

# Readout Error Mitigation for Mid-Circuit Measurements and Feedforward

Jin Ming Koh<sup>1,\*</sup>, Dax Enshan Koh<sup>2,3,4,†</sup> and Jayne Thompson<sup>5,6,3,‡</sup>

<sup>1</sup>*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*

<sup>2</sup>*Quantum Innovation Centre (Q.InC), Agency for Science, Technology and Research (A\*STAR),*

*2 Fusionopolis Way, Innovis #08-03, Singapore 138634, Singapore*

<sup>3</sup>*Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A\*STAR),*

*1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Singapore*

<sup>4</sup>*Science, Mathematics and Technology Cluster, Singapore University of Technology and Design,*

*8 Somapah Road, Singapore 487372, Singapore*

<sup>5</sup>*College of Computing and Data Science, Nanyang Technological University, Singapore 639798, Singapore*

<sup>6</sup>*Centre for Quantum Technologies, Nanyang Technological University, Singapore 637371, Singapore*



(Received 13 August 2024; revised 30 September 2025; accepted 10 December 2025; published 26 January 2026)

Current quantum computing platforms suffer from readout errors, where faulty measurement outcomes are reported by the device. These errors are particularly harmful in quantum programs that rely on branch statements, where operations in later parts of the program are dynamically determined by mid-circuit measurements. We propose a general protocol for mitigating mid-circuit measurement errors. Our solution works for any number of mid-circuit measurement and feedforward layers without increasing circuit depth or two-qubit gate counts, is efficient under structural assumptions on error channels, and is suitable for immediate application on noisy intermediate-scale quantum devices. We experimentally demonstrate up to a  $\sim 60\%$  reduction in error on superconducting quantum processors across several practically relevant feedforward circuits, including dynamic qubit resets, shallow-depth Greenberger–Horne–Zeiling state preparation, and multistage quantum teleportation. This work paves the way for more error-resilient adaptive quantum circuits, crucial for current and future quantum computing applications.

DOI: [10.1103/cj89-4h5t](https://doi.org/10.1103/cj89-4h5t)

## I. INTRODUCTION

In quantum computation, mid-circuit measurements and feedforward represent the analog of classical `if-else` branching statements in code. They allow us to make measurements on certain qubits during computation and decide which subsequent quantum operations to apply depending on the measurement outcomes. This capability is of immediate relevance in the present-day noisy intermediate-scale quantum (NISQ) era—enabling ancillary qubits to be reset mid-circuit so they can be reused during computation [1] and driving resource-saving protocols such as circuit knitting and stitching [2,3]. It also underpins quantum error correction [4–7], which functions by way of repeated syndrome measurements and conditioned recovery operations.

Indeed, such operations underlie many iconic quantum primitives and algorithms across diverse domains (e.g., resource distillation and injection [8–10], state and gate teleportation [11–13], quantum simulation [14–16], optimization algorithms [1], and efficient state preparation [17–22]).

However, this technology has an Achilles heel: it is highly vulnerable to readout errors, where a quantum device reports incorrect measurement outcomes [23,24]. In a program with mid-circuit measurements and feedforward, such errors in the measurement record induce the application of incorrect feedforward operations [Fig. 1(a)], thus initiating a cascade wherein the intermediary state of the unmeasured qubits is subsequently fed into an incorrect branch of the program. Standard readout error mitigation (REM) techniques assume all measurements happen at the end of the circuit and are only concerned with retrieving correct conglomerate measurement statistics through classical postprocessing [27–29]—they are not designed to handle this type of branching error occurring in the middle of computation.

In this work, we present *probabilistic readout error mitigation* (PROM), a general readout error mitigation method for quantum circuits containing mid-circuit measurements

\*Contact author: [jkoh@fas.harvard.edu](mailto:jkoh@fas.harvard.edu)

†Contact author: [dax\\_koh@a-star.edu.sg](mailto:dax_koh@a-star.edu.sg)

‡Contact author: [thompson.jayne2@gmail.com](mailto:thompson.jayne2@gmail.com)

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

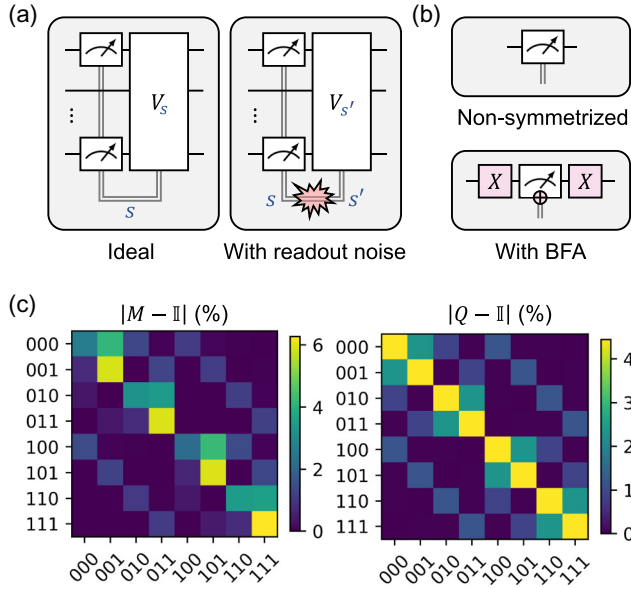


FIG. 1. Readout errors. (a) In an ideal noiseless setting, a layer of mid-circuit measurements produces an outcome (bitstring)  $s$ , which determines a subsequent feedforward quantum operation  $V_s$ . But in practice, readout errors can cause a measurement outcome  $s'$  to be reported instead, resulting in an incorrect operation  $V_{s'}$  applied. Such errors occur on a per-shot basis and their effect cannot be corrected in postprocessing. (b) Bit-flip averaging involves randomly sandwiching each measurement between a pair of inserted  $X$  gates (shaded pink) and applying a classical bit-flip to the outcome of the measurement. This symmetrizes the readout error channel [25,26]. (c) Without symmetrization, readout error channels are characterized by a confusion matrix  $M$  recording the probabilities of misreporting measurement outcomes (see Sec. II). Under bit-flip averaging, the confusion matrix  $M$  is symmetrized into  $Q$ , which is completely defined by its first column  $\mathbf{q}$  with entries describing readout error syndrome probabilities (see Sec. III A). The matrices displayed here are calibrated from superconducting quantum device *ibmq\_kolkata*.

and feedforward. Our protocol works by modifying the feedforward operations—we probabilistically sample from an engineered ensemble of feedforward trajectories and then perform an averaging process in postprocessing. The method accommodates multiple layers of mid-circuit measurements and feedforward and can be implemented using only additional single-qubit quantum gates and a modest increase in the number of samples as overhead. We provide analyses on resource requirements in terms of the number of extra shots of the circuit required, sensitivity to miscalibration of the device, and we experimentally demonstrate the effectiveness of the protocol on superconducting quantum processors in several settings of practical interest, namely qubit resets, shallow-depth Greenberger–Horne–Zeilinger (GHZ) state preparation, and multistage quantum state teleportation. Naturally, aside from mid-circuit measurements, PROM can subsume readout error mitigation on terminal measurements as a special case.

Our method integrates directly with probabilistic error cancellation (PEC) [30,31], which neglects readout errors, for complementary mitigation of quantum errors on dynamic circuits. Indeed, PROM is operationally similar to PEC, making them natural to combine: both symmetrize error channels by twirling, utilize quasi-probability sampling to effectively invert the error channel over an ensemble of shots, and incur analogous sampling overheads. More broadly, any quantum error suppression or mitigation technique (e.g., see Refs. [32–39]) can be layered atop PROM for a comprehensive targeting of hardware noise. Also PEC-like in structure, recent works have examined approximate order-by-order error channel inversion [40] and basic mitigation of uncorrelated errors in a single layer of mid-circuit measurements (Appendix C of Ref. [41]). By developing a readout error mitigation framework for general mid-circuit measurement and feedforward circuits supporting exact (closed-form) error channel inversion and methodological improvements in resource costs, we facilitate systematized error mitigation across generic dynamic circuit applications.

## II. FRAMEWORK

Our setup begins with the fundamental rationale of quantum error mitigation—that in most quantum computational settings, what one ultimately cares about is the measurement outcome statistics produced by the program. That is, many quantum algorithms involve (i) enacting a quantum process  $\mathcal{P}$  on  $n$  qubits to produce some state  $\rho_{\text{out}}$ , (ii) measuring a collection of observables  $\mathbf{O} = [O_1 \ O_2 \ \dots \ O_B]$  of bounded spectra on  $\rho_{\text{out}}$ , and (iii) repeating this process over  $N$  trials to obtain estimates of the expectation values  $\langle \mathbf{O} \rangle$  for some sufficiently large  $N$ . In such settings, the computation is considered successful provided one can estimate  $\langle \overline{O}_b \rangle$  for every  $b \in [B]$  to a designated accuracy  $\epsilon$  with some tolerable error probability  $\delta$ . Quantum error mitigation describes a family of protocols to successfully compute in NISQ settings where our capacity to precisely implement  $\mathcal{P}$  is compromised, at the expense of requiring a larger number  $N' > N$  of trials.

In our context, we focus on  $\mathcal{P}$  enacted through *mid-circuit measurements* and *feedforward*. For simplicity, we begin with the setting of a single layer of feedforward. Such a process involves:

- (1) Application of a unitary quantum circuit resulting in an intermediary quantum state  $\rho$ .
- (2) Single-qubit measurements on a subset of  $m \leq n$  qubits in the computational basis [42] resulting in a measurement outcome (bitstring)  $s \in \mathcal{S} = \{0, 1\}^m$ .
- (3) Application of an  $s$ -dependent unitary operation  $V_s$  as dictated by a feedforward function that designates one unitary to apply for each  $s$ .

We refer to step (2) as the mid-circuit measurements and step (3) as the feedforward operation.

In this work, we are concerned with *readout errors* occurring during mid-circuit measurements [see Fig. 1(a)], wherein the measurement apparatus reports a different outcome  $s' \in \mathcal{S}$  when the true outcome is  $s \in \mathcal{S}$  with some confusion probability  $M_{s's}$ . We call  $M$  the confusion matrix. This misreporting of measurement outcome causes two correlated sources of error:

- (1) While the true measurement outcome is  $s$  and the quantum state after measurement is collapsed to  $\rho_s \propto \Pi_s \rho \Pi_s$  for  $\Pi_s$  the projector associated with outcome  $s$ , the measurement outcome reported by the device is  $s'$ .
- (2) The incorrect feedforward operation  $V_{s'}$  is applied in place of  $V_s$ .

Such errors break the deterministic causal relation between the measurement outcome  $s$  and the feedforward operation  $V_s$ , resulting in sequences of operations which deviate from the intended program, where  $V_{s'}$  acts on states  $\rho_{s \neq s'}$ . The quantum state of the qubits is fed into an incorrect conditional branch of the circuit. Ultimately, this mangles the measured expectation values  $\langle \mathbf{O} \rangle$  such that they differ from the ideal values  $\langle \bar{\mathbf{O}} \rangle$ . We aim to develop techniques that mitigate the effects of these mid-circuit readout errors, so that the measured  $\langle \mathbf{O} \rangle$  accurately estimates  $\langle \bar{\mathbf{O}} \rangle$ .

Standard REM techniques are concerned only with mitigating readout errors for terminal measurements occurring at the end of the circuit [27–29]. Formally, suppose  $\mathbf{c}$  is a vector of counts whose element  $c_s$  corresponds to the number of times we observe outcome  $s$  over  $N$  shots of the circuit, and let  $\mathbf{c}'$  be the hypothetical ideal vector of counts were there no readout errors. Then for  $N \gg 1$ , we have the relation  $\mathbf{c} \approx M\mathbf{c}'$  where  $M$  is referred to as the confusion matrix [43] [see Fig. 1(c)]. Various REM techniques aim to estimate  $\mathbf{c}'$  when given  $\mathbf{c}$ , assuming  $M$  has been calibrated through device benchmarking [44].

Applying terminal REM schemes to our context of mid-circuit measurements with feedforward encounters immediate problems. Approximating the noiseless counts of the terminal measurements and thereby  $\langle \bar{\mathbf{O}} \rangle$  in the fashion above, without attention to the erroneous branching of the program, cannot lead to a correct estimation of the true  $\langle \bar{\mathbf{O}} \rangle$  regardless of the number of shots we take. At the same time, correcting the erroneous combinations of  $V_{s'}$  applied on  $\rho_s$  naïvely requires knowledge of the true measurement outcomes  $s$  at the single-shot level, which is inaccessible information. Fundamentally, *standard readout error mitigation does not work for mid-circuit measurement errors*.

We must therefore go beyond terminal readout error mitigation to develop a paradigm that functions for mid-circuit measurements and feedforward. Naturally, such a paradigm subsumes terminal measurements as a

special case. In what follows, we present such a class of methods. Section III begins with a general solution for the context introduced above involving a single layer of mid-circuit measurements and feedforward. We then extend our solution to arbitrarily many layers in Sec. IV. We also discuss a variant of the protocol for mitigating readout errors on both mid-circuit and terminal measurements in Appendix B 5.

### III. CORE PROTOCOL

#### A. Main results

We first begin by leveraging bit-flip averaging (BFA) [25,26]. This involves sandwiching each measurement between a pair of Pauli- $X$  gates with probability  $1/2$ , while simultaneously flipping the outcome [see Fig. 1(a)]. This symmetrizes readout errors in our system, such that the probability of any error occurring is rendered independent of the input state. The readout error channels are then completely characterized by a  $2^m$ -element probability vector  $\mathbf{q}$  in place of the more cumbersome  $2^m \times 2^m$  confusion matrix  $M$  [see Fig. 1(c)]. Each element  $q_s$  represents the probability that error syndrome  $s$  occurs—i.e., if the true measurement outcome is  $a \in \mathcal{S}$ , then  $q_s$  is the probability of reporting  $(a \oplus s)$ , where  $\oplus$  denotes the modulo-2 addition, or XOR, of two bitstrings. The element  $q_0$  is then the probability of the all-zero syndrome  $\mathbf{0}$  occurring, or equivalently readout being error-free. Note that, other than enabling a succinct description of the readout error channels, this symmetrization does not affect the operational behavior of the mid-circuit measurements. We assume that the characterization  $\mathbf{q}$  has been made available through prior calibration or noise learning, through any suitable method.

Our key idea is to intervene in the feedforward operations on a shot-by-shot basis. We facilitate this by introducing a bitmask  $f \in \mathcal{S}$  in each shot to be applied to the outcome of the  $m$  qubits we measure mid-circuit. When the mid-circuit measurements report an outcome  $s' \in \mathcal{S}$ , we invoke the feedforward operation  $V_{s' \oplus f}$  instead of  $V_{s'}$ . Our first key result is that there exists a systematic procedure to choose a distribution for  $f$  such that the effect of readout errors on the measured  $\langle \mathbf{O} \rangle$  is mitigated.

Specifically, to determine the bitmask to pick in each shot of  $\mathcal{P}$ , we introduce a vector of weights  $\boldsymbol{\alpha} = [\alpha_0 \dots \alpha_1]^T$  with norm  $\xi = \|\boldsymbol{\alpha}\|_1 = \sum_{f \in \mathcal{S}} |\alpha_f|$ . These weights are not necessarily non-negative and can be regarded as a quasi-probability distribution after normalization. The weights  $\boldsymbol{\alpha}$  can be systematically computed as a function of our readout error model (see below). We are now equipped to introduce our PROM protocol in Algorithm 1.

Our key observation is that  $\boldsymbol{\alpha}$  can be systematically chosen such that the protocol described in Algorithm 1 produces an unbiased estimator  $\hat{\mathbf{O}}$  for the observables  $\mathbf{O}$ .

## ALGORITHM 1. Probabilistic readout error mitigation for a single feedforward layer.

**Inputs:** Circuit  $\mathcal{P}$  containing a layer of mid-circuit measurements and feedforward [see Fig. 2(a)]. Probability vector  $\mathbf{q}$  from prior calibration characterizing readout error channels affecting mid-circuit measurements. Arbitrary measured observables  $\mathbf{O}$ . Number of shots  $N'$  to be executed.

**Output:** Estimate  $\langle \hat{\mathbf{O}} \rangle$  of expectation values.

We construct a bit-flip-averaged [see Fig. 2(b)] version of  $\mathcal{P}$  and an appropriate weight vector  $\alpha$  from  $\mathbf{q}$  (see Result 1). For shot  $k = 1$  to  $N'$ , we

1. Sample a bitmask  $f \in \mathcal{S}$  with probability  $|\alpha_f|/\xi$ .
2. Execute the mid-circuit measurements with bitmask  $f$  applied to the feedforward.
3. Measure  $\mathbf{O}$  to obtain outcome  $\omega^{(k)}$ . Accordingly each entry  $\omega_b^{(k)}$  is an eigenvalue of  $O_b$ .
4. Record  $\mathbf{o}^{(k)} = \omega^{(k)} \text{sgn}(\alpha_f)$ .

Output the estimate  $\langle \hat{\mathbf{O}} \rangle = \xi \cdot \frac{1}{N'} \sum_{k=1}^{N'} \mathbf{o}^{(k)}$ .

*Result 1 (Correctness).* Consider Algorithm 1 with readout error channels characterized by  $\mathbf{q}$ . If we choose

$$\alpha = \frac{1}{2^m} \mathcal{W} \left( \frac{\mathbf{1}}{\mathcal{W}(\mathbf{q})} \right), \quad (1)$$

where  $\mathcal{W}(\cdot)$  denotes the unnormalized Walsh-Hadamard transform, then the  $\langle \hat{\mathbf{O}} \rangle$  obtained from the protocol is an unbiased estimate of the ideal noiseless observable expectation values  $\langle \mathbf{O} \rangle$ . That is,  $\langle \hat{\mathbf{O}} \rangle = \langle \mathbf{O} \rangle$  in the limit of a large number of shots.

We give the intuition behind this result in Sec. III B and proof in Appendix B 1 a. In practice, standard numerical algorithms [45,46] can be used to efficiently compute the Walsh-Hadamard transforms required above.

As Algorithm 1 modifies the input quantum circuit only through introducing the feedforward bitmasks  $f$  and bit-flip averaging, no additional two-qubit gate or circuit depth

cost is incurred. In addition to being unbiased, statistical analysis shows that this error mitigation protocol carries a sampling overhead of  $\xi^2$  compared to a hypothetical noiseless experiment. This is made precise by the following result.

*Result 2 (Convergence).* Using Algorithm 1 with  $\alpha$  chosen for an unbiased estimator (as in Result 1), to obtain the same precision of measured  $\langle \mathbf{O} \rangle$  at the same confidence as  $N$  hypothetical noiseless shots on the input circuit  $\mathcal{P}$ , a number of noisy shots scaling as  $N' = \xi^2 N$  is sufficient.

In fact, under general conditions, for a total readout error probability  $\eta = 1 - q_0 < 1/2$  (recall that  $q_0$  is the probability of readout being error-free), the overhead factor can be shown to be bounded by  $\xi < 1/(1 - 2\eta)$ . In the noiseless limit  $\alpha_0 = \delta_{f,0}$  and  $\xi = 1$ . We refer readers to Appendix B 3 for the analysis underlying these results.

Last, we remark on the resource costs of the protocol. Result 2 implies that a factor of  $\xi^2$  more shots of

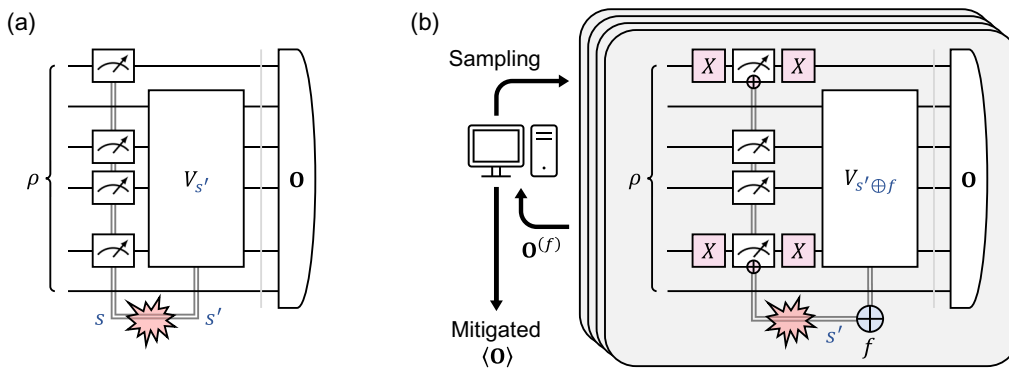


FIG. 2. Probabilistic readout error mitigation for a single feedforward layer. (a) Structure of the quantum circuit considered, comprising state preparation, a layer of mid-circuit measurements and feedforward, and terminal measurements of observables  $\mathbf{O}$ . From shot to shot, the correct measurement outcome (bitstring)  $s$  is corrupted into  $s'$  by readout errors, and the reported outcome  $s'$  conditions the feedforward unitary  $V_{s'}$  that is applied. (b) Schematic of our readout error mitigation protocol. Bit-flip averaging is applied by randomly twirling measurements with quantum and classical  $X$  gates (shaded pink), and the XOR between the reported measurement outcome and a bitmask  $f$  is used for feedforward. By sampling  $f$  over a suitable probability distribution and with some classical postprocessing, mitigated expectation values of  $\mathbf{O}$  are produced.

TABLE I. Summary of characteristics of the probabilistic readout error mitigation protocol studied in this paper. Here, on a circuit with  $L$  layers of mid-circuit measurements and feedforward,  $m = m^{[1]} + m^{[2]} + \dots + m^{[L]}$  denotes the total number of measurements,  $m^{[l]}$  is the number of qubits measured in layer  $l$ , and  $\bar{m}$  is the maximum number of measurements in any single layer. As a special case, when there is only a single layer ( $L = 1$ ),  $m$  is the number of qubits measured mid-circuit. Under independence assumptions,  $r_\ell \in [0, 1/2]$  are characteristic readout error probabilities for measurement  $\ell \in [m]$ , and under uniformity assumptions  $r_\ell = r$  identically. The “Init. time” and “Init. space” columns refer to the initial classical time and memory costs of computing coefficients for the mitigated estimator, “Sampling time” refers to the classical sampling time to generate a bitmask per experiment shot (see Sec. III), and “Overhead” refers to a multiplicative factor in the number of experiment shots required for the mitigated estimator to converge to the same precision at the same confidence as hypothetical noiseless circuit execution. Calibration costs to obtain the error channel characterization  $\mathbf{q}$  are separate from error mitigation and are not included here.

| Layers | Assumptions                       | Classical resources                |                                    |                  | Overhead $\xi^2$                      |
|--------|-----------------------------------|------------------------------------|------------------------------------|------------------|---------------------------------------|
|        |                                   | Init. time                         | Init. space                        | Sampling time    |                                       |
| 1      | Arbitrary correlations in errors  | $\mathcal{O}(m \cdot 2^m)$         | $\mathcal{O}(2^m)$                 | $\mathcal{O}(1)$ | $\ \alpha\ _1^2$                      |
|        | Independent errors                | $\mathcal{O}(m)$                   | $\mathcal{O}(m)$                   | $\mathcal{O}(m)$ | $\prod_{\ell=1}^m (1 - 2r_\ell)^{-2}$ |
|        | Uniform error                     | $\mathcal{O}(1)$                   | $\mathcal{O}(1)$                   | $\mathcal{O}(m)$ | $(1 - 2r)^{-2m}$                      |
| $L$    | Arbitrary correlations in errors  | $\mathcal{O}(m \cdot 2^m)$         | $\mathcal{O}(2^m)$                 | $\mathcal{O}(1)$ | $\ \alpha\ _1^2$                      |
|        | Independent errors between layers | $\mathcal{O}(m \cdot 2^{\bar{m}})$ | $\mathcal{O}(L \cdot 2^{\bar{m}})$ | $\mathcal{O}(L)$ | $\prod_{l=1}^L (\xi^{[l]})^2$         |
|        | Fully independent errors          | $\mathcal{O}(m)$                   | $\mathcal{O}(m)$                   | $\mathcal{O}(m)$ | $\prod_{\ell=1}^m (1 - 2r_\ell)^{-2}$ |
|        | Uniform error                     | $\mathcal{O}(1)$                   | $\mathcal{O}(1)$                   | $\mathcal{O}(m)$ | $(1 - 2r)^{-2m}$                      |

the noisy circuit is needed to achieve the same precision as in a noiseless scenario. This sampling overhead factor is exponential in the number of measurements—in particular,  $\xi^2 \sim (1 - 2r)^{-2m}$  for  $m$  mid-circuit measurements each with characteristic readout error probability  $r$  (more details in Sec. III D)—and is likely to be the chief concern in practical applications.

Of separate consideration are the classical resource costs. In the general setting, there are  $\sim 2^m$  independent entries of  $\alpha$  to be computed, which demand classical resources scaling exponentially in  $m$ . But, thereafter, using standard algorithms [47–49] to sample from the computed probability distribution, the drawing of bitmasks  $f \in \mathcal{S}$  can be performed in constant time per shot. By taking advantage of structure in the readout error channels, the exponential classical cost in  $m$  can be removed, as we describe in Sec. III D. These sampling and classical processing complexities are summarized in Table I. Note that the calibration costs of  $\mathbf{q}$  are separate and not reflected here; however, once characterized, the same  $\mathbf{q}$  can be reused to error-mitigate different experiments [50].

## B. Underlying principles

We now outline the underlying mechanism behind our protocol. We start from the quantum state after the mid-circuit measurements and feedforward in the absence of readout errors, which can be written as

$$\bar{\sigma} = \sum_{s \in \mathcal{S}} p_s V_s \rho_s V_s^\dagger, \quad (2)$$

where the measurement outcome  $s$  defines the quantum trajectory that the system has taken and results in a normalized quantum state  $\rho_s = (\Pi_s \rho \Pi_s) / p_s$  when  $p_s \neq 0$ . Here  $\Pi_s$  is the measurement projector associated with the outcome  $s$ , and the probability of the trajectory occurring is  $p_s = \text{tr}(\Pi_s \rho)$ . We define the tensor

$$T_{bss'} = p_s \text{tr} \left( O_b V_{s'} \rho_s V_{s'}^\dagger \right), \quad (3)$$

such that the entry  $T_{bss'}$  is the expectation value of observable  $O_b$  measured on quantum trajectories that had measurement outcome  $s$  but feedforward unitary  $V_{s'}$  applied, weighted by the probability  $p_s$  of those trajectories occurring. Thus, the diagonal and off-diagonal entries of the matrix  $T_b$  describe trajectories in which feedforward is correctly and incorrectly performed, respectively. Then

$$\langle \bar{O}_b \rangle = \text{tr} (O_b \bar{\sigma}) = \text{tr} T_b. \quad (4)$$

Now we consider the presence of readout errors. On a per-shot basis, a true mid-circuit measurement outcome  $s$  is corrupted into  $s'$ , and in addition we impose a bitmask  $f \in \mathcal{S}$  such that feedforward unitary  $V_{s' \oplus f}$  is applied, in accordance with the PROM protocol. The quantum state after the mid-circuit measurements and feedforward is

$$\sigma^{(f)} = \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} q_{s \oplus s'} p_s V_{s' \oplus f} \rho_s V_{s' \oplus f}^\dagger, \quad (5)$$

where  $q_{s \oplus s'}$  is the probability of the readout error occurring (as described in Sec. III A). Accordingly

$$\langle O_b^{(f)} \rangle = \text{tr} (O_b \sigma^{(f)}) = \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} q_{s \oplus s'} T_{bs(s' \oplus f)}. \quad (6)$$

We consider an estimator that is a weighted average of observable measurements over bitmasks,

$$\hat{\mathbf{O}} := \sum_{f \in \mathcal{S}} \alpha_f \mathbf{O}^{(f)} = \xi \sum_{f \in \mathcal{S}} \frac{|\alpha_f|}{\xi} \text{sgn}(\alpha_f) \mathbf{O}^{(f)}. \quad (7)$$

Desiring an unbiased estimator, we demand  $\langle \hat{\mathbf{O}} \rangle = \langle \mathbf{O} \rangle$ . Then, substituting Eqs. (4) and (6) into this equality, we obtain

$$\sum_{f \in \mathcal{S}} \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \alpha_f q_{s \oplus s' \oplus f} T_{bss'} = \text{tr} T_b. \quad (8)$$

We do not assume further knowledge of the problem at hand; thus, the tensor  $T$  is unknown and the mitigation protocol must work for arbitrary  $T$ . Then the equality necessitates that the coefficient of every  $T_{bss'}$  match on both sides, yielding the linear system

$$\sum_{f \in \mathcal{S}} \alpha_f q_{s \oplus s' \oplus f} = \delta_{ss'} \quad \forall s, s' \in \mathcal{S}. \quad (9)$$

Inverting this linear system gives a unique closed-form solution for the weights  $\alpha$ . In particular,  $\alpha$  is the first column of  $Q^{-1}$ , where  $Q$  is the symmetrized confusion matrix defined by  $Q_{sf} = q_{s \oplus f}$  for  $s, f \in \mathcal{S}$  [see Fig. 1(c)]. The inverse of  $Q$  is easily computed due to the symmetric structure of  $Q$ —in particular,  $Q$  and every recursively halved quadrant of  $Q$  is symmetric about its diagonals and antidiagonals. This means  $Q \in \text{span}(\{\mathbb{I}, X\}^{\otimes m})$  and has closed-form eigenvalues,

$$\lambda_k = \sum_{s \in \mathcal{S}} (-1)^{k \cdot s} q_s = 1 - 2 \sum_{\substack{s \in \mathcal{S} \\ k \cdot s \neq 0}} q_s, \quad (10)$$

for each  $k \in \mathcal{S}$ —see Appendix A 3 for technical elaboration. Thus a closed-form solution for the weights is

$$\alpha_f = \frac{1}{2^m} \sum_{k \in \mathcal{S}} (-1)^{f \cdot k} \left( \frac{1}{\lambda_k} \right). \quad (11)$$

Equivalently we can write  $\lambda = \mathcal{W}(\mathbf{q})$  where  $\mathcal{W}$  is the unnormalized Walsh-Hadamard transform [45,46] and similarly  $\alpha = \mathcal{W}(\mathbf{1}/\lambda)/2^m$ , as presented in Result 1.

