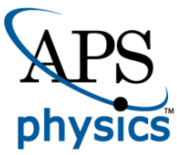

Computationally efficient zero-noise extrapolation for quantum-gate-error mitigation

Vincent R. Pascuzzi, Andre He, Christian W. Bauer, Wibe A. de Jong, and Benjamin Nachman

Phys. Rev. A **105**, 042406 — Published 5 April 2022

DOI: 10.1103/PhysRevA.105.042406

Computationally Efficient Zero Noise Extrapolation for Quantum Gate Error Mitigation

Vincent R. Pascuzzi[#,1,*] Andre He[#,2,†] Christian W. Bauer,[2,‡] Wibe A. de Jong,[2,§] and Benjamin Nachman[2,¶]

[1] *Computational Science Initiative, Brookhaven National Laboratory, Upton, NY 11972, USA*
[2] *Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*

(Dated: March 17, 2022)

Zero noise extrapolation (ZNE) is a widely used technique for gate error mitigation on near term quantum computers because it can be implemented in software and does not require knowledge of the quantum computer noise parameters. Traditional ZNE requires a significant resource overhead in terms of quantum operations. A recent proposal using a targeted (or random) instead of fixed identity insertion method (RIIM versus FIIM) requires significantly fewer quantum gates for the same formal precision. We start by showing that RIIM can allow for ZNE to be deployed on deeper circuits than FIIM, but requires many more measurements to achieve the same level of statistical uncertainty. We develop two extensions to FIIM and RIIM. The List Identity Insertion Method (LIIM) allows to mitigate the error from certain CNOT gates, typically those with the largest error. Set Identity Insertion Method (SIIM) naturally interpolates between the measurement-efficient FIIM and the gate-efficient RIIM, allowing to trade off fewer CNOT gates for more measurements. Finally, we investigate a way to boost the number of measurements, namely to run ZNE in parallel, utilizing as many quantum devices as are available. We explore the performance of RIIM in a parallel setting where there is a non-trivial spread in noise across sets of qubits within or across quantum computers.

I. INTRODUCTION

Noisy intermediate-scale quantum (NISQ) [1] computers are promising tools for performing certain calculations more efficiently than can be computed with classical computers. This may allow for the evaluation of currently intractable calculations. A fundamental challenge facing NISQ computation is that there is significant noise in the instruction sets (gates) and state readout. The ultimate computational performance (towards fault tolerance) will be achieved with quantum error correction (see e.g. Ref. [2]). However, quantum error correction typically requires a many-to-one physical-to-logical mapping of quantum bits (qubits) and small enough gate errors. Current NISQ devices do not allow for the implementation of fault tolerant algorithms.

A variety of error mitigation strategies have been proposed on current and near-term quantum computers. One widely used strategy is zero noise extrapolation (ZNE) [3–11]; additional approaches include estimation circuits [12–17], quasi-probability methods [6, 7, 18, 19], and others – see Ref. [20] for a recent review. In the ZNE protocol, measurements are made of an observable from a given circuit and a set of equivalent (auxiliary) circuits with amplified noise but the same zero-noise value. The noise is amplified in a controlled way so that measurements with different levels of noise can be used to extrapolate to the zero-noise limit. A hardware-agnostic approach to ZNE can be implemented by replacing a particular gate U_i by $U_i(U_i^\dagger U_i)^{n_i}$ for non-negative integer n_i (see Fig. 1). These n_i identity insertions do not change the measured expectation value of the circuit, but since U_i is noisy, the total error is increased. The standard Fixed Identity Insertion Method (FIIM) [5, 6] uses the same $n_i = n$ for every gate so that each gate is replaced by

$$r = 2n + 1\,, \tag{1}$$

copies of itself. Data are generated with $n = 0, 1, 2...$ and then the target observable is extrapolated to $n = -\frac{1}{2}$, corresponding to $r = 0$. This approach effectively mitigates gate errors, but at the expense of requiring a large number of quantum gates. ZNE is typically applied only to controlled-NOT (CNOT) gates which have a significantly higher error rate than single qubit gates. For a circuit with n_c CNOT gates, FIIM requires $2n \times n_c$ additional gates for each auxiliary circuit.

A quantum gate efficient alternative ZNE called the Random Identity Insertion Method (RIIM) was proposed recently [9]. Instead of using a global $n_i = n$, a non-uniform number of identity insertions is added for each gate. In particular, the first order correction is achieved by considering an input circuit having n_c CNOT gates (and hence n_c auxiliary circuits) where each auxiliary circuit has a different CNOT gate augmented with an identity insertion. Instead of requiring $2n_c$ gates for the lowest order correction as in FIIM, RIIM requires only two additional gates for each auxiliary circuit. In this way, RIIM requires fewer gates for auxiliary circuits and so has

* Work done while at Lawrence Berkeley National Laboratory; pascuzzi@bnl.gov
† andrehe@lbl.gov
‡ cwbauer@lbl.gov
§ wadejong@lbl.gov
¶ bpnachman@lbl.gov
# These authors contributed equally.

* Future work can consider the application of this protocol to different sets of single- and multi-qubit basis gates.

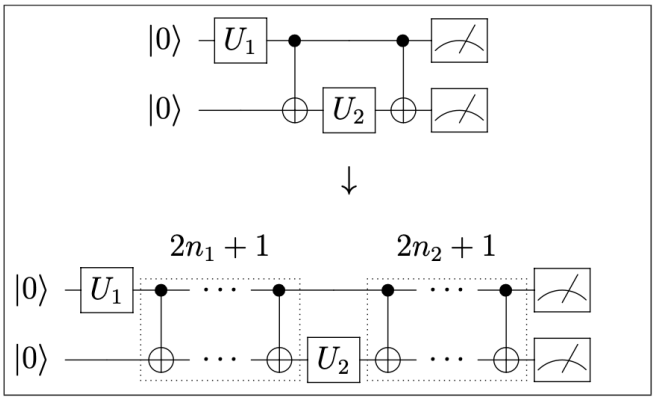

FIG. 1: An illustration of identity insertion for a generic controlled unitary operation with two qubits. The U_i represent unitary matrices and the n_i are non-negative integers.

the potential to enable gate error mitigation on deeper circuits than FIIM. As we will discuss, RIIM not only needs more circuits, it also necessitates the need for more measurements per circuit to achieve the same statistical uncertainty as FIIM.

In this paper we develop methods to improve our ability to mitigate noise via ZNE. We develop variations of FIIM and RIIM, and study how one can parallelize the running of the required RIIM circuits to obtain higher statistics. In particular, we propose new ZNE techniques called List and Set Identity Insertion Method (LIIM and SIIM) which replace only gates from a list (LIIM) or sets of gates (SIIM) with the same number of identity insertions instead of single gates (RIIM) or all gates at once (FIIM). We furthermore study a parallelization strategy for ZNE across quantum computers to generate a large number of measurements. In particular, we investigate the performance of RIIM applied in parallel across computers with a non-trivial distribution of errors.

