4(1)	0.57(1)	0.41(1)	0.81(4)	1.80(8)	2.08(3)	2.77(8)	1.40(7)

described in Eq. (11),

$$q_{AB}(\epsilon = 0) = \frac{9}{2}\underline{00}_{AB} - \frac{3}{2}\underline{01}_{AB} - \frac{3}{2}\underline{10}_{AB} - \frac{1}{2}\underline{11}_{AB}, \quad \text{(D3)}$$

fails to mitigate errors:

$$q_{AB}(0) \oplus p_{AB}(0) = \frac{31}{16}\underline{00}_{AB} - \frac{3}{16}\underline{01}_{AB} - \frac{3}{16}\underline{10}_{AB}$$
$$- \frac{9}{16}\underline{11}_{AB}. \quad \text{(D4)}$$

However—and this is the essence of the solution—it is possible to adapt the quasiprobabilistic inversion to address qubits A and B individually before mitigating them in tandem. Let us consider the error distributions marginalized over individual qubits:

$$p_A(\epsilon) = \left(\frac{3}{4} - \epsilon\right)\underline{0}_A + \left(\frac{1}{4} + \epsilon\right)\underline{1}_A, \quad \text{(D5a)}$$

$$p_B(\epsilon) = \left(\frac{3}{4} - \epsilon\right)\underline{0}_B + \left(\frac{1}{4} + \epsilon\right)\underline{1}_B. \quad \text{(D5b)}$$

Note that the error probabilities marginalized on the two subsystems are below $1/3$ as long as ϵ remains below $1/12$. From there, consider the first-order local quasiprobabilistic inversions obtained by using Eq. (11) with the marginalized distributions $p_A(\epsilon)$ and $p_B(\epsilon)$:

$$q_A(\epsilon) = \frac{2}{1 - 4\epsilon}\left[\left(\frac{3}{4} - \epsilon\right)\underline{0}_A - \left(\frac{1}{4} + \epsilon\right)\underline{1}_A\right], \quad \text{(D6a)}$$

$$q_B(\epsilon) = \frac{2}{1 - 4\epsilon}\left[\left(\frac{3}{4} - \epsilon\right)\underline{0}_B - \left(\frac{1}{4} + \epsilon\right)\underline{1}_B\right]. \quad \text{(D6b)}$$

Let us apply those local inversions to the total distribution $p_{AB}(\epsilon)$. Unsurprisingly, in the case in which $\epsilon = 0$, the error distribution is locally independent [i.e., $p_{AB}(0) = p_A(0)p_B(0)$] and we obtain perfect error mitigation:

$$q_A(0)q_B(0) \oplus p_{AB}(0) = \underline{00}_{AB}. \quad \text{(D7)}$$

In the more general case, we obtain

$$p'_{AB}(\epsilon) = q_A(\epsilon)q_B(\epsilon) \oplus p_{AB}(\epsilon) = \left(\frac{3 + (1 - 4\epsilon)^{-2}}{4}\right)\underline{00}_{AB}$$
$$+ \left(\frac{1 - (1 - 4\epsilon)^{-2}}{4}\right)\underline{01}_{AB} + \left(\frac{1 - (1 - 4\epsilon)^{-2}}{4}\right)\underline{10}_{AB}$$
$$+ \left(\frac{-1 + (1 - 4\epsilon)^{-2}}{4}\right)\underline{11}_{AB}. \quad \text{(D8)}$$

Instead of further carrying out heavy expressions in ϵ, for the sake of simplicity (especially since this a toy example), let us pick ϵ small enough such that we can ignore $\mathcal{O}(\epsilon^2)$. In that case, we obtain

$$p'_{AB}(\epsilon) = (1 + 2\epsilon)\underline{00}_{AB} - 2\epsilon\underline{01}_{AB} - 2\epsilon\underline{10}_{AB} + 2\epsilon\underline{11}_{AB}$$
$$+ \mathcal{O}(\epsilon^2). \quad \text{(D9)}$$