A full derivation of the algorithm, and alternate forms for  $\alpha$ , are given in Appendix B 1 a.

### C. Robustness to calibration errors

A pertinent further question concerns the sensitivity or well-conditionedness of the error mitigation protocol. That is, given differing  $\mathbf{q}$  and  $\mathbf{q}'$  inputs characterizing the readout errors present, how large are the differences in the mitigated expectation values  $\langle \hat{\mathbf{O}} \rangle$  and  $\langle \hat{\mathbf{O}}' \rangle$ , and secondarily

the overhead factors  $\xi$  and  $\xi'$ ? This question is relevant as in reality, calibration of  $\mathbf{q}$  on a quantum device inevitably carries small errors up to a finite experimental precision; moreover, it may be desirable to employ approximations on  $\mathbf{q}$ , for example as described next in Sec. III D. We show that small errors in inputs propagate mildly through the protocol.

*Result 3 (Robustness to calibration errors).* Given two different characterizations of readout errors  $\mathbf{q}$  and  $\mathbf{q}'$  distinguished by total variation distance  $d$ , the respective mitigated expectation values  $\langle \hat{\mathbf{O}} \rangle$  and  $\langle \hat{\mathbf{O}}' \rangle$  produced by Algorithm 1 and the overhead factors  $\xi$  and  $\xi'$  satisfy

$$|\langle \hat{\mathbf{O}}_b \rangle - \langle \hat{\mathbf{O}}'_b \rangle| \leq \frac{2\xi^2 d}{1 - 2\xi d} \cdot \|\mathbf{O}_b\|_2, \quad (12a)$$

$$|\xi' - \xi| \leq \frac{2\xi^2 d}{1 - 2\xi d} \quad (12b)$$

for  $2\xi d < 1$  and every  $b \in [B]$ , where  $\|\cdot\|_2$  denotes the spectral norm.

In particular, invoking Result 3 and taking  $\mathbf{q}'$  and  $\mathbf{q}$  as the (unknown) exact characterization and the characterization used in the error mitigation protocol respectively, the error in mitigated observables scales linearly as  $\sim 2\xi^2 d$  to leading order where  $d$  is small given sufficiently good calibration of readout error probabilities on the device.

Secondarily, suppose that the characterized readout errors  $\mathbf{q}$  on the quantum device contain small correlations, but the errors are otherwise close to being independent. Then Eq. (12b) implies that the overhead factor of the mitigation protocol is close to that assuming independent errors, which is easy to evaluate, as analyzed next in Sec. III D. We refer readers to Appendix B 4 for derivations of the sensitivity bounds above.

### D. Simplified solutions for uncorrelated channels

In a setting where readout errors affecting the  $m$  mid-circuit measurements are assumed to be independent, the solution for  $\alpha$  factorizes,

$$\alpha = \bigotimes_{\ell=1}^m \alpha^{(\ell)}, \quad \alpha^{(\ell)} = \frac{1}{1 - 2r_\ell} \begin{bmatrix} 1 - r_\ell \\ -r_\ell \end{bmatrix}, \quad (13)$$

where  $r_\ell \in [0, 1/2)$  is the readout error probability for the  $\ell$ th measurement. Writing  $\xi^{(\ell)} = \|\alpha^{(\ell)}\|_1$ , the overhead factor may be expressed simply as

$$\xi = \prod_{\ell=1}^m \xi^{(\ell)} = \prod_{\ell=1}^m \frac{1}{1 - 2r_\ell}. \quad (14)$$

The factorized structure of  $\alpha$  implies that  $|\alpha|/\xi$  is a product over individual  $|\alpha^{(\ell)}|/\xi^{(\ell)}$  probability distributions. This enables a reduction of classical resources

required to initialize the mitigation protocol to  $\mathcal{O}(m)$  time and space, with  $\mathcal{O}(m)$  sampling time for each shot. In an even simpler setting of uniform readout channels ( $r_\ell = r$ ), all  $\alpha^{(\ell)}$  are identical and classical initialization costs are trivially reduced to  $\mathcal{O}(1)$  time and space. These complexities are reflected in Table I, and further technical details are provided in Appendix B 1 b.

In practice, readout errors on a quantum device expectedly exhibit at least a small amount of correlations or nonuniformity. How large would the consequences of assuming independence or uniformity be? Our prior discussion surrounding Eq. (12a) has established a mild scaling of errors in the mitigated observables versus ideal values, with respect to the (small) total variation distance between the true  $\mathbf{q}$  and that used in the protocol. For small distances, errors made by an approximate mitigation assuming independence or uniformity are small, and importantly, the approximation enables significant resource complexity savings against the general setting. The utility of such simplifications thus depends on the quantum device and the size of the quantum circuit. We discuss this further in our experimental study in Sec. V.

#### IV. GENERALIZATION TO MULTIPLE FEEDFORWARD LAYERS

Many operational scenarios involve multiple layers of mid-circuit measurements and feedforward, thus extending beyond the single-layer setting introduced in Sec. II and analyzed in Sec. III. This multilayer setting is illustrated in Fig. 3(a) and is described by

- (1) Application of a unitary quantum circuit resulting in an intermediary quantum state  $\rho$ .
- (2) For layers  $l = 1, \dots, L$ :

- (a) Measurement of  $m^{[l]} \leq n$  qubits in the computational basis [51] resulting in a measurement outcome (bitstring)  $s^{[l]} \in \mathcal{S}^{[l]} = \{0, 1\}^{m^{[l]}}$ .
- (b) Application of an  $s^{[l]}$ -dependent unitary operation  $V_{s^{[l]}}^{[l]}$  as dictated by a feedforward function that designates one unitary to apply in this layer for each  $s^{[l]}$ .

In this generalized setting readout errors can occur in every mid-circuit measurement and feedforward layer—in each layer  $l$  readout errors can cause a measurement outcome  $s'^{[l]}$  to be reported while the true outcome is  $s^{[l]}$ , resulting in the application of an incorrect feedforward unitary  $V_{s'^{[l]}}^{[l]}$  in that layer [see Fig. 3(a)]. We denote by  $m$  the total number of measurements, such that  $m = \sum_{l=1}^L m^{[l]}$  for  $m^{[l]}$  the number of measurements in layer  $l$ . In the special case of a single layer,  $m$  reduces to the number of measurements in the circuit, as used in Sec. III.

#### A. Multilayer error mitigation protocol

While multilayer feedforward incurs readout errors in each layer, which can cause a sequence of incorrect feedforward unitaries to be applied, we show that our error mitigation scheme continues to work with slight modifications. Here, operationally, symmetrized readout error channels are still characterized by a vector  $\mathbf{q}$ , where each element  $q_s$  represents the probability that the multilayer error syndrome  $s \in \mathcal{S} = \{0, 1\}^m$  occurs. That is, if the true measurement outcome is  $a = a^{[1]}a^{[2]} \dots a^{[L]} \in \mathcal{S}$  concatenated across the layers, then  $q_s$  is the probability that the quantum device reports  $(a \oplus s)$ , where  $s = s^{[1]}s^{[2]} \dots s^{[L]}$  is the syndrome across layers; here  $a^{[l]} \in \mathcal{S}^{[l]}$  and  $s^{[l]} \in \mathcal{S}^{[l]}$  are the true measurement outcome and syndrome in layer  $l$  respectively.

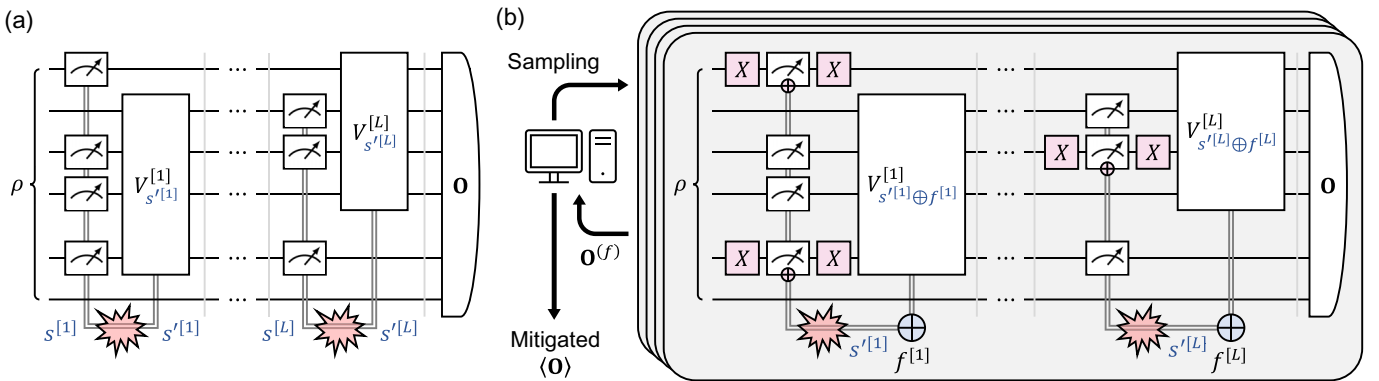


FIG. 3. Probabilistic readout error mitigation for multiple feedforward layers. (a) Structure of the quantum circuit considered, comprising state preparation,  $L$  layers of mid-circuit measurements and feedforward, and terminal measurements of observables  $\mathbf{O}$ . From shot to shot, in layer  $l$ , the correct measurement outcome (bitstring)  $s^{[l]}$  is corrupted into  $s'^{[l]}$  by readout errors, and the reported outcome  $s'^{[l]}$  determines a feedforward unitary  $V_{s'^{[l]}}^{[l]}$  that is applied. (b) Schematic of our readout error mitigation protocol. Bit-flip averaging is applied by randomly twirling measurements with quantum and classical  $X$  gates (shaded pink), and in layer  $l$  the XOR between the reported measurement outcome and a bitmask  $f^{[l]}$  is used for feedforward. By sampling  $f = f^{[1]}f^{[2]} \dots f^{[L]}$  over a suitable probability distribution and with some classical postprocessing, mitigated expectation values of  $\mathbf{O}$  are produced.

## ALGORITHM 2. Probabilistic readout error mitigation for multiple feedforward layers.

**Inputs:** Circuit  $\mathcal{P}$  containing  $L$  layers of mid-circuit measurements and feedforward [see Fig. 3(a)]. Probability vector  $\mathbf{q}$  from prior calibration characterizing readout error channels affecting mid-circuit measurements. Arbitrary measured observables  $\mathbf{O}$ . Number of shots  $N'$  to be executed.

**Output:** Estimate  $\langle \hat{\mathbf{O}} \rangle$  of expectation values.

We construct a bit-flip-averaged [see Fig. 3(b)] version of  $\mathcal{P}$  and an appropriate weight vector  $\alpha$  from  $\mathbf{q}$  (see Result 4). For shot  $k = 1$  to  $N'$ , we

1. Sample a bitmask  $f = f^{[1]}f^{[2]} \dots f^{[L]} \in \mathcal{S}$  with probability  $|\alpha_f|/\xi$ .
2. For layers  $l = 1, \dots, L$ , execute the mid-circuit measurements in layer  $l$  with bitmask  $f^{[l]}$  applied to the feedforward.
3. Measure  $\mathbf{O}$  to obtain outcome  $\omega^{(k)}$ .
4. Record  $\mathbf{o}^{(k)} = \omega^{(k)} \text{sgn}(\alpha_f)$ .

Output the estimate  $\langle \hat{\mathbf{O}} \rangle = \xi \cdot \frac{1}{N'} \sum_{k=1}^{N'} \mathbf{o}^{(k)}$ .

The idea remains to sample a bitmask  $f \in \mathcal{S}$  in each experiment shot. We then partition  $f = f^{[1]}f^{[2]} \dots f^{[L]} \in \mathcal{S}$  into layerwise bitmasks  $f^{[l]} \in \mathcal{S}^{[l]}$ , and we apply  $f^{[l]}$  to the feedforward in layer  $l$ —that is, when the mid-circuit measurements report an outcome  $s^{[l]} \in \mathcal{S}^{[l]}$ , we apply the feedforward unitary  $V_{s^{[l]} \oplus f^{[l]}}^{[l]}$  instead of  $V_{s^{[l]}}^{[l]}$ . We detail this protocol below.

Our key result in this generalized setting is that the weights  $\alpha$  can be systematically computed with a similar approach as in the single-layer setting, to produce an unbiased estimator  $\hat{\mathbf{O}}$  for the observables  $\mathbf{O}$ . Moreover, the same characterization of the sampling overhead factor  $\xi^2 = \|\alpha\|_1^2$  of the protocol applies.

*Result 4 (Multilayer correctness, convergence).* Consider Algorithm 2 with readout error channels characterized by  $\mathbf{q}$ . Then Results 1 and 2 originally stated for Algorithm 1 continue to hold for the multilayer case under the condition that  $\mathbf{q}$  is the multilayer probability vector.

As before, when the total readout error probability  $\eta < 1/2$ , the general bound  $\xi < 1/(1 - 2\eta)$  applies. The same robustness to calibration errors (Result 3) also holds with  $\mathbf{q}$  and  $\mathbf{q}'$  being multilayer probability vectors.

The single-layer context discussed in Sec. III is thus subsumed as a special case with  $L = 1$ . We give the intuition for this extension of the error mitigation scheme to multiple feedforward layers in Sec. IV B below and proof in Appendix B 2. Similarly to the single-layer setting, we continue to enjoy no increase in two-qubit gate count or depth in the circuits executed in the protocol relative to the input circuit (other than a few single-qubit gates required for bit-flip averaging). We summarize classical processing and sampling costs for multilayer scenarios in Table I.

## B. Principles underlying multilayer extension

The main ideas leading to the solution  $\alpha$  follow closely the structure laid out in the single-layer context in

Sec. III B. Here, with  $s = s^{[1]}s^{[2]} \dots s^{[L]} \in \mathcal{S}$  the concatenation of measurement outcomes across all layers and  $\Pi_{s^{[l]}}^{[l]}$  the measurement projector in layer  $l$  associated with the outcome  $s^{[l]}$ , we write the sequence  $W$  of mid-circuit measurements and feedforward unitaries, and likewise define the tensor  $T$ ,

$$W_{ss'} = \prod_{l=L}^1 V_{s^{[l]} \oplus f^{[l]}}^{[l]} \Pi_{s^{[l]}}^{[l]}, \quad T_{bss'} = \text{tr} \left[ O_b W_{ss'} \rho W_{ss'}^\dagger \right]. \quad (15)$$

Directly analogous to Eq. (3), the entry  $T_{bss'}$  is the expectation value of  $O_b$  measured on quantum trajectories that had true measurement outcomes  $s$  but feedforward unitaries applied according to  $s'$ , weighted by the probability of those trajectories occurring.

Then the ideal expectation value of  $O_b$  in a hypothetical noiseless setting,  $\langle \overline{O}_b \rangle$ , is of exactly the same form as in Eq. (4) but with  $T$  defined above. Moreover, the expectation value of  $O_b$  with readout errors and with bitmask  $f \in \mathcal{S}$  applied in the feedforward,  $\langle O_b^{(f)} \rangle$ , is of the same form as in Eq. (6). Just as in Sec. III B, we demand  $\langle \hat{\mathbf{O}} \rangle = \langle \overline{\mathbf{O}} \rangle$  and thus equate  $\langle \overline{O}_b \rangle$  and  $\langle O_b^{(f)} \rangle$  through the definition of estimator  $\hat{\mathbf{O}}$ , leading to a requirement that the coefficient of every  $T_{bss'}$  match on both sides.

As  $\langle \overline{O}_b \rangle$  and  $\langle O_b^{(f)} \rangle$  have remained identical in form, we have the same solution for  $\alpha$ —namely, that presented in Result 1. We provide an explicit derivation and further technical details in Appendix B 2 a. The classical resource costs are likewise unchanged (listed in Table I).

## C. Simplified solutions for uncorrelated channels

As the measurements in different layers of the circuit are separated in time, a plausible assumption is that readout errors affecting them are independent, consistent with standard Markovian assumptions on noise channels [52,53]. Measurements within the same layer, which may occur simultaneously or in overlapping time intervals, may still exhibit correlations in errors. In such a setting, the confusion vector  $\mathbf{q}$  factorizes and so do the solutions for  $\lambda$

and  $\alpha$ ,

$$\mathbf{q} = \bigotimes_{l=1}^L \mathbf{q}^{[l]}, \quad \lambda = \bigotimes_{l=1}^L \lambda^{[l]}, \quad \alpha = \bigotimes_{l=1}^L \alpha^{[l]}, \quad (16)$$

where  $k = k^{[1]}k^{[2]} \dots k^{[L]} \in \mathcal{S}$  and  $f = f^{[1]}f^{[2]} \dots f^{[L]} \in \mathcal{S}$  partition the bitstrings over the layers, and  $\mathbf{q}^{[l]}$ ,  $\lambda^{[l]}$  and  $\alpha^{[l]}$  are the confusion vector, eigenvalues and mitigator coefficients for layer  $l$  respectively. Explicitly, the latter two are given by Eqs. (10) and (11) but with  $(\mathcal{S}, m, \mathbf{q}, k, f)$  replaced by their layer counterparts  $(\mathcal{S}^{[l]}, m^{[l]}, \mathbf{q}^{[l]}, k^{[l]}, f^{[l]})$ . Writing layerwise overheads  $\xi^{[l]} = \|\alpha^{[l]}\|_1$ , the overhead factor likewise factorizes,

$$\xi = \prod_{l=1}^L \xi^{[l]}. \quad (17)$$

Then  $|\alpha|/\xi$  is a product over the layerwise  $|\alpha^{[l]}|/\xi^{[l]}$  probability distributions. This allows a reduction of the classical resources required to  $\mathcal{O}(m \cdot 2^{\bar{m}})$  time and  $\mathcal{O}(L \cdot 2^{\bar{m}})$  space for solution computation, where  $\bar{m} = \max_{l \in [L]} m^{[l]}$  is the maximum number of measurements in a layer, and  $\mathcal{O}(L)$  classical sampling time per shot, as summarized in Table I. Further analysis details are provided in Appendix B 2 b.

For yet further simplification, a stronger assumption is to impose independence on readout errors across all measurements regardless of their location in layers, and even stronger still, to assume that all readout error channels are uniform. These settings enforce additional structure in the solutions and land us in the context of Sec. III D. The results for  $\alpha$  and classical resource costs there directly apply.

## V. HARDWARE EXPERIMENTS

To demonstrate the effectiveness of our PROM strategy, we executed a series of experiments on 27- and 127-qubit IBM transmon-based superconducting quantum processors. These devices support mid-circuit measurements and feedforward and have characteristic readout error rates of 0.5%–3% per measurement. As there is clear nonuniformity in readout error rates between qubits, we examine PROM both in the general case of arbitrarily correlated as well as independent readout errors, but we omit further uniformity assumptions. We provide more details on the quantum hardware used in Appendix C 1, with device specifications listed in Table III.

Our experiments span dynamic qubit reset (Sec. V A), shallow-depth GHZ state preparation (Sec. V B) and repeated quantum state teleportation (Sec. V C), thereby probing a variety of use-cases of practical interest. These three benchmarks are designed to cover increasingly complex adaptive circuit scenarios: namely, (i) a single layer

of mid-circuit measurements with a separable input state, (ii) a single layer of mid-circuit measurements with an entangled input state, and (iii) multiple sequential layers of mid-circuit measurements and feedforward, respectively. The former two cases exercise Algorithm 1, while the last exercises Algorithm 2.

We employ PROM for mid-circuit measurements and feedforward (Algorithms 1 and 2), in conjunction with standard REM for terminal measurements (see Appendix C 2). (This split implementation was for technical convenience; PROM can also subsume REM for terminal measurements at identical sample complexity and classical resource costs, as analyzed in Appendix B 5.) We clearly distinguish results with terminal REM only, and with both PROM and terminal REM. We implement dynamical decoupling on all circuits to suppress decoherence on idle qubits (see Appendix C 3).

For comparison, we examine also a repetition-based REM strategy for feedforward. The strategy works by performing each mid-circuit measurement  $r > 1$  times consecutively and taking a consensus outcome for feedforward, such that the readout error rate perceived by the feedforward is suppressed. This general approach is commonly invoked in quantum error correction and fault-tolerant quantum computation [4, 54]. For our purposes, we consider two variants—Rep-MAJ, which takes the majority outcome for odd  $r$ , and Rep-ALL, which requires agreement between all measurement outcomes for feedforward to go through and discards the experiment shot otherwise (i.e., postselection). While repetition-based strategies significantly increase the number of measurements and depth, and therefore runtime, of the quantum circuit, PROM does not modify the circuit beyond modifying the classical feedforward logic and inserting single-qubit  $X$  gates, which can often be merged with preceding and subsequent gates. This difference has nontrivial implications in practice, as we shall observe.

### A. Mid-circuit dynamic qubit reset

To start, we demonstrate PROM in the context of dynamic qubit resets. A reset is a (nonunitary) operation that reinitializes a qubit to  $|0\rangle$  regardless of the input state [55, 56]; on digital quantum processors, it can be implemented as a mid-circuit measurement followed by an  $X$  gate conditioned on a 1 outcome. As they enable arbitrary reuse of qubits, resets are of pivotal utility across quantum algorithms, simulation, optimization, and metrology domains [16, 57, 58]. Concretely, qubit resets have enabled long-time simulations of open quantum systems [14, 15], execution of wide quantum approximate optimization algorithm (QAOA) circuits on present hardware [1] and have been leveraged in phase estimation [59] and spectroscopy techniques [60]. More broadly, resets serve as an entropy sink in measurement-free quantum error correction

protocols [61,62] and are generally useful in quantum algorithms that repeatedly access (and discard into) an ancillary reservoir [63–66].

The circuit structure we examine comprises  $n$  system qubits that are prepared in  $(U_1 \otimes U_2 \otimes \cdots \otimes U_n) |0\rangle^{\otimes n}$ , dynamically reset, and then terminally measured to verify their post-reset state [see Fig. 4(a)]. A pair of spectator qubits sandwiches the system register and are each prepared in  $V|0\rangle$ . After witnessing the resets on the system qubits, they are back-rotated by  $V^\dagger$  and measured. In an ideal noiseless setting, the terminal states of the system and spectator qubits should all be  $|0\rangle$ . To probe the fidelity of the reset operations, our observables  $\mathbf{O}$  thus constitute  $|0\rangle^{\otimes n} \langle 0|^{\otimes n}$  on the system qubits and  $|0\rangle^{\otimes 2} \langle 0|^{\otimes 2}$  on the spectator qubits. In our experiments, we place all qubits on a contiguous chain on the quantum processor, and we take  $U_1 = \cdots = U_n = V = H$ .

We present in Fig. 4(b) system infidelities against the ideal  $|0\rangle^{\otimes n}$  post-reset state, for  $n \in \{1, 2, 3, 4\}$  qubits. We compare data without PROM, with general PROM for arbitrarily correlated readout errors and PROM under an assumption of tensored readout error channels, and with Rep-ALL at  $r=2$ . The drastic effect of PROM is evident—our protocol reduces system infidelities to  $\sim 30\%$ – $40\%$  of their unmitigated values. In contrast to PROM, the repetition-based Rep-ALL scheme does not appreciably improve system fidelities, and in fact worsens them for  $n > 1$ , across the board.

While PROM addresses readout errors affecting feedforward, on hardware a variety of other noise sources are present (e.g., gate errors and decoherence), which account for the observed residual infidelities. As discussed, in sensitive applications, PROM can be used in conjunction with other error mitigation techniques that complementarily address these other noise sources. In Fig. 4(b) we have also reported data without terminal REM, drawn as translucent bars. The observed fidelity improvements provided by terminal REM alone (without PROM) and by the addition of PROM are comparable, indeed expected as there are equal numbers of mid-circuit and terminal measurements on the system qubits ( $n$  each).

An additional layer of contrast is unveiled when examining spectator qubit infidelities against their ideal  $|0\rangle^{\otimes 2}$  postwitness state, as presented in Fig. 4(c). Expectedly, PROM has negligible effect on spectator qubits, as the protocol does not modify the circuit structure. In comparison the repetition-based schemes introduce additional consecutive layers of mid-circuit measurements. On hardware, measurements are typically long operations—on the superconducting devices used here they are 2–8 times the duration of a controlled-NOT (CX) gate (see Table III)—thus waiting for mid-circuit measurements can incur considerable decoherence; moreover there may be measurement cross-talk, wherein the readout of a qubit disturbs the state of surrounding qubits. These effects manifest in the

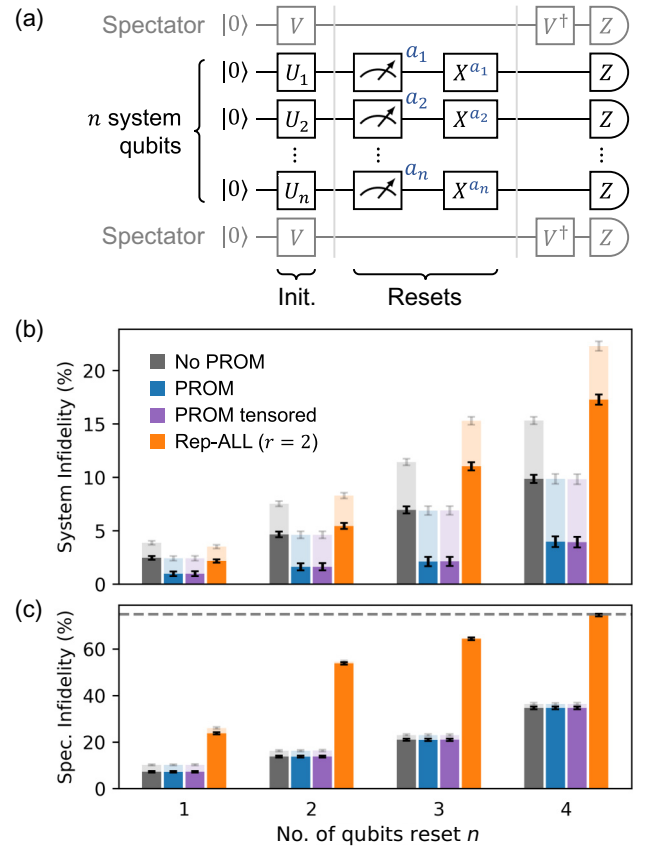


FIG. 4. Dynamic qubit resets. (a) Circuit structure to probe fidelity of qubit reset operations. A register of  $n$  system qubits (black) undergo evolution by  $U_1 \otimes \cdots \otimes U_n$  and are then reset into the  $|0\rangle^{\otimes n}$  state. A pair of adjacent spectator qubits (gray) sandwiching the system register witness the reset events. Dynamic resets are implemented as mid-circuit measurements followed by  $X$  gates conditioned on outcomes (blue labels). (b) Infidelities of system qubits after dynamic resets against  $|0\rangle^{\otimes n}$ , comparing probabilistic readout error mitigation, probabilistic readout error mitigation assuming tensored readout error channels, Rep-ALL at  $r=2$ , and without readout error mitigation for mid-circuit measurements and feedforward. (c) Infidelities of spectator qubits against  $|0\rangle^{\otimes 2}$  after witnessing reset events, conditioned on the system qubits being reset correctly. Dashed line demarcates 75% infidelity expected for a maximally mixed spectator state. In all plots, solid bars are with terminal readout error mitigation; translucent bars without. Experiments performed on superconducting quantum device *ibmq\_kolkata*. Error bars are standard deviations across 1000 trials per qubit chain, each with 10 000 shots; data averaged over  $\sim 8$  randomly chosen qubit chains.

poor observed spectator fidelities of Rep-ALL. In particular, at  $n = 4$  the spectator infidelity approaches that expected from a maximally mixed state, suggesting significant depolarization. This is despite the inclusion of dynamical decoupling in our experiments (see Appendix C 3) that conventionally suppresses background decoherence.

Last, we comment that, as observed in Figs. 4(b) and 4(c), PROM with and without independent readout error

assumptions exhibit similar mitigation performance. The former generally produced system infidelities  $\sim 1\%$ – $2\%$  worse relative to the latter. We checked that the total variation distance between characterized confusion vectors  $\mathbf{q}$  and their tensored counterparts assuming independent readout errors are  $\lesssim 1\%$ , thus this similarity is consistent with the sensitivity properties of the protocol as stated in Eq. (12a). On hardware exhibiting larger error correlations, imposing independence assumptions will introduce a larger mitigation quality penalty than observed here.

We report additional data acquired on two other superconducting quantum devices in Appendix D 1, which are qualitatively similar to those shown here in Figs. 4(b) and 4(c). There, we also show system and spectator infidelities for Rep-ALL and Rep-MAJ at  $r = 3$ —which, as could be expected from the  $r = 2$  data, are much too large to be of practical relevance.

### B. Shallow-depth GHZ state preparation

Next we demonstrate PROM in the context of preparing GHZ states, which are entangled states of foundational importance across quantum information, metrology, and communications domains [67–70]. On linear nearest-neighbor qubit connectivity, the unitary preparation of an  $n$ -qubit GHZ state ( $n$  even) necessarily requires  $n - 1$  CX gates with  $n/2$  depth [71]. The optimal unitary circuit comprises a qubit starting in the  $|+\rangle$  state followed by a bidirectional fan-out of CX gates to grow the GHZ state [see Fig. 9(b)].