This paper is organized as follows. Section II briefly reviews FIIM and RIIM, providing an explicit example where RIIM achieves the same fidelity for a deeper circuit than FIIM. Then, Sec. III introduces extensions of RIIM and FIIM, including SIIM and LIIM. The parallelization of ZNE is then studied in Sec. IV for a synthetic distribution of noise and then in Sec. V for a realistic distribution of noise. The paper ends with conclusions and outlook in Sec. VI.

II. FIIM AND RIIM OVERVIEW

The basic idea of ZNE methods is to measure a given observable at varying levels of noise, and using the measured dependence on the noise to extrapolate to the expected noiseless value. The dominant source of instruction-level noise in current digital quantum computers arises from the 2-qubit entangling CNOT gate, and the dominant noise channel is the 2-qubit depolarizing channel. In a two-qubit scenario, the depolarizing channel is given by the quantum operation acting on the system's density matrix, ρ:

$$\mathcal{E}(\rho) = (1-\epsilon)\rho + \frac{\epsilon}{4} I\,, \tag{2}$$

where I is the 2×2 identity matrix and the noise parameter ϵ is of order a percent on current NISQ machines. The action of a single noisy (depolarizing) CNOT gate on a general density matrix ρ can therefore be written as

$$\text{CNOT}_{kl}[\rho] = (1-\epsilon)U_C^{(kl)}\rho U_C^{(kl)} + \frac{\epsilon}{4}\rho_{\not kl}\otimes I_{kl}\,, \tag{3}$$

where $\rho_{\not kl}$ represents the density matrix after tracing over the k and l qubits and U_C is the unitary operator corresponding to the CNOT gate. As the action of two CNOT gates gives the identity, the application of an odd number r of CNOT gates to the same kl qubits produces

$$\text{CNOT}^r_{kl}[\rho] = (1-r\epsilon)U_C\rho U_C + \frac{r\epsilon}{4}\rho_{\not kl}\otimes I_{kl} + \mathcal{O}(\epsilon^2)\,. \tag{4}$$

Given Eq. (4), one can analyze the result of the action of a given quantum circuit C containing n_C CNOT gates and a universal depolarizing error rate ϵ. This circuit creates a density matrix, which to first order in ϵ can be written as [9]

$$C[\rho] = (1-n_c\epsilon)\rho_{\text{ex}} + \epsilon\sum_{i=1}^{n_c}\rho_i + \mathcal{O}(\epsilon^2)\,, \tag{5}$$

where ρ_{ex} is the density matrix that would be obtained from a noiseless circuit, and ρ_i denotes the density matrix obtained if the i^{th} CNOT gate in the circuit is replaced by the depolarizing channel. In other words, ρ_i is constructed by replacing the 2-qubit system that the i^{th} CNOT gate acts on with the completely mixed state $I/4$. Defining a circuit $C_{r_1,\ldots r_{n_c}}$, which replaces the i^{th} n_c CNOT gate by r_i copies of the same CNOT gate, one can write the action of this circuit as

$$C_{\{r_1,\ldots r_{n_c}\}}[\rho] = \left(1-\epsilon\sum_{i=1}^{n_c} r_i\right)\rho_{\text{ex}} + \epsilon\sum_{i=1}^{n_c} r_i\,\rho_i + \mathcal{O}(\epsilon^2)\,. \tag{6}$$

Given these expressions, one can now derive the expressions for the ZNE versions FIIM and RIIM, as introduced in Ref. [9]. In FIIM, one constructs from the nominal circuit, C_{nom}, an auxiliary circuit, C_{FIIM} in which each CNOT is replaced by $r_i = 3$, $\forall\ i$ CNOT gates. Given Eqs. (5) and (6), one can show that

$$C_{\text{FIIM}}[\rho] \equiv \frac{3}{2}C[\rho] - \frac{1}{2}C_{\{3,3,\ldots,3\}}[\rho] = \rho_{ex} + \mathcal{O}(\epsilon^2)\,. \tag{7}$$

The exact density matrix can therefore be obtained from a linear superposition of the nominal circuit and the

auxiliary circuit, up to errors that are quadratic in the depolarizing noise parameter ϵ. In RIIM, one replaces only a single CNOT gate by three CNOT gates, but then adds the n_c possible density matrices. Symbolically, this can be expressed as

$$C_{\text{RIIM}}[\rho] \equiv \frac{2+n_c}{2} C[\rho] - \frac{1}{2} \sum_{\sigma \in S_{n_c}} C_{\sigma\{3,1,\ldots,1\}}[\rho] = \rho_{ex} + \mathcal{O}(\epsilon^2)\,, \tag{8}$$

where

$$\sum_{\sigma \in S_{n_c}} C_{\sigma\{3,1,\ldots,1\}}[\rho] = C_{3,1,\ldots 1}[\rho] + C_{1,3,1,\ldots 1}[\rho] + \ldots + C_{1,\ldots 1,3}[\rho]\,, \tag{9}$$

for permutation matrices $\sigma \in S_{n_c}$.

The above analysis can be extended to higher orders in ϵ; in fact, by computing the $\mathcal{O}(\epsilon^2)$ correction term one can show that RIIM has a correction that is a factor of 3 smaller than that of FIIM [9]. In what follows, we will mostly focus on first-order corrections in ϵ.

While FIIM and RIIM both remove the linear depolarizing noise, they use different computational resources to achieve this goal. FIIM requires a circuit that multiplies the total CNOT count by a factor of three, whereas RIIM adds only two CNOT gates to its auxiliary circuit. This means that the FIIM auxiliary circuit is significantly deeper than in RIIM and one can therefore expect RIIM to outperform FIIM especially when the nominal circuit contains many CNOT gates. This is illustrated in Fig. 2 for a simple 2-qubit circuit (Fig. 3) with $2n+1$ CNOT gates in the absence of statistical noise. Note that the simulation in Fig. 2 includes the effect of amplitude damping, for which $T_1 = 50$ μs, $T_{\text{CNOT}} = 200$ ns, and the damping constant $\gamma = 1 - e^{-T_{\text{CNOT}}/T_1}$; this is why the data for which $\epsilon = 0$ are not unity. In addition to mitigating depolarizing noise, ZNE also reduces the impact of amplitude damping.