Let us now apply the quasiprobabilistic inversion technique prescribed in Sec. III on that new (quasiprobabilistic) distribution. Using

$$q'_{AB}(\epsilon) = \frac{1}{1 + 4\epsilon}\left((1 + 2\epsilon)\underline{00}_{AB} + 2\epsilon\underline{01}_{AB} + 2\epsilon\underline{10}_{AB}\right.$$
$$\left. - 2\epsilon\underline{11}_{AB}\right), \quad \text{(D10)}$$

we obtain

$$q'_{AB}(\epsilon) \oplus p'_{AB}(\epsilon) = \underline{00}_{AB} + O(\epsilon^2). \quad \text{(D11)}$$

In other words, performing the local mitigation $q_A(\epsilon)q_B(\epsilon)$ has changed the error distribution and brought us to a point at which we can effectively apply a joint mitigation operation.

Note that in this toy example, we have applied the quasiprobabilistic inversions sequentially, but they can easily be compiled into a single operation:

$$q_{AB}(\epsilon) := q'_{AB}(\epsilon) \oplus q_A(\epsilon)q_B(\epsilon)$$
$$= \left(\frac{9}{4} + 4\epsilon\right)\underline{00}_{AB} - \frac{3}{4}\underline{01}_{AB} - \frac{3}{4}\underline{10}_{AB} + \left(\frac{1}{4} - 4\epsilon\right)\underline{11}_{AB}$$
$$+ O(\epsilon^2). \quad \text{(D12)}$$

This is relevant to readout correction of midcircuit measurements, where the quasiprobability $q_{AB}(\epsilon)$ is used as input for sampling circuits (see Sec. IV).

In the light of the toy example, the generalization of the mitigation strategy for larger systems is fairly straightforward. Given an error distribution p_S over a set of qubits S, subdivide the system into disjoint partitions $S_1, \ldots, S_k$ such that $\bigcup_i S_i = S$ and obtain the marginal error probabilities over those partitions, $p_{S_1}, \ldots, p_{S_k}$. The goal is to choose partitioning such that the marginal error probability within every partition is lower than $1/2$ (or $1/3$ if using

first-order inversions). From these marginal error distributions, apply quasiprobabilistic corrections according to the method described in Sec. III:

$$p_S \rightarrow q_{S_1} \cdots q_{S_k} \oplus p_S = p_S'. \tag{D13}$$

p_S' should be closer to the identity. Repeat the process for increasingly larger partitions. Once the resulting total
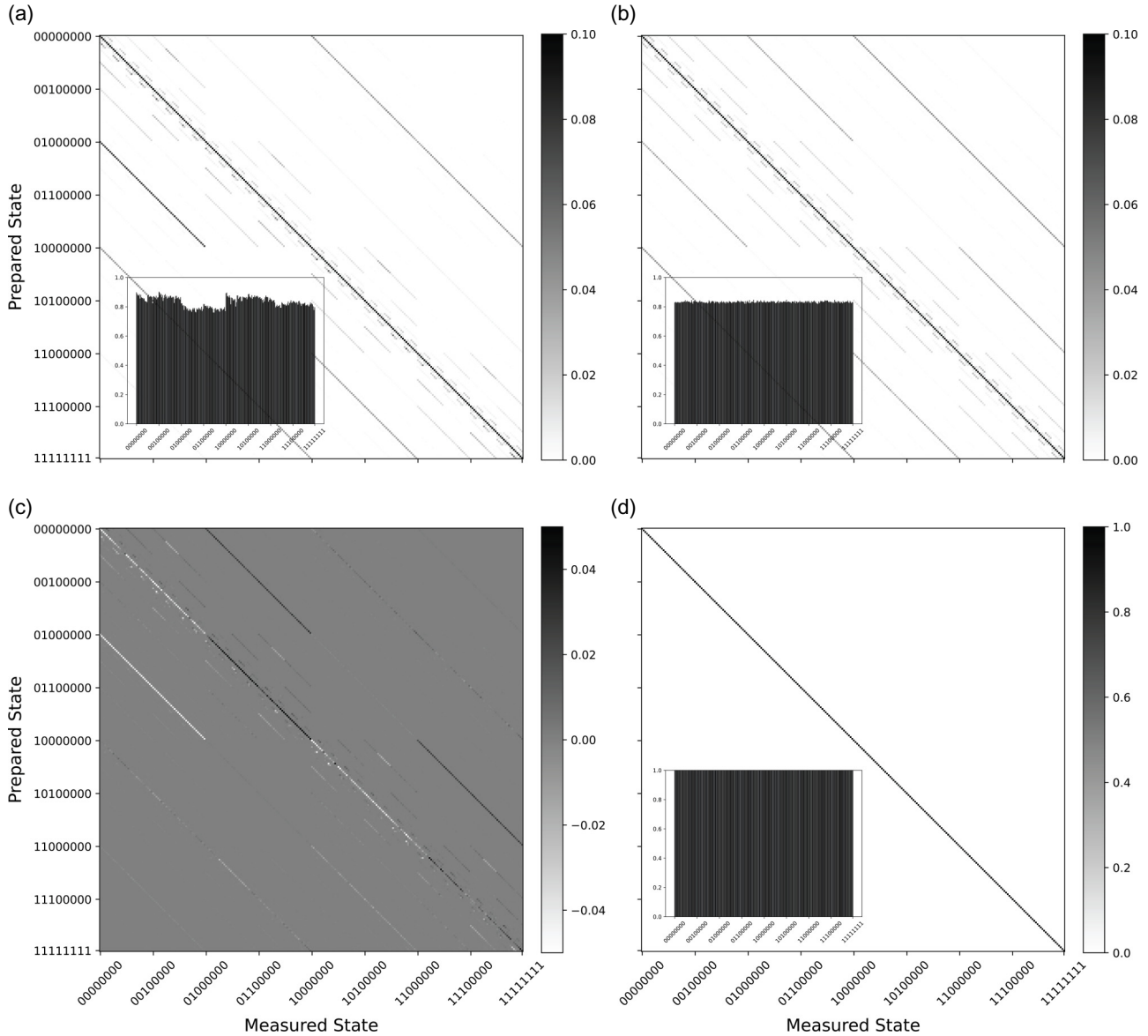

FIG. 5.   MRC and QPRC on eight qubits. (a) The eight-qubit confusion matrix. The maximum value of the color map is set to 0.1 for better contrast of the off-diagonal elements. The inset shows diagonal values of the confusion matrix (i.e., readout fidelities for different states). (b) The eight-qubit confusion matrix measured with MRC. The maximum value of the color map is set to 0.1 for better contrast of the off-diagonal elements. The inset shows the diagonal values of the confusion matrix (i.e., readout fidelities for different states). (c) The data from (b) minus the data from (a). This difference shows which entries are increased (black) or decreased (white) by using MRC. (d) An eight-qubit confusion matrix reconstruction from (b) by applying QPRC to the measured data for each different state preparation. The inset shows the diagonal values of the confusion matrix (i.e., effective readout fidelities for different states). The perfect correction for all states demonstrates that QPRC can correct readout errors equally well regardless of which input state was initially characterized.

error quasidistribution is close enough to the identity (e.g., once $|1 - p_0| < 1/2$), use the total distribution to infer the quasiprobabilistic inverse.

