Unifying and benchmarking state-of-the-art quantum error mitigation techniques

Daniel Bultrini[1,2], Max Hunter Gordon[3], Piotr Czarnik[1,4], Andrew Arrasmith[1,5], M. Cerezo[6,5], Patrick J. Coles[1,5], and Lukasz Cincio[1,5]

[1]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[2]Theoretische Chemie, Physikalisch-Chemisches Institut, Universität Heidelberg, INF 229, D-69120 Heidelberg, Germany
[3]Instituto de Física Teórica, UAM/CSIC, Universidad Autónoma de Madrid, Madrid, Spain
[4]Institute of Theoretical Physics, Jagiellonian University, Krakow, Poland.
[5]Quantum Science Center, Oak Ridge, TN 37931, USA
[6]Information Sciences, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

arXiv:2107.13470v2 [quant-ph] 22 May 2023

Error mitigation is an essential component of achieving a practical quantum advantage in the near term, and a number of different approaches have been proposed. In this work, we recognize that many state-of-the-art error mitigation methods share a common feature: they are data-driven, employing classical data obtained from runs of different quantum circuits. For example, Zero-noise extrapolation (ZNE) uses variable noise data and Clifford-data regression (CDR) uses data from near-Clifford circuits. We show that Virtual Distillation (VD) can be viewed in a similar manner by considering classical data produced from different numbers of state preparations. Observing this fact allows us to unify these three methods under a general data-driven error mitigation framework that we call UNIfied Technique for Error mitigation with Data (UNITED). In certain situations, we find that our UNITED method can outperform the individual methods (i.e., the whole is better than the individual parts). Specifically, we employ a realistic noise model obtained from a trapped ion quantum computer to benchmark UNITED, as well as other state-of-the-art methods, in mitigating observables produced from random quantum circuits and the Quantum Alternating Operator Ansatz (QAOA) applied to Max-Cut problems with various numbers of qubits, circuit depths and total numbers of shots. We find that the performance of different techniques depends strongly on shot budgets, with more powerful methods requiring more shots for optimal performance. For our largest considered shot budget (10^{10}), we find that UNITED gives the most accurate mitigation. Hence, our work represents a benchmarking of current error mitigation methods and provides a guide for the regimes when certain methods are most useful.

Daniel Bultrini: The first two authors contributed equally to this work.
Max Hunter Gordon: The first two authors contributed equally to this work.

1 Introduction

As new generations of quantum computers become larger and less noisy, the practical application of quantum algorithms is just around the corner. It is likely that any quantum algorithm that can achieve an advantage over its classical counterpart will have to contend with substantial hardware errors. While quantum error correction may eventually allow one to eliminate hardware noise, doing so will require many more qubits and significantly lower error rates. Although full error correction will not be practical in the near term, it is possible to partially mitigate errors with far lower resource requirements. A number of such error mitigation techniques have been proposed [6, 14], each with its own benefits and trade-offs.

One of the most popular approaches to error mitigation is zero noise extrapolation (ZNE) [45]. ZNE involves collecting data at different noise levels and fitting an extrapolation to the noiseless case. A variety of methods have been introduced to amplify the noise as well as to perform the extrapolation [13, 45]. ZNE has been shown to work well for small problem sizes [24] but has faced challenges for larger problems [10]. Due to the uncertainty of the extrapolation, ZNE does not have performance guarantees for regimes of high noise. This is particularly noticeable for circuits involving many gates. Here, the lowest available error points are likely to be too noisy for such fits to be helpful.

Alternatively, classically simulable circuits can be used to inform the experimenter about noise in a quantum device [10, 33, 43, 47, 48]. The Clifford Data Regression (CDR) method [10] makes use of the Gottesman-Knill theorem [18] which guarantees that

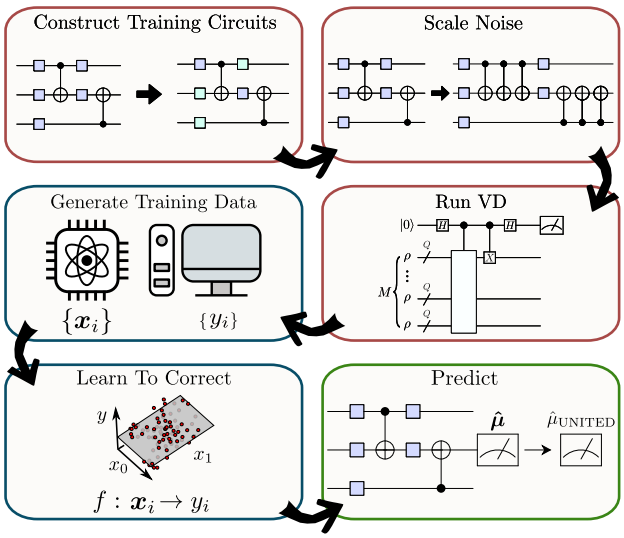

Figure 1: **Schematic summary of UNIfied Technique for Error mitigation with Data.** For a given circuit of interest near-Clifford training quantum circuits similar to the mitigated circuit are constructed. For these circuits, expectation values of the observable of interest are obtained using a noisy quantum computer of interest for multiple noise strengths. Furthermore, Virtual Distillation (VD) mitigated expectation values are obtained for different numbers of state copies. The expectation values of the training circuits are also evaluated classically. These expectation values are then used to fit an ansatz that maps the noisy and VD-mitigated expectation values to the exact expectation value. The trained ansatz is then applied to analogous noise-scaled and VD-mitigated expectation values produced by the circuit of interest to mitigate the effects of the noise improving on the VD-mitigated observables.

quantum circuits containing only Clifford gates can be classically simulated. CDR constructs a training set using data obtained from classically simulable near-Clifford circuits similar to the generic quantum circuit of interest. This set of data is then used to perform a regressive fit on an ansatz mapping noisy to exact expectation values. The fitted ansatz is used to estimate the value of the noiseless result. Furthermore, the combination of CDR and ZNE, which can outperform either method separately, has been considered and is called variable-noise CDR (vnCDR) [28]. vnCDR uses training data involving near-Clifford training circuits evaluated at multiple noise levels and can be thought of as using Clifford data to inform the extrapolation to the zero noise limit.

More recently, error mitigation via virtual distillation (VD) has been introduced [22, 26]. VD mitigates noise by simultaneously preparing multiple copies of a noisy state, entangling them, and making measurements to virtually prepare a state which is purer than the original noisy states. Further improvements have been made and analyses have been performed [5, 9, 23, 27]. While VD could achieve exponential error suppression in some regime [9, 22, 26], it has been noted that VD requires a number of shots that scales exponentially in the number of qubits used for sufficiently large noise rates [9].

Given the success of merging ZNE and CDR into vnCDR, in this paper, we consider whether or not it is beneficial to unify these other error mitigation methods with VD. By considering the number of copies used in VD to be a noise-control parameter, we use Clifford data to guide a combination of VD-mitigated observables, much like vnCDR. While also adding ZNE-like noise amplification we arrive at a framework unifying CDR, ZNE, and VD, which we call UNIfied Technique for Error mitigation with Data (UNITED). A summary of the steps used in UNITED is shown in Figure 1.