With mid-circuit measurements and feedforward, however, a GHZ state can be prepared in constant depth independent of  $n$  [17,71,72]. Here we examine a circuit structure that is an interpolation [72] between the unitary and feedforward circuit extremes [see Fig. 5(a)], such that the number of CX gates and measurements in the circuit and circuit depth become adjustable. We consider splitting the  $n$  qubits into  $b$  blocks each containing  $p$  system qubits and an ancillary qubit; smaller  $p$ -qubit GHZ states are unitarily prepared in each of the blocks, and the blocks are then merged in constant depth using mid-circuit measurements and feedforward. Conceptually, the ancillae perform parity checks on adjacent pairs of blocks, and detected domain walls in the quantum state are annihilated through feedforward [blue-shaded gates in Fig. 5(a)]. Last the ancillae are reset (yellow-shaded gates) and unitarily absorbed into the GHZ state. The constant-depth feedforward extreme discussed above corresponds to  $(b, p) = (n/2, 1)$ . For  $b \geq 2$ , a  $(b, p)$ -circuit contains  $b(p + 2) - 2$  CX gates,  $b - 1$  mid-circuit measurements, and is of depth  $\lfloor p/2 \rfloor + 4$  including the layer of mid-circuit measurements on the ancillae.

First, we report results at  $n = 4, 6$  in Fig. 5(b), using  $(b, p) = (2, 1), (3, 1)$  circuit structures respectively. We compare the observed infidelity of the prepared state

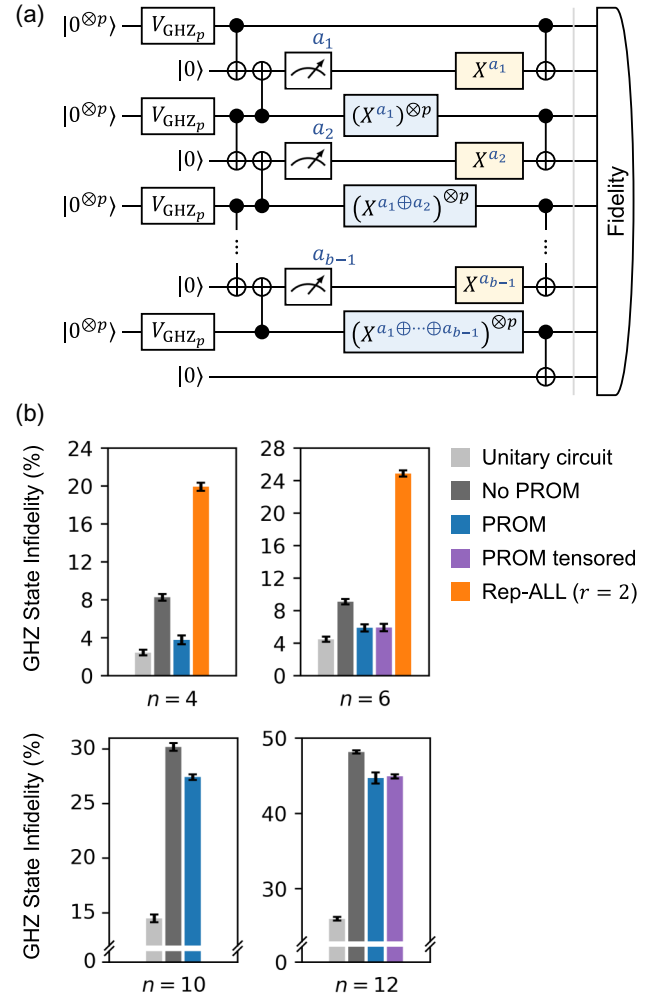


FIG. 5. GHZ state preparation using feedforward. (a) Dynamic circuit to prepare an  $n$ -qubit GHZ state. The  $n$  qubits are split into  $b$  blocks each comprising  $p$  qubits and an ancilla, the former prepared in a GHZ state through the standard unitary  $V_{\text{GHZ}_p}$  comprising a fan-out of CX gates—see Fig. 9(b). Thereafter the small GHZ states are merged in constant depth through mid-circuit measurements and feedforward operations (shaded blue), and finally the ancillae are reset (shaded yellow) and absorbed into the state. (b) Infidelities of the prepared state against the ideal  $n$ -qubit GHZ state, comparing probabilistic readout error mitigation, probabilistic readout error mitigation assuming tensored readout channels, Rep-ALL at  $r = 2$ , and without readout error mitigation for mid-circuit measurements and feedforward. Also shown are infidelities on a unitary circuit ( $V_{\text{GHZ}_n}$ ). Readout error mitigation for terminal measurements was applied for all experiments. The  $n = 4, 6, 10, 12$  experiments employed  $(b, p) = (2, 1), (3, 1), (2, 4), (3, 3)$  circuit structures respectively, executed on the superconducting quantum device *ibm\_osaka*. Error bars are standard deviations across 100 trials each with 10 000 shots.

against the ideal GHZ state without PROM, with general PROM and PROM assuming tensored readout error channels, and with Rep-ALL at  $r = 2$ . We include also

for comparison the unitary GHZ state preparation circuit [drawn in Fig. 9(b)] run on the same  $n$ -qubit chain. We used a stabilizer method to measure the fidelity of the prepared state against the GHZ state—see Appendix C 5 for details. The application of PROM is evidently effective, reducing state infidelities to  $\sim 45\%$ – $65\%$  of their unmitigated values. Residual infidelities are attributed to quantum errors and decoherence that are not targeted by PROM. Just as in the qubit reset experiments (Sec. V A), the Rep-ALL mitigation scheme is not competitive.

The qualitative trend that unitary state preparation outperforms unmitigated dynamic circuits is consistent with a prior study [17]; as discussed there and in other dynamic circuit studies [73], this is generally attributable to decoherence, crosstalk, and imperfect quantum-non-demolition (QND) properties of measurements on hardware. In our experiments, with PROM, the infidelities on dynamic circuits become fairly competitive with but do not surpass unitary circuits.

We next examined larger  $n = 10, 12$  settings in Fig. 5(c), therein utilizing an interpolation between unitary and feed-forward circuit extremes by choosing  $(b, p) = (2, 4), (3, 3)$  circuit structures respectively. Here PROM provided reductions in state infidelities comparable to the  $n = 4, 6$  experiments, but residual infidelities were expectedly higher. As it has become clear that the repetition-based REM schemes are uncompetitive, we have omitted them at  $n = 10, 12$ . On circuits with  $b > 2$ , in particular  $n = 6, 12$ , general PROM yielded relative state infidelities  $\sim 1$ – $2\%$  lower than PROM assuming independent readout errors—similar to the qubit reset experiments in Sec. V A, this small difference is consistent with our protocol sensitivity bounds as the characterized readout error channels were close to being independent.

We report additional data on a second superconducting quantum device in Appendix D 2 at  $n = 4, 6$ , which are qualitatively similar. There, we also report infidelities measured with Rep-ALL and Rep-MAJ at  $r = 3$ , which were entirely uncompetitive and omitted here.

### C. Staged quantum state teleportation

Last we investigate repeated quantum state teleportation [see Fig. 6(a)]. In standard state teleportation [74–76], a Bell pair is shared between a second and third qubit; the first qubit hosts the quantum state  $|\psi\rangle$  to be transported. A Bell-basis measurement is performed on the first and second qubits and the outcomes feedforward into operations on the third qubit, which ends in state  $|\psi\rangle$ . As this procedure requires CX gates between adjacent pairs of qubits, the 3 qubits must be connected to each other—thus  $|\psi\rangle$  is transported a distance of 2 qubits. Generically, repeating the teleportation  $k$  times transports  $|\psi\rangle$  a distance of  $2k$  qubits [see Fig. 6(a)]. Scenarios that require the transport of a qubit state arise commonly

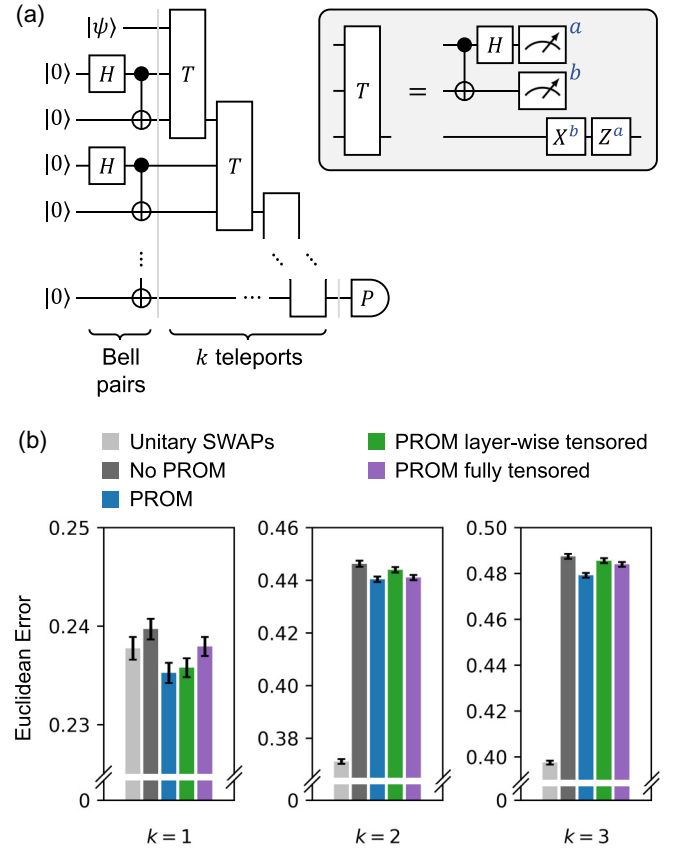


FIG. 6. Repeated quantum state teleportation. (a) Circuit schematic. An input quantum state  $|\psi\rangle$  is transported to a qubit a distance  $2k$  away by  $k$  teleportations, consuming  $k$  shared Bell pairs. On the destination qubit, Pauli expectation values ( $P$ ) are measured and the deviations  $\delta P$  against ideal expectation values of the input state  $|\psi\rangle$  assessed, for  $P \in \{X, Y, Z\}$ . (b) Euclidean errors  $\Delta = (\delta X^2 + \delta Y^2 + \delta Z^2)^{1/2}$  after  $k \in \{1, 2, 3\}$  teleportations, comparing probabilistic readout error mitigation, probabilistic readout error mitigation assuming layerwise and fully tensored readout channels, and without readout error mitigation for mid-circuit measurements and feedforward. Also shown are errors on a unitary circuit with SWAP gates that transport  $|\psi\rangle$  across the same qubits. Readout error mitigation for terminal measurements was applied for all cases. Experiments were performed on the superconducting quantum device *ibm\_osaka*. Error bars are standard deviations across 100 trials each with 1 000 000 shots.

in quantum computation and simulation, necessitated by device connectivity or classical control constraints [77,78]. The straightforward alternative to teleportation is to connect starting and ending qubits with a path of SWAP gates, thus unitarily transporting the qubit state [see Fig. 9(c)].

In our experiment, we prepare a generic single-qubit state  $|\psi\rangle = e^{-i\phi_z Z} e^{-i\phi_x X} |0\rangle$  with arbitrarily chosen  $(\phi_x, \phi_z) = (\pi/8, 3\pi/8)$ , and teleport the state  $k$  times, as drawn in Fig. 6(a). We terminally measure Pauli expectation values  $\langle X \rangle, \langle Y \rangle, \langle Z \rangle$  on the ending qubit and assess their deviations  $\delta X, \delta Y, \delta Z$  with respect to ideal values

on the  $|\psi\rangle$  starting state. To eliminate dependence on the measurement basis, we examine the Euclidean error  $\Delta = (\delta X^2 + \delta Y^2 + \delta Z^2)^{1/2}$ . Thus  $\Delta$  provides an averaged measure of the degradation of the state  $|\psi\rangle$  as it undergoes transport. We employ the multilayer formulation of PROM, with each teleportation stage placed in a layer (totaling  $k$  layers).

In Fig. 6(b) we report results for  $k \in \{1, 2, 3\}$  stages of teleportation, comparing the measured  $\Delta$  without PROM, with general PROM, and with PROM assuming layerwise and fully tensored readout error channels. We include also results on a unitary SWAP circuit transporting  $|\psi\rangle$  across the same qubits. Across all  $k$ , PROM results in lower  $\Delta$  than without mitigation. In particular at  $k = 1$  this improvement is sufficient for teleportation to be comparable in error to unitary SWAPS, but at higher  $k$  the unitary circuit is better. Similar to prior experiments (Secs. V A and V B), general PROM provides a relative advantage of  $\sim 1\%$ – $2\%$  over PROM with independence assumptions, which is consistent with our sensitivity bounds.

We report additional data on a second quantum device in Appendix D 3, which are qualitatively similar. There we also show results with Rep-ALL at  $r = 2$ , which were uncompetitive and omitted here.

## VI. DISCUSSION

In this work we have described a general REM method called PROM targeting mid-circuit measurements and feed-forward in dynamic circuits, for which existing standard techniques for terminal measurements cannot address. In addition to laying the theoretical groundwork, we demonstrated the effectiveness of our method in a series of experiments on superconducting quantum processors. These experiments spanned multiple use cases of practical interest with several performance measures—in each case the observed error metric significantly decreased with our REM method. In the following, we discuss several facets of PROM in a broader context, relating to the resource costs of the protocol, modifications and generalizations, and connections to other literature.

**Resource costs.** The sample complexity of error mitigation is typically a primary concern in practice as it pertains to wall-time usage of the quantum device. Generically, the application of PROM requires a number of shots scaling exponentially with the number of mid-circuit measurements (see Table I). In particular, for a characteristic probability of readout error  $r$  in each measurement, mitigating a circuit with  $m$  total mid-circuit measurements carries a sample complexity overhead of  $\xi^2 = (1 - 2r)^{-2m}$ . This exponential overhead in the number of operations mitigated and error rate is generally inevitable and can be understood as a consequence of universal lower bounds [79–82] on the cost of quantum error mitigation (applied here to readout error channels). While readout error rates

are typically larger than those of single- and two-qubit unitary gates on quantum devices, the number of mid-circuit measurements is small compared to the number of gates in most circuits of practical interest [74,83,84]. Thus, we typically expect the sample cost of PROM to be comparable, or subdominant, with that of other quantum error mitigation methods (e.g., PEC, zero-noise extrapolation (ZNE)) that one may use in addition to PROM in an experiment.

A separate consideration is the classical processing costs of the error mitigation protocol. Accommodating arbitrarily correlated readout errors demands classical resources scaling exponentially in the total number of measurements  $m$  to construct the mitigated estimator (see Table I). This operational mode of PROM is feasible for  $m \lesssim 25$  on personal computers and  $m \lesssim 35$  on a cluster. Assuming independence of readout errors, the classical resources required reduces to scale only polynomially, in fact linearly, with  $m$ . On quantum devices exhibiting small error correlations, this provides a cheap, high-quality mitigated estimator (see Result 3 informing robustness). The sensitivity and overhead bounds we provide could aid in identifying the most suitable variant of the mitigation protocol to utilize in an experiment given resource budgets.

**Modifications to PROM.** First, we comment that while our formulation of PROM uses BFA (recall Figs. 2 and 3) to symmetrize readout error channels, symmetrization is not strictly necessary and a modified algorithm can function on nonsymmetrized readout error channels, albeit with increased classical processing costs. Broadly, symmetrization affords three advantages: (1) reduction of the characterization of readout error channels from a matrix problem ( $M$  in Sec. II) to a vector problem ( $\mathbf{q}$  in Sec. III and Fig. 3), thereby simplifying the calibration process prior to PROM and the theory underlying PROM; (2) reduction of the quasi-probability distribution  $\alpha$  from which bitmasks are sampled to be independent of the measurement outcomes in the circuit; and (3) the ability to exploit the fast Walsh-Hadamard transform (FWHT), in place of plain matrix inversion, in calculating  $\alpha$  to reduce processing costs. All of these advantages are lost without symmetrization.

Second, we reiterate that PROM is applicable also to readout errors on terminal measurements. This is because terminal measurements are a special case of mid-circuit ones: namely, ones not feeding into feedforward and whose resultant (postcollapse) quantum states are irrelevant. This equivalence can be formalized, and PROM thereby shown to be a general REM framework encompassing both cases (see Appendixes B 5 a and B 5 b). The sampling overhead and classical processing overhead of PROM and conventional REM methods on terminal measurements are identical, so long as the conventional methods also utilize BFA and exploit the resulting symmetrization of readout error channels [85]. For convenience, we provide explicitly a version of the PROM protocol integrating readout error mitigation for terminal measurements in Appendix B 5 c.

Finally, we anticipate that PROM can be extended to accommodate multi-qubit positive operator-valued measures (POVMs) of rank above one, such as native ZZ parity measurements possible on certain platforms. Given an appropriate characterization of the readout error channels and a twirling group that symmetrizes the readout errors, PROM can be applied with natural generalizations to its components [86].

**Relations to other work.** Our protocol is conceptually related to other error mitigation and virtual distillation schemes. First, PROM is similar in structure to PEC [30,31]: both take a premise of noise channels characterized in some manner, utilize twirling to reshape error channels to a convenient form, and sample from a quasi-probability distribution plus simple classical postprocessing to yield an unbiased estimator. The exponential overhead factor  $\xi^2$  in PROM is analogous to  $\gamma^2$  in the formalism of PEC. The difference is that PEC targets quantum error channels acting on unitary gates, whereas PROM treats readout error channels on measurements, which are nonunitary channels. In principle, a quasi-probability sampling-based algorithm that unifies both PEC and PROM can be formalized, with error channels acting on isometries  $\mathcal{H} \otimes \mathbb{F}_2$  where  $\mathcal{H}$  is the Hilbert space of the qubits and  $\mathbb{F}_2$  the space of classical measurement outcomes. We remark also that the quasi-probabilistic sampling to correct the effects of broken measurement-feedforward correlations in PROM is reminiscent of the classical sampling to restore coherence in channel and resource distillation protocols [87–90].

PROM can be hybridized with other twirling and sampling schemes, or used alongside other error mitigation methods, to address different errors besides readout. Randomized compiling of measurements [91], which twirls measurement channels under more general groups, could be used in place of BFA to symmetrize readout errors. No further modification to the protocol is necessary when twirling with larger groups (e.g., Pauli twirling [40]): the same processing of  $\mathbf{q}$  and sampling of feedforward bit-masks apply. An advantage of generalized twirling is that coherent quantum errors on the measured qubits, in addition to readout errors, could be symmetrized and reduced to stochastic errors, which may benefit error resilience or further layers of error mitigation downstream. While coherent errors appear small for terminal measurements in superconducting devices [92,93], they may be relevant in mid-circuit contexts where the postmeasurement quantum state must be preserved. Separately, in place of exact error channel inversion, alternative approximate methods may be considered [40] and techniques to robustly construct POVMs [41] could be useful at the algorithmic level.

While PROM does not target and is currently ineffective against non-QND errors, combining it with such broader techniques can aid in closing this gap. However, we expect that certain types of errors are beyond the ability of error

mitigation protocols in general to completely treat. In particular, leakage errors that bring qubits outside of their computational subspace will likely require hardware and control interventions, such as leakage suppression and resets, in tandem with error mitigation to be effectively addressed.

**Outlook.** In closing the lacuna in error mitigation strategies for circuits with mid-circuit measurements and feed-forward, our study paves the way for effectively utilizing dynamic circuit capabilities on both near-term and future hardware. This supports pivotal quantum resource advantages across a broad range of domains and is imperative for realizing useful quantum computation on near-future quantum hardware.

## ACKNOWLEDGMENTS

The authors express gratitude to Lorcán O. Conlon, Bujiao Wu, Jun Ye, Yunlong Xiao, Syed M. Assad, Mile Gu, Tianqi Chen, and Ching Hua Lee for helpful discussions. The authors acknowledge the use of IBM Quantum services for this work. The views expressed are those of the authors and do not reflect the official policy or position of IBM or the IBM Quantum team. Access to IBM quantum devices was enabled by the Quantum Engineering Programme (QEP) 2.0. This research is supported by the National Research Foundation, Singapore and the Agency for Science, Technology and Research (A\*STAR) under its Quantum Engineering Programme (NRF2021-QEP2-02-P03, NRF2021-QEP2-02-P06), its Centre for Quantum Technologies Funding Initiative (S24Q2d0009), A\*STAR C230917003, A\*STAR under the Central Research Fund (CRF) Award for Use-Inspired Basic Research (UIBR), and the Quantum Innovation Centre (Q.InC) Strategic Research and Translational Thrust (SRTT). J.T. is supported by FQxI foundation under Grant No. FQXi-RFP-IPW-1903 (“Are quantum agents more energetically efficient at making predictions?”).

## DATA AVAILABILITY

The data that support the findings of this article are not publicly available upon publication because it is not technically feasible and/or the cost of preparing, depositing, and hosting the data would be prohibitive within the terms of this research project. The data are available from the authors upon reasonable request.

## APPENDIX A: PRELIMINARIES

### 1. Notation and basic facts

Throughout our work some symbols such as  $s, t, f$  may represent a bitstring or bitvector when so defined. We do not distinguish between bitstrings and bitvectors. We use the following notation:

- (1) The length of a bitstring  $s$ , that is, the number of bits in  $s$ , is denoted  $|s|$ .
- (2) The Hamming weight of a bitstring  $s$ , that is, the number of nonzero bits in  $s$ , is denoted  $wt(s)$ .
- (3) The  $j$ th-bit of a bitstring  $s$  is denoted  $s_j \in \{0, 1\}$ .
- (4) The bit-wise sum, equivalently difference, of bitstrings  $s$  and  $t$  of equal length is  $s \oplus t$ , which is a bitstring.
- (5) The dot product of bitstrings  $s$  and  $t$  of equal length is  $s \cdot t = \sum_{j=1}^{|s|} s_j t_j \pmod{2}$ , which is a bit.
- (6) Where unambiguous by context, the concatenation of two bitstrings  $s^{(1)}$  and  $s^{(2)}$  is written simply  $s^{(1)}s^{(2)}$ .
- (7) The all-zero bitstring is denoted  $\mathbf{0}$ , which has all entries zero. The all-one counterpart is denoted  $\mathbf{1}$ .
- (8) Bitstrings can be used as indices. The basis for the vector, matrix or tensor is simply regarded as being enumerated (labelled) by the bitstrings. Alternatively one can implicitly convert the bitstrings to their decimal values for conventional integer enumeration.

We write  $[n] = \{1, 2, \dots, n\}$  for  $n \in \mathbb{N}$ . We use the following notation for vectors, matrices and tensors:

- (1) Vectors are written with boldface, for example  $\mathbf{v}$ . Matrices and tensors are not bolded, for example  $A$ .
- (2) Entries are indexed with subscripts. For example  $v_j$  is the  $j$ th entry of a vector  $\mathbf{v}$ , and  $A_{ij}$  is the  $(i, j)$ th entry of a matrix  $A$ .
- (3)  $\|\mathbf{v}\|_p$  is the  $p$ -norm of a vector  $\mathbf{v}$ .
- (4)  $\|A\|_p$  for a matrix  $A$  denotes the matrix norm induced by the vector  $p$ -norm. For example  $\|A\|_2$  is the spectral norm of  $A$  and is the largest singular value of  $A$ .
- (5)  $|\mathbf{v}|$  and  $|A|$  are the entrywise absolute values of a vector  $\mathbf{v}$  and a matrix  $A$ , themselves a vector and a matrix respectively, not to be confused with the norms of  $\mathbf{v}$  and  $A$ .
- (6)  $\mathbf{v} > \mathbf{v}'$  means that  $\mathbf{v}$  is entrywise larger than  $\mathbf{v}'$ , that is,  $v_j > v'_j$  for every index  $j$ . Likewise for  $\geq, \leq, <$ .
- (7) The binary convolution between two vectors  $\mathbf{u}$  and  $\mathbf{v}$  of equal length is a vector with entries,

$$(\mathbf{u} * \mathbf{v})_s = \sum_{t \in \mathcal{S}} u_t v_{s \oplus t}, \quad (\text{A1})$$

for each  $s \in \mathcal{S}$ , where  $\mathcal{S}$  is the set of bitstrings indexing  $\mathbf{u}$  and  $\mathbf{v}$ .

We use conventional notation for Pauli matrices and the Hadamard and  $S$  gates,

$$\begin{aligned} X &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, & Y &= \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, & Z &= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \\ H &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, & S &= \begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}, \end{aligned} \quad (\text{A2})$$

and the CX gate with the first qubit as control and the second qubit as target is written  $|0\rangle\langle 0| \otimes \mathbb{I} + |1\rangle\langle 1| \otimes X$ . We denote the identity as  $\mathbb{I}$ , with dimensions inferred from context. The  $d$ -fold tensor product of an object  $A$ , for instance quantum state, vector, or matrix, is denoted  $A^{\otimes d}$ . More generally we write

$$A^{(s)} = \bigotimes_{j=1}^{|s|} A^{s_j} \quad (\text{A3})$$

for a bitstring  $s$ , which is a  $d$ -fold tensor product over  $\{A, \mathbb{I}\}$  as determined by the bitstring  $s$ . The diagonal entries of tensor products of  $Z$  Pauli matrices (which are diagonal) satisfy

$$[Z^{(s)}]_{kk} = (-1)^{k \cdot s}, \quad (\text{A4})$$

where  $k$  is a bitstring index. The entries of tensor products of  $H$  satisfy

$$\sqrt{2^m} [H^{\otimes m}]_{fk} = [Z^{(f)}]_{kk} = (-1)^{f \cdot k}, \quad (\text{A5})$$

where  $f$  and  $k$  are bitstring indices.

## 2. Properties of confusion matrices $M$ , $Q$ and vector $\mathbf{q}$

As written in Sec. II, in a setting with  $m$  measurements, each entry  $M_{ts} \in [0, 1]$  records the probability of the quantum device reporting measurement outcome  $t \in \mathcal{S} = \{0, 1\}^m$  when the true outcome is  $s \in \mathcal{S}$ . Correspondingly  $M$  is stochastic, i.e.,

$$\sum_{t \in \mathcal{S}} M_{ts} = 1 \quad \forall s \in \mathcal{S}, \quad (\text{A6})$$

and consequently,

$$\|M\|_1 = \max_{s \in \mathcal{S}} \underbrace{\sum_{t \in \mathcal{S}} |M_{ts}|}_1 = 1. \quad (\text{A7})$$

Each entry  $q_s \in [0, 1]$  of the symmetrized confusion vector  $\mathbf{q}$  records the probability of a readout error syndrome  $s \in \mathcal{S}$  occurring (with BFA applied)—that is, the probability that any true measurement outcome  $a \in \mathcal{S}$  is reported as  $a \oplus s$ . This probability is uniform across  $a \in$

$\mathcal{S}$  due to the symmetrization of the readout channels. Correspondingly  $\mathbf{q}$  is a probability distribution,

$$\sum_{s \in \mathcal{S}} q_s = 1. \quad (\text{A8})$$

The symmetrized confusion matrix  $Q$ , with entries  $Q_{sf} = q_{s \oplus f} \in [0, 1]$  for  $s \in \mathcal{S}$  and  $f \in \mathcal{S}$ , is doubly stochastic, i.e.,

$$\begin{aligned} \sum_{s \in \mathcal{S}} Q_{sf} &= \sum_{s \in \mathcal{S}} q_{s \oplus f} = \sum_{s' \in \mathcal{S}} q_{s'} = 1 \quad \forall f \in \mathcal{S}, \\ \sum_{f \in \mathcal{S}} Q_{sf} &= \sum_{f \in \mathcal{S}} q_{s \oplus f} = \sum_{s' \in \mathcal{S}} q_{s'} = 1 \quad \forall s \in \mathcal{S}, \end{aligned} \quad (\text{A9})$$

and consequently,

$$\|Q\|_1 = \max_{f \in \mathcal{S}} \underbrace{\sum_{s \in \mathcal{S}} |Q_{sf}|}_1 = 1, \quad \|Q\|_\infty = \max_{s \in \mathcal{S}} \underbrace{\sum_{f \in \mathcal{S}} |Q_{sf}|}_1 = 1. \quad (\text{A10})$$

For a more concrete visualization, below we give explicit forms of  $Q$  for  $m = 1$ ,  $m = 2$  and  $m = 3$  cases,

$$\begin{aligned} Q &= \begin{bmatrix} q_0 & q_1 \\ q_1 & q_0 \end{bmatrix}, \quad Q = \begin{bmatrix} q_{00} & q_{01} & q_{10} & q_{11} \\ q_{01} & q_{00} & q_{11} & q_{10} \\ q_{10} & q_{11} & q_{00} & q_{01} \\ q_{11} & q_{10} & q_{01} & q_{00} \end{bmatrix}, \\ Q &= \begin{bmatrix} q_{000} & q_{001} & q_{010} & q_{011} & q_{100} & q_{101} & q_{110} & q_{111} \\ q_{001} & q_{000} & q_{011} & q_{010} & q_{101} & q_{100} & q_{111} & q_{110} \\ q_{010} & q_{011} & q_{000} & q_{001} & q_{110} & q_{111} & q_{100} & q_{101} \\ q_{011} & q_{010} & q_{001} & q_{000} & q_{111} & q_{110} & q_{101} & q_{100} \\ q_{100} & q_{101} & q_{110} & q_{111} & q_{000} & q_{001} & q_{010} & q_{011} \\ q_{101} & q_{100} & q_{111} & q_{110} & q_{001} & q_{000} & q_{011} & q_{010} \\ q_{110} & q_{111} & q_{100} & q_{101} & q_{010} & q_{011} & q_{000} & q_{001} \\ q_{111} & q_{110} & q_{101} & q_{100} & q_{011} & q_{010} & q_{001} & q_{000} \end{bmatrix}. \end{aligned} \quad (\text{A11})$$

In an ideal setting without readout errors,  $M_{ts} = \delta_{ts}$  or equivalently  $M = \mathbb{I}$  is the identity matrix. In the symmetrized picture, in the ideal setting the probability distribution  $\mathbf{q}$  has support only on the trivial syndrome,  $q_s = \delta_{s,0}$ , and correspondingly  $Q = \mathbb{I}$  is the identity matrix.