On the other hand, RIIM requires many more measurements to obtain the same statistical accuracy as FIIM. This can be seen from the linear combinations in Eqs. (7) and (8) taken in FIIM and RIIM. Assume we know the value of $C[\rho]$ with statistical accuracy $\delta\rho_1$, while we know the value of each $C_{\{r_1,\ldots,r_{n_c}\}}[\rho]$ with accuracy $\delta\rho_2$. By taking the appropriate linear combinations, the standard deviation across measurements is given by

$$\sigma\left(C_{\text{FIIM}}[\rho]\right) = \sqrt{\frac{9}{4}\delta\rho_1^2 + \frac{1}{4}\delta\rho_2^2}$$
$$\sigma\left(C_{\text{RIIM}}[\rho]\right) = \sqrt{\frac{(1+2n_c)^2}{4}\delta\rho_1^2 + n_c\frac{1}{4}\delta\rho_2^2}\,, \tag{10}$$

where the n_c multiplying the $\delta\rho_2$ term in $C_{\text{RIIM}}[\rho]$ arises from the fact that $\sum_{\sigma\in S_{n_c}} C_{\sigma\{3,1,\ldots,1\}}[\rho]$ contains n_c terms. This implies that in order to get the same statistical precision in the two approaches requires a much more precise knowledge of $\delta\rho_1$ and $\delta\rho_2$ in RIIM compared to FIIM. In particular, one requires

$$[\delta\rho_1]_{\text{RIIM}} \approx \frac{3}{2}\frac{[\delta\rho_1]_{\text{FIIM}}}{n_c}\,, \qquad [\delta\rho_2]_{\text{RIIM}} = \frac{[\delta\rho_2]_{\text{FIIM}}}{\sqrt{n_c}}\,. \tag{11}$$

Given that the statistical error scales with the square of the number of measurements, Eq. (11) shows that for RIIM to match the FIIM precision, one needs n_c^2 more measurements for the nominal circuit and n_c more measurements for each of each of the $C_{\{r_1,\ldots r_{n_c}\}}[\rho]$ circuits. Especially for large number of CNOT gates, for which RIIM is especially expected to outperform FIIM, this is a potential drawback of RIIM (but it is of course not an in principle limitation). In the next section we will discuss a variant of RIIM which allows for improvements over FIIM while keeping the number of measurements required more manageable. After that we discuss how one can parallelize the execution of the required quantum circuits, such that one can obtain the required number of measurements in a shorter amount of time.

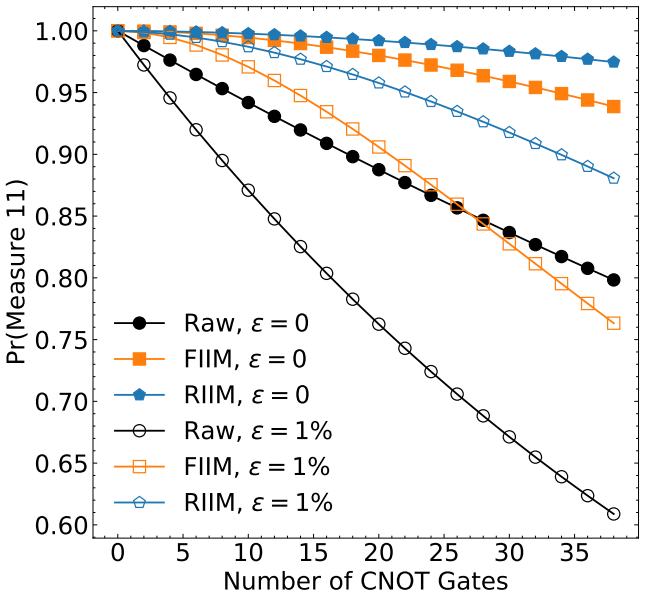

FIG. 2: An illustration of the FIIM and RIIM protocols in the absence of statistical noise for a two-qubit circuit with an even number of CNOT gates as specified by the horizontal axis. The noise model includes depolarizing noise as specified in the legend and decoherence noise modeled by amplitude damping with a T_1 of 50 μs and a CNOT gate time of 200 ns. The two qubits are prepared in the $|1\rangle$ state. Simulations performed with Cirq [21].

III. EXTENDING FIIM AND RIIM

In this section, we explore strategies for ZNE that use fewer quantum resources than FIIM and fewer measurements than RIIM.

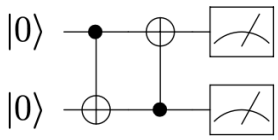

FIG. 3: A simple two-qubit circuit comprising two CNOT gates followed by a measurement of each qubit.

A. Correcting for individual CNOT noise: The List Identity Insertion Method (LIIM)

In general, CNOT errors are different for each pair of qubits, and for systems in which some CNOT errors are much larger than the rest, one may be able to mitigate only the dominant CNOT errors with fewer resources than mitigating all errors. Using a separate depolarizing noise value (ϵ_i) for each CNOT gate, the leading order expression Eq. (6) of the depolarizing noise becomes

$$C_{\{r_1,\dots r_{n_c}\}}[\rho] = \left(1-\sum_{i=1}^{n_c}\epsilon_i r_i\right)\rho_{\rm ex} + \sum_{i=1}^{n_c}\epsilon_i r_i\,\rho_i + \mathcal{O}(\epsilon_i\epsilon_j)\,. \tag{12}$$

We can define a circuit $C_{\{3\}_L^{\rm all}}[\rho]$ that replaces all CNOTs of a given list L by 3 CNOTs, and the sum of circuits $C_{\{3\}_L^{\rm sum}}[\rho]$ that replace one CNOT from each in this list by 3 CNOTs. Then, one can construct a FIIM version that removes the linear terms of the large CNOT errors

$$\begin{aligned} C_{{\rm FIIM},L}[\rho] &\equiv \frac{3}{2}C[\rho] - \frac{1}{2}C_{\{3\}_L^{\rm all}}[\rho] \\ &= \rho_{ex} + \mathcal{O}(\{\epsilon_k\}_{k\notin L}, \{\epsilon_i\epsilon_j\}_{i,j\in L})\,. \end{aligned} \tag{13}$$

The correction remains linear in the ϵ terms that were not part of the list L, but those that were part of the list are multiplied by one other value of ϵ. Similarly, one can obtain an analogous version of RIIM:

$$\begin{aligned} C_{{\rm RIIM},L}[\rho] &\equiv \frac{2+|L|}{2}C[\rho] - \frac{1}{2}C_{\{3\}_L^{\rm sum}}[\rho] \\ &= \rho_{ex} + \mathcal{O}(\{\epsilon_k\}_{k\notin L}, \{\epsilon_i\epsilon_j\}_{i,j\in L})\,, \end{aligned} \tag{14}$$

where $|L|$ denotes the number of elements in the list. An illustration of the list-based ZNE is presented in Fig. 4.

B. Interpolating between FIIM and RIIM: The Set Identity Insertion Method SIIM

In the case that all of the CNOT errors are comparable and need to be mitigated, it is still possible to selectively replace gates by viewing FIIM and RIIM as special cases of a more general approach, which we coin the Set Identity Insertion Method (SIIM). In SIIM one divides the n_c CNOT gates into n_s sets[†] containing the same number of CNOT gates $m = n_c/n_s$. One is then free to choose some sets in n_s to replace each CNOT gate by e.g. three CNOT, while keeping the other sets untouched. We denote this by $C_{\{1\},\dots,\{3\},\dots\{1\}}[\rho]$. Adding all different sets results in

$$C_{\{3\}^m}[\rho] \equiv C_{\{3\},\{1\},\dots\{1\}}[\rho] + C_{\{1\},\{3\},\{1\},\dots\{1\}}[\rho] + \dots\,, \tag{15}$$

which contains a total of n_s terms, with m CNOT gates replaced. Following the same steps as for FIIM and RIIM one can now define the linear combination

$$C^{(n_s)}_{\rm SIIM}[\rho] \equiv \frac{2+n_s}{2}C[\rho] - \frac{1}{2}C_{\{3\}^{n_s}}[\rho] = \rho_{ex} + \mathcal{O}(\epsilon^2)\,. \tag{16}$$

Note that FIIM is recovered by using a single set $n_s = 1$, while RIIM is obtained by using as many sets as CNOT gates $n_s = n_c$.