In order to get a realistic idea of how ZNE, vnCDR, VD, and UNITED compare we carry out a numerical study benchmarking the performance of each technique when mitigating a local observable produced by a random quantum circuit, and one term of a Max-Cut problem cost function for a circuit obtained when optimizing the Quantum Alternating Operator Ansatz (QAOA). We explore their performance on a simulator based on the current error rates of an ion-trap quantum computer. We compare performance across various system sizes and circuit depths, involving systems of up to 10 qubits and depths of up to 128 Molmer-Sørensøn entangling gate layers. Furthermore, we systematically compare the necessary shot cost for each method in its mitigation task, providing a fair comparison to guide experimenters in this field.

Overall, for the largest considered shot budget (10^{10}) our new unified method UNITED gives an advantage over all other methods considered in mitigating observables produced by both random quantum circuits (RQC) and QAOA applied to the Max-Cut problem. We observe that ZNE outperforms the learning-based methods for the 10^5 shot budget but fails to improve significantly when the total number of shots used is increased. This is in contrast to the learning-based methods, such as vnCDR and UNITED. These outperform ZNE when using 10^6 and 10^8 shots respectively. VD performs similarly to ZNE in the RQC task but performs significantly better in the QAOA task being the best method for $10^5, 10^6$, and 10^8 shot budgets. That is expected, as solutions of the Max-Cut problems are eigenstates of each term in the cost function which is beneficial to the performance of VD [27]. Nevertheless, for the highest shot budget we consider (10^{10}), UNITED improves on the best-performing mitigation methods for QAOA and the RQC experiments. These results suggest that the unified method successfully leverages the strengths of each of its constituent parts. Hence, we believe that is it crucial to explore such unified methods.

2 State-of-the-art error mitigation methods

2.1 Zero Noise Extrapolation

ZNE [45] uses noisy expectation values of an observable of interest X obtained from the quantum circuit of interest U run at increased noise to extrapolate to the noise-free limit. Originally, Richardson extrapolation [38] was proposed to estimate the noise-free observable [45]. The method was applied successfully to mitigate variational ground state quantum simulations with IBM quantum computers [12, 24].

Let us now explain the ZNE setting more formally. Here, the goal is to estimate the noiseless expectation value $\mu = \langle 0|^{\otimes Q} U^\dagger X U |0\rangle^{\otimes Q}$ from measurements on a noisy quantum device, where $|0\rangle^{\otimes Q}$ denotes the input state to the quantum circuit. In a realistic scenario, one cannot experimentally access μ due to the presence of hardware noise. Instead, one measures in the quantum computer the noisy observable $\hat{\mu}$. In the ZNE framework one usually assumes that the noise in the device can be characterized by a single parameter λ which is the noise strength. Moreover, while we assume that the noise cannot be decreased past a certain value λ_0, there always exist ways to further increase the noise strength (see below). Thus, in ZNE one wishes to estimate the noisy observable for increased noise levels, and then leverage extrapolation techniques to obtain the noiseless value. To perform the extrapolation, we measure the observable for $n+1$ noise strengths $\lambda_j = c_j \lambda_0$, where $1 = c_0 < c_1 < \cdots < c_n$ coefficients are noise levels. This allows us to create a dataset composed of expectation values estimated at different noise levels $\mathcal{T}^{\mathrm{ZNE}} = \{\hat{\mu}_j^{\mathrm{ZNE}}\}_{j=0}^n$. The Richardson extrapolation estimates the noise-free value μ via $\mathcal{T}^{\mathrm{ZNE}}$ as

$$\mu_n^{\mathrm{ZNE}} = \sum_{j=0}^{n} \gamma_j \hat{\mu}_j^{\mathrm{ZNE}}, \tag{1}$$

where

$$\sum_{j=0}^{n} \gamma_j = 1, \quad \sum_{j=0}^{n} \gamma_j c_j^k = 0 \ \text{ for } k = 1, \ldots, n. \tag{2}$$

Under the standard assumption, Richardson extrapolation approximates the true noiseless value μ with an error scaling as $|\mu^{\mathrm{ZNE}} - \mu| \in \mathcal{O}(\lambda_0^{n+1})$ [45], see Appendix A for details. Nevertheless, the Richardson extrapolation in the presence of shot noise leads to the exponential growth of the uncertainty in μ^{ZNE} with increasing n [17]. Richardson extrapolation is equivalent to fitting a polynomial of order n on the various noisy expectation values, treating the noise level c_j as an independent variable [21]. To address problems with the reliability of Richardson extrapolation for large n, lower order polynomial fits and exponential fits have been proposed [3, 12, 17]. Furthermore, approaches for noise characterized by multiple parameters have been postulated [36].

Regarding how one can increase the noise strength, the original proposal stretches pulses used to implement gates to increase the hardware noise [45]. Nevertheless, it is often more convenient to increase noise performing so-called identity insertions [12, 17, 21]. These methods insert groups of gates equivalent to the identity into the circuit of interest. This does not change a state encoded by the noiseless circuit U but increases the number of gates in the circuit, scaling the noise.

ZNE provides a very simple framework to mitigate errors introduced by noise on some observable of interest. However, there are several technical challenges that limit its power. Firstly, ZNE assumes that low enough noise levels are experimentally accessible to enable successful extrapolation to the zero noise limit. Furthermore, the best choice of the ansatz used to perform the extrapolation is not known a priori [17]. Also, a controlled increase of the noise strength is hardware dependent and may be challenging to implement in practice [24].

2.2 Clifford Data Regression

Let us now introduce the CDR framework which provides background for the more general vnCDR and UNITED methods. CDR [10] uses classically simulable training circuits consisting mainly of Clifford gates to learn the effects of the noise on an observable of interest. In particular, given the circuit of interest U, one constructs a set of circuits $\{V_i\}_{i=1}^{N_t}$, which are chosen to be similar to U. The simplest strategy to construct such a training set is to replace most of the non-Clifford gates in the circuit of interest with Clifford gates [10]. Here we note that to ensure classical simulability of the training circuits, they contain a small number N of non-Clifford gates. This is due to the fact that the cost of their classical simulation grows exponentially with N.

For each training circuit one computes via classical simulations [18] the noiseless expectation value of the observable of interest $\mu_i = \langle 0|^{\otimes Q} V_i^\dagger X V_i |0\rangle^{\otimes Q}$, as well as uses a quantum computer to estimate the noisy expectation value $\hat{\mu}_i^{\mathrm{CDR}}$ of X. Here, we will denote as $\mathcal{E}_{\mathrm{noiseless}}$ and $\mathcal{E}_{\mathrm{noisy}}$ the space of noiseless and noisy expectation values, respectively, so that $\mu_i \in \mathcal{E}_{\mathrm{noiseless}}$ and $\hat{\mu}_i^{\mathrm{CDR}} \in \mathcal{E}_{\mathrm{noisy}}$. Then, a dataset $\mathcal{T}^{\mathrm{CDR}} = \{(\mu_i, \hat{\mu}_i^{\mathrm{CDR}})\}_{i=1}^{N_t}$ is used to train a function $f^{\mathrm{CDR}} : \mathcal{E}_{\mathrm{noisy}} \to \mathcal{E}_{\mathrm{noiseless}}$, i.e., a function mapping the noisy to the exact expectation values. The simplest form for f^{CDR} which approximates a relation between the noisy and exact expectation values is

$$f^{\mathrm{CDR}}(x) = a_1 x + a_2\,, \tag{3}$$

Here, the values of parameters a_1, a_2 are found by least-squares linear regression on $\mathcal{T}^{\mathrm{CDR}}$. Once the

function is trained, CDR estimates the noiseless expectation value for U as $\mu^{\mathrm{CDR}} = f^{\mathrm{CDR}}(\hat{\mu})$. This method has been successfully applied to quantum simulations of condensed matter systems with IBM quantum computers [10, 41].