In general, the following properties about the diagonal entries and trace of  $Q$  hold:

$$Q_{ss} = q_{s \oplus s} = q_0 \quad \forall s \in \mathcal{S}, \quad \text{tr} Q = 2^m q_0. \quad (\text{A12})$$

The determinant of  $Q$  can also be analytically evaluated—see Eq. (A20). This is most compactly done via the diagonalization of  $Q$ , which we describe in Appendix A 3.

Last, given an  $M$ , one can in fact compute the corresponding  $\mathbf{q}$  and  $Q$  that characterizes the symmetrized

readout error channels, that would have been measured had BFA been used during calibration, and which would be needed for readout error mitigation in subsequent experiment circuits executed with BFA. In particular,

$$\begin{aligned} q_s &= \frac{1}{2^m} \sum_{\substack{t, t' \in \mathcal{S} \\ t \oplus t' = s}} M_{tt'} = \frac{1}{2^m} \sum_{t \in \mathcal{S}} M_{t(t \oplus s)}, \\ Q_{sf} &= q_{s \oplus f} = \frac{1}{2^m} \sum_{t \in \mathcal{S}} M_{t(t \oplus s \oplus f)}, \end{aligned} \quad (\text{A13})$$

that is,  $q_s$  is the average of all entries  $M_{tt'}$  consistent with the syndrome  $t \oplus t' = s$ , and the construction of  $Q$  follows from  $\mathbf{q}$  by definition. The explicit meaning that  $\mathbf{q}$  and  $Q$  describe the readout error channels under symmetrization is thus made clear. We clarify, however, that in practice  $\mathbf{q}$  and  $Q$  can be directly obtained by including BFA in the calibration circuits, and there is no need to go through  $M$ . See Appendix C 2 for details on the calibration process as implemented in our experiments. This direct calibration with BFA provides resource and precision advantages [25,26].

### 3. Diagonalization of symmetrized confusion matrix $Q$ and the inverse of $Q$

As noted in Sec. III B of the main text,  $Q$  is hugely symmetric. Its entries  $Q_{sf}$  are invariant under the index shift  $(s, f) \rightarrow (s \oplus s', f \oplus s')$  for any  $s' \in \mathcal{S}$ . In particular,  $Q$  is symmetric about both its diagonal and antidiagonal, and so is every of its recursively halved quadrants—see, for example, Eq. (A11) for visualization. Recall that  $m$ -fold tensor products of the Pauli matrices form a basis for  $2^m \times 2^m$  complex matrices, and of the Pauli matrices, only  $\mathbb{I}$  and  $X$  are symmetric about both their diagonal and antidiagonal. Thus it follows that  $Q$  must be of the form

$$Q = \sum_{s \in \mathcal{S}} A_s X^{(s)} \quad (\text{A14})$$

for coefficients  $A_s \in \mathbb{R}$ , in order for it to possess the symmetries exhibited. To fix the coefficients, one examines the first column of  $Q$ , which must match  $\mathbf{q}$  by definition. Note that the first column of  $X^{(s)}$  has a unit entry at row  $s$  and is zero everywhere else,  $[X^{(s)}]_{r0} = \delta_{rs}$ . Thus

$$Q_{r0} = \sum_{s \in \mathcal{S}} A_s [X^{(s)}]_{r0} = \sum_{s \in \mathcal{S}} A_s \delta_{rs} = A_r = q_r. \quad (\text{A15})$$

That is,

$$Q = \sum_{s \in \mathcal{S}} q_s X^{(s)}. \quad (\text{A16})$$

Now note that  $X^{(s)}$  is diagonalized by  $H^{\otimes m}$ , that is,  $H^{\otimes m} X^{(s)} H^{\otimes m} = Z^{(s)}$ . Then  $Q$ , which is a linear combination of  $X^{(s)}$ , is also diagonalized by  $H^{\otimes m}$ ,

$$H^{\otimes m} Q H^{\otimes m} = \sum_{s \in \mathcal{S}} q_s Z^{(s)}. \quad (\text{A17})$$

The eigenvalues of  $Q$  are thus given by

$$\lambda_k = \sum_{s \in \mathcal{S}} q_s [Z^{(s)}]_{kk} = \sum_{s \in \mathcal{S}} (-1)^{k \cdot s} q_s, \quad (\text{A18})$$

where we have invoked the identity in Eq. (A4). Recalling the normalization  $\sum_{s \in \mathcal{S}} q_s = 1$ , one can obtain an alternate form

$$\begin{aligned} \lambda_k &= q_0 + \sum_{\substack{s \in \mathcal{S} \\ s \neq 0}} (-1)^{k \cdot s} q_s = \left( 1 - \sum_{\substack{s \in \mathcal{S} \\ s \neq 0}} q_s \right) + \sum_{\substack{s \in \mathcal{S} \\ s \neq 0}} (-1)^{k \cdot s} q_s \\ &= 1 - \sum_{\substack{s \in \mathcal{S} \\ s \neq 0}} [1 - (-1)^{k \cdot s}] q_s \\ &= 1 - 2 \sum_{\substack{s \in \mathcal{S} \\ k \cdot s \neq 0}} q_s, \end{aligned} \quad (\text{A19})$$

which makes the special case  $\lambda_0 = 1$  clear. The solutions for the eigenvalues in  $Q$  in Eqs. (A18) and (A19) above are reported in Eq. (10) of the main text. In general, as  $Q$  is doubly stochastic, as in Eq. (A9), we are guaranteed that  $|\lambda_k| \leq 1$  for all  $k \in \mathcal{S}$ . In the absence of readout errors ( $q_s = \delta_{s0}$ ), we straightforwardly have  $\lambda = \mathbf{1}$ , that is,  $\lambda_k = 1$  uniformly for all  $k \in \mathcal{S}$ . Last we comment that the trace and determinant of  $Q$  yield the properties

$$\begin{aligned} \text{tr} Q &= \sum_{k \in \mathcal{S}} \lambda_k = 2^m q_0, \\ \det Q &= \prod_{k \in \mathcal{S}} \lambda_k = \prod_{k \in \mathcal{S}} \left[ \sum_{s \in \mathcal{S}} (-1)^{k \cdot s} q_s \right] \\ &= \prod_{k \in \mathcal{S}} \left[ 1 - 2 \sum_{\substack{s \in \mathcal{S} \\ k \cdot s \neq 0}} q_s \right], \end{aligned} \quad (\text{A20})$$

where in the first line we have used our prior result for the trace of  $Q$  in Eq. (A12). As  $Q$  is stochastic we have that  $|\det Q| \leq 1$ , and Eq. (A20) implies  $|\det Q| = 1$  only when no readout errors are present.

In anticipation of developments in Appendix B, we examine  $Q^{-1}$ . Trivially, we remind that in an ideal setting without readout errors,  $Q = \mathbb{I}$  and thus  $Q^{-1} = \mathbb{I}$ . Now, the diagonalization of  $Q$  from Eq. (A17) gives

$$Q^{-1} = H^{\otimes m} \begin{bmatrix} \frac{1}{\lambda_1} & & & \\ & \frac{1}{\lambda_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_m} \end{bmatrix} H^{\otimes m}, \quad (\text{A21})$$

and thus entries of  $Q^{-1}$  can be written as

$$\begin{aligned} [Q^{-1}]_{sf} &= \sum_{k \in \mathcal{S}} [H^{\otimes m}]_{sk} \left( \frac{1}{\lambda_k} \right) [H^{\otimes m}]_{kf} \\ &= \frac{1}{2^m} \sum_{k \in \mathcal{S}} (-1)^{s \cdot k} (-1)^{k \cdot f} \left( \frac{1}{\lambda_k} \right), \end{aligned} \quad (\text{A22})$$

where we have used the identities for entries of tensor products of  $H$  matrices in Eq. (A5). For future use, we note that the first column of  $Q^{-1}$  is given by

$$[Q^{-1}]_{s0} = \frac{1}{2^m} \sum_{k \in \mathcal{S}} (-1)^{s \cdot k} \left( \frac{1}{\lambda_k} \right). \quad (\text{A23})$$

Moreover, that  $Q$  is hugely symmetric implies the same for  $Q^{-1}$ —namely, that the entries  $[Q^{-1}]_{sf}$  are invariant under the index shift  $(s, f) \rightarrow (s \oplus s', f \oplus s')$  for any  $s' \in \mathcal{S}$ , indeed also explicitly observable in Eq. (A22). In particular, we observe  $[Q^{-1}]_{sf} = [Q^{-1}]_{(s \oplus f)0} = [Q^{-1}]_{0(s \oplus f)}$ , and thus each column and row of  $Q^{-1}$  contains the same entries but arranged in different orders. This means

$$\begin{aligned} \|Q^{-1}\|_1 &= \max_{f \in \mathcal{S}} \sum_{s \in \mathcal{S}} |[Q^{-1}]_{sf}| = \sum_{s \in \mathcal{S}} |[Q^{-1}]_{s0}|, \\ \|Q^{-1}\|_\infty &= \max_{s \in \mathcal{S}} \sum_{f \in \mathcal{S}} |[Q^{-1}]_{sf}| = \sum_{f \in \mathcal{S}} |[Q^{-1}]_{0f}|, \end{aligned} \quad (\text{A24})$$

and indeed  $\|Q^{-1}\|_1 = \|Q^{-1}\|_\infty$ . Furthermore, from  $QQ^{-1} = Q^{-1}Q = \mathbb{I}$ , we observe that row and column sums of  $Q^{-1}$  evaluate uniformly to unity,

$$\begin{aligned}
\sum_{k \in \mathcal{S}} Q_{sk} [Q^{-1}]_{kf} = \delta_{sf} &\implies \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{S}} Q_{sk} [Q^{-1}]_{kf} = \sum_{k \in \mathcal{S}} \underbrace{\left( \sum_{s \in \mathcal{S}} Q_{sk} \right)}_1 [Q^{-1}]_{kf} = \sum_{k \in \mathcal{S}} [Q^{-1}]_{kf} = \sum_{s \in \mathcal{S}} \delta_{sf} = 1 \quad \forall f \in \mathcal{S}, \\
\sum_{k \in \mathcal{S}} [Q^{-1}]_{sk} Q_{kf} = \delta_{sf} &\implies \sum_{f \in \mathcal{S}} \sum_{k \in \mathcal{S}} [Q^{-1}]_{sk} Q_{kf} = \sum_{k \in \mathcal{S}} [Q^{-1}]_{sk} \underbrace{\left( \sum_{f \in \mathcal{S}} Q_{kf} \right)}_1 = \sum_{k \in \mathcal{S}} [Q^{-1}]_{sk} = \sum_{f \in \mathcal{S}} \delta_{sf} = 1 \quad \forall s \in \mathcal{S},
\end{aligned} \tag{A25}$$

where we have used the fact that  $Q$  is doubly stochastic, as in Eq. (A9). Despite this property, we clarify that  $Q^{-1}$  is not generally doubly stochastic as its entries are not guaranteed to be non-negative. In fact  $QQ^{-1} = Q^{-1}Q = \mathbb{I}$  implies that each row and column of  $Q^{-1}$  must contain at least one negative entry when readout errors are present (when  $Q \neq \mathbb{I}$ ).

#### 4. Tensorized readout channels

##### a. General setting

Consider a setting wherein we partition the  $m$  measurements into subsets of sizes  $\{m^{(1)}, m^{(2)}, \dots, m^{(p)}\}$ , where  $m^{(\ell)} \geq 1$  for each  $\ell$  and  $m^{(1)} + m^{(2)} + \dots + m^{(p)} = m$ , and we make the assumption that there are no correlations in readout errors between measurements in different subsets. That is, readout errors in different subsets are independent; but within a subset correlations in readout error can be present. One then has the decomposition

$$M = \bigotimes_{\ell=1}^p M^{(\ell)}, \tag{A26}$$

where  $M^{(\ell)}$  is a  $2^{m^{(\ell)}} \times 2^{m^{(\ell)}}$ -sized confusion matrix characterizing readout errors affecting the  $\ell$ th subset of measurements. This is equivalent to a statement of factorization of probabilities

$$M_{ts} = \prod_{\ell=1}^p M_{t^{(\ell)}s^{(\ell)}}^{(\ell)}, \tag{A27}$$

where  $s = s^{(1)}s^{(2)} \dots s^{(p)} \in \mathcal{S}$  and  $t = t^{(1)}t^{(2)} \dots t^{(p)} \in \mathcal{S}$  are the corresponding partitionings of the bitstrings over the  $p$  subsets,  $s^{(\ell)} \in \{0, 1\}^{m^{(\ell)}}$  and  $t^{(\ell)} \in \{0, 1\}^{m^{(\ell)}}$  for each  $\ell \in [p]$ . Likewise,

$$\begin{aligned}
\mathbf{q} &= \bigotimes_{\ell=1}^p \mathbf{q}^{(\ell)}, \quad q_s = \prod_{\ell=1}^p q_{s^{(\ell)}}^{(\ell)}, \\
Q &= \bigotimes_{\ell=1}^p Q^{(\ell)}, \quad Q_{sf} = \prod_{\ell=1}^p Q_{s^{(\ell)}f^{(\ell)}}^{(\ell)},
\end{aligned} \tag{A28}$$

with  $s = s^{(1)}s^{(2)} \dots s^{(p)} \in \mathcal{S}$  and  $f = f^{(1)}f^{(2)} \dots f^{(p)} \in \mathcal{S}$  being partitionings of the bitstrings, and where  $\mathbf{q}^{(\ell)}$  is a  $2^{m^{(\ell)}}$ -length probability vector and  $Q^{(\ell)}$  is the corresponding  $2^{m^{(\ell)}} \times 2^{m^{(\ell)}}$ -sized symmetrized confusion matrix characterizing readout errors affecting the  $\ell$ th subset of measurements. It follows from this tensor product structure that the eigenvalues of  $Q$  factorize,

$$\lambda = \bigotimes_{\ell=1}^p \lambda^{(\ell)}, \quad \lambda_k = \prod_{\ell=1}^p \lambda_{k^{(\ell)}}^{(\ell)}, \tag{A29}$$

where  $\lambda^{(\ell)}$  are the eigenvalues of  $Q^{(\ell)}$  and are given by the same solutions as in Eqs. (A18) and (A19), but with  $\mathbf{q}^{(\ell)}$  taking the place of  $\mathbf{q}$ . Likewise the inverse of  $Q$  factorizes

$$\begin{aligned}
Q^{-1} &= \bigotimes_{\ell=1}^p Q^{(\ell)-1}, \quad [Q^{-1}]_{sf} = \prod_{\ell=1}^p [Q^{(\ell)-1}]_{s^{(\ell)}f^{(\ell)}}, \\
[Q^{-1}]_{s0} &= \prod_{\ell=1}^p [Q^{(\ell)-1}]_{s^{(\ell)}0},
\end{aligned} \tag{A30}$$

where  $Q^{(\ell)-1}$  is the inverse of  $Q^{(\ell)}$ , whose general and first-column entries are given in Eqs. (A22) and (A23), but now with  $\lambda^{(\ell)}$  taking the place of  $\lambda$ .

##### b. Layerwise tensorized readout channels

In a feedforward circuit with  $L > 1$  layers (i.e., the setting of Sec. IV of the main text), the discussion above straightforwardly applies when an assumption of layerwise independence of readout errors is made. That is, readout errors on measurements in different layers of the circuit are taken to be independent, but we accommodate arbitrary correlations between errors in the same layer. Then in the notation of Appendix A 4a above, one takes  $p = L$ , and the subsets correspond to the measurements in each layer,  $\{m^{[1]}, m^{[2]}, \dots, m^{[L]}\}$ . The factorization of  $\mathbf{q}$ ,  $Q$ ,  $Q^{-1}$  and eigenvalues  $\lambda$  then follow. In this setting we use square brackets for superscripts, for example  $\mathbf{q}^{[l]}$  instead of  $\mathbf{q}^{(l)}$ , to emphasize the physical meaning that the indices enumerate the layers of the feedforward circuit, as is consistent with our writing in the main text.

### c. Fully tensored readout channels

The setup in Appendix A 4a likewise applies in a setting where every measurement is assumed to be independent—then  $p = m$  and the subsets  $\{m^{(1)}, m^{(2)}, \dots, m^{(p)}\}$  are single measurements. To be explicit, in the factorization of  $M$  in Eq. (A26), we have in this setting

$$M^{(\ell)} = \begin{bmatrix} 1 - M_{10}^{(\ell)} & M_{01}^{(\ell)} \\ M_{10}^{(\ell)} & 1 - M_{01}^{(\ell)} \end{bmatrix}, \quad (\text{A31})$$

which are  $2 \times 2$  confusion matrices each characterizing the readout error affecting the  $\ell$ th measurement, for  $\ell \in [m]$ . The diagonal elements record the probabilities of 0 and 1 outcomes being correctly reported, whereas off-diagonals record probabilities of bit-flip errors. With symmetrization, one has

$$\mathbf{q}^{(\ell)} = \begin{bmatrix} 1 - r_\ell \\ r_\ell \end{bmatrix}, \quad (\text{A32})$$

where  $r_\ell$  is the probability of a readout error occurring on the  $\ell$ th measurement. Concretely, the symmetrization means that  $r_\ell = (M_{01}^{(\ell)} + M_{10}^{(\ell)})/2$ , which is expressed more generally in Eq. (A13). Likewise, the symmetrized confusion matrices are

$$Q^{(\ell)} = \begin{bmatrix} q_0^{(\ell)} & q_1^{(\ell)} \\ q_1^{(\ell)} & q_0^{(\ell)} \end{bmatrix} = \begin{bmatrix} 1 - r_\ell & r_\ell \\ r_\ell & 1 - r_\ell \end{bmatrix}, \quad (\text{A33})$$

whose diagonal and off-diagonals elements record probabilities of the measurement outcome being correctly and incorrectly reported respectively. The eigenvalues of  $Q^{(\ell)}$  are easily found to be

$$\lambda^{(\ell)} = \begin{bmatrix} 1 \\ 1 - 2r_\ell \end{bmatrix}, \quad (\text{A34})$$

which coincides with our general solution for  $\lambda_k$  in Eqs. (A18) and (A19) or Eq. (10) of the main text, setting  $m = 1$  for a single independent qubit. The inverse of  $Q^{(\ell)}$  is also easily found,

$$Q^{(\ell)-1} = \frac{1}{1 - 2r_\ell} \begin{bmatrix} 1 - r_\ell & -r_\ell \\ -r_\ell & 1 - r_\ell \end{bmatrix}, \quad (\text{A35})$$

which is consistent with the general solution for entries of  $Q^{-1}$  in Eq. (A22) upon substitution of the present setting.

## APPENDIX B: FURTHER TECHNICAL DETAILS ON PROBABILISTIC READOUT ERROR MITIGATION

### 1. Theory for a single feedforward layer

#### a. General solution of $\alpha$ for arbitrarily correlated readout channels

Here we provide an elaboration of the derivation in Sec. III B of the main text, leading to the general solution

for  $\alpha$  in Eq. (11). First, in the absence of readout errors, the quantum state after the mid-circuit measurements and feedforward is straightforwardly

$$\bar{\sigma} = \sum_{s \in \mathcal{S}} V_s \Pi_s \rho \Pi_s V_s^\dagger, \quad (\text{B1})$$

where  $\Pi_s$  is the measurement projector associated with the outcome  $s \in \mathcal{S}$ . Equivalently, in the picture of quantum trajectories,

$$\bar{\sigma} = \sum_{s \in \mathcal{S}} p_s V_s \rho_s V_s^\dagger, \quad \rho_s = \frac{\Pi_s \rho \Pi_s}{p_s}, \quad p_s = \text{tr}(\Pi_s \rho), \quad (\text{B2})$$

which was presented in Eq. (2) of the main text. Technically, here  $\rho_s$  is well-defined so long as  $p_s \neq 0$ ; the  $p_s = 0$  trajectories are immaterial to  $\bar{\sigma}$  and can be disregarded. In the presence of readout errors characterized by  $\mathbf{q}$ , and imposing a bitmask  $f \in \mathcal{S}$  in feedforward, the quantum state after the mid-circuit measurements and feedforward is

$$\begin{aligned} \sigma^{(f)} &= \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} q_{s \oplus s'} V_{s' \oplus f} \Pi_s \rho \Pi_s V_{s' \oplus f}^\dagger \\ &= \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} q_{s \oplus s'} p_s V_{s' \oplus f} \rho_s V_{s' \oplus f}^\dagger, \end{aligned} \quad (\text{B3})$$

where the second equality was presented in Eq. (5) of the main text. The definition of the tensor  $T$  follows:

$$T_{bss'} = \text{tr} \left( O_b V_{s'} \Pi_s \rho \Pi_s V_{s'}^\dagger \right) = p_s \text{tr} \left( O_b V_{s'} \rho_s V_{s'}^\dagger \right), \quad (\text{B4})$$

as presented in Eq. (3) of the main text. Then from Eq. (4), one has

$$\langle \bar{O}_b \rangle = \text{tr} T_b = \sum_{s \in \mathcal{S}} T_{bss} = \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \delta_{ss'} T_{bss'}. \quad (\text{B5})$$

At the same time, from Eq. (6),

$$\langle O_b^{(f)} \rangle = \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} q_{s \oplus s'} T_{bs(s' \oplus f)} = \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} q_{s \oplus s' \oplus f} T_{bss'}. \quad (\text{B6})$$

Thus, by the definition of  $\hat{\mathbf{O}}$  in Eq. (7),

$$\langle \hat{O}_b \rangle = \sum_{f \in \mathcal{S}} \alpha_f \langle O_b^{(f)} \rangle = \sum_{f \in \mathcal{S}} \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \alpha_f q_{s \oplus s' \oplus f} T_{bss'}. \quad (\text{B7})$$

We desire  $\langle \hat{\mathbf{O}} \rangle = \langle \bar{\mathbf{O}} \rangle$  independent of the tensor  $T$ , which is unknown to us. This guarantees that the mitigation protocol works for arbitrary quantum circuits of the structure

considered, with no additional problem-specific information required. The equality necessitates that the coefficients of every  $T_{bs'}$  match on both sides, in particular, between Eqs. (B5) and (B7) for every  $b \in [B]$ , yielding

$$\sum_{f \in \mathcal{S}} \alpha_f q_{s \oplus s' \oplus f} = \delta_{ss'} \quad \forall s, s' \in \mathcal{S}. \quad (\text{B8})$$

Noting that  $\delta_{ss'} = \delta_{(s \oplus s')\mathbf{0}}$ , the above reduces to

$$\sum_{f \in \mathcal{S}} \alpha_f q_{s \oplus f} = \delta_{s\mathbf{0}} \quad \forall s \in \mathcal{S}. \quad (\text{B9})$$

This linear system can be written as

$$\mathcal{Q}\alpha = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \iff \alpha = \mathcal{Q}^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \quad (\text{B10})$$

That is,  $\alpha$  is the first column of  $\mathcal{Q}^{-1}$ . Then by Eq. (A23), we have

$$\alpha_f = [\mathcal{Q}^{-1}]_{f\mathbf{0}} = \frac{1}{2^m} \sum_{k \in \mathcal{S}} (-1)^{f \cdot k} \left( \frac{1}{\lambda_k} \right), \quad (\text{B11})$$

which is the solution reported in Eq. (11) of the main text. This exactly coincides with and can be equivalently written as  $\alpha = \mathcal{W}(\mathbf{1}/\lambda)/2^m$  for  $\mathcal{W}$  the unnormalized Walsh-Hadamard transform [45,46], as also reported in the main text. Noting that  $\lambda_{\mathbf{0}} = 1$ , as discussed in Appendix A3, we obtain also the alternate expression

$$\begin{aligned} \alpha_f &= \frac{1}{2^m} \left[ \frac{1}{\lambda_{\mathbf{0}}} + \sum_{\substack{k \in \mathcal{S} \\ k \neq \mathbf{0}}} (-1)^{f \cdot k} \left( \frac{1}{\lambda_k} \right) \right] \\ &= \frac{1}{2^m} \left[ 1 + \sum_{\substack{k \in \mathcal{S} \\ k \neq \mathbf{0}}} (-1)^{f \cdot k} \left( \frac{1}{\lambda_k} \right) \right]. \end{aligned} \quad (\text{B12})$$

For illustration, in Table II we provide explicitly written solutions for  $\alpha$  at small  $m$ . For brevity we have used the form of  $\alpha$  in Eq. (B12), and the form of  $\lambda$  in Eq. (A19).

*Resource complexity.* In this general setting, a naïve procedure of obtaining  $\alpha$  is first to compute  $\lambda$  from  $\mathbf{q}$  using Eq. (A18) or Eq. (A19), which comprises  $2^m$  entries each involving a sum over extensively many entries of  $\mathbf{q}$ . Then  $\alpha$  is computed from  $\lambda$  using Eq. (B11) or Eq. (B12). Each of these steps requires  $\mathcal{O}(2^{2m})$  time and  $\mathcal{O}(2^m)$  space in classical resources, for the same overall complexity. But noting that  $\mathbf{q}$ ,  $\lambda$  and  $\alpha$  are related by Walsh-Hadamard

TABLE II. Explicitly written general solutions for  $\alpha$  at small  $m$  for illustrative purposes. In consideration of space and readability,  $\alpha_k$  entries for  $m = 3$  and  $k \neq \mathbf{0}$  are left out; they follow the same form as  $\alpha_{\mathbf{0}}$  but with sign flips on individual terms.

| $m$ | Solution                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1   | $\alpha_{\mathbf{0}} = \frac{1}{2} \left( 1 + \frac{1}{1 - 2q_1} \right)$ $\alpha_1 = \frac{1}{2} \left( 1 - \frac{1}{1 - 2q_1} \right)$                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| 2   | $\alpha_{\mathbf{00}} = \frac{1}{4} \left( 1 + \frac{1}{1 - 2q_{01} - 2q_{10}} + \frac{1}{1 - 2q_{01} - 2q_{11}} + \frac{1}{1 - 2q_{10} - 2q_{11}} \right)$ $\alpha_{01} = \frac{1}{4} \left( 1 - \frac{1}{1 - 2q_{01} - 2q_{10}} - \frac{1}{1 - 2q_{01} - 2q_{11}} + \frac{1}{1 - 2q_{10} - 2q_{11}} \right)$ $\alpha_{10} = \frac{1}{4} \left( 1 - \frac{1}{1 - 2q_{01} - 2q_{10}} + \frac{1}{1 - 2q_{01} - 2q_{11}} - \frac{1}{1 - 2q_{10} - 2q_{11}} \right)$ $\alpha_{11} = \frac{1}{4} \left( 1 + \frac{1}{1 - 2q_{01} - 2q_{10}} - \frac{1}{1 - 2q_{01} - 2q_{11}} - \frac{1}{1 - 2q_{10} - 2q_{11}} \right)$ |
| 3   | $\alpha_{\mathbf{000}} = \frac{1}{8} \left( 1 + \frac{1}{1 - 2q_{010} - 2q_{011} - 2q_{100} - 2q_{101}} + \frac{1}{1 - 2q_{001} - 2q_{011} - 2q_{100} - 2q_{110}} \right.$ $+ \frac{1}{1 - 2q_{001} - 2q_{010} - 2q_{101} - 2q_{110}} + \frac{1}{1 - 2q_{001} - 2q_{010} - 2q_{100} - 2q_{111}}$ $+ \frac{1}{1 - 2q_{001} - 2q_{011} - 2q_{101} - 2q_{111}} + \frac{1}{1 - 2q_{010} - 2q_{011} - 2q_{110} - 2q_{111}} \left. \right)$ $\alpha_{001} = \frac{1}{8} (1 - \dots)$                                                                                                                                       |

transforms, as noted in the main text and in the discussion above, one can in fact utilize the FWHT algorithm [46,94], which runs in  $\mathcal{O}(N \log N)$  time and  $\mathcal{O}(N)$  space for an  $N$ -length input vector, for a more efficient solution computation procedure. This reduces the time complexity of solution computation to  $\mathcal{O}(m \cdot 2^m)$  and maintains space complexity at  $\mathcal{O}(2^m)$ . Explicitly, the procedure goes as follows:

- (1) Perform unnormalized FWHT to compute  $\lambda = \mathcal{W}(\mathbf{q})$ , requiring  $\mathcal{O}(m \cdot 2^m)$  time and  $\mathcal{O}(2^m)$  space.
- (2) Compute in-place element-wise division  $\mathbf{1}/\lambda$ , requiring  $\mathcal{O}(2^m)$  time.
- (3) Perform unnormalized FWHT to compute  $\alpha = \mathcal{W}(\mathbf{1}/\lambda)/2^m$ , requiring  $\mathcal{O}(m \cdot 2^m)$  time and  $\mathcal{O}(2^m)$  space.

Thereafter, computing  $\xi$  and  $|\alpha|/\xi$  take subdominant  $\mathcal{O}(2^m)$  time, the latter doable in-place. Last, standard algorithms to sample from an explicit probability distribution, such as alias and table methods [47–49,95], allow the drawing of bitmasks  $f \in \mathcal{S}$  from the distribution  $|\alpha|/\xi$  in  $\mathcal{O}(1)$  time per shot, with an initial preprocessing time linear in the number of entries in the distribution, a subdominant  $\mathcal{O}(2^m)$  in the context here. The overall resource requirements of  $\mathcal{O}(m \cdot 2^m)$  time and  $\mathcal{O}(2^m)$  space for solution computation and  $\mathcal{O}(1)$  time per shot are reported in Table I of the main text.