The SIIM keeps much of the advantage of RIIM, while greatly reducing its main disadvantage. In particular, the the extra number of gates required in SIIM is $2n_c/n_s$, while the extra number of measurements scales with n_s^2 instead

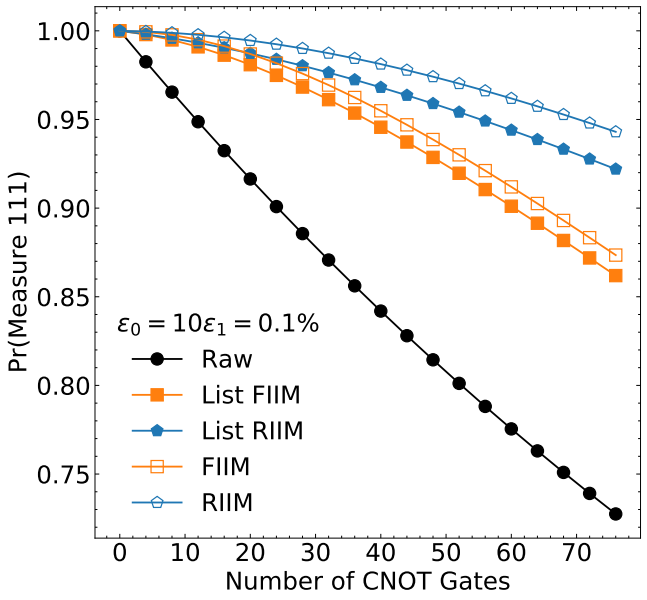

FIG. 4: An illustration of the LIIM protocol for the FIIM and RIIM variants in the absence of statistical noise for a three-qubit circuit with an even number of CNOT gates between the first two and second two qubits as specified by the horizontal axis. The noise model only includes depolarizing noise. The three qubits are prepared in the $|1\rangle$ state. Simulations performed with CIRQ [21].

[†] We are assuming here that n_c/n_s is an integer to simplify the notation. The general approach still works if this is not true, but is a little more complicated to explain.

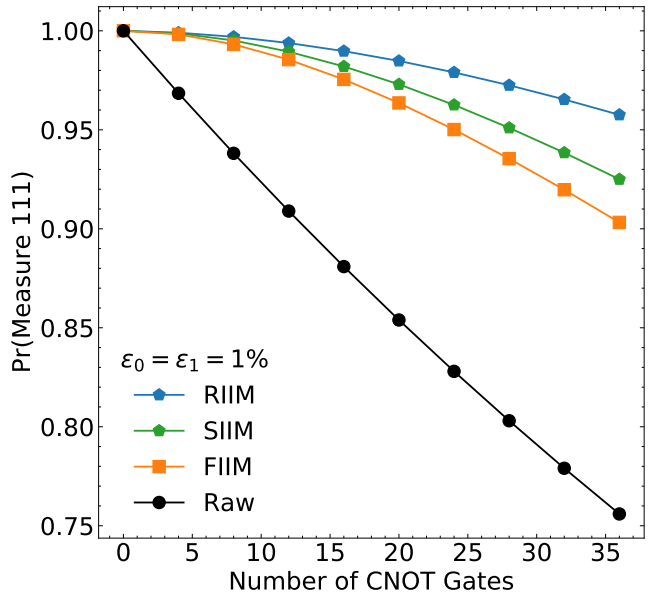

FIG. 5: An illustration of the SIIM approach (with two sets) in the absence of statistical noise for a three-qubit circuit with an even number of CNOT gates between the first two and second two qubits as specified by the horizontal axis. The noise model only includes depolarizing noise. The three qubits are prepared in the $|1\rangle$ state. Simulations performed with CIRQ [21].

of n_c^2 as for RIIM. Figure 5 illustrates how SIIM interpolates between RIIM and FIIM. For purely depolarizing noise, the coefficient of the remaining $(n_c\epsilon)^2$ term in SIIM is smaller by a factor of $(2 + n_c/n_s)/3n_c$ compared to FIIM. This is summarized in the following table.

Approach	# of CNOT	error (rel to FIIM)	# of measurements	
			Nominal	Correction
FIIM	$3n_c$	1	N_{nom}	N_{corr}
RIIM	$n_c + 2$	$\frac{1}{3}$	$N_{\text{nom}}\frac{(1+2n_c)^2}{9}$	$N_{\text{corr}}n_c$
SIIM	$n_c + 2\frac{n_c}{n_s}$	$\frac{\delta}{1+2n_s}$	$N_{\text{nom}}\frac{(1+2n_s)^2}{9}$	$N_{\text{corr}}n_s$

TABLE I: Summary of the three different ZNE approaches. As discussed in the text, the RIIM approach requires almost a factor of three fewer CNOT gates to achieve a uncertainty that is smaller by a factor of 3. However, it does require many more events to reach the same statistical precision. The SIIM approach interpolates between the two methods, allowing to reach a trade off between the number of CNOT gates and number of events needed.

IV. PARALLELING ZNE: SYNTHETIC ERROR MODELS

The results in the previous section focused on the case of zero statistical noise from a finite number of measurements. This is of course not realistic, and to reduce the statistical noise one will need to perform a large number of measurements. One way to rapidly accumulate a large number of measurements is to parallelize across computers. This section will explore this idea first by considering a simple model in which depolarizing errors are normally distributed across a batch of quantum computers and then second with a realistic distribution of errors.

A. Analytical Results

Suppose that the depolarizing error is constant within quantum computer i and that this value across computers is normally distributed with $\epsilon_i \sim \mathcal{N}(\epsilon, \sigma^2)$. There is no unique way to parallelize the ZNE approaches introduced earlier. One possibility is to run the nominal circuit and the noise-amplified auxiliary circuits on every computer and then average the results:

$$\begin{aligned}\langle C_{\text{ZNE}}[\rho]\rangle &= \rho_{ex} + \mathcal{O}(\langle\epsilon_i^2\rangle) \\ &= \rho_{ex} + \mathcal{O}(\epsilon^2 + \sigma^2)\,, \end{aligned} \tag{17}$$

where we assume that $\epsilon \ll 1$ and $\sigma \lesssim \epsilon$ so that higher-order terms can be neglected. Another strategy would be to run different parts of the ZNE calculation on different computers, where in general $\epsilon_i \neq \epsilon_j$. From Eqs. (5) and (6), this would lead to

$$\begin{aligned}\langle C_{\text{ZNE}}[\rho]\rangle &= \rho_{ex} + \mathcal{O}(\langle\epsilon_i - \epsilon_j\rangle) + \mathcal{O}(\langle\epsilon_i^2\rangle) \\ &= \rho_{ex} + \mathcal{O}(\epsilon^2 + \sigma^2)\,, \end{aligned} \tag{18}$$

which is an equivalent expression to Eq. (17). However, executing the nominal and auxiliary circuits on every computer—as opposed to spreading the computations across different computers—has the advantage of a smaller variance since the difference $\epsilon_i - \epsilon_j$ in Eq. (18) does not in general cancel.