One feature of near-Clifford circuits constructed in this manner is the clustering of their exact expectation values around 0. This can lead to a large clustering of the expectation values in a near-Clifford training set which is detrimental to the shot efficiency of such an approach [11]. Therefore, to improve the shot efficiency, strategies have been proposed to ensure sufficient variance of the exact observable of interest among the training circuits [11]. Alternatively, the usage of Markov Chain Monte Carlo was proposed to construct training sets, where the training circuits have a low value of a cost function in the case of variational quantum algorithms [10].

The primary challenge in CDR is the construction of the near-Clifford training circuits as the optimal procedure to construct the training set is not known in general. Furthermore, the method requires evaluation of the observable of interest for the training circuits which may result in a large shot cost of the mitigation for large training sets.

2.3 Variable noise Clifford Data Regression

Variable noise Clifford Data Regression (vnCDR) conceptually unifies ZNE with CDR by using a training set involving near-Clifford circuits evaluated at several noise levels [28]. The dataset for vnCDR is constructed as follows. First, just like in CDR, one constructs a dataset of circuits $\{V_i\}_{i=1}^{N_t}$ which are chosen to be similar to U. Subsequently, one uses a classical computer to obtain the noiseless expectation value of the observable of interest μ_i. Then, for each V_i and each noise level c_j one uses a quantum computer to estimate the noisy expectation values $\hat{\mu}_{i,j}^{\mathrm{vnCDR}}$, just like in ZNE. This allows one to obtain a training dataset of the form $\mathcal{T}^{\mathrm{vnCDR}} = \{(\hat{\boldsymbol{\mu}}_i^{\mathrm{vnCDR}}, \mu_i)\}_{i=1}^{N_t}$, where $\hat{\boldsymbol{\mu}}_i^{\mathrm{vnCDR}} = (\hat{\mu}_{i,0}^{\mathrm{vnCDR}}, \ldots, \hat{\mu}_{i,n}^{\mathrm{vnCDR}})$. We note that for $n = 0$ one recovers the CDR training set $\mathcal{T}^{\mathrm{CDR}}$. The dataset $\mathcal{T}^{\mathrm{vnCDR}}$ is then used to fit a linear function $f^{\mathrm{vnCDR}} : (\mathcal{E}_{\mathrm{noisy}})^{(n+1)} \to \mathcal{E}_{\mathrm{noiseless}}$, of the form

$$f^{\mathrm{vnCDR}}(\boldsymbol{x}) = \sum_{j=0}^{n} a_j x_j \,. \tag{4}$$

This function is trained using least-squares regression which leads to the best-fit parameters a_j^*. These fitted parameters a_j^* are then used to obtain the vnCDR estimate of μ, as $\mu^{\mathrm{vnCDR}} = f^{\mathrm{vnCDR}}(\hat{\boldsymbol{\mu}}^{\mathrm{ZNE}})$. Here, we have defined $\hat{\boldsymbol{\mu}}^{\mathrm{ZNE}} = (\hat{\mu}_0^{\mathrm{ZNE}}, \ldots, \hat{\mu}_n^{\mathrm{ZNE}})$ as the vector comprised of the $n+1$ noisy expectations for the circuit of interest (that is, $\hat{\boldsymbol{\mu}}^{\mathrm{ZNE}}$ is a vector of the data in the ZNE dataset $\mathcal{T}^{\mathrm{ZNE}}$). The vnCDR method has been successfully implemented for quantum dynamics simulations with IBM quantum computers [41].

As we can see, vnCDR shares several common features with CDR and ZNE. The functional form of the ansatz is a linear combination of noisy estimates which resembles Richardson extrapolation. Therefore, vnCDR can be understood as performing an extrapolation guided by near-Clifford circuits with a similar structure to the circuit of interest. Nevertheless, the fitted ansatz does not require explicit knowledge of the noise levels, unlike the ZNE ansatzes, potentially easing ZNE requirements for precise control of the noise strength. Both CDR and vnCDR use training on data produced by near-Clifford quantum circuits. Consequently, vnCDR can also be treated as adding additional relevant data features to the CDR training set. We note that the ansatz for vnCDR does not include a constant term unlike that of CDR.

As in the case of CDR, the construction of an informative training set is the primary technical challenge. Furthermore, adding additional noise levels to the vnCDR training set increases the shot cost of the mitigation. Therefore, the best choice of noise levels is not known a priori.

2.4 Virtual Distillation

Virtual distillation (VD) [22, 26] takes a slightly different approach than ZNE and CDR. While the former suppresses error via post-processing, VD attempts to mitigate the effect of noise coherently. Specifically, let us denote by ρ the noisy state obtained at the output of the circuit of interest U. In VD, one uses $M > 1$ copies of a noisy state ρ to obtain expectation values of an observable of interest X for a purer, "distilled" state. Specifically, VD mitigates the expectation value of X by computing the quantity

$$\hat{\mu}^{\mathrm{VD}} = \frac{\mathrm{Tr}[\rho^M X]}{\mathrm{Tr}[\rho^M]}. \tag{5}$$

To better understand VD, consider the eigenvalue decomposition of ρ

$$\rho = \sum_{i=0}^{2^Q-1} p_i |\psi_i\rangle\langle\psi_i|, \tag{6}$$

with p_i decreasing with increasing i, and Q being the number of the qubits. Then,

$$\hat{\mu}^{\mathrm{VD}} = \frac{\langle\psi_0|X|\psi_0\rangle + \sum_{i=1}^{2^Q-1}(p_i/p_0)^M \langle\psi_i|X|\psi_i\rangle}{1 + \sum_{i=0}^{2^Q-1}(p_i/p_0)^M}. \tag{7}$$

Therefore, we can see that VD exponentially suppresses contributions to $\hat{\mu}^{\mathrm{VD}}$ from eigenvectors of ρ other than the dominant one $|\psi_0\rangle$. While increasing M suppresses the errors, for large error rates it also exponentially increases the shot cost of the estimation of $\hat{\mu}^{\mathrm{VD}}$ with a given accuracy [9].

VD implicitly assumes that the dominant eigenvector $|\psi_0\rangle$ is close to the noiseless state $|\psi_{\mathrm{exact}}\rangle =$

$U|0\rangle^{\otimes Q}$. Therefore, the quality of the VD correction is limited by a noise floor

$$\epsilon = \langle\psi_0|X|\psi_0\rangle - \langle\psi_{\text{exact}}|X|\psi_{\text{exact}}\rangle. \tag{8}$$

The noise floor can be bounded from above using a quantity called the coherent mismatch

$$c = 1 - |\langle\psi_{\text{exact}}|\psi_0\rangle|^2, \tag{9}$$

i.e.,

$$|\epsilon| \leq 2\sqrt{c}||X||_\infty, \tag{10}$$

where $||X||_\infty$ is the absolute value of the largest eigenvalue of X [27]. When $|\psi_0\rangle$ is an eigenvector of X, the bound (10) is replaced [27] by

$$|\epsilon| \leq 2c||X||_\infty \tag{11}$$

leading to better-than-expected performance of VD. For realistic noise, ϵ is usually larger than 0 [26]. Nevertheless, it was shown that for sufficiently small error rates the noise floor does not prevent VD from significantly reducing the impact of noise [27].