### b. Simplified solution of $\alpha$ for tensored readout channels

As outlined in Sec. III D, making an assumption that readout errors on each of the  $m$  measurements are independent, equivalently that the readout error channel acting on the  $m$  measurements factorizes into a tensor product of channels acting on individual measurements, allows a simplification of the  $\alpha$  solution. In particular, under such a setting,  $\mathbf{q}$ ,  $\mathbf{Q}$ , and  $\mathbf{Q}^{-1}$  each factorizes into a tensor product of single-measurement counterparts, as worked through in Appendix A 4 c. Then from Eqs. (A30) and (A35), we have

$$\alpha = \bigotimes_{\ell=1}^m \alpha^{(\ell)}, \quad \alpha^{(\ell)} = \frac{1}{1 - 2r_\ell} \begin{bmatrix} 1 - r_\ell \\ -r_\ell \end{bmatrix}, \quad (\text{B13})$$

where  $\alpha^{(\ell)}$  are coefficients for the  $\ell$ -measurement. This is presented as Eq. (13) in the main text. Equivalently,

$$\alpha_f = \prod_{\ell=1}^m \alpha_{f_\ell}^{(\ell)}, \quad \alpha_b^{(\ell)} = \frac{r_\ell^b (r_\ell - 1)^{1-b}}{2r_\ell - 1}, \quad (\text{B14})$$

where  $b \in \{0, 1\}$ . The overhead factor associated with mitigating the  $\ell$ -measurement is easily computed,

$$\xi^{(\ell)} = \|\alpha^{(\ell)}\|_1 = \frac{1}{1 - 2r_\ell}, \quad (\text{B15})$$

and Eq. (B13) implies that

$$\xi = \|\alpha\|_1 = \prod_{\ell=1}^m \|\alpha^{(\ell)}\|_1 = \prod_{\ell=1}^m \xi^{(\ell)} = \prod_{\ell=1}^m \frac{1}{1 - 2r_\ell}, \quad (\text{B16})$$

which is presented as Eq. (14) of the main text.

*Resource complexity.* The product structure of  $\alpha$  and  $\xi$  above imply that  $|\alpha|/\xi$  is a product over the individual  $|\alpha^{(\ell)}|/\xi^{(\ell)}$  probability distributions,

$$\frac{|\alpha|}{\xi} = \bigotimes_{\ell=1}^m \frac{|\alpha^{(\ell)}|}{\xi^{(\ell)}} \iff \frac{|\alpha_f|}{\xi} = \prod_{\ell=1}^m \frac{|\alpha_{f_\ell}^{(\ell)}|}{\xi^{(\ell)}}. \quad (\text{B17})$$

The probability distributions  $|\alpha^{(\ell)}|/\xi^{(\ell)}$  and overall rescaling factor  $\xi$  to be applied at the end of the mitigation protocol can be computed in  $\mathcal{O}(m)$  time and space, as there are  $2m$  entries across all  $\alpha^{(\ell)}$  and  $m$  overhead factors  $\xi^{(\ell)}$  in total, as explicitly written above. Then, for each experiment shot, a round of sampling can be used to draw  $f_\ell$  from the distribution  $|\alpha^{(\ell)}|/\xi^{(\ell)}$  for each  $\ell \in [m]$ , thus producing a bitmask  $f \in \mathcal{S}$  from the product distribution  $|\alpha|/\xi$  in  $\mathcal{O}(m)$  time. These complexities are reflected in Table I of the main text.

In the simpler setting of uniform readout channels,  $r_\ell = r$  for all  $\ell \in [m]$  and all the  $|\alpha^{(\ell)}|/\xi^{(\ell)}$  distributions are identical. Thus, one needs to only compute a single  $|\alpha^{(\ell)}|/\xi^{(\ell)}$ , which requires  $\mathcal{O}(1)$  space and time. Then, for each experiment shot,  $m$  rounds of sampling can be used to draw  $f_\ell$  from the  $|\alpha^{(\ell)}|/\xi^{(\ell)}$  distribution repeatedly, thus building up the bitmask  $f \in \mathcal{S}$  for the shot in  $\mathcal{O}(m)$  time. This is also reflected in Table I of the main text.

## 2. Multiple feedforward layers

### a. General solution of $\alpha$ for arbitrarily correlated readout channels

Similar to Appendix B 1 a, here we provide an elaboration of the discussion in Sec. IV B of the main text. To start, we write the quantum state after the  $L$  layers of mid-circuit measurements and feedforward in the absence of readout errors,

$$\bar{\sigma} = \sum_{s \in \mathcal{S}} W_{ss} \rho W_{ss}^\dagger, \quad W_{ss'} = \prod_{l=L}^1 V_{s^{[l]}}^{[l]} \Pi_{s^{[l]}}, \quad (\text{B18})$$

where  $W$  is the sequence of mid-circuit measurements and feedforward unitaries applied, as also defined in Eq. (15) of the main text. Here we remind that  $\mathcal{S} = \{0, 1\}^m$  is the set of possible measurement outcomes over the  $m$  total mid-circuit measurements in the quantum circuit, and  $s = s^{[1]}s^{[2]} \dots s^{[L]} \in \mathcal{S}$  is the partitioning of bitstring  $s$  over

the  $L$  layers,  $s^{[l]} \in \mathcal{S}^{[l]} = \{0, 1\}^{m^{[l]}}$ . As the computational basis ( $Z$  basis) measurement projectors  $\Pi_{s^{[l]}}^{[l]}$  are Hermitian,

$$W_{ss'}^\dagger = \prod_{l=1}^L \Pi_{s^{[l]}}^{[l]} V_{s'^{[l]}}^{[l]\dagger}. \quad (\text{B19})$$

Taking  $L = 1$  and dropping the layer superscripts that are unneeded, Eq. (B18) recovers Eq. (B1) for a single feedforward layer. It is straightforward also to rewrite Eq. (B18) in a picture of quantum trajectories,

$$\bar{\sigma} = \sum_{s \in \mathcal{S}} p_s \rho_s, \quad \rho_s = \frac{W_{ss} \rho W_{ss}^\dagger}{p_s}, \quad p_s = \text{tr}[W_{ss} \rho W_{ss}^\dagger], \quad (\text{B20})$$

where  $\rho_s$  is the normalized quantum state at the end of the quantum circuit for quantum trajectories defined by the measurement outcome  $s \in \mathcal{S}$ , and  $p_s$  is the probability of those trajectories occurring. Taking  $L = 1$  and dropping layer superscripts, this recovers Eq. (2) of the main text for a single feedforward layer. Reproducing Eq. (15), we define

$$T_{bss'} = \text{tr} \left[ O_b W_{ss'} \rho W_{ss'}^\dagger \right], \quad (\text{B21})$$

which is the expectation value of  $O_b$  measured on quantum trajectories that had true measurement outcomes  $s$  but feedforward unitaries applied according to  $s'$ , weighted by the probability of those trajectories occurring. This is directly analogous to Eq. (3) in the single-layer context. Then, using Eq. (B18) and the definition of  $T_{bss'}$ , we have

$$\langle \bar{O}_b \rangle = \text{tr}(O_b \bar{\sigma}) = \sum_{s \in \mathcal{S}} T_{bss} = \text{tr} T_b. \quad (\text{B22})$$

Next we consider the presence of readout errors. We write the quantum state after the  $L$  layers of mid-circuit measurements and feedforward in a case where readout errors have corrupted measurement outcomes  $s$  into  $s'$ , and a bitmask  $f \in \mathcal{S}$  has been imposed on the feedforward,

$$\sigma^{(f)} = \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} q_{s \oplus s'} W_{s(s' \oplus f)} \rho W_{s(s' \oplus f)}^\dagger. \quad (\text{B23})$$

Then

$$\langle O_b^{(f)} \rangle = \text{tr}(O_b \sigma^{(f)}) = \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} q_{s \oplus s'} T_{bs(s' \oplus f)}. \quad (\text{B24})$$

Note that Eqs. (B22) and (B24) are identical to Eqs. (B5) and (B6) in Appendix B 1 a. Moreover, as discussed in Sec. IV A, our definition of the mitigated estimator  $\hat{\mathbf{O}}$  is identical for both the single-layer and multiple-layer contexts. Thus, the remainder of the derivation in

Appendix B 1 a follows through, leading to the solution for  $\alpha$  in Eqs. (B11) and (B12), the former presented in Eq. (11) of the main text and referenced in Sec. IV B accordingly. The only differences to keep in mind, which are only interpretational and do not change the form of the equations and solutions, are the generalized definitions of  $\mathcal{S}$  being the set of possible measurement outcomes across all layers, and  $T$  likewise being expectation values on trajectories across all layers, as written above. The classical resource requirements analyzed in Appendix B 1 a likewise straightforwardly apply.

### b. Simplified solution of $\alpha$ for layerwise tensored readout channels

As discussed in Sec. IV C, one may examine a setting where readout errors on measurements in different layers are assumed to be independent. This setting falls exactly into that considered in Appendix A 4 b. Picking up the discussion there, from Eqs. (A28) to (A30), we have

$$\mathbf{q} = \bigotimes_{l=1}^L \mathbf{q}^{[l]}, \quad \lambda = \bigotimes_{l=1}^L \lambda^{[l]} \iff \lambda_k = \prod_{l=1}^L \lambda_{k^{[l]}}^{[l]},$$

$$\alpha = \bigotimes_{l=1}^L \alpha^{[l]} \iff \alpha_f = \prod_{l=1}^L \alpha_f^{[l]}, \quad (\text{B25})$$

where  $k = k^{[1]}k^{[2]} \dots k^{[L]} \in \mathcal{S}$  and  $f = f^{[1]}f^{[2]} \dots f^{[L]} \in \mathcal{S}$  are partitionings of the bitstrings over the layers, with  $k^{[l]} \in \mathcal{S}^{[l]}$  and  $f^{[l]} \in \mathcal{S}^{[l]}$ , and  $\mathbf{q}^{[l]}$ ,  $\lambda^{[l]}$  and  $\alpha^{[l]}$  are respectively the confusion probability vector, eigenvalues and mitigator coefficients for layer  $l \in [L]$ . Part of the above is presented in Eq. (16) of the main text. The tensor product structure of  $\alpha$  also implies that

$$\xi = \|\alpha\|_1 = \prod_{l=1}^L \|\alpha^{[l]}\|_1 = \prod_{l=1}^L \xi^{[l]}, \quad (\text{B26})$$

which is presented as Eq. (17) of the main text. To be explicit,  $\lambda^{[l]}$  and  $\alpha^{[l]}$  are given by the same solutions as in Eqs. (A18), (A19), (B11), and (B12) or equivalently Eqs. (10) and (11) of the main text, but with quantities replaced by their layerwise counterparts—that is,  $(\mathcal{S}, m, \mathbf{q}, k, f)$  replaced by  $(\mathcal{S}^{[l]}, m^{[l]}, \mathbf{q}^{[l]}, k^{[l]}, f^{[l]})$ .

*Resource complexity.* The product structure of  $\alpha$  and  $\xi$  above imply that  $|\alpha|/\xi$  is a product over the layerwise  $|\alpha^{[l]}|/\xi^{[l]}$  probability distributions,

$$\frac{|\alpha|}{\xi} = \bigotimes_{l=1}^L \frac{|\alpha^{[l]}|}{\xi^{[l]}} \iff \frac{|\alpha_f|}{\xi} = \prod_{l=1}^m \frac{|\alpha_f^{[l]}|}{\xi^{[l]}}. \quad (\text{B27})$$

Thus, using the solution computation method in Appendix B 1 a, one can compute  $|\alpha^{[l]}|/\xi^{[l]}$  for every

$l \in [L]$  in total  $\mathcal{O}\left(\sum_{l=1}^L m^{[l]} \cdot 2^{m^{[l]}}\right)$  time and  $\mathcal{O}\left(\sum_{l=1}^L 2^{m^{[l]}}\right)$  space. After a subdominant preprocessing time of  $\mathcal{O}\left(\sum_{l=1}^L 2^{m^{[l]}}\right)$ , for each experiment shot, a round of constant-time sampling can be run to draw  $f^{[l]}$  from the  $|\alpha^{[l]}|/\xi^{[l]}$  distribution for each  $l \in [L]$ , thus building up the bitmask  $f = f^{[1]}f^{[2]} \dots f^{[L]} \in \mathcal{S}$  for the shot in  $\mathcal{O}(L)$  time. Thus the overall complexities are  $\mathcal{O}\left(\sum_{l=1}^L m^{[l]} \cdot 2^{m^{[l]}}\right)$  time and  $\mathcal{O}\left(\sum_{l=1}^L 2^{m^{[l]}}\right)$  space for solution computation, and  $\mathcal{O}(L)$  sampling time per shot.

Noting that  $\mathcal{O}\left(\sum_{l=1}^L m^{[l]} \cdot 2^{m^{[l]}}\right) \leq \mathcal{O}\left(\sum_{l=1}^L m^{[l]} \cdot 2^{\bar{m}}\right) \leq \mathcal{O}(m \cdot 2^{\bar{m}})$  and  $\mathcal{O}\left(\sum_{l=1}^L 2^{m^{[l]}}\right) \leq \mathcal{O}(L \cdot 2^{\bar{m}})$ , where  $\bar{m} = \max_{l \in [L]} m^{[l]}$  is the maximum number of measurements in a layer, we can more succinctly bound the time and space complexities as  $\mathcal{O}(m \cdot 2^{\bar{m}})$  and  $\mathcal{O}(L \cdot 2^{\bar{m}})$  respectively, which are reported in Table I of the main text.

### c. Simplified solution of $\alpha$ for fully tensored readout channels

This setting is exactly equivalent to that considered in Appendix B 1 b; thus, we direct readers to the discussion provided therein.

### 3. Sampling overhead of mitigated estimator $\hat{\mathbf{O}}$

We reproduce the definition of our estimator,

$$\hat{\mathbf{O}} := \sum_{f \in \mathcal{S}} \alpha_f \mathbf{O}^{(f)} = \xi \underbrace{\sum_{f \in \mathcal{S}} \frac{|\alpha_f|}{\xi} \text{sgn}(\alpha_f) \mathbf{O}^{(f)}}_{\tilde{\mathbf{O}}}, \quad (\text{B28})$$

as in Eq. (7) of the main text. We have labeled the estimator before the final rescaling by  $\xi = \|\alpha\|_1$  as  $\tilde{\mathbf{O}}$ . Here we generically regard multiple feedforward layers, of which a single feedforward layer is a special case, and thus the results obtained apply to both Secs. III and IV of the main text as referenced therein.

Recall the general PROM protocol, as described in Algorithms 1 and 2, which is to draw a bitstring  $f$  from the  $|\alpha|/\xi$  probability distribution in each shot of the experiment circuit and measure  $\text{sgn}(\alpha_f) \mathbf{O}^{(f)}$  on that experiment shot. The observables measured in each shot are  $\mathbf{O} = [O_1 \ O_2 \ \dots \ O_B]$ . Thus, the shot yields a measurement result that comprises an eigenvalue of each  $O_b$  measured up to sign, which is bounded between  $[-\|O_b\|_2, +\|O_b\|_2]$ , where the spectral norm  $\|O_b\|_2$  reflects the largest singular value of the observable  $O_b$ . Then their convex linear combination  $\tilde{O}_b$  are all bounded within  $[-\|O_b\|_2, +\|O_b\|_2]$ . In a hypothetical ideal scenario without readout errors, not employing PROM, in each shot the measurement of  $\mathbf{O}$  likewise gives a result with entries bounded within  $[-\|O_b\|_2, +\|O_b\|_2]$ .

We follow a standard line of analysis, similar to, for example, Ref. [96] albeit in a different context. We

compare the mitigated estimator  $\hat{\mathbf{O}}$  and the ideal estimator  $\bar{\mathbf{O}}$ , the latter accessible only in a hypothetical noiseless setting, averaged over  $N$  independent experiment shots,

$$\hat{\mathbf{O}}^{(N)} = \frac{1}{N} \sum_{j=1}^N \hat{\mathbf{O}}, \quad \bar{\mathbf{O}}^{(N)} = \frac{1}{N} \sum_{j=1}^N \bar{\mathbf{O}}, \quad (\text{B29})$$

and we likewise define

$$\tilde{\mathbf{O}}^{(N)} = \frac{1}{N} \sum_{j=1}^N \tilde{\mathbf{O}}. \quad (\text{B30})$$

We demand precision  $\epsilon$  and confidence  $1 - \delta$  that holds over all observables, explicitly

$$\begin{aligned} \Pr \left[ \left| \hat{O}_b^{(N)} - \langle \bar{O}_b \rangle \right| \geq \epsilon \right] &\leq \delta \quad \forall b \in [B], \\ \Pr \left[ \left| \bar{O}_b^{(N)} - \langle \bar{O}_b \rangle \right| \geq \epsilon \right] &\leq \delta \quad \forall b \in [B], \end{aligned} \quad (\text{B31})$$

and we want to compute the scaling of the number of shots  $N$  required with these parameters. Standard application of Hoeffding's inequality gives a tail bound

$$\begin{aligned} \Pr \left[ \left| \tilde{O}_b^{(N)} - \langle \tilde{O}_b \rangle \right| \geq \chi \right] &= \Pr \left[ \left| \xi \tilde{O}_b^{(N)} - \xi \langle \tilde{O}_b \rangle \right| \geq \xi \chi \right] \\ &= \Pr \left[ \left| \hat{O}_b^{(N)} - \langle \hat{O}_b \rangle \right| \geq \xi \chi \right] \\ &\leq 2 \exp \left( -\frac{N \chi^2}{2 \|O_b\|_2^2} \right), \end{aligned} \quad (\text{B32})$$

and setting  $\xi \chi = \epsilon$ , we have

$$\begin{aligned} \Pr \left[ \left| \hat{O}_b^{(N)} - \langle \hat{O}_b \rangle \right| \geq \epsilon \right] &\leq 2 \exp \left( -\frac{N \epsilon^2}{2 \xi^2 \|O_b\|_2^2} \right) \\ &\leq \delta \implies N \propto \frac{2 \xi^2 \|O_b\|_2^2}{\epsilon^2} \log \left( \frac{2}{\delta} \right). \end{aligned} \quad (\text{B33})$$

On the other hand, Hoeffding's inequality applied to the ideal estimator gives

$$\begin{aligned} \Pr \left[ \left| \bar{O}_b^{(N)} - \langle \bar{O}_b \rangle \right| \geq \epsilon \right] &\leq 2 \exp \left( -\frac{N \epsilon^2}{2 \|O_b\|_2^2} \right) \\ &\leq \delta \implies N \propto \frac{2 \|O_b\|_2^2}{\epsilon^2} \log \left( \frac{2}{\delta} \right). \end{aligned} \quad (\text{B34})$$

Comparing Eqs. (B33) and (B34), we observe that to reach the same precision ( $\epsilon$ ) at the same confidence (failure probability not exceeding  $\delta$ ) as in a hypothetical setting of noiseless circuit execution, the number of shots of the noisy circuit required is larger by a multiplicative factor of  $\xi^2$ . We thus associate  $\xi^2$  with the sample complexity overhead of mitigating readout errors with PROM.

#### 4. Sensitivity of error mitigation to differences in inputs

Consider  $\mathbf{q}$  and  $Q$  inputs with their corresponding  $\alpha$  solution for the PROM protocol, and a different  $\mathbf{q}'$  and  $Q'$  with their  $\alpha'$  solution. We denote the overhead factors of the two solutions  $\xi = \|\alpha\|_1$  and  $\xi' = \|\alpha'\|_1$ . We analyze the relation between the difference in  $\mathbf{q}$  and  $\mathbf{q}'$  and that in  $\alpha$  and  $\alpha'$ , and thus the mitigated estimators produced by the protocol. Recall that  $\alpha$  is the solution of a linear system specified by  $Q$ —i.e., Eq. (B10). Thus, we start from a standard result from the theory of sensitivity of linear systems [97,98],

$$\frac{\|\alpha' - \alpha\|_p}{\|\alpha\|_p} \leq \frac{\kappa_p(Q)r(Q)}{1 - \kappa_p(Q)r(Q)}, \quad \kappa_p(Q) = \|Q^{-1}\|_p \|Q\|_p, \\ r(Q) = \frac{\|Q' - Q\|_p}{\|Q\|_p}, \quad (\text{B35})$$

valid provided  $Q$  is nonsingular and  $\kappa_p(Q)r(Q) < 1$ , for any vector  $p$ -norm and the associated induced matrix norm, and where  $\kappa_p(Q)$  is the normwise condition number of  $Q$  under that norm. Here for our purposes we examine the 1-norm. Note from Eq. (A10) that  $\|Q\|_1 = 1$ . On the other hand

$$\|Q' - Q\|_1 = \max_{f \in S} \sum_{s \in S} |Q'_{sf} - Q_{sf}| \\ = \max_{f \in S} \underbrace{\sum_{s \in S} |q'_{s \oplus f} - q_{s \oplus f}|}_{2d(\mathbf{q}', \mathbf{q})} = 2d(\mathbf{q}', \mathbf{q}), \quad (\text{B36})$$

where  $0 \leq d(\mathbf{q}', \mathbf{q}) \leq 1$  is the total variation distance between the probability distributions  $\mathbf{q}'$  and  $\mathbf{q}$ . Moreover, recall that  $\alpha$  is the first column of  $Q^{-1}$  as derived in Eqs. (B10) and (B11), and that  $\|Q^{-1}\|_1$  reduces to the 1-norm of its first column (or indeed any of its columns), as shown in Eq. (A24). Thus,

$$\|Q^{-1}\|_1 = \|\alpha\|_1 = \xi, \quad \kappa_1(Q) = \|Q^{-1}\|_1 \|Q\|_1 = \xi. \quad (\text{B37})$$

Putting the above together,

$$\frac{\|\alpha' - \alpha\|_1}{\xi} \leq \frac{2\xi d(\mathbf{q}', \mathbf{q})}{1 - 2\xi d(\mathbf{q}', \mathbf{q})}, \quad (\text{B38})$$

valid provided  $2\xi d(\mathbf{q}', \mathbf{q}) < 1$ . Then, using the reverse triangle inequality, we observe the following sensitivity bound on the overhead factor  $\xi'$  given differing  $\mathbf{q}$  and  $\mathbf{q}'$

inputs,

$$\frac{|\xi' - \xi|}{\xi} = \frac{|\|\alpha'\|_1 - \|\alpha\|_1|}{\xi} \leq \frac{\|\alpha' - \alpha\|_1}{\xi} \\ = \frac{2\xi d(\mathbf{q}', \mathbf{q})}{1 - 2\xi d(\mathbf{q}', \mathbf{q})}. \quad (\text{B39})$$

Next we examine the mitigated estimators

$$\hat{\mathbf{O}} = \sum_{f \in S} \alpha_f \mathbf{O}^{(f)}, \quad \hat{\mathbf{O}}' = \sum_{f \in S} \alpha'_f \mathbf{O}^{(f)}. \quad (\text{B40})$$

We observe the following sensitivity bound on the mitigated expectation values  $\langle \hat{\mathbf{O}}_b \rangle$  given differing  $\mathbf{q}$  and  $\mathbf{q}'$  inputs,

$$|\langle \hat{\mathbf{O}}_b \rangle - \langle \hat{\mathbf{O}}'_b \rangle| = \left| \sum_{f \in S} (\alpha'_f - \alpha_f) \langle \mathbf{O}_b^{(f)} \rangle \right| \\ \leq \sum_{f \in S} |\alpha'_f - \alpha_f| |\langle \mathbf{O}_b^{(f)} \rangle| \\ \leq \sum_{f \in S} |\alpha'_f - \alpha_f| \cdot \|\mathbf{O}_b\|_2 \\ = \|\alpha' - \alpha\|_1 \cdot \|\mathbf{O}_b\|_2 \\ \leq \frac{2\xi^2 d(\mathbf{q}', \mathbf{q})}{1 - 2\xi d(\mathbf{q}', \mathbf{q})} \cdot \|\mathbf{O}_b\|_2 \quad (\text{B41})$$

for each  $b \in [B]$ , where spectral norm  $\|\mathbf{O}_b\|_2$  corresponds to the largest singular value of the observable  $\mathbf{O}_b$ .

##### a. Corollary—General bound on sampling overheads

Consider the sensitivity bound in Eq. (B39) but with primed and unprimed quantities swapped, and take  $\mathbf{q}'$  to correspond to the ideal scenario without readout errors—such that  $q'_s = \delta_{s0}$  and thus  $\alpha'_f = \delta_{f0}$  and  $\xi' = 1$ . We leave  $\mathbf{q}$ , with corresponding solution  $\alpha$  and sampling overhead  $\xi$ , to be completely arbitrary. That is,  $\mathbf{q}$  accommodates arbitrary correlations in readout errors. Suppose the total readout error probability is at most  $\eta$ , that is,  $q_0 \geq 1 - \eta$ . Then

$$d(\mathbf{q}, \mathbf{q}') = d(\mathbf{q}', \mathbf{q}) = \frac{1}{2} \sum_{s \in S} |q_s - q'_s| \\ = \frac{1}{2} \left( \underbrace{|q_0 - q'_0|}_{\leq 1 - (1 - \eta)} + \underbrace{\sum_{\substack{s \in S \\ s \neq 0}} |q_s - q'_s|}_{\leq \eta} \right) \leq \eta, \quad (\text{B42})$$

and by the sensitivity bound,

$$\begin{aligned} \frac{|\xi - \xi'|}{\xi'} &\leq \frac{2\xi'd(\mathbf{q}, \mathbf{q}')}{1 - 2\xi'd(\mathbf{q}, \mathbf{q}')} \implies |\xi - 1| \\ &\leq \frac{2\eta}{1 - 2\eta} \implies \xi \leq \frac{1}{1 - 2\eta}, \end{aligned} \quad (\text{B43})$$

valid for  $\eta \in [0, 1/2)$ .

**b. Corollary—General bound for errors in observables due to readout errors in feedforward**

Now we consider the sensitivity bound in Eq. (B41). We take  $\mathbf{q}'$  to be the exact characterization of readout errors on the device, and  $\mathbf{q}$  to correspond to a scenario without readout errors ( $q_s = \delta_{s0}$ ), such that  $\alpha_f = \delta_{f0}$  and  $\xi = 1$ . That is, employing the PROM protocol with  $\mathbf{q}$  is tantamount to performing no readout error mitigation at all—since only  $f = 0$  is ever sampled. Thus

$$\langle \widehat{O}_b \rangle = \langle O_b^{\text{raw}} \rangle, \quad \langle \widehat{O}'_b \rangle = \langle \overline{O}_b \rangle, \quad (\text{B44})$$

where  $\langle O_b^{\text{raw}} \rangle$  is the expectation value of observable  $O_b$  measured in a setting with readout errors described by  $\mathbf{q}'$  but with no readout error mitigation applied, and  $\langle \overline{O}_b \rangle$  is the ideal expectation value as would be measured in a noiseless setting, the latter as written with the same notation in Secs. III and IV of the main text.

Suppose the total readout error probability is at most  $\eta$ , that is,  $q'_0 \geq 1 - \eta$ . Then Eq. (B42) identically applies. The error in the expectation value of observable  $O_b$  measured in a setting with readout errors (characterized by  $\mathbf{q}'$ ) that impact the mid-circuit measurements and feedforward, but with no readout error mitigation applied, with respect to an ideal setting without readout errors, is then bounded by

$$\begin{aligned} |\langle O_b^{\text{raw}} \rangle - \langle \overline{O}_b \rangle| &= |\langle \widehat{O}'_b \rangle - \langle \widehat{O}_b \rangle| \leq \frac{2\xi'^2 d(\mathbf{q}', \mathbf{q})}{1 - 2\xi' d(\mathbf{q}', \mathbf{q})} \cdot \|O_b\|_2 \\ &\leq \frac{2\eta}{1 - 2\eta} \cdot \|O_b\|_2, \end{aligned} \quad (\text{B45})$$

valid for  $\eta \in [0, 1/2)$ .

## 5. Terminal readout error mitigation

Here, we discuss the special case of readout error mitigation for terminal measurements. While the PROM protocol was motivated by the need for readout error mitigation on mid-circuit measurements with feedforward, the protocol can also be used for terminal measurements, as we describe in this section.

First, in Appendix B 5 a, we establish the theoretical setting of readout error mitigation for terminal measurements and examine sample complexity and classical processing costs for conventional readout error mitigation methods.

Then, in Appendix B 5 b, we cast terminal readout error mitigation into a problem class treated by PROM, thereby establishing PROM as a general framework for readout error mitigation on both mid-circuit and terminal measurements. This discussion is theoretically motivated and explains that the description of PROM in the main text, despite being phrased in the language of mid-circuit measurements, without loss of generality accommodates also terminal measurements. We observe that the sample complexity and classical processing costs of PROM applied to terminal measurements are equivalent to those of conventional methods, assuming both exploit symmetrization of readout error channels. Last, in Appendix B 5 c we describe a version of the PROM protocol with integrated terminal readout error mitigation, which is suitable for straightforward implementation in experiments.

**a. Framework and resource requirements of conventional terminal readout error mitigation**

The conventional framework for readout error mitigation on terminal measurements is described in Appendix C 2, in particular Appendix C 2 c with symmetrization of readout channels—we refer readers to these sections for more detailed background. To summarize, conceptually, the conventional mitigation works by way of linear inversion. In an experiment where a subset of  $h \leq n$  qubits are terminally measured in the computational basis,

$$\mathbf{c}' = Q^{-1} \mathbf{c} \iff \mathbf{p}' = Q^{-1} \mathbf{p}, \quad (\text{B46})$$

where  $\mathbf{c}$  is the vector of observed bitstring counts, in which each entry  $c_s$  records the number of occurrences of terminal measurement outcome  $s \in \mathcal{S} = \{0, 1\}^h$  in the experiment,  $\mathbf{c}'$  is the vector of mitigated bitstring counts, and  $Q$  is the symmetrized confusion matrix characterizing readout error channels on the terminal measurements. We denote the observed and mitigated bitstring probabilities as  $\mathbf{p} = \mathbf{c}/N'$  and  $\mathbf{p}' = \mathbf{c}'/N'$  respectively, where  $N'$  is the number of experiment shots executed.