B. Numerical Results

As shown in the previous section, relaxing the assumption of uniform gate errors introduces an additional error term to the overall extrapolation error which, in the case of a normally-distributed set of gate errors, is dependent on the standard deviation of the distribution. In Fig. 7, the 4-CNOT circuit shown in Fig. 6 with a initial state $|10\rangle$ was executed across an ensemble of simulated quantum devices in QISKIT [22]. When interpreting the output string as a integer, the ideal result in the absence of noise is 3 for a given measurement The 2-qubit gate errors in each of these simulated devices are drawn from a normal distribution with $\mu = 0.1$, corresponding to a mean CNOT error of 10%, while the standard deviation of the gate error distribution is increased to study the

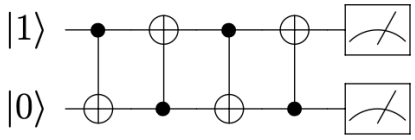

FIG. 6: A simple two-qubit test circuit with four CNOT gates, followed by measurement of each qubit.

scaling of the error incurred by a wider distribution of errors. These errors are larger than typical uncertainties on existing machines, but the large mean ensures that the samples values are positive and to clearly demonstrate the scaling behavior. The circuit was simulated across each of these devices, and the observable we use is the average value of the state, interpreting the bitstring in binary[‡]. These are averaged across devices. Results from non-error-mitigated circuits are included, as well as results from error-mitigating using first and second order FIIM. The additional error is the excess error induced by the $\mathcal{O}(\epsilon^2 + \sigma^2)$ term over the $\mathcal{O}(\epsilon^2)$ extrapolation error seen when assuming a single error rate across all devices and gates. The experiments shown in Fig. 7 indicate that although there is a slight increase in the additional error as σ is increased, it remains fairly small in magnitude. Figure 8 provides a visualization of the error distributions used to generate these simulated ensembles.

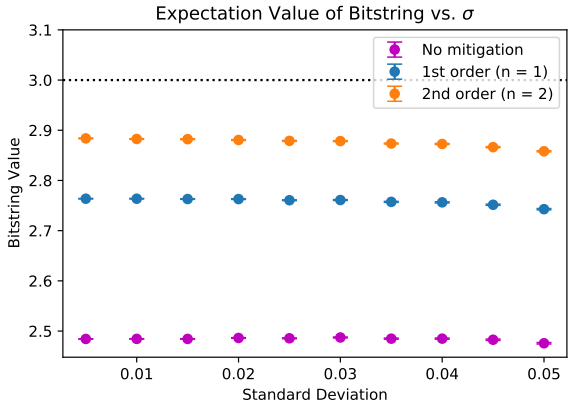

FIG. 7: A demonstration of the scaling of extrapolation error in the 4-CNOT circuit as the standard deviation of the gate error is increased. Simulations were run using Qiskit. A total of 10,000 different error rates were sampled, and each instance of the circuit was run with 10,000 shots.

[‡] Note that while this observable has been studied in other contexts, it has the feature that averaging integer values can artificially enhance the apparent fidelity when migrations in opposite directions cancel.

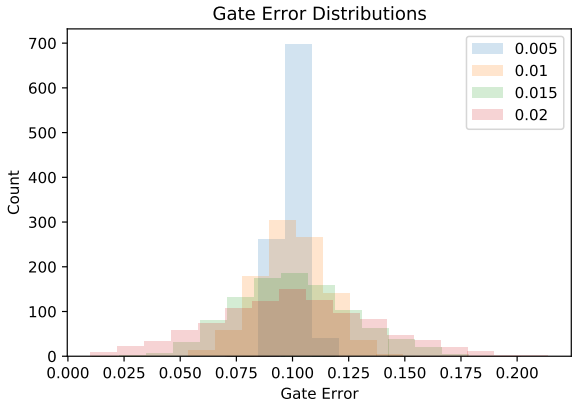

FIG. 8: Distribution of the error rates used to generate the data in Fig. 7.

V. PARALLELING ZNE: REALISTIC ERROR MODELS

We turn now to simulations using two-qubit depolarizing noise parameters, ϵ_i, extracted from devices currently available through IBM Quantum (IBM Q). At the time of writing, IBM Q "advanced access" offers 14 unique systems (excluding simulators) with various properties such as depolarizing error parameters, T1 and T2 relaxation time constants, and gate execution times. As retrieved from IBM Q backend (*system*) properties, depolarizing error parameters, $\epsilon_{i,(jk)}$, where i enumerates the systems themselves and (jk) refers to the coupling between qubits j and k, are symmetric about the qubits; that is, $\epsilon_{i,(jk)} = \epsilon_{i,(kj)}$. In the following studies, we consider depolarizing error channels in simulations performed using Qiskit to evaluate the utility of parallelization across multiple systems.

In the case of RIIM, where the number of required measurements (shots), n_{shots}, is proportional to the square of the number of CNOT gates in the circuit, error mitigation can be computationally time-consuming; this applies to both simulation and execution on real systems. Moreover, significant latency can be incurred if one desires the highest fidelity system available to execute their circuits, as cloud-based systems typically use a first in, first out queuing mechanism unless the preferred system is otherwise reserved. Additionally, current IBM cloud quantum systems have limitations on the number of shots per circuit – typically 8192 – and the total number of circuits – ranging from 75 to 900 – a single job submission can contain. Since extensive allocations to the ideal system may not always be available for a given experiment, spreading large workloads across multiple (manifestly error-prone) systems is beneficial in terms of higher throughput and reliability of measurements.

For our experiment, we consider the same circuit from the previous section (Fig. 6). One may artificially deepen the four-CNOT circuit to gain some insight into the per-

formance of a given error mitigation technique by, for example, replacing each of the four CNOT gates with an odd integer number of them. The application of RIIM would result in n_c auxiliary circuits which in practice could be greater than the circuit limit on a cloud-based quantum system. Furthermore, the required number of measurements to perform the protocol could far exceed the per-circuit shot limit of a system. This motivates the use of multiple systems for parallelizing the error mitigation method.

Shown in Fig. 9 are simulation results of executing the previously described circuit at various depths, ranging between 4–124 total CNOT gates, with only depolarizing noise applied; thermal relaxation and readout errors are disabled in the simulations. The observable measured, made in the standard basis, is the classical bit string value of the final state. In the top panel are the unmitigated and RIIM mitigated results using `ibmq_guadalupe` with depolarizing error parameter $\epsilon_{01} = 0.0104$. Each circuit was executed 8192 times. The bottom panel shows the unmitigated and RIIM output when executing the sets of RIIM circuits across multiple, arbitrarily chosen, systems. The number of systems used in the calculation of each data point is equal to $2n + 1$ (*i.e.*, the number of CNOT gates replacing a single CNOT in a circuit) with at most 14 unique systems (the maximum available from IBM Q at the time of writing) executing in parallel. For cases when $2n+1 > 14$, systems were selected – again arbitrarily – to be recycled and used for executing multiple sets of RIIM circuits. The mean depolarizing error parameter value of the 14 systems is $\epsilon_{01} = 0.0158 \pm 0.0035$, and includes the device used in the upper panel. The distribution of machine CNOT error rates are shown in Fig. 10.