APPENDIX E: MEASUREMENT RC AND QUASIPROBABILISTIC READOUT CORRECTION ON EIGHT QUBITS

In the main text, we show the impact of MRC on a confusion matrix for only three qubits (see Fig. 1). To demonstrate that MRC scales equally well to larger numbers of qubits, we apply MRC to a full eight-qubit confusion matrix, shown in Fig. 5. In Fig. 5(a), we plot the raw eight-qubit confusion matrix reconstructed for all qubits on our quantum processor. We observe an asymmetry in the diagonal and off-diagonal elements, indicating the presence of state and context-dependent errors. In Fig. 5(b), we plot the eight-qubit confusion matrix reconstructed with MRC. We observe symmetry in both the diagonal and off-diagonal elements, demonstrating that MRC works equally well in tailoring readout noise for eight qubits as it did for three qubits. In Fig. 5(c), we plot the difference between the confusion matrix with MRC and the raw confusion matrix, showing where elements of the confusion matrix are increased or decreased with MRC. Finally, in Fig. 5(d), we perform QPRC on the data in Fig. 5(b), to demonstrate that our effective readout fidelity is perfect across all states with QPRC. More specifically, for each preparation state, we apply QPRC to the measured data using the measured data itself to compute the inverse distribution, similar to the analysis performed in Fig. 2(b). This demonstrates that we can perform the QPRC protocol using readout errors characterized for any initial input state.

---

[1] D. Gottesman and I. L. Chuang, Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations, Nature **402**, 390 (1999).

[2] K. S. Chou, J. Z. Blumoff, C. S. Wang, P. C. Reinhold, C. J. Axline, Y. Y. Gao, L. Frunzio, M. Devoret, L. Jiang, and R. Schoelkopf, Deterministic teleportation of a quantum gate between two logical qubits, Nature **561**, 368 (2018).

[3] R. Raussendorf and H. J. Briegel, A one-way quantum computer, Phys. Rev. Lett. **86**, 5188 (2001).

[4] H. J. Briegel, D. E. Browne, W. Dür, R. Raussendorf, and M. Van den Nest, Measurement-based quantum computation, Nat. Phys. **5**, 19 (2009).

[5] M. Foss-Feig, A. Tikku, T.-C. Lu, K. Mayer, M. Iqbal, T. M. Gatterman, J. A. Gerber, K. Gilmore, D. Gresh, A. Hankin, *et al.*, Experimental demonstration of the advantage of adaptive quantum circuits, ArXiv:2302.03029.

[6] A. Hashim, M. Yuan, P. Gokhale, L. Chen, C. Juenger, N. Fruitwala, Y. Xu, G. Huang, L. Jiang, and I. Siddiqi, Efficient generation of multi-partite entanglement between non-local superconducting qubits using classical feedback, ArXiv:2403.18768.

[7] P. W. Shor, Scheme for reducing decoherence in quantum computer memory, Phys. Rev. A **52**, R2493 (1995).

[8] E. Knill and R. Laflamme, Theory of quantum error-correcting codes, Phys. Rev. A **55**, 900 (1997).

[9] J. Chiaverini, D. Leibfried, T. Schaetz, M. D. Barrett, R. Blakestad, J. Britton, W. M. Itano, J. D. Jost, E. Knill, C. Langer, *et al.*, Realization of quantum error correction, Nature **432**, 602 (2004).

[10] N. Ofek, A. Petrenko, R. Heeres, P. Reinhold, Z. Leghtas, B. Vlastakis, Y. Liu, L. Frunzio, S. Girvin, L. Jiang, *et al.*, Extending the lifetime of a quantum bit with error correction in superconducting circuits, Nature **536**, 441 (2016).

[11] S. Rosenblum, P. Reinhold, M. Mirrahimi, L. Jiang, L. Frunzio, and R. J. Schoelkopf, Fault-tolerant detection of a quantum error, Science **361**, 266 (2018).

[12] W. P. Livingston, M. S. Blok, E. Flurin, J. Dressel, A. N. Jordan, and I. Siddiqi, Experimental demonstration of continuous quantum error correction, Nat. Commun. **13**, 2307 (2022).

[13] J. Z. Blumoff, K. Chou, C. Shen, M. Reagor, C. Axline, R. Brierley, M. Silveri, C. Wang, B. Vlastakis, S. E. Nigg, *et al.*, Implementing and characterizing precise multiqubit measurements, Phys. Rev. X **6**, 031041 (2016).