The desired outputs from the experiment are generically expectation values of some observables  $\mathbf{O}$ . These expectation values are obtained from a linear transformation of  $\mathbf{p}'$ . That is, the expectation values are reported as

$$\langle \widehat{\mathbf{O}} \rangle = A \mathbf{p}' \quad (\text{B47})$$

for a matrix  $A$  associated with  $\mathbf{O}$ . We give a few examples to illustrate this setup:

- (1) Suppose the desired output is the complete set of bitstring probabilities  $\mathbf{p}'$  over the  $h$  terminally measured qubits. This occurs, for example, in random circuit sampling or quantum optimization (e.g., QAOA) experiments. This corresponds to taking  $\mathbf{O}$

to be the complete set of projectors onto computational basis states,  $O_s = \Pi_s$  for  $s \in \mathcal{S}$ , and the associated  $A = \mathbb{I}$ .

- (2) Suppose  $h = 1$ , and the observables of interest comprise the single Pauli operator  $\sigma^z$  on the terminally measured qubit. Then, in the bitstring basis  $(0, 1)$  over the columns,

$$A = \begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix}. \quad (\text{B48})$$

- (3) Suppose  $h = 2$ , and the observables of interest are  $\mathbf{O} = [\sigma_1^z \otimes \sigma_2^z, \sigma_1^z (\mathbb{I} + \sigma_2^z)/2]^\top$ , where the subscripts denote the terminally measured qubits. Then, in the bitstring basis  $(00, 01, 10, 11)$  over the columns,

$$A = \begin{bmatrix} +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 \\ +1 & 0 & +1 & 0 \end{bmatrix}. \quad (\text{B49})$$

*Sampling overhead of terminal readout error mitigation.* To discuss the sampling overhead of terminal readout error mitigation, it is analytically most expedient to cast the framework above, i.e., as described in Eqs. (B46) and (B47), as an equivalent sampling procedure, where equivalence here follows from linearity of expectation values. This procedure is as follows:

- (1) For shot  $k = 1$  to  $N'$ :
  - (a) Perform the terminal measurements and obtain an outcome bitstring  $s \in \mathcal{S}$ , with the associated one-hot counts vector  $\mathbf{c}^{(k)}$  containing a 1 entry at bitstring index  $s$  and is zero everywhere else.
  - (b) Compute the linear inversion  $\mathbf{c}'^{(k)} = Q^{-1} \mathbf{c}^{(k)}$ .
  - (c) Record the observable sample  $\mathbf{o}^{(k)} = A \mathbf{c}'^{(k)}$ .
- (2) Output estimate  $\langle \hat{\mathbf{O}} \rangle = \frac{1}{N'} \sum_{k=1}^{N'} \mathbf{o}^{(k)}$  as the average over all shots.

We consider the range of values  $\mathbf{o}^{(k)}$  can take. In shot  $k$  producing the outcome bitstring  $s \in \mathcal{S}$  from the terminal measurements, we have for each observable  $O_b$  in  $\mathbf{O}$ ,

$$o_b^{(k)} = \mathbf{a}_b^\top \mathbf{c}'^{(k)} = \mathbf{a}_b^\top Q^{-1} \mathbf{c}^{(k)} = \sum_{l \in \mathcal{S}} A_{bl} [Q^{-1}]_{ls}, \quad (\text{B50})$$

where  $\mathbf{a}_b$  is the  $b$ th row of  $A$ , and thus we observe

$$\begin{aligned} |o_b^{(k)}| &\leq \sum_{l \in \mathcal{S}} |A_{bl}| |[Q^{-1}]_{ls}| \\ &\leq \left( \max_{l \in \mathcal{S}} |A_{bl}| \right) \left( \max_{s \in \mathcal{S}} \sum_{l \in \mathcal{S}} |[Q^{-1}]_{ls}| \right) \\ &= \|\mathbf{a}_b\|_\infty \|Q^{-1}\|_1 = \xi \|\mathbf{a}_b\|_\infty, \end{aligned} \quad (\text{B51})$$

where we have noted  $\|Q^{-1}\|_1 = \|Q^{-1}\|_\infty = \xi$ . The bound in Eq. (B51) is tight and is saturated, for example, when  $Q = \mathbb{I}$  and when  $\mathbf{a}_b$  has its maximal entry in magnitude at bitstring index  $s$ . For completeness, we comment that another natural bound, which follows from an application of Cauchy-Schwarz inequality, is  $|o_b^{(k)}| \leq \|\mathbf{a}_b\|_1 \|Q^{-1}\|_1$ , but this bound is ultimately superfluous as  $\|\mathbf{a}_b\|_\infty \leq \|\mathbf{a}_b\|_1$  for any  $\mathbf{a}_b$ .

Equation (B51) indicates that each shot yields an observable sample  $o_b^{(k)}$  that is bounded between  $[-\xi \|\mathbf{a}_b\|_\infty, +\xi \|\mathbf{a}_b\|_\infty]$ . Then applying Hoeffding's inequality as a concentration bound on the estimate  $\langle O_b \rangle$  averaged over shots, in an identical fashion as in Appendix B3, we see that to reach precision  $\epsilon$  at confidence  $1 - \delta$ , a number of shots

$$N' \propto \frac{2\xi^2 \|\mathbf{a}_b\|_\infty^2}{\epsilon^2} \log\left(\frac{2}{\delta}\right) \quad (\text{B52})$$

is required, in comparison to

$$N \propto \frac{2 \|\mathbf{a}_b\|_\infty^2}{\epsilon^2} \log\left(\frac{2}{\delta}\right) \quad (\text{B53})$$

for a hypothetical noiseless experiment (where  $Q = Q^{-1} = \mathbb{I}$  and  $\xi = 1$ ). Thus, we conclude that in conventional terminal readout error mitigation, to reach the same precision at the same confidence as in a setting of noiseless circuit execution, the number of extra shots of the noisy circuit required is larger by an overhead factor of  $\xi^2$ .

*Classical processing costs.* Last, we remark on the classical computational cost of running the sampling procedure described above for conventional terminal readout error mitigation. The dominant cost lies in computing the linear inversion  $\mathbf{c}'^{(k)} = Q^{-1} \mathbf{c}^{(k)}$ , as  $Q^{-1}$  and  $\mathbf{c}^{(k)}$  are exponentially large in the number of qubits measured (of size  $2^h \times 2^h$  and of length  $2^h$  respectively). Naïve multiplication incurs a time complexity of  $\mathcal{O}(2^{2h})$ .

We can, however, take advantage of the highly symmetric structure in  $Q^{-1}$  to reduce the classical computational cost of terminal readout error mitigation, in the same fashion as was done for our PROM protocol (described in Appendixes A3 and B1 a), to reduce this complexity. As  $\mathbf{c}^{(k)}$  is one-hot in the bitstring index  $s$  produced in shot  $k$ , we have

$$c_f'^{(k)} = \sum_{l \in \mathcal{S}} [Q^{-1}]_{fl} \delta_{ls} = [Q^{-1}]_{fs} = [Q^{-1}]_{(f \oplus s)\mathbf{0}} = \alpha_{(f \oplus s)} \quad (\text{B54})$$

for each bitstring index  $f \in \mathcal{S}$ , where we recall the index shift symmetry  $[Q^{-1}]_{sf} = [Q^{-1}]_{(s \oplus f)\mathbf{0}}$  and the definition of the weights vector  $\alpha$  in PROM as the first column of  $Q^{-1}$ . That is,  $\mathbf{c}'^{(k)}$  is precisely  $\alpha$  but with entries shifted by the index offset  $s$ . Then our algorithm using the fast

Walsh-Hadamard transform to compute  $\alpha$  and the associated complexity analysis, as described in the main text and in Appendixes B 1 a and B 1 b, apply. In particular, for arbitrarily correlated readout error channels, a classical time complexity of  $\mathcal{O}(h \cdot 2^h)$  for computing  $\mathbf{c}^{(k)}$  in each shot is achieved, and with assumptions of independence and uniformity, this complexity is reduced  $\mathcal{O}(h)$  and  $\mathcal{O}(1)$  respectively.

### b. Terminal readout error mitigation as a special case of probabilistic readout error mitigation

Our formulation of PROM in the main text [i.e., in Secs. III and IV, or see Figs. 2(b) and 3(b)] concerns quantum circuits comprising layers of mid-circuit measurements subject to readout noise and feedforward, followed by noiseless terminal measurement of the observables of interest  $\mathbf{O}$ . This formulation is agnostic to the exact procedure of measuring  $\mathbf{O}$ , and we presume that a standard terminal readout error mitigation scheme is separately applied (i.e., on top of PROM). A question is whether this formulation of PROM is sufficiently general to accommodate also readout error mitigation on terminal measurements as part of PROM—to which we answer in the affirmative.

Again, to discuss terminal readout error mitigation, it is necessary to make explicit the measurement process of  $\mathbf{O}$ . To be concrete and to allow a direct comparison, we consider the general setup of the previous subsection (Appendix B 5 a). Suppose a subset of  $h \leq n$  qubits of the circuit, which we denote  $(q_1, q_2, \dots, q_h)$ , are terminally measured in the computational basis, and the confusion probability vector  $\mathbf{q}$  characterizes the readout error channels. To review, each entry  $q_s$  for  $s \in \mathcal{S} = \{0, 1\}^h$  records the probability that a true terminal measurement outcome

$a \in \mathcal{S}$  is reported as  $a \oplus s$ . After running the circuit in an experiment, the probabilities of terminal measurement outcomes  $\mathbf{p}$  are then used to compute expectation values of observables  $\langle \mathbf{O} \rangle$  through a linear transformation  $A$ , as described in Appendix B 5 a.

*Circuit recasting.* Conceptually, we recast the noisy terminal measurements as noiseless terminal measurements preceded by mid-circuit ones with feedforward, so that our formulation of PROM can be applied without modification. A simple way of doing so is by introducing  $h$  fictitious qubits, denoted  $(a_1, a_2, \dots, a_h)$  and which are initialized in  $|0\rangle^{\otimes h}$ , performing mid-circuit measurements in the computational basis on them, and applying a conditional  $X$  gate on the system qubit  $q_l$  when the measurement on  $a_l$  produces a 1 outcome for each  $l \in [h]$ . We stipulate that the readout error channels afflicting the mid-circuit measurements on these  $h$  fictitious qubits are characterized by  $\mathbf{q}$ . Last, noiseless terminal measurements are performed on the system qubits  $(q_1, q_2, \dots, q_h)$ . This restructuring of the quantum circuit is illustrated in Fig. 7. Observe that the recast circuit is exactly equivalent to the original in behavior. In the original circuit, for any terminal measurement outcome  $s \in \mathcal{S}$  on the system qubits  $(q_1, q_2, \dots, q_h)$  that would have hypothetically been obtained in the absence of readout noise, there is a  $q_a$  probability that  $s' = s \oplus a$  is observed for syndrome  $a \in \mathcal{S}$ . Likewise, in the recast circuit, for each hypothetical  $s \in \mathcal{S}$ , there is a  $q_a$  probability that the mid-circuit measurements on  $(a_1, a_2, \dots, a_h)$  produce an outcome  $a \in \mathcal{S}$ , which is fedforward to ultimately produce the outcome  $s' = a \oplus s$  during the terminal measurement.

*Sampling overhead and classical processing costs.* In effect, in this restructured circuit, we have transplanted readout errors from terminal measurements to mid-circuit

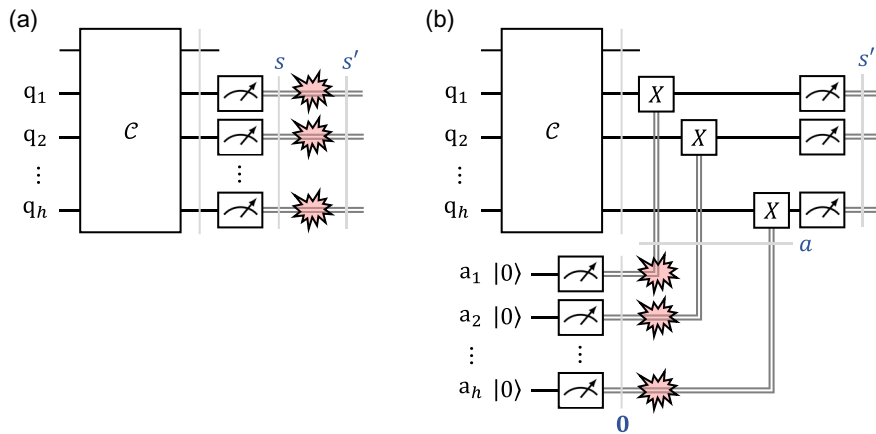


FIG. 7. Recasting noisy terminal measurements. (a) An arbitrary circuit  $C$  is followed by noisy terminal measurements. Readout errors corrupt a hypothetical measurement outcome (bitstring)  $s$  that would have been obtained without noise into  $s'$ . (b) A circuit producing an equivalent distribution of terminal measurement outcomes. The terminal measurements are now noiseless. Ancillary qubits all initialized in  $|0\rangle$  are introduced and measured mid-circuit; the measurement outcomes feedforward into  $X$  gates. Readout error channels affecting the mid-circuit measurements in (b) and those affecting terminal measurements in (a) are both identically characterized by syndrome probability vector  $\mathbf{q}$ .

ones with feedforward, leaving the terminal measurements noiseless. Then, PROM can be applied as-is to the circuit to mitigate the readout errors on the mid-circuit measurements (on the qubits  $a_1, a_2, \dots, a_h$ ), which have error channels characterized by  $\mathbf{q}$  by construction. Our analysis of the costs of PROM, as detailed in Result 2 and Appendixes B 1 and B 3, then carry through. In particular, the sampling overhead factor is  $\xi^2$  with respect to the hypothetical noiseless experiment, where  $\xi = \|\alpha\|_1$  for the weights solution  $\alpha$  computed from  $\mathbf{q}$ . Likewise, the classical processing costs follow Table I and depend on assumptions made about the structure of  $\mathbf{q}$ . These are equivalent to the costs of conventional terminal readout error mitigation as described in Appendix B 5 a, assuming symmetrization (bit-flip averaging) and the Walsh-Hadamard trick to computing  $\alpha$  are used there as well.

### c. Integrating terminal readout error mitigation into the PROM protocol

In the previous subsection we established that our formulation of PROM is sufficiently general to treat readout error mitigation on terminal measurements. The proof lies in restructuring a circuit with noisy terminal measurements into an equivalent one with noisy mid-circuit measurements but noiseless terminal measurements, which can be used to program PROM. This approach is viable and will work correctly.

Here we present an alternative approach which performs readout error mitigation for both mid-circuit and terminal measurements, and which does not involve modifications of the circuit used to program PROM but rather extends the protocol to natively accommodate terminal measurements. This modified algorithm is likely more convenient for

implementation in experiments. Operationally, we combine the PROM protocol (Algorithm 2) and the conventional terminal readout error mitigation procedure laid out in Appendix B 5 a. We consider the same general setting of measuring observables as in Appendix B 5 a, namely that a linear transform matrix  $A$  defines the calculation of expectation values from observed measurement outcome statistics. On top of the arbitrary mid-circuit measurements and feedforward present on the circuit, we suppose a subset of  $h \leq n$  qubits of the circuit are terminally measured in the computational basis [see Fig. 8(a) for circuit diagram], and the confusion probability vector  $\mathbf{q}^{\text{term}}$  characterizes the readout error channels impacting those terminal measurements. Let  $\alpha^{\text{term}}$  be the corresponding weights computed from  $\mathbf{q}^{\text{term}}$  using the same formula as in Result 4. Then we describe this protocol as follows [see Fig. 8(b) for circuit schematic]:

## APPENDIX C: TECHNICAL DETAILS ON HARDWARE DEMONSTRATION AND BENCHMARKING

Here we provide implementation and technical details on our experiments on quantum hardware, as presented in Sec. V of the main text.

### 1. Quantum hardware

We utilized IBM transmon-based superconducting quantum devices in our experiments. These included 27-qubit devices  $\{ibm\_algiers, ibmq\_kolkata, ibm\_mumbai\}$  hosting Falcon-family processors, and 127-qubit devices  $\{ibm\_osaka, ibm\_kyoto\}$  hosting Eagle-family processors. The basis gate sets of all devices comprise 1-qubit gates  $\{X, \sqrt{X}, RZ\}$ , with the  $\sigma^z$ -rotation (RZ) gate implemented

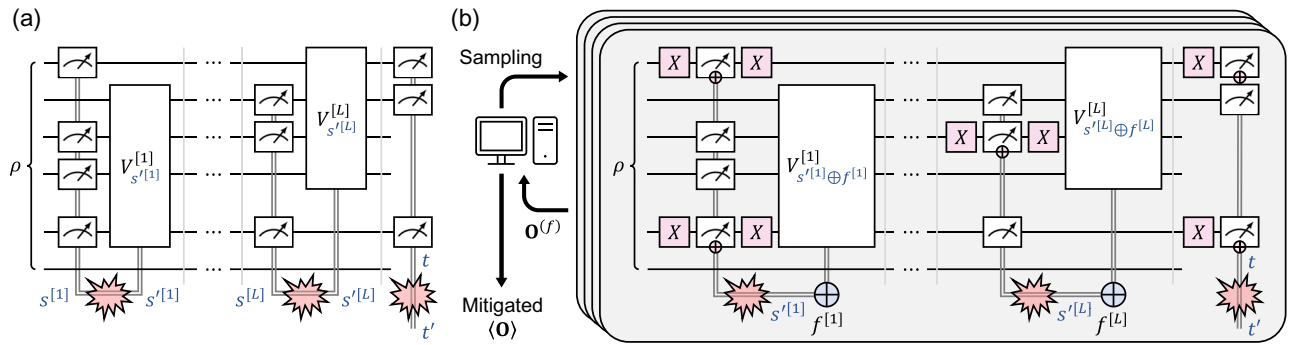


FIG. 8. Probabilistic readout error mitigation for multiple feedforward layers and terminal measurements. (a) Structure of the quantum circuit considered, comprising state preparation,  $L$  layers of mid-circuit measurements and feedforward, and terminal measurements on a subset of (or all) qubits. From shot to shot, in layer  $l$ , the correct measurement outcome (bitstring)  $s^{[l]}$  is corrupted into  $s'^{[l]}$  by readout errors, and the reported outcome  $s'^{[l]}$  determines a feedforward unitary  $V_{s'^{[l]}}^{[l]}$  that is applied. On the terminal measurements, the correct measurement outcome  $t$  is corrupted into  $t'$  by readout errors. (b) Schematic of the readout error mitigation protocol (Algorithm 3). Bit-flip averaging is applied by randomly twirling all measurements with quantum  $X$  gates and classical bit-flips on the outcome; for terminal measurements the trailing  $X$  gates are not necessary. In layer  $l$  the XOR between the reported measurement outcome and a bitmask  $f^{[l]}$  is used for feedforward. The bitmasks  $f = f^{[1]}f^{[2]} \dots f^{[L]}$  are sampled over a suitable probability distribution.

ALGORITHM 3. Probabilistic readout error mitigation for multiple feedforward layers with integrated mitigation of terminal readout error.

**Inputs:** Circuit  $\mathcal{P}$  containing  $L$  layers of mid-circuit measurements and feedforward. Probability vector  $\mathbf{q}$  characterizing readout error channels on mid-circuit measurements. Probability vector  $\mathbf{q}^{\text{term}}$  characterizing readout error channels on terminal measurements. Linear transform matrix  $A$  for recovering expectation values of observables [see Eq. (B47)]. Number of shots  $N'$  to be executed.

**Output:** Estimate  $\langle \hat{\mathbf{O}} \rangle$  of expectation values.

We construct a bit-flip-averaged version of  $\mathcal{P}$ , and appropriate weight vectors  $\alpha$  and  $\alpha^{\text{term}}$  from  $\mathbf{q}$  and  $\mathbf{q}^{\text{term}}$  respectively (using Result 4). For shot  $k = 1$  to  $N'$ , we

1. Sample a bitmask  $f = f^{[1]}f^{[2]} \dots f^{[L]} \in \mathcal{S}$  with probability  $|\alpha_f|/\xi$ .
2. For layers  $l = 1, \dots, L$ , execute the mid-circuit measurements in layer  $l$  with bitmask  $f^{[l]}$  applied to the feedforward.
3. Perform the terminal measurements to obtain an outcome bitstring  $t'$ .
4. Record the observable sample  $\mathbf{o}^{(k)} = A\alpha_{\oplus t'}^{\text{term}}$ , where  $\alpha_{\oplus y}^{\text{term}}$  is  $\alpha^{\text{term}}$  with entries index-shifted by  $y$ , that is,  $(\alpha_{\oplus y}^{\text{term}})_x = \alpha_{x \oplus y}^{\text{term}}$  for all indices  $x$ .

Output the estimate  $\langle \hat{\mathbf{O}} \rangle = \xi \cdot \frac{1}{N'} \sum_{k=1}^{N'} \mathbf{o}^{(k)}$ .

virtually in hardware via framechanges at zero error and duration [99], and 2-qubit gate CX for Falcon processors and echoed cross-resonance (ECR) for Eagle processors [100]. The CX and ECR gates are equivalent up to 1-qubit rotations. We constructed all experiment circuits using arbitrary 1-qubit and CX gates, as shown in Figs. 4(a), 5(a), and 6(a) of the main text and Fig. 9, and transpiled to the native gate set of each device before execution.

Typical performance metrics of the devices, such as gate errors, gate times, readout errors, readout times, relaxation  $T_1$  and dephasing  $T_2$  times, are provided in Table III.

## 2. Readout error mitigation for terminal measurements

### a. Background

In addition to implementing the proposed PROM in our experiments, we utilized also a standard readout error mitigation scheme for terminal measurements—i.e., measurements for the observables of interest at the end of the quantum circuits—to correct for readout errors on those measurements. This improves accuracy of results on the measured observables, which would otherwise be distorted

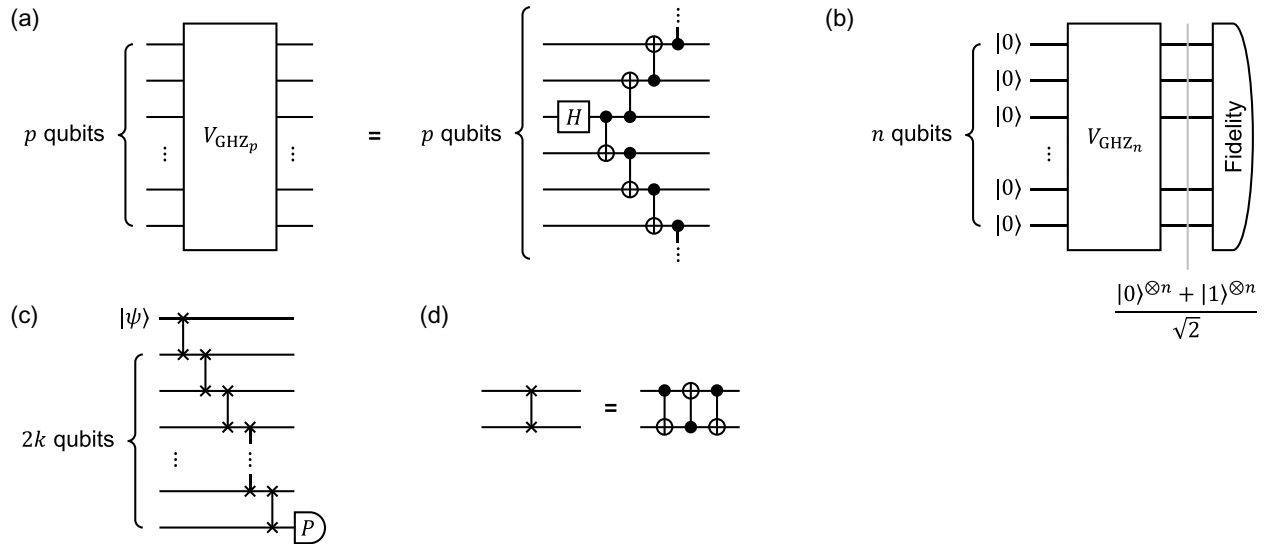


FIG. 9. Additional circuit diagrams. (a) Structure of the  $V_{\text{GHZ}_p}$  block acting on  $p$  qubits, which unitarily prepares a  $p$ -qubit GHZ state on an  $|0\rangle^{\otimes p}$  input. A Hadamard gate is applied on a qubit, and a fan-out of CX gates spanning all  $p$  qubits follow. Thus  $p - 1$  CX gates are used, with depth  $\lceil p/2 \rceil$ . (b) Unitary GHZ state preparation circuit as used in experiments (Sec. VB of the main text) for comparison. (c) Unitary circuit that transports an input qubit state  $|\psi\rangle$  a distance of  $2k$  qubits, as used in staged quantum state teleportation experiments (Sec. VC of the main text) for comparison.  $P \in \{X, Y, Z\}$  denotes a Pauli-basis terminal measurement. (d) Standard decomposition of a SWAP gate into 3 CX gates, which is optimal.

TABLE III. Performance characteristics of quantum devices used in experiments. For quantities that vary between qubits on the same device, we provide 10th and 90th percentile and median values evaluated over all qubits on the device. Characteristics summarized here are based on routine ( $\sim$ daily) calibration data of the machines, pulled over a duration of two weeks in the middle of experiments.

| Device              | Qubits | 2Q gate | 1Q gate error ( $\times 10^{-4}$ ) |      |      | 2Q gate error ( $\times 10^{-3}$ ) |      |      | Readout error ( $\times 10^{-2}$ ) |      |      |
|---------------------|--------|---------|------------------------------------|------|------|------------------------------------|------|------|------------------------------------|------|------|
|                     |        |         | P10                                | P90  | Med. | P10                                | P90  | Med. | P10                                | P90  | Med. |
| <i>ibm_osaka</i>    | 127    | ECR     | 1.23                               | 12.5 | 2.41 | 4.03                               | 19.3 | 7.15 | 0.65                               | 8.94 | 1.99 |
| <i>ibm_kyoto</i>    | 127    | ECR     | 1.46                               | 10.5 | 2.68 | 4.56                               | 19.4 | 8.01 | 0.61                               | 8.15 | 1.51 |
| <i>ibm_algiers</i>  | 27     | CX      | 1.83                               | 8.23 | 2.84 | 5.88                               | 15.3 | 8.90 | 0.59                               | 3.59 | 1.01 |
| <i>ibmq_kolkata</i> | 27     | CX      | 1.90                               | 4.75 | 2.55 | 4.96                               | 17.2 | 8.62 | 0.56                               | 14.5 | 1.23 |
| <i>ibmq_mumbai</i>  | 27     | CX      | 1.58                               | 3.75 | 2.08 | 5.66                               | 11.5 | 7.52 | 1.24                               | 5.87 | 1.70 |

| Device              | 1Q Gate time (ns) | Readout time (ns) | 2Q gate times (ns) |     |      | $T_1$ ( $\mu$ s) |     |      | $T_2$ ( $\mu$ s) |     |      |
|---------------------|-------------------|-------------------|--------------------|-----|------|------------------|-----|------|------------------|-----|------|
|                     |                   |                   | P10                | P90 | Med. | P10              | P90 | Med. | P10              | P90 | Med. |
| <i>ibm_osaka</i>    | 60                | 1400              | 660                | 660 | 660  | 127              | 396 | 275  | 19.8             | 322 | 151  |
| <i>ibm_kyoto</i>    | 60                | 1400              | 660                | 660 | 660  | 124              | 320 | 223  | 28.2             | 237 | 115  |
| <i>ibm_algiers</i>  | 35.6              | 910               | 263                | 622 | 370  | 93.1             | 184 | 136  | 44.3             | 252 | 144  |
| <i>ibmq_kolkata</i> | 35.6              | 640               | 391                | 556 | 462  | 58.5             | 139 | 96.4 | 22.8             | 119 | 63.8 |
| <i>ibmq_mumbai</i>  | 35.6              | 3513              | 277                | 640 | 405  | 54.6             | 130 | 108  | 58.4             | 251 | 149  |

by readout noise, as is appropriate for our purpose of investigating the effectiveness of our PROM protocol.

As described in Secs. III and IV of the main text, PROM targets readout errors for mid-circuit measurements and feedforward, and is not precisely intended for terminal measurements. The protocol nonetheless works when applied, for terminal measurements are a special case of mid-circuit ones with no further quantum operations, but is not as resource-efficient as more restrictive schemes specialized for terminal measurements. In practice, it is best to employ PROM in conjunction with a separate readout error mitigation scheme for terminal measurements, as was done here.

To set the stage and for completeness, we first describe a general procedure in Appendix C 2 b without symmetrization. Then we describe in particular the procedure when readout channels are symmetrized, as in our present setting, and specific implementation details in our experiments in Appendix C 2 c.

### b. Standard procedure without symmetrization of readout error channels

We first summarize a standard prescription for terminal readout error mitigation [27–29]. We consider the confusion matrix  $M$ , as described in Sec. II of the main text and Appendix A 2, over  $m$  qubits. Recall that each entry  $M_{ts}$  records the probability of the quantum device reporting measurement outcome  $t \in \mathcal{S} = \{0, 1\}^m$  when the true outcome is  $s \in \mathcal{S}$ . Prior or alongside the experiment, calibration circuits can be run to estimate  $M$ . In particular, in a calibration quantum circuit, the register of  $m$  qubits is prepared in computational basis state  $|s\rangle$ , which requires simply the application of  $X$  gates on the initial state  $|0\rangle^{\otimes m}$  on appropriate qubits, and then measured

in the computational basis. The probability of observing measurement bitstring  $t$  gives  $M_{ts}$  by definition. Accordingly  $|\mathcal{S}| = 2^m$  distinct quantum circuits need to be run, to enumerate over all  $s \in \mathcal{S}$ .