Here we recall that $\text{Tr}[\rho^M X]$ can be computed using M copies of the noisy state ρ and an ancillary qubit with a controlled permutation of the copies which changes the position of each copy (so-called controlled derangement). Consequently, $QM+1$ qubits are required to mitigate a Q-qubit state. This number can be reduced to $2Q+1$ by employing qubit resets and increasing the depth of the circuit performing the controlled derangement [9]. The controlled derangement can be implemented as a circuit of QM controlled swaps. Therefore, the derangement introduces errors to the estimates of $\text{Tr}[\rho^M X]$ and $\text{Tr}[\rho^M]$ which may require correction with other error mitigation methods [26].

Alternative approaches utilizing state verification to implement VD while further reducing required qubit counts were proposed recently [5, 23]. Furthermore, approaches conceptually unifying VD with both ZNE and subspace expansion techniques were proposed [4, 29, 51, 52].

3 Unifying variable noise Clifford Data Regression methods with Virtual Distillation

As we have discussed in the previous section, results obtained with CDR and vnCDR show that it is beneficial to train an error mitigation scheme with training data obtained from near-Clifford circuits, as well as from data arising from running circuits under different noise levels. Here, we propose to additionally use data from VD with multiple numbers of copies. This conceptually unifies VD, ZNE, and CDR into one mitigation method which we call UNIfied Technique for Error mitigation with Data (UNITED).

Technique	Noise Levels	Copies	Training Circuits
ZNE [45]	✓	✗	✗
VD [22, 26]	✗	✓	✗
CDR [10]	✗	✗	✓
vnCDR [28]	✓	✗	✓
CGVD [Here]	✗	✓	✓
UNITED [Here]	✓	✓	✓

Table 1: **Resource requirements for the different error mitigation strategies explored in this work.** In this table we show the necessary ingredients to employ the different error mitigation strategies that we unify and benchmark in this work.

Let us now explain how the dataset for UNITED is generated. Given a set of near-Clifford circuits similar to the circuit of interest U, $\{V_i\}_{i=1}^{N_t}$, we first use a classical computer to calculate the noiseless expectation values (denoted by μ_i) of the observable of interest X for each circuit. Then, we estimate the noisy and VD mitigated expectation values of X for different noise levels c_j with $j = 0, \ldots, n$ (as in ZNE), and for different numbers of copies $M = 1, \ldots, M_{\max}$. These estimates are made using a noisy quantum computer. Note that $M = 1$ labels the standard noisy estimates as they are obtained using a single copy. Putting all this together, we obtain a training set $\mathcal{T}^{\text{UNITED}} = \{(\hat{\boldsymbol{\mu}}_i^{\text{UNITED}}, \mu_i)\}_{i=1}^{N_t}$, where $\hat{\boldsymbol{\mu}}_i^{\text{UNITED}} = (\hat{\mu}_{i,0,1}, \ldots, \hat{\mu}_{i,n,M_{\max}})$ is a tensor of the noisy and VD mitigated estimates for V_i, and $\hat{\mu}_{i,j,M}$ is the expectation value obtained at noise level c_j and with M copies. We use $\mathcal{T}^{\text{UNITED}}$ to train a linear function $f^{\text{UNITED}} : (\mathcal{E}_{\text{VD}})^{M_{\max}(n+1)} \to \mathcal{E}_{\text{noiseless}}$ of the form

$$f^{\text{UNITED}}(\boldsymbol{x}) = \sum_{j=0}^{n} \sum_{M=1}^{M_{\max}} d_{j,M} x_{j,M}\,. \tag{12}$$

Here we use $\mathcal{E}_{\text{VD}}$ to denote a space of the VD-mitigated expectation values. For simplicity of notation, $\mathcal{E}_{\text{VD}}$ also includes the noisy expectation values ($M = 1$).

We use least-squares regression to find the values $d^*_{j,M}$ that correspond to the best possible fit of $\mathcal{T}^{\text{UNITED}}$. This fitted ansatz is then used to obtain an error-mitigated expectation value of X for the circuit of interest as $\mu^{\text{UNITED}} = f^{\text{UNITED}}(\hat{\boldsymbol{\mu}}^{\text{UNITED}})$. Here, $\hat{\boldsymbol{\mu}}^{\text{UNITED}} = (\hat{\mu}_{0,1}, \ldots, \hat{\mu}_{n,M_{\max}})$ is a tensor of the VD-mitigated and the noisy expectation values of X for U, where $\hat{\mu}_{j,M}$ is obtained at the noise level c_j with M copies. It is not hard to see that under this framework, vnCDR is a special case of UNITED. Indeed, for the special case of $M_{\max} = 1$ $\mathcal{T}^{\text{UNITED}}$ becomes the vnCDR dataset $\mathcal{T}^{\text{vnCDR}}$. For the reader's benefit, we summarize the similarities and differences between UNITED and other error mitigation methods in Table 1.

While we have presented the UNITED method, we still need to show that it is well-motivated. In what

follows we will study special cases of UNTIED that illustrate how its different moving parts help to mitigate errors.

3.1 Error mitigation from VD data with multiple copy numbers

In this section, we consider a special case of the ansatz used in the UNITED method where we do not use multiple noise levels ($n = 0$). That is, we show that one can use a combination of the VD mitigated observables with different numbers of copies as an estimate of the noiseless observable that has an error smaller than each VD mitigated observable. For this purpose, we consider a toy noise model where for each gate an error occurs with probability λ, and every different sequence of errors results in an orthogonal state. Additionally, at the end of the circuit we place a coherent error given by a unitary $U(\theta) = e^{-i\theta E}$, where E is a Hermitian operator. Such an error produces a non-zero noise floor for $\theta \neq 0$, while preserving the orthogonality of states corresponding to different sequences of the gate errors. We describe this noise model in more detail in Appendix B.2. Furthermore, in the following, we neglect the errors caused by an imperfect derangement operation and the shot noise.

First, let's consider taking a linear combination of VD-mitigated expectation values of X for U obtained using various M up to $M_{\max}$. We use such a linear combination to construct an estimate of μ that improves on all the VD mitigated estimates. That is, we take

$$\mu'_{M_{\max}} = \sum_{M=1}^{M_{\max}} b_M \hat{\mu}_M^{\mathrm{VD}} \,, \tag{13}$$

where b_M are some coefficients to be found. In Appendix B we show that for this noise model the orthogonal error scales as

$$|\mu_M^{\mathrm{VD}} - \mu - \epsilon| \in \mathcal{O}(\lambda^M) \,, \tag{14}$$

while one can find coefficients b_M (see Appendix B.2) which give

$$|\mu'_{M_{\max}} - \mu - \epsilon| \in \mathcal{O}(\lambda^{M_{\max}+1}) \,. \tag{15}$$

Therefore, $\mu'_{M_{\max}}$ can be used to suppress the orthogonal error beyond the error of the VD estimate with the largest number of copies. We note that the ansatz used here is similar to the one used by the Richardson extrapolation (1) with higher noise levels $c > 1$ replaced by larger copy numbers $M > 1$.