The utility of parallelization permits larger numbers of measurements, executing a given set of circuits multiple times, and thus greater efficacy when employing RIIM. While the expected value in the lower panel cannot be better than the upper panel (the errors of the extra machines used for the additional data were all larger than the value of $\epsilon = 0.0104$ used in upper panel of Fig. 9), a significantly larger number of shots results in a much more stable result that often closer to the right answer than on the single best machine. Additionally, the entire RIIM protocol throughput increases, potentially reducing the time-to-solution by a factor of 14 (*i.e.*, the number of available systems) on real hardware.

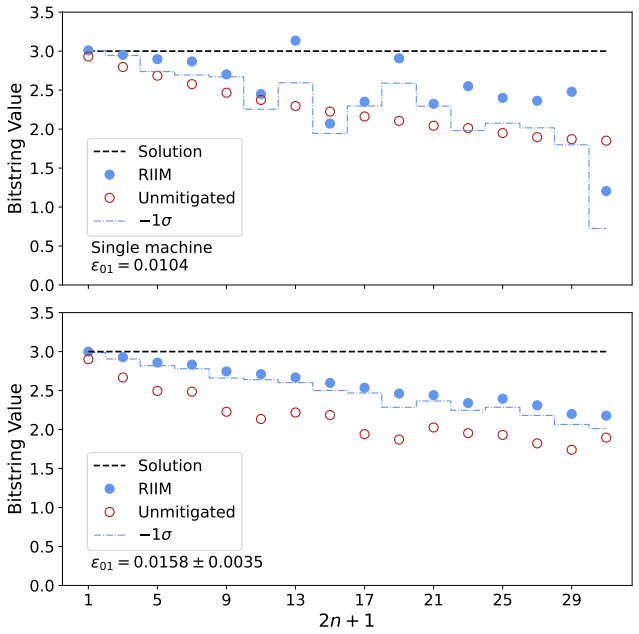

FIG. 9: Simulation results of the 4-CNOT circuit using a single-device (top) and multiple devices (bottom) with depolarizing gate noise. The top figure consists of a single machine executing each circuit 8192 times, and the bottom an arbitrary set of $4 \leq n_{\mathrm{NM}} \leq 14$ noise models executing each circuit $8192\, n_{\mathrm{NM}}$ times. The gate noise in both cases is taken from IBM Q backend properties data.

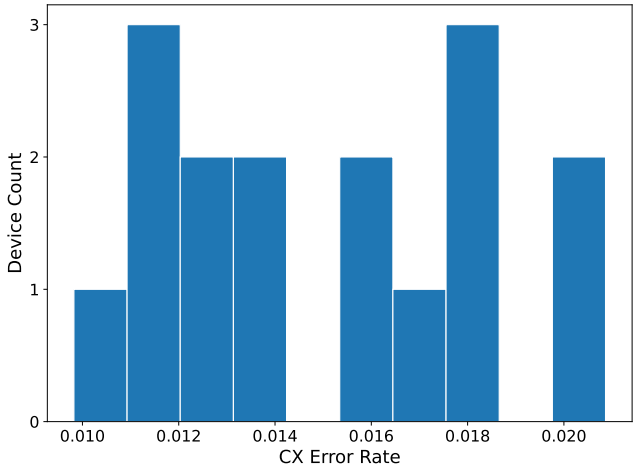

FIG. 10: Distribution of CNOT error rates applied in noisy simulations of the 4-CNOT circuit. Bins are centered at the error rate. Data obtained from IBM Q backend properties.

VI. CONCLUSIONS

Zero noise extrapolation is a critical technique for error mitigation and is being used for a variety of experimental demonstrations on near term quantum devices (see e.g. Ref. [15, 23]). While the core idea of ZNE is simple, there are many variations that can lead to improvements in resource usage and fidelity. In this paper, we have introduced multiple approaches to ZNE. First, we have developed a partial ZNE that applies to only a subset of qubits (List Identity Insertion Method). Second, we showed that it is possible to generalize the Fixed and Random Identity Insertion Methods (FIIM and RIIM) by replacing sets of qubits instead of all or single qubits. This Set Identity Insertion Method (SIIM) can trade-off

the demanding gate resources of FIIM for the demanding measurement resources of RIIM. Lastly, we studied the parallelization of RIIM in order to cope with the extensive measurement resources required to reach precision.

In this paper, we have focused on software and hardware agnostic methods for gate error mitigation. Noise is amplified by means of identity insertions. All of the methods introduced here could also be used with pulse-level gate lengthening for error amplification [24]. Additionally, the ZNE methods introduced here can be combined with other gate error mitigation methods as well as readout error mitigation approaches [25–43] for the ultimate error mitigation strategy.

ACKNOWLEDGMENTS

We thank Aniruddha Bapat and Plato Deliyannis for detailed feedback on the manuscript. This work is supported by the U.S. Department of Energy, Office of Science under contract DE-AC02-05CH11231. In particular, support comes from Quantum Information Science Enabled Discovery (QuantISED) for High Energy Physics (KA2401032) and the Office of Advanced Scientific Computing Research (ASCR) through the Accelerated Research for Quantum Computing Program. This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