Given an  $M$ , linear inversion can be performed to mitigate measured experiment counts. In particular, suppose an experiment yielded counts  $\mathbf{c}$ , where entry  $c_s$  records the number of shots of the quantum circuit that produced terminal measurement outcome  $s \in \mathcal{S}$ . Then  $\mathbf{c}$  and the true counts  $\mathbf{c}'$  that would hypothetically be observed without readout errors are related by

$$c_t = \sum_{s \in \mathcal{S}} M_{ts} c'_s \iff \mathbf{c} = M \mathbf{c}', \quad (\text{C1})$$

as mentioned in Sec. II of the main text, and accordingly the true counts can be approximately recovered as

$$\mathbf{c}' = M^{-1} \mathbf{c}, \quad (\text{C2})$$

where  $M^{-1}$  is numerically computed. Such a linear inversion does not guarantee that the recovered  $\mathbf{c}'$  is non-negative, as would be expected for a physical counts vector. In cases where non-negativity is to be enforced, an additional mapping step [101] can be performed to find the closest non-negative counts distribution to the recovered  $\mathbf{c}'$ , or alternatively a least-squares fit with non-negative constraints (and possibly other physicality constraints) can be used in place of a direct numerical inverse [29, 93, 102, 103].

The need to execute  $|\mathcal{S}| = 2^m$  quantum circuits to calibrate  $M$ , which is of size  $|\mathcal{S}| \times |\mathcal{S}| = 2^m \times 2^m$ , and the subsequent classical computation of its inverse, becomes problematic in resource requirements at large  $m$ , say exceeding  $\sim 14$  qubits. For this reason some studies

(e.g., Refs. [102,104–107]) employ a tensored readout error mitigation scheme, wherein the full register of  $m$  qubits is partitioned into  $p$  subregisters comprising  $m^{(1)}, m^{(2)}, \dots, m^{(p)}$  qubits each, and one assumes that there are no correlations in readout errors between measurements in different subregisters. This is precisely the setting described in Appendix A 4 a. In particular, in such a setting the confusion matrix factorizes,  $M = M^{(1)} \otimes M^{(2)} \otimes \dots \otimes M^{(p)}$ , and  $M^{(\ell)}$  for each subregister can be calibrated separately, each requiring  $2^{m_\ell}$  calibration quantum circuits. As the qubits in the calibration circuits for different  $\ell \in [p]$  are disjoint, the circuits can in fact be merged breadthwise, thus requiring  $2^{\max(m_1, m_2, \dots, m_p)}$  calibration circuits in total instead of  $2^{m_1} + 2^{m_2} + \dots + 2^{m_p}$ . The inverses of  $M^{(\ell)}$ , each a  $2^{m_\ell} \times 2^{m_\ell}$  matrix, can also be computed separately, since  $M^{-1} = [M^{(1)}]^{-1} \otimes [M^{(2)}]^{-1} \otimes \dots \otimes [M^{(p)}]^{-1}$ . In practice, it is not necessary to explicitly construct  $M^{-1}$ , which is expensive. Rather one simply applies each  $[M^{(\ell)}]^{-1}$  to the relevant sectors of  $\mathbf{c}$ .

In certain scenarios, one may have calibrated an  $M$  covering measurements on a larger set of qubits than that to be mitigated. The larger than needed  $M$  can be used after marginalizing out the unused measurements. We label the calibrated qubits in  $M$  by  $\mathcal{X} = [m]$ , and we retain the notation  $\mathcal{S} = \{0, 1\}^m$  for the set of possible measurement outcomes on the  $m$  measurements. Let the terminally measured qubits present on the quantum circuit be  $\mathcal{X}' \subseteq \mathcal{X}$ . The circuit contains  $m' = |\mathcal{X}'|$  measurements. Then the confusion matrix characterizing readout errors on those  $m'$  measurements, to be used for terminal readout error mitigation on the circuit, is

$$M'_{t's'} = \frac{1}{2^{m-m'}} \sum_{\substack{t \in \mathcal{S} \\ t_{\mathcal{X}'=t'}}} \sum_{\substack{s \in \mathcal{S} \\ s_{\mathcal{X}'=s'}}} M_{ts}, \quad (\text{C3})$$

where  $s_{\mathcal{X}'}$  denotes the substring of  $s$  formed by entries at indices  $\mathcal{X}'$ . Thus, in principle, in an experiment with multiple quantum circuits containing terminal measurements on different sets of qubits, Eq. (C3) allows one to calibrate a larger  $M$  over the union of measured qubits across all circuits; then for each circuit the marginalized  $M'$  is used for error mitigation. In practice, error characteristics of a quantum device drifts over time, thus recalibrations are typically required for long experiments.

### c. Procedure with symmetrization of readout error channels

Applying BFA [25,26] symmetrizes the readout error channels, which are then characterized by the confusion vector  $\mathbf{q}$  and associated symmetrized confusion matrix  $Q$ , as described in the main text (Secs. II and III B) and in Appendix A 2. Recall that each entry  $q_s$  records the probability of a readout error syndrome  $s \in \mathcal{S}$  on the

quantum device, that is, the probability of reporting outcome  $a \oplus t \in \mathcal{S}$  when the true outcome is some  $a \in \mathcal{S}$ . By definition,  $Q_{sf} = q_{s \oplus f}$  for  $s \in \mathcal{S}$  and  $f \in \mathcal{S}$ . Then in an experiment, the observed counts  $\mathbf{c}$  on terminal measurements and the true counts  $\mathbf{c}'$  that would hypothetically be observed without readout errors are related by

$$c_t = \sum_{a \in \mathcal{S}} q_{a \oplus t} c'_a \iff \mathbf{c} = \mathbf{c}' * \mathbf{q}, \quad (\text{C4})$$

where  $*$  denotes convolution (see Appendix A 1 for definition), or equivalently

$$c_t = \sum_{a \in \mathcal{S}} Q_{at} c'_a \iff \mathbf{c} = Q \mathbf{c}'. \quad (\text{C5})$$

Accordingly the true counts are recovered as

$$\mathbf{c}' = Q^{-1} \mathbf{c}. \quad (\text{C6})$$

In our implementation, we additionally enforce non-negativity of  $\mathbf{c}'$  by mapping to the closest non-negative counts distribution [101] in Euclidean norm, a method that was also outlined in Appendix C 2 b above.

To calibrate  $\mathbf{q}$ , a single calibration circuit preparing the  $|0\rangle^{\otimes m}$  computational basis state and then measuring all qubits (with BFA twirling applied) suffices [25,26]. The probability of observing a measurement outcome  $s \in \mathcal{S}$  gives  $q_s$  by definition. This is in contrast to Appendix C 2 b, which was without symmetrization and required  $2^m$  calibration circuits preparing all  $2^m$  computational basis states. The prior discussion on tensored calibration and terminal readout error mitigation schemes in Appendix C 2 b directly carries over but with  $Q$  replacing  $M$ .

In our experiments, we calibrated  $\mathbf{q}$  and thus  $Q$  on the measured qubits present on the quantum circuits. When  $\mathbf{q}$  and  $Q$  on subsets of qubits were required, for example when the mid-circuit measurements supporting feedforward or terminal measurements occurred on a subset of qubits, or when layerwise or fully tensored assumptions were to be imposed on our PROM protocol (as in Sec. V of the main text), we performed marginalization on the larger than needed  $\mathbf{q}$  and  $Q$ , identical to Eq. (C3) but with  $Q$  replacing  $M$ . As the error characteristics of quantum devices drift over time, we interleaved multiple rounds of calibration circuits with experiment circuits throughout our experiments.

## 3. Dynamical decoupling

Inevitably, in quantum circuits of practical interest there exist periods of time during which certain qubits are idle as they wait for operations on other qubits to be completed. For example, sequences of gates may need to be executed on other qubits before a qubit is itself involved in 2-qubit operations with those qubits, or the qubit could be waiting

for mid-circuit measurements for feedforward operations to follow. These idle periods can be observed in the circuit diagrams of Figs. 4(a), 5(a), and 6(a).

During idle periods, qubits are still subject to decoherence (e.g., thermal relaxation and dephasing) noise. Dynamical decoupling sequences [108], which run through an alternating schedule of basis changes that overall multiply to the identity, can be implemented to suppress the effect of such decoherence. At a basic level, dynamical decoupling refocuses the dephasing noise experienced by the qubits, such that the accumulated contributions cancel to a certain order at the end of the sequence.

In our experiments, we employ a combination of commonly used XY4 and XY8 dynamical decoupling sequences [29,109,110], the former inserted in idle periods approximately up to the duration of a CX gate on the device ( $\sim 400$  ns) and the latter for longer idle durations, such as those induced by waiting for mid-circuit measurements. This selection is not specifically fine-tuned, and we do not expect device performance to differ significantly for minor modifications to the exact sequences. We remark that there exist more sophisticated dynamical decoupling sequences that cancel noise to higher orders or offer other theoretical advantages [111,112]; however, empirical results [113] indicate that simpler sequences can perform better in practice, owing in part to the presence of errors on the gates introduced by the dynamical decoupling sequence, and the fact that noise backgrounds often differ from theoretical assumptions (i.e., in form and temporal uniformity).

While dynamical decoupling has been broadly employed to great effect on quantum devices (see, in addition to the references above, e.g., Refs. [114–116]), we comment on a couple of limitations that are relevant in the present context. First, in reality perfect cancellation of noise incurred during idle periods does not happen, as the noise background is not uniform over time. This may be especially so when measurements are being performed on the device, which are generally considered nonsmooth processes. Second, the superconducting quantum processors we utilized do not yet support pulse operations during the feedforward latency [73], that is, during the short period of time ( $\sim 200$  ns) that the mid-circuit measurement outcomes are processed by classical control hardware and feedforward operations decided. Therefore in our experiments, dynamical decoupling sequences are not applied during these latency periods. These factors in general limit the effectiveness of dynamical decoupling on circuits with feedforward. Engineering improved dynamical decoupling methods for such contexts is an active area of research [73,117].

#### 4. Measuring Pauli expectation values

As a preliminary to subsequent discussions, we describe explicitly the primitive used to measure Pauli expectation values in our experiments. As with most quantum devices, the superconducting processors we used perform all measurements in the computational basis, that is, the Z basis. Hence to measure in other Pauli bases a change of basis is

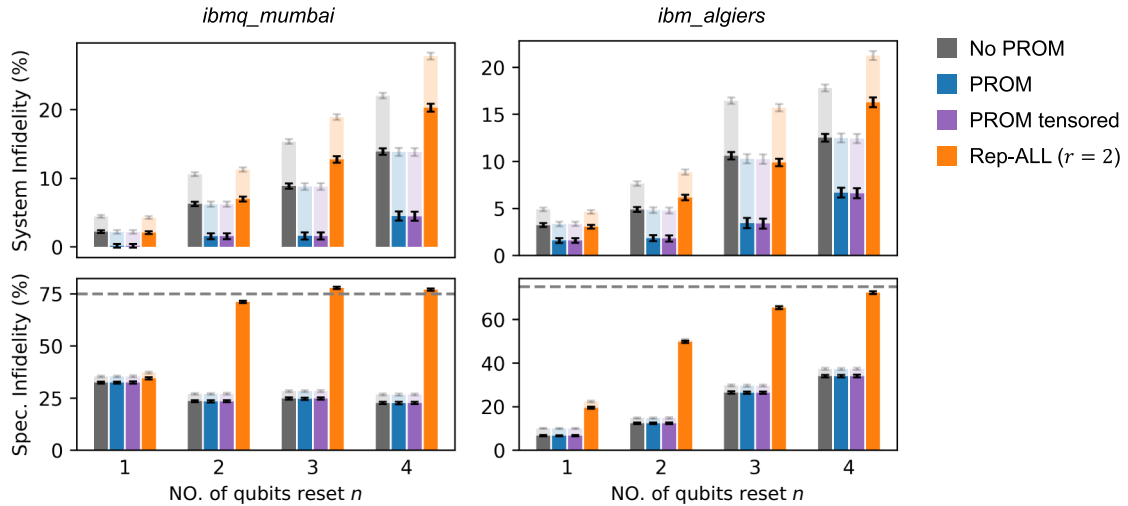


FIG. 10. Additional dynamic qubit reset experiment results. Infidelities of system qubits after dynamic resets against the ideal  $|0\rangle^{\otimes n}$  post-reset state, and infidelities of spectator qubits against  $|0\rangle^{\otimes 2}$  after witnessing reset events conditioned on system qubits being correctly reset, comparing probabilistic readout error mitigation, probabilistic readout error mitigation assuming tensored readout channels, Rep-ALL at  $r = 2$ , and without readout error mitigation for mid-circuit measurements and feedforward. Dashed line demarcates 75% infidelity expected for a maximally mixed spectator state. In all plots, solid bars are with terminal readout error mitigation; translucent bars without. Experiment performed on superconducting quantum device *ibmq\_mumbai* and *ibmq\_algiers* as labeled, complementing results on *ibmq\_kolkata* as presented in Figs. 4(b) and 4(c) of the main text. Error bars are standard deviations across 1000 trials per qubit chain, each with 10 000 shots; data averaged over  $\sim 8$  randomly chosen qubit chains.

required. To measure the expectation value of an  $n$ -qubit Pauli operator  $P \in \{\mathbb{I}, X, Y, Z\}^n$ , the following operation comprising single-qubit rotations is performed,

$$B_P = \bigotimes_{k=1}^n U_{P_k}, \quad U_g = \begin{cases} \mathbb{I}, & g \in \{\mathbb{I}, Z\}, \\ H, & g = X, \\ HS^\dagger, & g = Y, \end{cases} \quad (C7)$$

before measuring all qubits in the support of  $P$  (i.e., qubits on which  $P$  acts nontrivially) in the computational basis. In an experiment shot, these measurements yield a bitstring  $b$ . The measurement outcome of  $P$  yielded by that shot is  $+1$  when  $wt(b)$  is even and  $-1$  when  $wt(b)$  is odd.

### 5. Measuring fidelity of GHZ states via stabilizer expectation values

In Sec. VB of the main text, we are interested in measuring the fidelity  $\mathcal{F}$  of the prepared  $n$ -qubit state at the end of the quantum circuit  $\rho_c$  against the ideal GHZ state  $\rho_{\text{GHZ}}$ ,

$$\mathcal{F} = \left[ \text{tr}(\sqrt{\sqrt{\rho_{\text{GHZ}}} \rho_c \sqrt{\rho_{\text{GHZ}}}}) \right]^2 = \langle \psi_{\text{GHZ}} | \rho_c | \psi_{\text{GHZ}} \rangle, \quad (C8)$$

$$\rho_{\text{GHZ}} = |\psi_{\text{GHZ}}\rangle \langle \psi_{\text{GHZ}}|, \quad |\psi_{\text{GHZ}}\rangle = \frac{|0\rangle^{\otimes n} + |1\rangle^{\otimes n}}{\sqrt{2}}.$$

In general, the fidelity between two quantum states can be measured through the SWAP test [118,119] and related variants. But the size of the circuit required, in particular, the need to prepare both quantum states in parallel qubit registers and the hugely expensive controlled-SWAP operation between both registers, makes this direct measurement

impractical on present hardware. The need to prepare an ideal GHZ state to SWAP against is also fundamentally problematic in our context.

In our work, we used a method based on measuring Pauli expectation values to calculate the fidelity  $\mathcal{F}$ , as introduced in Refs. [120,121] and used, for example, in Refs. [17,122] which we follow. We refer readers to these references for the broader theory, which in fact applies for more general quantum states, not only the GHZ state. In the present context, the fidelity  $\mathcal{F}$  can be decomposed as

$$\mathcal{F} = \frac{1}{2^n} \sum_{P \in \mathcal{P}} \langle P \rangle_{\rho_c} = \frac{1}{2^n} \sum_{P \in \mathcal{P}} \text{tr}[P \rho_c], \quad (C9)$$

where  $\mathcal{P}$  are the stabilizers of the GHZ state, comprising exactly  $2^n$  elements. That is, each element  $P \in \mathcal{P}$  satisfies  $P |\psi_{\text{GHZ}}\rangle = |\psi_{\text{GHZ}}\rangle$  and is a Pauli operator over  $n$  qubits. Explicitly, the set of stabilizers can be written as

$$\mathcal{P} = \{Z^{(x)} : x \in \{0, 1\}^n, wt(x) \text{ even}\} \cup \{(-1)^{wt(x)/2} X^{(1 \oplus x)} Y^{(x)} : x \in \{0, 1\}^n, wt(x) \text{ even}\}. \quad (C10)$$

To measure  $\langle P \rangle$  for each  $P \in \mathcal{P}$ , we use the basic method described in Appendix C4, accounting for the possible minus sign of  $P$  thereafter. We comment that a random sampling of  $P \in \mathcal{P}$  suffices to estimate the fidelity  $\mathcal{F}$  up to sampling error [17,120–122], but we were able to measure all stabilizers in our experiments, which also eliminated potential considerations about sampling noise in the interpretation of data.

A further reduction in quantum resources, in particular the number of quantum circuits to be executed, can

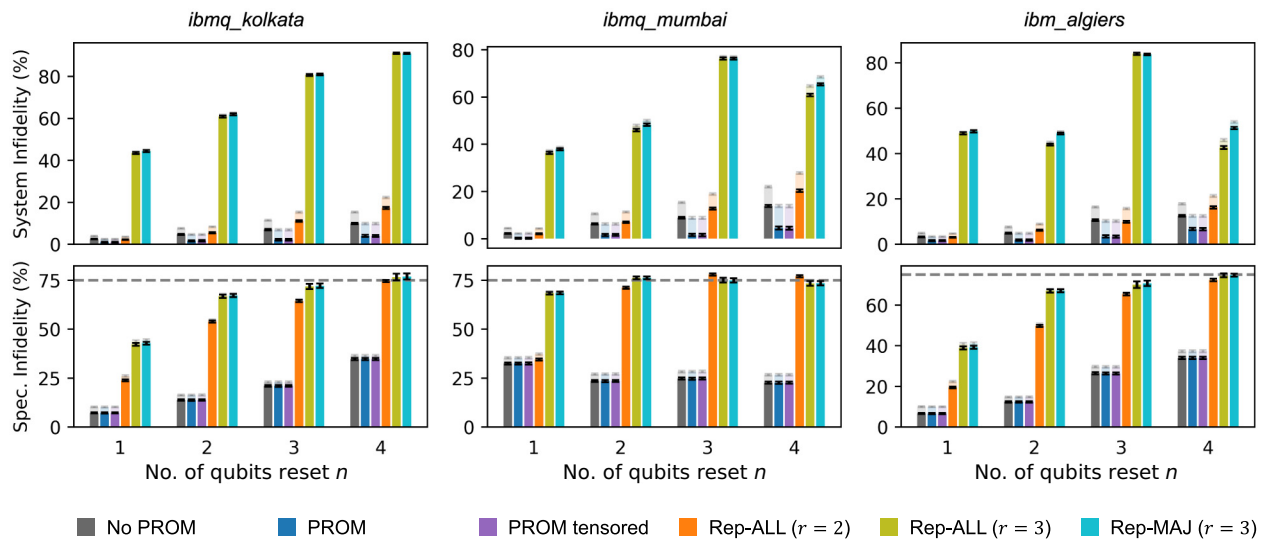


FIG. 11. Additional dynamic qubit reset experiment results. Zoomed-out version of Figs. 4(b) and 4(c) of the main text and Fig. 10 that includes repetition-based readout error mitigation methods Rep-ALL and Rep-MAJ at  $r = 3$ .

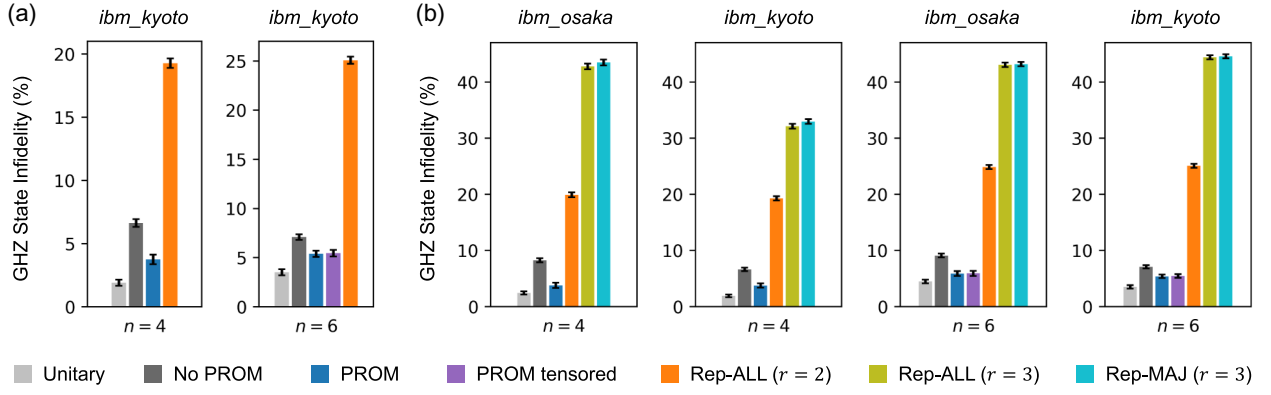


FIG. 12. Additional GHZ state preparation experiment results. (a) Infidelities of the prepared state against the ideal  $n$ -qubit GHZ state, comparing probabilistic readout error mitigation, probabilistic readout error mitigation assuming tensored readout channels, Rep-ALL at  $r=2$ , and without readout error mitigation for mid-circuit measurements and feedforward. Also shown are infidelities on a unitary circuit without feedforward ( $V_{\text{GHZ}_n}$ ) for comparison. Data acquired on the *ibmq\_kyoto* superconducting quantum device, complementing results on *ibmq\_osaka* as presented in Fig. 5(b) of the main text. (b) Zoomed-out version of Fig. 5(b) of the main text and (a) that includes repetition-based readout error mitigation methods Rep-ALL and Rep-MAJ at  $r=3$ . Readout error mitigation for terminal measurements was applied for all data. The  $n=4, 6$  experiments employ  $(b, p) = (2, 1), (3, 1)$  circuit structures respectively. Error bars are standard deviations across 100 trials each with 10 000 shots.

be achieved by simultaneously measuring a maximally sized subset of commuting Pauli operators on each circuit, for instance as employed in Refs. [106,123] in varying contexts. The basis change circuit components for such measurements can be solved for efficiently through a stabilizer tableau approach. But as these basis rotations are generally entangling and are no longer independent 1-qubit rotations, they require a number of CX gates to implement (generically scaling with the number of qubits). As our objective is to measure the fidelity of the prepared state against the GHZ state, we chose to avoid this technique to avoid introducing additional error in the measurements.

Last, we remark that there exist alternative methods to measure GHZ state fidelities. A commonly used method is that of parity oscillations, as used, for example, in Refs. [124–129]. Applied toward our present purpose, we highlight that care should be taken to suitably account for the effect of relative phases in the prepared state, that is, that a phase  $\phi \neq 0$  may be present in the prepared state  $(|0\rangle^{\otimes n} + e^{i\phi}|1\rangle^{\otimes n})/\sqrt{2}$ . A third method, multiple quantum coherences [130,131], has also been demonstrated in prior literature. We chose the stabilizer measurement approach in our experiment as it is straightforward and robust both theoretically and in implementation on the quantum device.

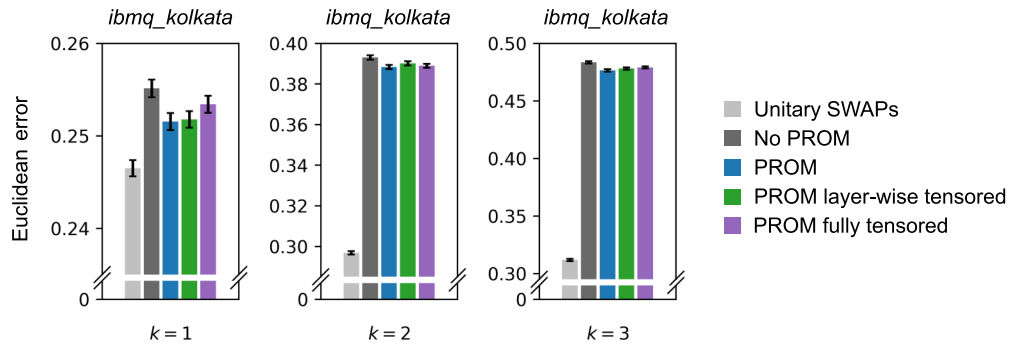


FIG. 13. Additional quantum state teleportation experiment results. Euclidean errors  $\Delta = (\delta X^2 + \delta Y^2 + \delta Z^2)^{1/2}$  after  $k \in \{1, 2, 3\}$  teleportations, comparing probabilistic readout error mitigation, probabilistic readout error mitigation assuming layerwise and fully tensored readout channels, and without readout error mitigation for mid-circuit measurements and feedforward. Also shown are errors on a unitary circuit with SWAP gates that transport the input state  $|\psi\rangle$  across the same qubits. Readout error mitigation for terminal measurements was applied for all data. Experiments performed on the superconducting quantum device *ibmq\_kolkata*, complementing results on *ibmq\_osaka* as presented in Fig. 6(b) of the main text. Error bars are standard deviations across 100 trials each with 1 000 000 shots.

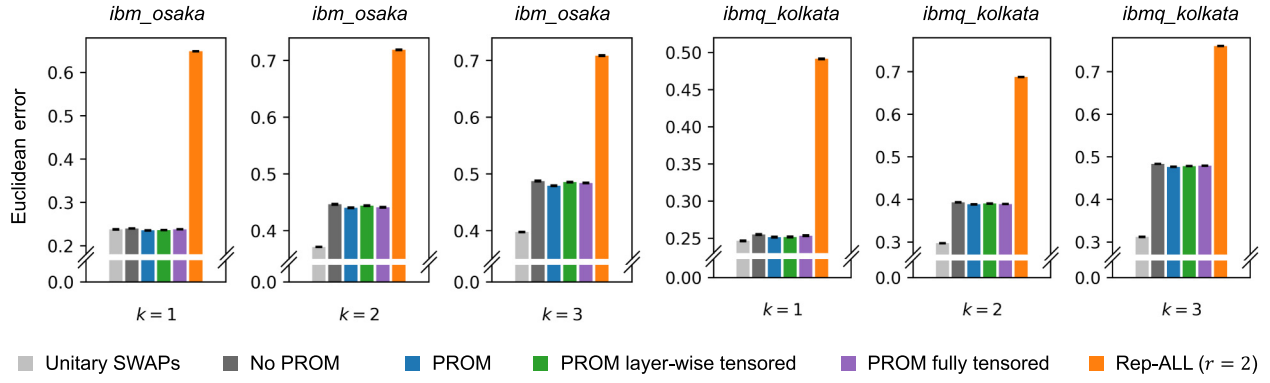


FIG. 14. Additional quantum state teleportation experiment results. Zoomed-out version of Fig. 6(b) of the main text and Fig. 13 that includes repetition-based readout error mitigation method Rep-ALL at  $r = 2$ .

TABLE IV. Readout error mitigation overheads in experiments. We summarize the probabilistic readout error mitigation sampling overhead factors  $\xi^2$  for the various hardware experiments performed in this work.

| Dynamic qubit reset experiments   |                         |                |                         |
|-----------------------------------|-------------------------|----------------|-------------------------|
| Experiment                        | No. of qubits reset $n$ | $\xi^2$ (PROM) | $\xi^2$ (PROM tensored) |
| ibmq_kolkata (Main Fig. 4)        | 1                       | 1.06           | 1.06                    |
|                                   | 2                       | 1.14           | 1.14                    |
|                                   | 3                       | 1.23           | 1.23                    |
|                                   | 4                       | 1.30           | 1.30                    |
| ibmq_mumbai (Figs. 10 and 11)     | 1                       | 1.10           | 1.10                    |
|                                   | 2                       | 1.23           | 1.23                    |
|                                   | 3                       | 1.38           | 1.38                    |
|                                   | 4                       | 1.54           | 1.55                    |
| ibmq_algiers (Figs. 10 and 11)    | 1                       | 1.08           | 1.08                    |
|                                   | 2                       | 1.14           | 1.14                    |
|                                   | 3                       | 1.37           | 1.37                    |
|                                   | 4                       | 1.32           | 1.32                    |
| GHZ state preparation experiments |                         |                |                         |
| Experiment                        | GHZ system size $n$     | $\xi^2$ (PROM) | $\xi^2$ (PROM tensored) |
| ibmq_osaka (Main Fig. 5)          | 4                       | 1.37           | -                       |
|                                   | 6                       | 1.35           | 1.36                    |
|                                   | 10                      | 1.33           | -                       |
|                                   | 12                      | 1.76           | 1.76                    |
| ibmq_kyoto (Fig. 12)              | 4                       | 1.21           | -                       |
|                                   | 6                       | 1.17           | 1.17                    |
|                                   | 10                      | 1.71           | -                       |
|                                   | 12                      | 1.74           | 1.74                    |
| Staged teleportation experiments  |                         |                |                         |
| Experiment                        | No. of teleports $k$    | $\xi^2$ (PROM) | $\xi^2$ (PROM tensored) |
| ibmq_osaka (Main Fig. 6)          | 1                       | 1.60           | 1.61                    |
|                                   | 2                       | 1.79           | 1.80                    |
|                                   | 3                       | 1.80           | 1.82                    |
| ibmq_kolkata (Figs. 13 and 14)    | 1                       | 1.35           | 1.35                    |
|                                   | 2                       | 1.36           | 1.36                    |
|                                   | 3                       | 1.45           | 1.45                    |

## 6. Additional circuit diagrams

For clarity we provide illustrations of a number of relevant quantum circuit structures and components in Fig. 9.