3.2 Error mitigation from VD data with multiple copy numbers guided by near-Clifford circuits

Here we consider a special case of UNITED which corresponds to taking $n = 0$ and $M_{\max} > 1$. This is equivalent to fitting the ansatz considered above (13) using a training set obtained from near-Clifford circuits. First, we start by noting that even in the case of the simple noise model considered in the previous section, one needs detailed knowledge of the noise model to determine the coefficients b_M (see Appendix B.2). However, such knowledge would require expensive process tomography in the case of a real device. Therefore, we propose to use data from near-Clifford quantum circuits to fit the coefficients b_M that lead to the desired orthogonal error suppression. We note that this corresponds to constructing a training set analogously to the case of vnCDR but replacing the $c > 1$ noisy expectation values with VD-mitigated ones. Furthermore, the same as in the case of vnCDR, we use a linear ansatz. We call this approach Clifford Guided Virtual Distillation (CGVD).

In more detail, the CGVD training set for an observable of interest X and a circuit of interest U is defined as $\mathcal{T}^{\mathrm{CGVD}} = \{(\hat{\boldsymbol{\mu}}_i^{\mathrm{CGVD}}, \mu_i)\}_{i=1}^{N_t}$, where $\hat{\boldsymbol{\mu}}_i^{\mathrm{CGVD}} = (\hat{\mu}_{i,1}^{\mathrm{CGVD}}, \ldots, \hat{\mu}_{i,M_{\max}}^{\mathrm{CGVD}})$ is a vector of the VD-mitigated and noisy expectation values of X for a near-Clifford circuit V_i, which are obtained at the base noise level $c = 1$ and with $M = 1, \ldots, M_{\max}$. The training set is fitted with a linear function $f^{\mathrm{CGVD}} : (\mathcal{E}_{\mathrm{VD}})^{M_{\max}} \to \mathcal{E}_{\mathrm{noiseless}}$, of the form

$$f^{\mathrm{CGVD}}(\boldsymbol{x}) = \sum_{M=1}^{M_{\max}} b_M x_M \,, \tag{16}$$

where the parameters b_M are found using least-squares regression. This leads to the best-fit parameters b_M^* which are then used to correct the noisy and VD-mitigated estimates of X expectation value for the circuit of interest, i.e. $\mu^{\mathrm{CGVD}} = f^{\mathrm{CDVD}}((\hat{\mu}, \ldots, \hat{\mu}_{M_{\max}}^{VD}))$. We note that CGVD perfectly mitigates global depolarizing noise as shown in Appendix C. Such perfect correction is possible as for this noise model the noise floor is zero.

3.3 UNITED ansatz for a simple noise model with orthogonal and coherent errors

In this section we demonstrate that by combining ZNE, VD and CDR, the UNITED method can suppress the error better than its individual parts. For this purpose, we consider again a simple noise model from Section 3.1. Let us construct for this model a dataset $\hat{\boldsymbol{\mu}}^{\mathrm{UNITED}}$, which is created using $n + 1$ noise levels and $M_{\max}$ copy numbers. For a noise level c_j we scale the noise strengths λ and θ as $\lambda_j = c_j \lambda_0$ and $\theta_j = c_j \theta_0$, where λ_0 and θ_0 are the base noise strengths. As described in detail in Appendix B.3, for each one of these noise levels we suppress the orthogonal gate errors with a combination of VD mitigated observables of form (13), i.e.

$$\mu'_{M_{\max},j} = \sum_{M=1}^{M_{\max}} b_M \hat{\mu}_{j,M} \,. \tag{17}$$

From Eq. (15) we obtain

$$|\mu'_{M_{\max},j} - \mu - \epsilon_j| \in \mathcal{O}(\lambda_j^{M_{\max}+1}), \tag{18}$$

where ϵ_j is the noise floor for the noise level c_j. As shown in Appendix B.3, for such obtained estimates $\{\mu'_{M_{\max},j}\}_{j=0}^{n}$ one can perform a Richardson extrapolation to additionally suppress the coherent error resulting in the noise floor.

By doing that one gets the final estimate of the expectation value of interest

$$\mu''_{M_{\max},n} = \sum_{j=0}^{n} \gamma_j \mu'_{M_{\max},j} = \sum_{j=0}^{n} \sum_{M=1}^{M_{\max}} \gamma_j b_M \hat{\mu}_{j,M}, \tag{19}$$

where γ_j coefficients are given by Eq. (2), that for large enough $M_{\max}$ has an error

$$|\mu''_{M_{\max},n} - \mu| \in \mathcal{O}(\theta_0^{n+1}). \tag{20}$$

For a more detailed explanation of the final error see Appendix B.3. This improved estimate has a form of the UNITED ansatz (12) motivating its choice. We note that, similarly to the case of CGVD, even in the case of our simple noise models determining the ansatz coefficients requires detailed knowledge of the noise model that is usually unavailable for real-word quantum computers. Therefore, in the UNITED methods we aim to find these coefficients by learning from near-Clifford training circuits.

3.4 General Remarks

It is important to note that although a combination of noisy observables inspired by Richardson extrapolation motivates the form of the ansatz in Eq. (12), it does not guarantee good performance of the new method for several reasons. Firstly, all methods to scale physical noise are only approximate so the noise levels will be imperfectly controlled, leading to imperfect extrapolated estimates. Nevertheless, successful applications of ZNE and vnCDR show that imperfect noise scaling control does not preclude successful error mitigation. Very importantly, training a model based on near-Clifford circuit data is not equivalent to performing the Richardson extrapolation, so the performance of UNITED depends on the ability to construct informative training sets. However, analogous to the case of vnCDR, we expect that fitting ansatz coefficients (12) instead of performing an extrapolation is beneficial as it does not require knowledge of noise levels, reducing the impact of imperfect noise control.

Overall, combining observables with different noise levels can be used to construct estimates with less error as shown by Richardson extrapolation. However, the current level of noise in quantum computers is significant. Due to the concentration of the expectation values of observables for high noise levels, a large number of shots is necessary to extract information from highly noisy data [42, 50]. This limits the range of the noise levels that can be used in practice. Therefore, at an intuitive level, using VD-mitigated observables is another way to provide data corresponding to different noise levels enhancing previous error mitigation methods using variable noise levels, like ZNE and vnCDR. From this perspective, the UNITED ansatz uses near-Clifford circuits to learn the best way to combine the various observables calculated under different noise levels.

Another important remark is that adding copy data to the training set further increases the shot cost of the mitigation. Therefore, the best choice of noise levels and the optimal number of copies is not known a priori. Despite this, we found that the method gives mitigated estimates, which remain stable with increasing n and $M_{\max}$ and that it either outperforms or matches vnCDR performance for sufficiently large shot numbers.

4 Numerical Results

In this section we present the performance of ZNE, vnCDR, VD, and UNITED applied to noisy outputs obtained from simulations of a trapped-ion quantum computer with a realistic noise model. We consider the task of mitigating the output of random quantum circuits and the output of the Quantum Alternating Operator Ansatz applied to Max-Cut problems. We use these two settings to benchmark the mitigation methods for various system sizes and different circuit depths. We study several total shot budgets per mitigated expectation value, highlighting the strengths and weaknesses of each method.

4.1 Random quantum circuits

We investigate random quantum circuits for various numbers of qubits $Q = 4, 6, 8, 10$ and several circuit depths. We consider random quantum circuits constructed using the native gates of the trapped-ion quantum computer we are modeling, namely of L layers of alternating nearest-neighbor Molmer-Sørensøn entangling gates decorated with general single qubit unitaries, see Figure 2 for details. Both the entangling gates and the single qubit unitaries are parametrized by randomly chosen angles. Here we simulate random quantum circuits with a number of layers $L = gQ$, taking $g = 1$ and $g = 16$.