[1] J. Preskill, Quantum Computing in the NISQ era and beyond, Quantum **2**, 79 (2018).
[2] B. M. Terhal, Quantum error correction for quantum memories, Rev. Mod. Phys. **87**, 307 (2015).
[3] L. F. Richardson and J. A. Gaunt, The deferred approach to the limit, part i, Philosophica Transactions of the RoyalSociety of London. Series A **226**, 636 (1927).
[4] 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**, 491–495 (2019).
[5] E. F. Dumitrescu, A. J. McCaskey, G. Hagen, G. R. Jansen, T. D. Morris, T. Papenbrock, R. C. Pooser, D. J. Dean, and P. Lougovski, Cloud quantum computing of an atomic nucleus, Phys. Rev. Lett. **120**, 210501 (2018).
[6] S. Endo, S. C. Benjamin, and Y. Li, Practical quantum error mitigation for near-future applications, Phys. Rev. X **8**, 031027 (2018).
[7] K. Temme, S. Bravyi, and J. M. Gambetta, Error mitigation for short-depth quantum circuits, Physical Review Letters **119**, 10.1103/physrevlett.119.180509 (2017).
[8] S. Endo, S. C. Benjamin, and Y. Li, Practical quantum error mitigation for near-future applications, Physical Review X **8**, 10.1103/physrevx.8.031027 (2018).
[9] W. A. d. J. A. He, B. Nachman and C. W. Bauer, Resource Efficient Zero Noise Extrapolation with Identity Insertions, Phys. Rev. A **102**, 10.1103/PhysRevA.102.012426 (2020), arXiv:2003.04941 [quant-ph].
[10] T. Giurgica-Tiron, Y. Hindy, R. LaRose, A. Mari, and W. J. Zeng, Digital zero noise extrapolation for quantum error mitigation, 2020 IEEE International Conference on Quantum Computing and Engineering (QCE) 10.1109/qce49297.2020.00045 (2020).
[11] Z. Cai, Multi-exponential error extrapolation and combining error mitigation techniques for nisq applications (2021), arXiv:2007.01265 [quant-ph].
[12] A. Strikis, D. Qin, Y. Chen, S. C. Benjamin, and Y. Li, Learning-based quantum error mitigation (2021), arXiv:2005.07601 [quant-ph].
[13] A. Zlokapa and A. Gheorghiu, A deep learning model for noise prediction on near-term quantum devices (2020), arXiv:2005.10811 [quant-ph].
[14] P. Czarnik, A. Arrasmith, P. J. Coles, and L. Cincio, Error mitigation with clifford quantum-circuit data (2021), arXiv:2005.10189 [quant-ph].
[15] M. Urbanek, B. Nachman, V. R. Pascuzzi, A. He, C. W. Bauer, and W. A. de Jong, Mitigating depolarizing noise on quantum computers with noise-estimation circuits (2021), arXiv:2103.08591 [quant-ph].
[16] A. Lowe, M. H. Gordon, P. Czarnik, A. Arrasmith, P. J. Coles, and L. Cincio, Unified approach to data-driven quantum error mitigation, Physical Review Research **3**, 10.1103/physrevresearch.3.033098 (2021).
[17] P. Czarnik, A. Arrasmith, L. Cincio, and P. J. Coles, Qubit-efficient exponential suppression of errors (2021), arXiv:2102.06056 [quant-ph].
[18] S. Zhang, Y. Lu, K. Zhang, W. Chen, Y. Li, J.-N. Zhang, and K. Kim, Error-mitigated quantum gates exceeding physical fidelities in a trapped-ion system, Nature Communications **11**, 10.1038/s41467-020-14376-z (2020).
[19] A. Mari, N. Shammah, and W. J. Zeng, Extending quantum probabilistic error cancellation by noise scaling (2021), arXiv:2108.02237 [quant-ph].
[20] S. Endo, Z. Cai, S. C. Benjamin, and X. Yuan, Hybrid quantum-classical algorithms and quantum error mitigation, Journal of the Physical Society of Japan **90**, 032001 (2021).
[21] Cirq, a python framework for creating, editing, and invoking noisy intermediate scale quantum (nisq) circuits, `https://github.com/quantumlib/Cirq`.
[22] G. Aleksandrowicz, T. Alexander, P. Barkoutsos, L. Bello, Y. Ben-Haim, D. Bucher, F. J. Cabrera-Hernández, J. Carballo-Franquis, A. Chen, C.-F. Chen, J. M. Chow, A. D. Córcoles-Gonzales, A. J. Cross, A. Cross, J. Cruz-Benito, C. Culver, S. D. L. P. González, E. D. L. Torre, D. Ding, E. Dumitrescu, I. Duran, P. Eendebak,

M. Everitt, I. F. Sertage, A. Frisch, A. Fuhrer, J. Gambetta, B. G. Gago, J. Gomez-Mosquera, D. Greenberg, I. Hamamura, V. Havlicek, J. Hellmers, Łukasz Herok, H. Horii, S. Hu, T. Imamichi, T. Itoko, A. Javadi-Abhari, N. Kanazawa, A. Karazeev, K. Krsulich, P. Liu, Y. Luh, Y. Maeng, M. Marques, F. J. Martín-Fernández, D. T. McClure, D. McKay, S. Meesala, A. Mezzacapo, N. Moll, D. M. Rodríguez, G. Nannicini, P. Nation, P. Ollitrault, L. J. O'Riordan, H. Paik, J. Pérez, A. Phan, M. Pistoia, V. Prutyanov, M. Reuter, J. Rice, A. R. Davila, R. H. P. Rudy, M. Ryu, N. Sathaye, C. Schnabel, E. Schoute, K. Setia, Y. Shi, A. Silva, Y. Siraichi, S. Sivarajah, J. A. Smolin, M. Soeken, H. Takahashi, I. Tavernelli, C. Taylor, P. Taylour, K. Trabing, M. Treinish, W. Turner, D. Vogt-Lee, C. Vuillot, J. A. Wildstrom, J. Wilson, E. Winston, C. Wood, S. Wood, S. Wörner, I. Y. Akhalwaya, and C. Zoufal, Qiskit: An Open-source Framework for Quantum Computing (2019).

[23] 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 (2021), arXiv:2108.09197 [quant-ph].

[24] 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).

[25] R. C. Bialczak, M. Ansmann, M. Hofheinz, E. Lucero, M. Neeley, A. D. O'Connell, D. Sank, H. Wang, J. Wenner, M. Steffen, A. N. Cleland, and J. M. Martinis, Quantum process tomography of a universal entangling gate implemented with Josephson phase qubits, Nature Physics **6**, 409 (2010).

[26] M. Neeley, R. C. Bialczak, M. Lenander, E. Lucero, M. Mariantoni, A. D. O'Connell, D. Sank, H. Wang, M. Weides, J. Wenner, Y. Yin, T. Yamamoto, A. N. Cleland, and J. M. Martinis, Generation of three-qubit entangled states using superconducting phase qubits, Nature **467**, 570 (2010).

[27] A. Dewes, F. R. Ong, V. Schmitt, R. Lauro, N. Boulant, P. Bertet, D. Vion, and D. Esteve, Characterization of a Two-Transmon Processor with Individual Single-Shot Qubit Readout, Physical Review Letters **108**, 057002 (2012).

[28] E. Magesan, J. M. Gambetta, A. Córcoles, and J. M. Chow, Machine Learning for Discriminating Quantum Measurement Trajectories and Improving Readout, Physical Review Letters **114**, 200501 (2015).

[29] S. Debnath, N. M. Linke, C. Figgatt, K. A. Landsman, K. Wright, and C. Monroe, Demonstration of a small programmable quantum computer with atomic qubits, Nature **536**, 63 (2016).

[30] C. Song, K. Xu, W. Liu, C.-p. Yang, S.-B. Zheng, H. Deng, Q. Xie, K. Huang, Q. Guo, L. Zhang, P. Zhang, D. Xu, D. Zheng, X. Zhu, H. Wang, Y.-A. Chen, C.-Y. Lu, S. Han, and J.-W. Pan, 10-Qubit Entanglement and Parallel Logic Operations with a Superconducting Circuit, Physical Review Letters **119**, 180511 (2017).

[31] M. Gong, M.-C. Chen, Y. Zheng, S. Wang, C. Zha, H. Deng, Z. Yan, H. Rong, Y. Wu, S. Li, F. Chen, Y. Zhao, F. Liang, J. Lin, Y. Xu, C. Guo, L. Sun, A. D. Castellano, H. Wang, C. Peng, C.-Y. Lu, X. Zhu, and J.-W. Pan, Genuine 12-qubit entanglement on a superconducting quantum processor, Physical Review Letters **122**, 110501 (2019), arXiv: 1811.02292.