## APPENDIX D: ADDITIONAL EXPERIMENT RESULTS

### 1. Mid-circuit dynamic qubit reset

As referenced in Sec. V A of the main text, in Fig. 10 we present additional experiment results acquired on two other superconducting quantum devices. The results are qualitatively similar to that shown in Figs. 4(b) and 4(c). In Fig. 11 we include repetition-based readout error mitigation protocols Rep-ALL and Rep-MAJ at  $r = 3$ , which are not at all competitive with PROM or even without readout error mitigation for mid-circuit measurements and feedforward.

### 2. Shallow-depth GHZ state preparation

As referenced in Sec. V B of the main text, in Fig. 12(a) we present additional experiment results acquired on a second superconducting quantum device for GHZ system sizes  $n = 4, 6$ . The results are qualitatively similar to that shown in Fig. 5(b). In Fig. 12(b) we include repetition-based readout error mitigation protocols Rep-ALL and Rep-MAJ at  $r = 3$ , which are not at all competitive with PROM or even without readout error mitigation for mid-circuit measurements and feedforward.

### 3. Staged quantum state teleportation

As referenced in Sec. V C of the main text, in Fig. 13 we present additional experiment results acquired on a second superconducting quantum device. The results are qualitatively similar to that shown in Fig. 6(b). In Fig. 14 we include repetition-based readout error mitigation protocol Rep-ALL at  $r = 2$ , which is not at all competitive with PROM or even without readout error mitigation for mid-circuit measurements and feedforward.

### 4. Overhead factors in experiments

We present in Table IV a summary of the PROM sampling overhead factors  $\xi^2 = \|\alpha\|_1^2$  recorded in experiments. The overhead factors ( $\leq 2$ ) were modest for the circuits we investigated. The observed differences between  $\xi^2$  of general PROM and PROM assuming uncorrelated (tensored) readout error channels are small, expected as readout errors on the devices used in our experiments are not overly correlated—this is consistent with the small observed differences in error mitigation accuracy of general and tensored PROM in the experiment results. The overhead factor  $\xi^2$  is expected to exponentially grow with the number of

mid-circuit measurements mitigated in the circuit; but fluctuations in readout error rates of the devices in experiments introduce a degree of spread in the recorded overheads.

- 
- [1] M. DeCross, E. Chertkov, M. Kohagen, and M. Foss-Feig, Qubit-reuse compilation with mid-circuit measurement and reset, *Phys. Rev. X* **13**, 041057 (2023).
  - [2] C. Piveteau and D. Sutter, Circuit knitting with classical communication, *IEEE Trans. Inf. Theory* **70**, 2734 (2024).
  - [3] A. Carrera Vazquez, C. Tornow, D. Ristè, S. Woerner, M. Takita, and D. J. Egger, Combining quantum processors with real-time classical communication, *Nature* **636**, 75 (2024).
  - [4] D. Gottesman, in *Quantum Information Science and its Contributions to Mathematics, Proceedings of Symposia in Applied Mathematics* (American Mathematical Society, Rhode Island, United States, 2010), Vol. 68, pp. 13–58.
  - [5] R. Acharya, I. Aleiner, R. Allen, T. I. Andersen, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. Atalaya, R. Babbush *et al.*, Suppressing quantum errors by scaling a surface code logical qubit, *Nature* **614**, 676 (2023).
  - [6] V. Sivak, A. Eickbusch, B. Royer, S. Singh, I. Tsioutsios, S. Ganjam, A. Miano, B. Brock, A. Ding, L. Frunzio *et al.*, Real-time quantum error correction beyond break-even, *Nature* **616**, 50 (2023).
  - [7] N. Sundaresan, T. J. Yoder, Y. Kim, M. Li, E. H. Chen, G. Harper, T. Thorbeck, A. W. Cross, A. D. Córcoles, and M. Takita, Demonstrating multi-round subsystem quantum error correction using matching and maximum likelihood decoders, *Nat. Commun.* **14**, 2852 (2023).
  - [8] E. Knill, Quantum computing with realistically noisy devices, *Nature* **434**, 39 (2005).
  - [9] S. Bartolucci, P. Birchall, H. Bombin, H. Cable, C. Dawson, M. Gimeno-Segovia, E. Johnston, K. Kieling, N. Nickerson, M. Pant *et al.*, Fusion-based quantum computation, *Nat. Commun.* **14**, 912 (2023).
  - [10] R. S. Gupta, N. Sundaresan, T. Alexander, C. J. Wood, S. T. Merkel, M. B. Healy, M. Hillenbrand, T. Jochym-O'Connor, J. R. Wootton, T. J. Yoder *et al.*, Encoding a magic state with beyond break-even fidelity, *Nature* **625**, 259 (2024).
  - [11] S. Pirandola, J. Eisert, C. Weedbrook, A. Furusawa, and S. L. Braunstein, Advances in quantum teleportation, *Nat. Photonics* **9**, 641 (2015).
  - [12] Y. Wan, D. Kienzler, S. D. Erickson, K. H. Mayer, T. R. Tan, J. J. Wu, H. M. Vasconcelos, S. Glancy, E. Knill, D. J. Wineland *et al.*, Quantum gate teleportation between separated qubits in a trapped-ion processor, *Science* **364**, 875 (2019).
  - [13] X.-M. Hu, Y. Guo, B.-H. Liu, C.-F. Li, and G.-C. Guo, Progress in quantum teleportation, *Nat. Rev. Phys.* **5**, 339 (2023).
  - [14] B. Rost, L. Del Re, N. Earnest, A. F. Kemper, B. Jones, and J. K. Freericks, Long-time error-mitigating simulation of open quantum systems on near term quantum computers, *npj Quantum Inf.* **11**, 10 (2025).

- [15] J. Han, W. Cai, L. Hu, X. Mu, Y. Ma, Y. Xu, W. Wang, H. Wang, Y. P. Song, C.-L. Zou, and L. Sun, Experimental simulation of open quantum system dynamics via Trotterization, *Phys. Rev. Lett.* **127**, 020504 (2021).
- [16] L. Del Re, B. Rost, M. Foss-Feig, A. F. Kemper, and J. K. Freericks, Robust measurements of  $n$ -point correlation functions of driven-dissipative quantum systems on a digital quantum computer, *Phys. Rev. Lett.* **132**, 100601 (2024).
- [17] E. Bäumer, V. Tripathi, D. S. Wang, P. Rall, E. H. Chen, S. Majumder, A. Seif, and Z. K. Mineev, Efficient long-range entanglement using dynamic circuits, *PRX Quantum* **5**, 030339 (2024).
- [18] G.-Y. Zhu, N. Tantivasadakarn, A. Vishwanath, S. Trebst, and R. Verresen, Nishimori's cat: Stable long-range entanglement from finite-depth unitaries and weak measurements, *Phys. Rev. Lett.* **131**, 200201 (2023).
- [19] N. Tantivasadakarn, R. Verresen, and A. Vishwanath, Shortest route to non-Abelian topological order on a quantum processor, *Phys. Rev. Lett.* **131**, 060405 (2023).
- [20] M. Foss-Feig, A. Tikku, T.-C. Lu, K. Mayer, M. Iqbal, T. M. Gatterman, J. A. Gerber, K. Gilmore, D. Gresh, A. Hankin, N. Hewitt, C. V. Horst, M. Matheny, T. Mengle, B. Neyenhuis, H. Dreyer, D. Hayes, T. H. Hsieh, and I. H. Kim, Experimental demonstration of the advantage of adaptive quantum circuits, [arXiv:2302.03029](https://arxiv.org/abs/2302.03029) [quant-ph].
- [21] K. C. Smith, E. Crane, N. Wiebe, and S. M. Girvin, Deterministic constant-depth preparation of the AKLT state on a quantum processor using fusion measurements, *PRX Quantum* **4**, 020315 (2023).
- [22] M. Iqbal, N. Tantivasadakarn, R. Verresen, S. L. Campbell, J. M. Dreiling, C. Figgatt, J. P. Gaebler, J. Johansen, M. Mills, S. A. Moses *et al.*, Non-Abelian topological order and anyons on a trapped-ion processor, *Nature* **626**, 505 (2024).
- [23] Z. Cai, R. Babbush, S. C. Benjamin, S. Endo, W. J. Huggins, Y. Li, J. R. McClean, and T. E. O'Brien, Quantum error mitigation, *Rev. Mod. Phys.* **95**, 045005 (2023).
- [24] J. Lin, J. J. Wallman, I. Hincks, and R. Laflamme, Independent state and measurement characterization for quantum computers, *Phys. Rev. Res.* **3**, 033285 (2021).
- [25] 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).
- [26] R. Hicks, C. W. Bauer, and B. Nachman, Readout rebalancing for near-term quantum computers, *Phys. Rev. A* **103**, 022407 (2021).
- [27] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, *Nature* **549**, 242 (2017).
- [28] 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).
- [29] P. Jurcevic, A. Javadi-Abhari, L. S. Bishop, I. Lauer, D. F. Bogorin, M. Brink, L. Capelluto, O. Günlük, T. Itoko, N. Kanazawa *et al.*, Demonstration of quantum volume 64 on a superconducting quantum computing system, *Quantum Sci. Technol.* **6**, 025020 (2021).
- [30] E. van den Berg, Z. K. Mineev, A. Kandala, and K. Temme, Probabilistic error cancellation with sparse Pauli–Lindblad models on noisy quantum processors, *Nat. Phys.* **19**, 1116 (2023).
- [31] R. S. Gupta, E. van den Berg, M. Takita, D. Ristè, K. Temme, and A. Kandala, Probabilistic error cancellation for dynamic quantum circuits, *Phys. Rev. A* **109**, 062617 (2024).
- [32] Y. Li and S. C. Benjamin, Efficient variational quantum simulator incorporating active error minimization, *Phys. Rev. X* **7**, 021050 (2017).
- [33] K. Temme, S. Bravyi, and J. M. Gambetta, Error mitigation for short-depth quantum circuits, *Phys. Rev. Lett.* **119**, 180509 (2017).
- [34] T. Giurgica-Tiron, Y. Hindy, R. LaRose, A. Mari, and W. J. Zeng, in *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)* (IEEE, Denver, CO, USA, 2020), pp. 306–316.
- [35] P. Czarnik, A. Arrasmith, P. J. Coles, and L. Cincio, Error mitigation with Clifford quantum-circuit data, *Quantum* **5**, 592 (2021).
- [36] H. Liao, D. S. Wang, I. Sitdikov, C. Salcedo, A. Seif, and Z. K. Mineev, Machine learning for practical quantum error mitigation, *Nat. Mach. Intell.* **6**, 1478 (2024).
- [37] J. J. Wallman and J. Emerson, Noise tailoring for scalable quantum computation via randomized compiling, *Phys. Rev. A* **94**, 052325 (2016).
- [38] 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).
- [39] S.-N. Sun, B. Marinelli, J. M. Koh, Y. Kim, L. B. Nguyen, L. Chen, J. M. Kreikebaum, D. I. Santiago, I. Siddiqi, and A. J. Minnich, Quantum computation of frequency-domain molecular response properties using a three-qubit iToffoli gate, *npj Quantum Inf.* **10**, 55 (2024).
- [40] A. Hashim, A. Carignan-Dugas, L. Chen, C. Jünger, N. Fruitwala, Y. Xu, G. Huang, J. J. Wallman, and I. Siddiqi, Quasiprobabilistic readout correction of midcircuit measurements for adaptive feedback via measurement randomized compiling, *PRX Quantum* **6**, 010307 (2025).
- [41] P. Ivashkov, G. Uchegara, L. Jiang, D. S. Wang, and A. Seif, High-fidelity, multiqubit generalized measurements with dynamic circuits, *PRX Quantum* **5**, 030315 (2024).
- [42] Mid-circuit measurements in other bases can be absorbed into the state preparation of  $\rho$  and the feedforward operations; thus this computational basis assumption can be made without loss of generality.
- [43] J. M. Chow, L. DiCarlo, J. M. Gambetta, A. Nunnenkamp, L. S. Bishop, L. Frunzio, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, Detecting highly entangled states with a joint qubit readout, *Phys. Rev. A* **81**, 062325 (2010).
- [44] Though certain recent protocols consider cases where our knowledge of  $M$  is partial, or indirectly constructs the data in  $M$  to reduce computational costs [132–134].
- [45] N. Ahmed and K. R. Rao, in *Orthogonal Transforms for Digital Signal Processing* (Springer, Berlin, 1975), pp. 99–152..
- [46] T. Beer, Walsh transforms, *Am. J. Phys.* **49**, 466 (1981).

- [47] G. Marsaglia, W. W. Tsang, and J. Wang, Fast generation of discrete random variables, *J. Stat. Softw.* **11**, 1 (2004).
- [48] M. D. Vose, A linear algorithm for generating random numbers with a given distribution, *IEEE Trans. Softw. Eng.* **17**, 972 (1991).
- [49] A. J. Walker, An efficient method for generating discrete random variables with general distributions, *ACM Trans. Math. Softw.* **3**, 253 (1977).
- [50] At least within the timescale over which noise parameters drift on the quantum device.
- [51] Mid-circuit measurements in other bases can be absorbed into the state preparation of  $\rho$  and the feedforward operations of the previous and current layers; thus this computational basis assumption is without loss of generality.
- [52] D. Suter and G. A. Alvarez, Colloquium: Protecting quantum information against environmental noise, *Rev. Mod. Phys.* **88**, 041001 (2016).
- [53] A. A. Clerk, M. H. Devoret, S. M. Girvin, F. Marquardt, and R. J. Schoelkopf, Introduction to quantum noise, measurement, and amplification, *Rev. Mod. Phys.* **82**, 1155 (2010).
- [54] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University, Cambridge, England, 2010).
- [55] D. P. DiVincenzo, The physical implementation of quantum computation, *Fortschr. Phys.* **48**, 771 (2000).
- [56] D. Ristè, J. G. van Leeuwen, H.-S. Ku, K. W. Lehnert, and L. DiCarlo, Initialization by measurement of a superconducting quantum bit circuit, *Phys. Rev. Lett.* **109**, 050507 (2012).
- [57] G. Di Bartolomeo, M. Vischi, T. Feri, A. Bassi, and S. Donadi, Efficient quantum algorithm to simulate open systems through a single environmental qubit, *Phys. Rev. Res.* **6**, 043321 (2024).
- [58] F. Hua, Y. Jin, Y. Chen, S. Vittal, K. Krsulich, L. S. Bishop, J. Lapeyre, A. Javadi-Abhari, and E. Z. Zhang, Exploiting qubit reuse through mid-circuit measurement and reset, [arXiv:2211.01925](https://arxiv.org/abs/2211.01925) [quant-ph].
- [59] A. D. Córcoles, M. Takita, K. Inoue, S. Lekuch, Z. K. Mineev, J. M. Chow, and J. M. Gambetta, Exploiting dynamic quantum circuits in a quantum algorithm with superconducting qubits, *Phys. Rev. Lett.* **127**, 100501 (2021).
- [60] J. Yirka and Y. Subaşı, Qubit-efficient entanglement spectroscopy using qubit resets, *Quantum* **5**, 535 (2021).
- [61] S. Heußen, D. F. Locher, and M. Müller, Measurement-free fault-tolerant quantum error correction in near-term devices, *PRX Quantum* **5**, 010333 (2024).
- [62] M. A. Perlin, V. N. Premakumar, J. Wang, M. Saffman, and R. Joynt, Fault-tolerant measurement-free quantum error correction with multiqubit gates, *Phys. Rev. A* **108**, 062426 (2023).
- [63] A. M. Childs and N. Wiebe, Hamiltonian simulation using linear combinations of unitary operations, *Quantum Inf. Comput.* **12**, 0901 (2012).
- [64] D. W. Berry, A. M. Childs, R. Cleve, R. Kothari, and R. D. Somma, Simulating Hamiltonian dynamics with a truncated Taylor series, *Phys. Rev. Lett.* **114**, 090502 (2015).
- [65] A. Gilyén, Y. Su, G. H. Low, and N. Wiebe, in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (Association for Computing Machinery, New York, United States, 2019), pp. 193–204.
- [66] Q. Dong, M. T. Quintino, A. Soeda, and M. Murao, Success-or-draw: A strategy allowing repeat-until-success in quantum computation, *Phys. Rev. Lett.* **126**, 150504 (2021).
- [67] D. M. Greenberger, M. A. Horne, A. Shimony, and A. Zeilinger, Bell's theorem without inequalities, *Am. J. Phys.* **58**, 1131 (1990).
- [68] D. Leibfried, M. D. Barrett, T. Schaetz, J. Britton, J. Chiaverini, W. M. Itano, J. D. Jost, C. Langer, and D. J. Wineland, Toward Heisenberg-limited spectroscopy with multiparticle entangled states, *Science* **304**, 1476 (2004).
- [69] C.-H. Liao, C.-W. Yang, and T. Hwang, Dynamic quantum secret sharing protocol based on GHZ state, *Quantum Inf. Process.* **13**, 1907 (2014).
- [70] F. Hahn, J. de Jong, and A. Pappa, Anonymous quantum conference key agreement, *PRX Quantum* **1**, 020325 (2020).
- [71] A. B. Watts, R. Kothari, L. Schaeffer, and A. Tal, in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (Association for Computing Machinery, New York, United States, 2019), pp. 515–526.
- [72] Y. Quek, E. Kaur, and M. M. Wilde, Multivariate trace estimation in constant quantum depth, *Quantum* **8**, 1220 (2024).
- [73] E. Bäumer, V. Tripathi, A. Seif, D. Lidar, and D. S. Wang, Quantum Fourier transform using dynamic circuits, *Phys. Rev. Lett.* **133**, 150602 (2024).
- [74] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels, *Phys. Rev. Lett.* **70**, 1895 (1993).
- [75] D. Bouwmeester, J.-W. Pan, K. Mattle, M. Eibl, H. Weinfurter, and A. Zeilinger, Experimental quantum teleportation, *Nature* **390**, 575 (1997).
- [76] J.-G. Ren, P. Xu, H.-L. Yong, L. Zhang, S.-K. Liao, J. Yin, W.-Y. Liu, W.-Q. Cai, M. Yang, L. Li *et al.*, Ground-to-satellite quantum teleportation, *Nature* **549**, 70 (2017).
- [77] S. Martiel and T. G. de Brugière, Architecture aware compilation of quantum circuits via lazy synthesis, *Quantum* **6**, 729 (2022).
- [78] F. Wagner, A. Bärmann, F. Liers, and M. Weissenböck, Improving quantum computation by optimized qubit routing, *J. Optim. Theory Appl.* **197**, 1161 (2023).
- [79] R. Takagi, H. Tajima, and M. Gu, Universal sampling lower bounds for quantum error mitigation, *Phys. Rev. Lett.* **131**, 210602 (2023).
- [80] R. Takagi, S. Endo, S. Minagawa, and M. Gu, Fundamental limits of quantum error mitigation, *npj Quantum Inf.* **8**, 114 (2022).
- [81] K. Tsubouchi, T. Sagawa, and N. Yoshioka, Universal cost bound of quantum error mitigation based on quantum estimation theory, *Phys. Rev. Lett.* **131**, 210601 (2023).
- [82] Y. Quek, D. Stilck França, S. Khatri, J. J. Meyer, and J. Eisert, Exponentially tighter bounds on limitations of quantum error mitigation, *Nat. Phys.* **20**, 1648 (2024).

- [83] S. Parker and M. B. Plenio, Efficient factorization with a single pure qubit and  $\log N$  mixed qubits, *Phys. Rev. Lett.* **85**, 3049 (2000).
- [84] K. Temme, T. J. Osborne, K. G. Vollbrecht, D. Poulin, and F. Verstraete, Quantum metropolis sampling, *Nature* **471**, 87 (2011).
- [85] Like PROM, without BFA and exploiting symmetrization, conventional REM methods for terminal measurements are inefficient in classical processing costs—see Appendix B 5 a.
- [86] For  $p$ -outcome measurements, all indices for vector and matrix quantities such as  $\mathbf{q}$ ,  $Q$ ,  $\alpha$ , and  $T$  are elevated from bitstrings to strings in the alphabet  $\mathbb{Z}_p$ ;  $\oplus$  accordingly denotes addition in  $\mathbb{Z}_p$  instead of  $\mathbb{Z}_2$ ; the bitmasks  $f$  to be sampled for feedforward are elevated to be strings in  $\mathbb{Z}_p$ ; computation of  $\alpha$  can take advantage of the generalized fast Walsh-Hadamard transform over larger rings.
- [87] X. Yuan, B. Regula, R. Takagi, and M. Gu, Virtual quantum resource distillation, *Phys. Rev. Lett.* **132**, 050203 (2024).
- [88] R. Takagi, X. Yuan, B. Regula, and M. Gu, Virtual quantum resource distillation: General framework and applications, *Phys. Rev. A* **109**, 022403 (2024).
- [89] B. Regula and R. Takagi, Fundamental limitations on distillation of quantum channel resources, *Nat. Commun.* **12**, 4411 (2021).
- [90] Y. Liu and X. Yuan, Operational resource theory of quantum channels, *Phys. Rev. Res.* **2**, 012035(R) (2020).
- [91] S. J. Beale and J. J. Wallman, Randomized compiling for subsystem measurements, [arXiv:2304.06599](https://arxiv.org/abs/2304.06599) [quant-ph].
- [92] 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).
- [93] 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).
- [94] Unified matrix treatment of the fast Walsh-Hadamard transform, *IEEE Trans. Comput.* **100**, 1142 (1976).
- [95] A. J. Walker, New fast method for generating discrete random numbers with arbitrary frequency distributions, *Electron. Lett.* **8**, 127 (1974).
- [96] X. Yuan, Y. Liu, Q. Zhao, B. Regula, J. Thompson, and M. Gu, Universal and operational benchmarking of quantum memories, *npj Quantum Inf.* **7**, 108 (2021).
- [97] A. Deif, in *Sensitivity Analysis in Linear Systems* (Springer, Berlin, 1986), pp. 1–43..
- [98] J. Rohn, New condition numbers for matrices and linear systems, *Computing* **41**, 167 (1989).
- [99] D. C. McKay, C. J. Wood, S. Sheldon, J. M. Chow, and J. M. Gambetta, Efficient Z gates for quantum computing, *Phys. Rev. A* **96**, 022330 (2017).
- [100] N. Sundaresan, I. Lauer, E. Pritchett, E. Magesan, P. Jurcevic, and J. M. Gambetta, Reducing unitary and spectator errors in cross resonance with optimized rotary echoes, *PRX Quantum* **1**, 020318 (2020).
- [101] J. A. Smolin, J. M. Gambetta, and G. Smith, Efficient method for computing the maximum-likelihood quantum state from measurements with additive Gaussian noise, *Phys. Rev. Lett.* **108**, 070502 (2012).
- [102] M. R. Geller, Rigorous measurement error correction, *Quantum Sci. Technol.* **5**, 03LT01 (2020).
- [103] B. Nachman, M. Urbanek, W. A. de Jong, and C. W. Bauer, Unfolding quantum computer readout noise, *npj Quantum Inf.* **6**, 84 (2020).
- [104] J. M. Koh, T. Tai, and C. H. Lee, Simulation of interaction-induced chiral topological dynamics on a digital quantum computer, *Phys. Rev. Lett.* **129**, 140502 (2022).
- [105] J. M. Koh, T. Tai, Y. H. Phee, W. E. Ng, and C. H. Lee, Stabilizing multiple topological fermions on a quantum computer, *npj Quantum Inf.* **8**, 16 (2022).
- [106] J. M. Koh, S.-N. Sun, M. Motta, and A. J. Minnich, Measurement-induced entanglement phase transition on a superconducting quantum processor with mid-circuit readout, *Nat. Phys.* **19**, 1314 (2023).
- [107] J. M. Koh, T. Tai, and C. H. Lee, Realization of higher-order topological lattices on a quantum computer, *Nat. Commun.* **15**, 5807 (2024).
- [108] L. Viola, E. Knill, and S. Lloyd, Dynamical decoupling of open quantum systems, *Phys. Rev. Lett.* **82**, 2417 (1999).
- [109] Y. Kim, C. J. Wood, T. J. Yoder, S. T. Merkel, J. M. Gambetta, K. Temme, and A. Kandala, Scalable error mitigation for noisy quantum circuits produces competitive expectation values, *Nat. Phys.* **19**, 752 (2023).
- [110] B. Pokharel, N. Anand, B. Fortman, and D. A. Lidar, Demonstration of fidelity improvement using dynamical decoupling with superconducting qubits, *Phys. Rev. Lett.* **121**, 220502 (2018).
- [111] K. Khodjasteh and D. A. Lidar, Fault-tolerant quantum dynamical decoupling, *Phys. Rev. Lett.* **95**, 180501 (2005).
- [112] G. S. Uhrig, Keeping a quantum bit alive by optimized  $\pi$ -pulse sequences, *Phys. Rev. Lett.* **98**, 100504 (2007).
- [113] N. Ezzell, B. Pokharel, L. Tewala, G. Quiroz, and D. A. Lidar, Dynamical decoupling for superconducting qubits: A performance survey, *Phys. Rev. Appl.* **20**, 064027 (2023).
- [114] B. Pokharel and D. A. Lidar, Better-than-classical Grover search via quantum error detection and suppression, *npj Quantum Inf.* **10**, 23 (2024).
- [115] B. Pokharel and D. A. Lidar, Demonstration of algorithmic quantum speedup, *Phys. Rev. Lett.* **130**, 210602 (2023).
- [116] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. Van Den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme *et al.*, Evidence for the utility of quantum computing before fault tolerance, *Nature* **618**, 500 (2023).
- [117] C. Tong, H. Zhang, and B. Pokharel, Empirical learning of dynamical decoupling on quantum processors, *PRX Quantum* **6**, 030319 (2025).
- [118] H. Buhrman, R. Cleve, J. Watrous, and R. de Wolf, Quantum fingerprinting, *Phys. Rev. Lett.* **87**, 167902 (2001).
- [119] A. Barenco, A. Berthiaume, D. Deutsch, A. Ekert, R. Jozsa, and C. Macchiavello, Stabilization of quantum computations by symmetrization, *SIAM J. Comput.* **26**, 1541 (1997).
- [120] S. T. Flammia and Y.-K. Liu, Direct fidelity estimation from few Pauli measurements, *Phys. Rev. Lett.* **106**, 230501 (2011).

- [121] M. P. da Silva, O. Landon-Cardinal, and D. Poulin, Practical characterization of quantum devices without tomography, *Phys. Rev. Lett.* **107**, 210404 (2011).
- [122] S. Cao, B. Wu, F. Chen, M. Gong, Y. Wu, Y. Ye, C. Zha, H. Qian, C. Ying, S. Guo *et al.*, Generation of genuine entanglement up to 51 superconducting qubits, *Nature* **619**, 738 (2023).
- [123] P. Gokhale, O. Angiuli, Y. Ding, K. Gui, T. Tomesh, M. Suchara, M. Martonosi, and F. T. Chong, Minimizing state preparations in variational quantum eigensolver by partitioning into commuting families, [arXiv:1907.13623](https://arxiv.org/abs/1907.13623) [quant-ph].
- [124] C. A. Sackett, D. Kielpinski, B. E. King, C. Langer, V. Meyer, C. J. Myatt, M. Rowe, Q. Turchette, W. M. Itano, D. J. Wineland *et al.*, Experimental entanglement of four particles, *Nature* **404**, 256 (2000).
- [125] D. Leibfried, E. Knill, S. Seidelin, J. Britton, R. B. Blakestad, J. Chiaverini, D. B. Hume, W. M. Itano, J. D. Jost, C. Langer *et al.*, Creation of a six-atom “Schrödinger cat” state, *Nature* **438**, 639 (2005).
- [126] T. Monz, P. Schindler, J. T. Barreiro, M. Chwalla, D. Nigg, W. A. Coish, M. Harlander, W. Hänsel, M. Hennrich, and R. Blatt, 14-qubit entanglement: Creation and coherence, *Phys. Rev. Lett.* **106**, 130506 (2011).
- [127] A. Omran, H. Levine, A. Keesling, G. Semeghini, T. T. Wang, S. Ebadi, H. Bernien, A. S. Zibrov, H. Pichler, S. Choi *et al.*, Generation and manipulation of Schrödinger cat states in Rydberg atom arrays, *Science* **365**, 570 (2019).
- [128] I. Pogorelov, T. Feldker, C. D. Marciniak, L. Postler, G. Jacob, O. Kriegelsteiner, V. Podlesnic, M. Meth, V. Negnevitsky, M. Stadler, B. Höfer, C. Wächter, K. Lakhmanskiy, R. Blatt, P. Schindler, and T. Monz, Compact ion-trap quantum computing demonstrator, *PRX Quantum* **2**, 020343 (2021).
- [129] P. Thomas, L. Ruscio, O. Morin, and G. Rempe, Efficient generation of entangled multiphoton graph states from a single atom, *Nature* **608**, 677 (2022).
- [130] G. J. Mooney, G. A. White, C. D. Hill, and L. C. Hollenberg, Generation and verification of 27-qubit Greenberger-Horne-Zeilinger states in a superconducting quantum computer, *J. Phys. Commun.* **5**, 095004 (2021).
- [131] K. X. Wei, I. Lauer, S. Srinivasan, N. Sundaresan, D. T. McClure, D. Toyli, D. C. McKay, J. M. Gambetta, and S. Sheldon, Verifying multipartite entangled Greenberger-Horne-Zeilinger states via multiple quantum coherences, *Phys. Rev. A* **101**, 032343 (2020).
- [132] 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).
- [133] G. S. Barron and C. J. Wood, Measurement error mitigation for variational quantum algorithms, [arXiv:2010.08520](https://arxiv.org/abs/2010.08520) [quant-ph].
- [134] P. D. Nation, H. Kang, N. Sundaresan, and J. M. Gambetta, Scalable mitigation of measurement errors on quantum computers, *PRX Quantum* **2**, 040326 (2021).