We expect that for this setup, random quantum circuits will output states that converge to random states sampled according to Haar measure [32], with increasing L. Therefore, the setup is general and challenging, making it relevant for testing error mitigation methods. Furthermore, it is worth noting that deep circuits implemented in a trapped-ion quantum computer present a favorable setting for VD as this archi-

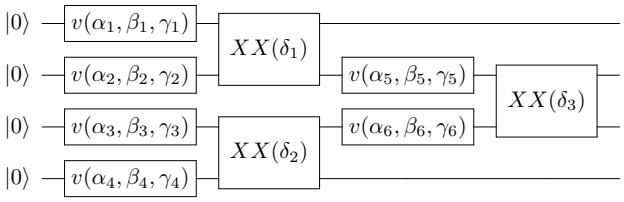

Figure 2: **Random quantum circuit structure used for our benchmark simulations.** Here we show a layer of an exemplary random quantum circuit for $Q = 4$. The layer is built from two layers of alternating nearest-neighbor Molmer-Sørensøn gates $XX(\delta) = e^{-i\delta\sigma_X^j\sigma_X^{j+1}}$ gates, where σ_X^j is a Pauli operator acting on qubit j. The XX gates are decorated with general single-qubit unitaries $v(\alpha,\beta,\gamma) = R_Z(\alpha)R_Y(\beta)R_Z(\gamma)$. Here $R_Z(\alpha) = e^{-i\alpha/2\sigma_Z}$, $R_Y(\beta) = e^{-i\beta/2\sigma_Y}$, and σ_Z, σ_Y are Pauli operators. We choose angles $\alpha,\beta,\gamma,\delta$ randomly for each random quantum circuit instance of the gates. We remark that R_Y, R_Z, XX are native gates of a trapped-ion quantum computer.

tecture enables all-to-all connectivity which decreases the depth of the controlled derangement circuit compiled to native gates [30]. Furthermore, it is expected that for sufficiently deep circuits and sufficiently small gate error rates, the coherent mismatch does not prevent a good quality mitigation with VD [27].

4.2 Random quantum circuit results

First, we analyze the case of $L = Q$ ($g = 1$). For this choice, we consider system sizes of $Q = 4, 6, 8, 10$. For each pair of L and Q values we consider 30 instances of random quantum circuits and for each random quantum circuit, we mitigate the expectation value of σ_Z acting on the first qubit of the circuit $\langle\sigma_Z^1\rangle$. We generate noisy data using a realistic trapped-ion computer noise model described in Appendix H. In general, deep RQCs will produce observables that are clustered around 0 and noise makes this clustering phenomenon even stronger [32, 50]. However, the RQCs we explore here are not sufficiently deep to observe this feature, see Appendix G for details.

We perform the mitigation with data obtained using a finite number of shots. For each mitigated expectation value and each error mitigation method, we assign the total number of shots used by the method to mitigate the expectation value. This number is defined as the total shot budget N_{tot}. We consider $N_{\text{tot}} = 10^5 - 10^{10}$.

Different error mitigation methods require estimates of observables from a different total number of circuits. In the case of ZNE, the circuit of interest needs to be evaluated for several different state preparation noise levels. In the case of vnCDR, estimates for the observable of interest are required for many training circuits evaluated at several noise levels. UNITED additionally requires performing VD mitigation for both the training circuits and the circuit of interest. Therefore, for the same shot budget, different mitigation methods have different numbers of shots per circuit used in the method. Consequently, for a given total number of shots, the accuracy of individual observable estimates decreases as the required total number of circuits for a given error mitigation method increases. We note that for fixed N_{tot} we can distribute shots between different observable evaluations in many ways. Here for the sake of simplicity for a given method, we assign the number of shots to be uniformly distributed across all the required circuits. That is the number of shots per circuit to run for a given mitigation method is the total number of shots divided by the total number of circuits necessary to run.

In Figure 3 we plot the mean and maximal absolute errors of $\langle\sigma_Z^1\rangle$. We find that the performance of the error mitigation methods depends strongly on the shot budget. The mitigated observables produced by ZNE and VD converge fastest with increasing N_{tot}. Furthermore, ZNE offers the best correction for the smallest total shot budget $N_{\text{tot}} = 10^5$. The convergence of vnCDR is slower but it outperforms ZNE and the other methods for $N_{\text{tot}} = 10^6$ up to $N_{\text{tot}} = 10^8$. We find that UNITED exhibits the slowest convergence with increasing N_{tot} but it gives the best results for $N_{\text{tot}} = 10^{10}$. The quality of the mitigated results decreases with increasing Q as the number of noisy gates increases. Nevertheless, even for the largest considered $Q = 10$, a factor of 20 improvement over the noisy result can be obtained with vnCDR and UNITED. The convergence of the results with N_{tot} is more explicitly shown for $g = 1$, $Q = 4$ in Figure 4(a).

Next, we analyze the case of $L = 16Q$ ($g = 16$). For this choice, we simulate $Q = 4, 6, 8$. Again, for each pair of L and Q values, we consider 30 instances of random quantum circuits and mitigate $\langle\sigma_Z^1\rangle$ for each of them. In this regime, the appearance of clustering of the exact observables, due to the properties of RQCs, becomes apparent in the largest system size, see Appendix G for details.

As in the case of $g = 1$, we perform the mitigation for $N_{\text{tot}} = 10^5 - 10^{10}$. For $Q = 6, L = 96$ and $Q = 8, L = 128$ we were unable to obtain improvement over the noisy results with any error mitigation method as the number of noisy gates is too large. For smaller systems of $Q = 4, L = 64$ we find that learning-based error mitigation methods outperform VD and ZNE even for $N_{\text{tot}} = 10^5$. For $N_{\text{tot}} = 10^7 - 10^{10}$ we find that both vnCDR and UNITED give a factor of $5 - 10$ improvement of the error in the noisy results. For $N_{\text{tot}} = 10^9 - 10^{10}$ we see an improvement of UNITED over vnCDR. Again we find that ZNE and VD estimates converge fastest with increasing N_{tot}. Again, the convergence of UNTIED is the slowest but it also provides the best mitigation quality for the largest shot budget. We show $g = 16$, $Q = 4$ convergence results in detail in Figure 4(b).