[14] J. Heinsoo, C. K. Andersen, A. Remm, S. Krinner, T. Walter, Y. Salathé, S. Gasparinetti, J.-C. Besse, A. Potočnik, A. Wallraff, *et al.*, Rapid high-fidelity multiplexed readout of superconducting qubits, Phys. Rev. Appl. **10**, 034040 (2018).

[15] Y. Chen, M. Farahzad, S. Yoo, and T.-C. Wei, Detector tomography on IBM quantum computers and mitigation of an imperfect measurement, Phys. Rev. A **100**, 052315 (2019).

[16] F. B. Maciejewski, Z. Zimborás, and M. Oszmaniec, Mitigation of readout noise in near-term quantum devices by classical post-processing based on detector tomography, Quantum **4**, 257 (2020).

[17] S. S. Elder, C. S. Wang, P. Reinhold, C. T. Hann, K. S. Chou, B. J. Lester, S. Rosenblum, L. Frunzio, L. Jiang, and R. J. Schoelkopf, High-fidelity measurement of qubits encoded in multilevel superconducting circuits, Phys. Rev. X **10**, 011001 (2020).

[18] M. A. Graydon, J. Skanes-Norman, and J. J. Wallman, Designing stochastic channels, ArXiv:2201.07156.

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

[20] A. Hashim, R. K. Naik, A. Morvan, J.-L. Ville, B. Mitchell, J. M. Kreikebaum, M. Davis, E. Smith, C. Iancu, K. P. O'Brien, I. Hincks, J. J. Wallman, J. Emerson, and I. Siddiqi, Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor, Phys. Rev. X **11**, 041039 (2021).

[21] S. J. Beale and J. J. Wallman, Randomized compiling for subsystem measurements, ArXiv:2304.06599.

[22] A. Hashim, S. Seritan, T. Proctor, K. Rudinger, N. Goss, R. Naik, J. M. Kreikebaum, D. Santiago, and I. Siddiqi, Benchmarking quantum logic operations relative to thresholds for fault tolerance, npj Quantum Inf. **9**, 109 (2023).

[23] D. McLaren, M. A. Graydon, and J. J. Wallman, Stochastic errors in quantum instruments, ArXiv:2306.07418.

[24] S. Bravyi, S. Sheldon, A. Kandala, D. C. Mckay, and J. M. Gambetta, Mitigating measurement errors in multiqubit experiments, Phys. Rev. A **103**, 042605 (2021).

[25] J. M. Günther, F. Tacchino, J. R. Wootton, I. Tavernelli, and P. K. Barkoutsos, Improving readout in quantum simulations with repetition codes, Quantum Sci. Technol. **7**, 015009 (2021).

[26] R. Hicks, B. Kobrin, C. W. Bauer, and B. Nachman, Active readout-error mitigation, Phys. Rev. A **105**, 012419 (2022).

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

[28] A. Winick, J. J. Wallman, D. Dahlen, I. Hincks, E. Ospadov, and J. Emerson, Concepts and conditions for error suppression through randomized compiling, ArXiv:2212.07500.

[29] N. Goss, S. Ferracin, A. Hashim, A. Carignan-Dugas, J. M. Kreikebaum, R. K. Naik, D. I. Santiago, and I. Siddiqi, Extending the computational reach of a superconducting qutrit processor, ArXiv:2305.16507.

[30] A. W. Smith, K. E. Khosla, C. N. Self, and M. Kim, Qubit readout error mitigation with bit-flip averaging, Sci. Adv. **7**, eabi8009 (2021).

[31] R. Hicks, C. W. Bauer, and B. Nachman, Readout rebalancing for near-term quantum computers, Phys. Rev. A **103**, 022407 (2021).

[32] While Pauli-*Z* gates can be implemented entirely via phase shifts in the following pulse, this will have no effect if the gate directly precedes a measurement. Instead, we implement Pauli-*Z* gates using the *ZXZXZ* decomposition of the gate, such that physical pulses are always played prior to measurement.

[33] E. Van Den Berg, Z. K. Minev, and K. Temme, Model-free readout-error mitigation for quantum expectation values, Phys. Rev. A **105**, 032620 (2022).