[32] 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 GHZ States via Multiple Quantum Coherences, Physical Review A **101**, 032343 (2020), arXiv: 1905.05720.

[33] V. Havlicek, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum enhanced feature spaces, Nature **567**, 209 (2019), arXiv: 1804.11326.

[34] Y. Chen, M. Farahzad, S. Yoo, and T.-C. Wei, Detector Tomography on IBM 5-qubit Quantum Computers and Mitigation of Imperfect Measurement, Physical Review A **100**, 052315 (2019), arXiv: 1904.11935.

[35] M.-C. Chen, M. Gong, X.-S. Xu, X. Yuan, J.-W. Wang, C. Wang, C. Ying, J. Lin, Y. Xu, Y. Wu, S. Wang, H. Deng, F. Liang, C.-Z. Peng, S. C. Benjamin, X. Zhu, C.-Y. Lu, and J.-W. Pan, Demonstration of Adiabatic Variational Quantum Computing with a Superconducting Quantum Coprocessor, arXiv:1905.03150 [quant-ph] (2019), arXiv: 1905.03150.

[36] 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), arXiv: 1907.08518.

[37] M. Urbanek, B. Nachman, and W. A. de Jong, Quantum error detection improves accuracy of chemical calculations on a quantum computer, Physical Review A **102**, 022427 (2020), arXiv: 1910.00129.

[38] B. Nachman, M. Urbanek, W. A. de Jong, and C. W. Bauer, Unfolding Quantum Computer Readout Noise, arXiv:1910.01969 [physics, physics:quant-ph] (2020), arXiv: 1910.01969.

[39] K. E. Hamilton and R. C. Pooser, Error-mitigated data-driven circuit learning on noisy quantum hardware, arXiv:1911.13289 [quant-ph] (2019), arXiv: 1911.13289.

[40] P. J. Karalekas, N. A. Tezak, E. C. Peterson, C. A. Ryan, M. P. da Silva, and R. S. Smith, A quantum-classical cloud platform optimized for variational hybrid algorithms, Quantum Science and Technology **5**, 024003 (2020), arXiv: 2001.04449.

[41] M. R. Geller and M. Sun, Efficient correction of multiqubit measurement errors, arXiv:2001.09980 [quant-ph] (2020), arXiv: 2001.09980.

[42] M. R. Geller, Rigorous measurement error correction, Quantum Science and Technology **5**, 03LT01 (2020).

[43] C. B. R. Hicks and B. Nachman, Readout Rebalancing for Near Term Quantum Computers, Phys. Rev. A **103**, 022407 (2021), arXiv:2010.07496 [quant-ph].

Appendix: IBM Q Data and Code Availability

Tables IIa and IIb contain backend properties for each system used in this work. Note that five of the fourteen systems have been retired since these studies were performed. As such, we were unable to retrieve backend properties for the following systems: `ibmq_16_melbourne`, `ibmq_athens`, `ibmq_manhattan`, `ibmq_paris` and `ibmq_rome`.

System	T1 [μs] (Q0, Q1)	T2 [μs] (Q0, Q1)	Frequency [GHz] (Q0, Q1)	$P(0\|1)$ (Q0, Q1)	$P(1\|0)$ (Q0, Q1)
`ibmq_belem`	1.144[2], 9.928[1]	1.144[2], 9.928[1]	5.090, 5.246	2.720[-2], 3.660[-2]	6.400[-3], 8.000[-3]
`ibmq_bogota`	8.777[1], 8.468[1]	8.777[1], 8.468[1]	5.000, 4.850	2.220[-2], 8.040[-2]	1.000[-2], 2.040[-2]
`ibmq_casablanca`	1.202[2], 1.029[2]	1.202[2], 1.029[2]	4.822, 4.760	1.010[-1], 3.440[-2]	1.560[-2], 8.600[-3]
`ibmq_guadalupe`	1.258[2], 1.097[2]	1.258[2], 1.097[2]	5.113, 5.161	1.820[-2], 2.140[-2]	5.800[-3], 5.400[-3]
`ibmq_lima`	9.090[1], 9.864[1]	9.090[1], 9.864[1]	5.030, 5.128	3.480[-2], 4.460[-2]	8.400[-3], 8.600[-3]
`ibmq_manila`	1.847[2], 1.396[2]	1.847[2], 1.396[2]	4.963, 4.838	3.160[-2], 3.060[-2]	1.140[-2], 1.400[-2]
`ibmq_montreal`	1.019[2], 1.017[2]	1.019[2], 1.017[2]	4.911, 4.835	1.160[-2], 2.020[-2]	6.800[-3], 8.000[-3]
`ibmq_quito`	7.571[1], 9.799[1]	7.571[1], 9.799[1]	5.301, 5.081	6.300[-2], 3.360[-2]	1.820[-2], 1.140[-2]
`ibmq_santiago`	9.181[1], 7.148[1]	9.181[1], 7.148[1]	4.833, 4.624	2.940[-2], 1.460[-2]	6.800[-3], 1.020[-2]

(a) Qubit properties

System	X-Gate Error (Q0, Q1)	X-Gate Length [ns] ($X_{(0)} = X_{(1)}$)	CX-Gate Error ($\epsilon_{(0,1)} = \epsilon_{(1,0)}$)	CX-Gate Length [ns] ($CX_{(0,1)} = CX_{(1,0)}$)
`ibmq_belem`	1.975[-4], 2.538[-4]	3.556[1]	1.801[-2]	8.107[2]
`ibmq_bogota`	1.947[-4], 2.464[-4]	3.556[1]	1.334[-2]	6.898[2]
`ibmq_casablanca`	2.672[-4], 2.804[-4]	3.556[1]	1.885[-2]	7.609[2]
`ibmq_guadalupe`	2.104[-4], 3.436[-4]	3.556[1]	1.437[-2]	3.342[2]
`ibmq_lima`	2.508[-4], 1.870[-4]	3.556[1]	1.172[-2]	3.058[2]
`ibmq_manila`	1.658[-4], 2.491[-4]	3.556[1]	2.120[-2]	2.773[2]
`ibmq_montreal`	1.835[-4], 1.605[-4]	3.556[1]	1.473[-2]	3.840[2]
`ibmq_quito`	6.078[-4], 2.839[-4]	3.556[1]	1.194[-2]	2.347[2]
`ibmq_santiago`	4.634[-4], 2.028[-4]	3.556[1]	1.838[-2]	5.262[2]

(b) Gate properties

TABLE II: Backend properties of the systems used in the simulations presented in Sec. V. These data correspond to the calibration data at the time our experiments were executed. Numerals in square brackets represent power of 10. Five systems—`ibmq_16_melbourne`, `ibmq_athens`, `ibmq_manhattan`, `ibmq_paris` and `ibmq_rome`—have since been retired and therefore are not included in this table.

The source codes implementing the algorithms described in this paper will be available at: https://github.com/vrpascuzzi/computationally-efficient-zne.