For this high noise regime, learning-based error mit-

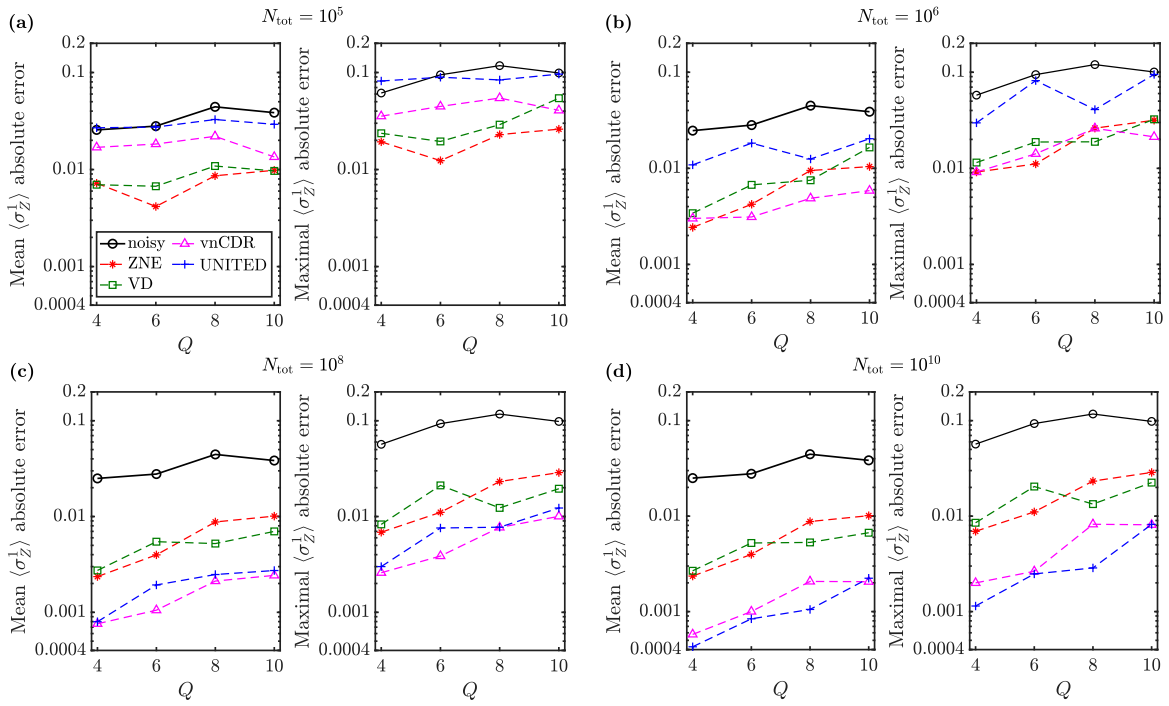

Figure 3: **Comparing error mitigation methods applied to RQCs for a number of layers $L = Q$.** We show the mean and maximal absolute errors of the expectation value of σ_Z measured at the first qubit of 30 RQC instances for the noisy estimates and estimates obtained with different error mitigation methods. In (a), the results for the total number of shots per mitigated (noisy) expectation value $N_{\rm tot} = 10^5$. In (b), $N_{\rm tot} = 10^6$. In (c), $N_{\rm tot} = 10^8$. In (d), $N_{\rm tot} = 10^{10}$. For $N_{\rm tot} = 10^5$, we find that ZNE gives the best results. For $N_{\rm tot} = 10^6$, 10^8 vnCDR gives the best results while for $N_{\rm tot} = 10^{10}$ UNITED gives the best results.

igation is particularly sensitive to shot noise [11] as noisy expectation values in the training set become more clustered with increasing noise strength. Different runs with different shot noise instances can lead to large differences in the mitigated observable. Therefore, one can expect that the behavior of the mean and maximal errors of the mitigated expectation values for a sample of RQCs are very sensitive to finite sample size effects. Consequently, the peak around $N_{\rm tot} = 10^8$ observed in the convergence of vnCDR with an increasing number of shots in Figure 4(b) is consistent with the expected behavior in this regime.

4.3 QAOA for Max-Cut problems

In this section, we investigate mitigating observables arising from the application of the Quantum Alternating Operator Ansatz (QAOA) [16, 19] to solve Max-Cut problems. We consider a graph $G = (V, E)$ with vertices in V, $|V| = Q$, and edges in E. The Max-Cut problem is to find a bipartition of the graph that maximizes the number of edges connecting the halves. It can be solved by minimizing the energy of a Hamiltonian

$$H_P = \sum_{(i,j)\in E} \sigma_Z^i \sigma_Z^j, \tag{21}$$

where we are summing over the edges. Here we consider Erdös-Renyi graphs [15] which connect each pair of nodes (i, j) with probability p_e. The QAOA ansatz is built with L layers

$$|\Psi(\boldsymbol{\alpha}, \boldsymbol{\beta})\rangle = e^{-i\beta_L H_M} e^{-i\alpha_L H_P} \ldots \\ \ldots e^{-i\beta_2 H_M} e^{-i\alpha_2 H_P} e^{-i\beta_1 H_M} e^{-i\alpha_1 H_P} |+\rangle \tag{22}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_L)$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_L)$ are parameters of the ansatz, $H_M = \sum_{i=1}^{Q} \sigma_X^i$, and $|+\rangle$ is a tensor product state of single-qubit $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ states.

Here we consider $Q = 4, 6, 8, 10$ and $L = Q$. For each Q we construct 28 Erdös-Renyi graphs, taking $p_e = 0.5$, obtained by the uniform sampling of their distribution. For each graph, we perform noiseless optimization of the Max-Cut cost function $\langle H_P\rangle$ to find optimal $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$. Then, we mitigate the expectation value $\langle\psi(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)| \sigma_Z^i \sigma_Z^j |\psi(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\rangle$ for one randomly selected $(i, j) \in E$. We compile $|\psi(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\rangle$ to the trapped-ion native gate set and obtain the noisy expectation value using the trapped-ion noisy simulator. Here we assume full device connectivity. Additional details on the compilation are given in Appendix I. For $Q = 10$ the compiled circuits have on average 96 layers of gates and 227 2-qubit Molmer-Sørensøn gates. We note that since the number of edges in Erdös-Renyi graphs is a random variable, the structure of the compiled circuits varies from instance to instance. Hence, we give average values of the depth

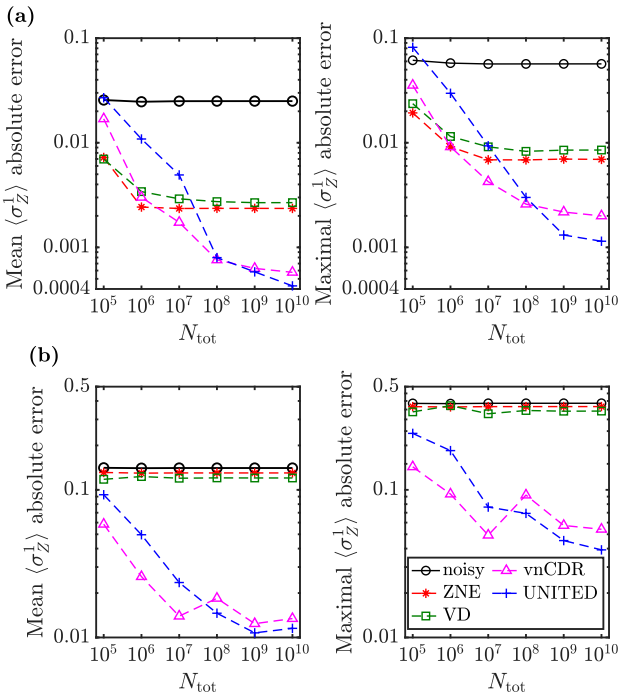

Figure 4: **Convergence of error mitigation methods applied to RQCs with increasing N_{tot}.** Mean and maximal absolute errors of the expectation value of $\langle\sigma_Z^1\rangle$ for random quantum circuits plotted versus the shot budget. We use 30 instances of random quantum circuits. In (a) we show results for $Q = 4, g = 1$. In (b) we display the results for $Q = 4, g = 16$, where the number of layers $L = gQ$. For $Q = 4, L = 4$ we find that ZNE and VD estimates converge first. The estimates produced by vnCDR converge slower but outperform those given by VD and ZNE for shot budgets of $10^7 - 10^{10}$. The convergence of UNITED is the slowest but it gives the best results for $N_{\text{tot}} = 10^9, 10^{10}$. For $Q = 4, L = 64$ estimates produced by ZNE and VD basically offer no mitigation and are outperformed by vnCDR and UNITED for all considered shot budgets. We see an improvement of UNITED over vnCDR for $N_{\text{tot}} = 10^9, 10^{10}$. A further discussion of the performance of VD for different shot budgets is explored in Appendix E.

and the gate count here.