[34] The accuracy of the estimate is also limited by state-preparation errors and single-qubit gate errors but these tend to be low compared to measurement errors.

[35] Y. Li and S. C. Benjamin, Efficient variational quantum simulator incorporating active error minimization, Phys. Rev. X **7**, 021050 (2017).

[36] K. Temme, S. Bravyi, and J. M. Gambetta, Error mitigation for short-depth quantum circuits, Phys. Rev. Lett. **119**, 180509 (2017).

[37] S. Endo, S. C. Benjamin, and Y. Li, Practical quantum error mitigation for near-future applications, Phys. Rev. X **8**, 031027 (2018).

[38] A. Kandala, K. Temme, A. D. Córcoles, A. Mezzacapo, J. M. Chow, and J. M. Gambetta, Error mitigation extends the computational reach of a noisy quantum processor, Nature **567**, 491 (2019).

[39] S. Ferracin, A. Hashim, J.-L. Ville, R. Naik, A. Carignan-Dugas, H. Qassim, A. Morvan, D. I. Santiago, I. Siddiqi, and J. J. Wallman, Efficiently improving the performance of noisy quantum computers, Quantum **8**, 1410 (2024).

[40] R. Takagi, S. Endo, S. Minagawa, and M. Gu, Fundamental limits of quantum error mitigation, npj Quantum Inf. **8**, 114 (2022).

[41] K. Rudinger, G. J. Ribeill, L. C. Govia, M. Ware, E. Nielsen, K. Young, T. A. Ohki, R. Blume-Kohout, and T. Proctor, Characterizing midcircuit measurements on a superconducting qubit using gate set tomography, Phys. Rev. Appl. **17**, 014014 (2022).

[42] Y. Xu, G. Huang, J. Balewski, R. Naik, A. Morvan, B. Mitchell, K. Nowrouzi, D. I. Santiago, and I. Siddiqi, QubiC: An open-source FPGA-based control and measurement system for superconducting quantum information processors, IEEE Trans. Quantum Eng. **2**, 1 (2021).

[43] Y. Xu, G. Huang, N. Fruitwala, A. Rajagopala, R. K. Naik, K. Nowrouzi, D. I. Santiago, and I. Siddiqi, QubiC 2.0: An extensible open-source qubit control system capable of mid-circuit measurement and feed-forward, ArXiv:2309.10333.

[44] R. S. Gupta, E. v. d. Berg, M. Takita, D. Risté, K. Temme, and A. Kandala, Probabilistic error cancellation for dynamic quantum circuits, Phys. Rev. A **109**, 062617 (2024).

[45] A. Erhard, J. J. Wallman, L. Postler, M. Meth, R. Stricker, E. A. Martinez, P. Schindler, T. Monz, J. Emerson, and R. Blatt, Characterizing large-scale quantum computers via cycle benchmarking, Nat. Commun. **10**, 5347 (2019).

[46] P. Ivashkov, G. Uchehara, L. Jiang, D. S. Wang, and A. Seif, High-fidelity, multi-qubit generalized measurements with dynamic circuits, PRX Quantum **5**, 030315 (2024).

[47] F. Mallet, F. R. Ong, A. Palacios-Laloy, F. Nguyen, P. Bertet, D. Vion, and D. Esteve, Single-shot qubit readout in circuit quantum electrodynamics, Nat. Phys. **5**, 791 (2009).

[48] N. Fruitwala, G. Huang, Y. Xu, A. Rajagopala, A. Hashim, R. K. Naik, K. Nowrouzi, D. I. Santiago, and I. Siddiqi, Distributed architecture for FPGA-based superconducting qubit control, ArXiv:2404.15260.

[49] N. R. Vora, Y. Xu, A. Hashim, N. Fruitwala, H. N. Nguyen, H. Liao, J. Balewski, A. Rajagopala, K. Nowrouzi, Q. Ji, *et al.*, ML-powered FPGA-based real-time quantum state discrimination enabling mid-circuit measurements, ArXiv:2406.18807.