4.4 QAOA results

We mitigate the expectation value of a random term of H_P for each instance of Erdös-Renyi graph. Once again, we employ the trapped-ion noise model described in Appendix H. In Figure 5 we show the mean and maximal absolute errors of the mitigated observables over the 28 graph instances. We again find (as in the case of random quantum circuits) that the performance of the error mitigation methods depends primarily on the shot budget. Interestingly, here we find that VD performs much better than in the case of random quantum circuits and we observe that it is the best-performing method for $N_{\text{tot}} = 10^5, 10^6, 10^8$ shot budgets. ZNE is the second best method for $N_{\text{tot}} = 10^5$ but fails to improve significantly for larger shot budgets. The learning-based methods improve systematically with increasing N_{tot}. For the largest considered $N_{\text{tot}} = 10^{10}$, UNITED is the best method for $Q = 6, 8, 10$ and vnCDR is the best for $Q = 4$.

Figure 6 analyzes the convergence of the various techniques for $Q = 8$ in more detail. Here VD is the best performer at the smallest shot budget ($N_{\text{tot}} = 10^5$), followed by ZNE, vnCDR, and UNITED. ZNE shows no improvement with an increased shot budget. UNITED, vnCDR, and VD, on the other hand, continue improving with UNITED becoming better than vnCDR at $N_{\text{tot}} = 10^8$ and vnCDR seeing no further improvement after this point. Similarly, VD sees continued improvements up to $N_{\text{tot}} = 10^8$, being the best technique up to that point. From $N_{\text{tot}} = 10^9$ and above, UNITED is the best method.

The reason for the better performance of VD when compared to the random quantum circuit case is likely due to the fact that the solutions to the Max-Cut problems are eigenstates of the mitigated observables. Therefore, we expect the noise floor to be reduced relative to the random quantum circuits (see (10) and (11)).

We note that the relative performance of VD with respect to vnCDR depends on the application. In particular, VD performance is better for QAOA than for the RQC experiments. This coincides with a larger improvement of UNITED over vnCDR for QAOA. Therefore, these two numerical studies highlight the utility of unifying error mitigation methods. Leveraging the strengths of ZNE, CDR, and VD into one method appears to allow for better performance across a range of tasks.

5 Implementation Details

5.1 Zero Noise Extrapolation

In the simulations described above, when implementing ZNE we use two noise levels $c \in \{1, 2\}$ and linear extrapolation. For the case of random quantum circuits, we compared this approach to exponential extrapolation (defined as in Ref. [17]) and linear extrapolation with $c \in \{1, 2, 3\}$ and found that the method we employ works best in general. We present details of this comparison in Appendix F.

When scaling the noise in our simulations we use factors $c = 2, 3$. To implement $c = 2$ we replace $XX(\theta)$ by $XX(\alpha)XX(\theta-\alpha)$ with α chosen randomly. We treat R_Z and R_Y in the same way. To implement $c = 3$ we replace $XX(\theta)$ by $XX(\theta)XX(\alpha)XX(-\alpha)$ performing an identity insertion. Again, we treat R_Z and R_Y in the same way. Both procedures scale the number of all types of gates in our random quantum circuits by the noise level.

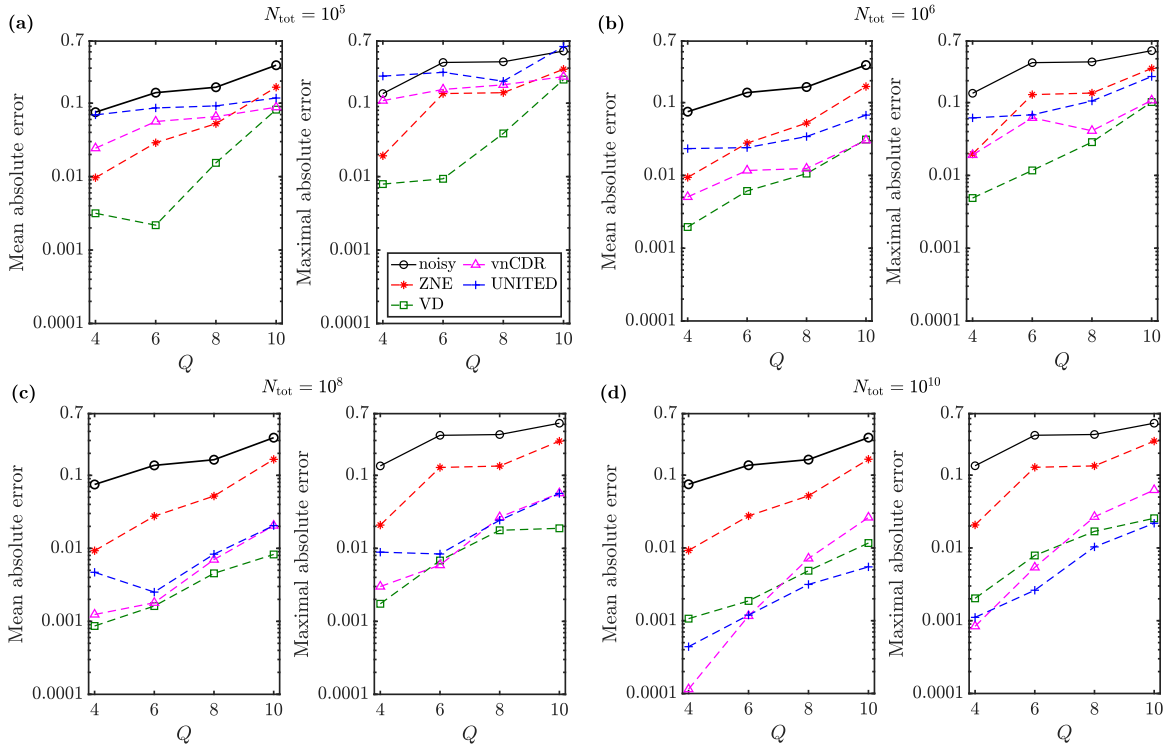

Figure 5: **Comparing error mitigation methods applied to Max-Cut QAOA.** Mean and maximal absolute errors of the expectation value of a randomly selected term of the Max-Cut Hamiltonian on 28 instances of Erdös-Renyi graphs. Here we compare the noisy estimates and the error-mitigated ones obtained with various error mitigation methods. In (a), the results for the total number of shots per mitigated (noisy) expectation value at $N_{\rm tot} = 10^5$. In (b) $N_{\rm tot} = 10^6$, (c) $N_{\rm tot} = 10^8$ and in (d) $N_{\rm tot} = 10^{10}$. For $N_{\rm tot} = 10^5,\ 10^6,\ 10^8$, VD gives the best result while for $N_{\rm tot} = 10^{10}$ UNITED gives the best results.

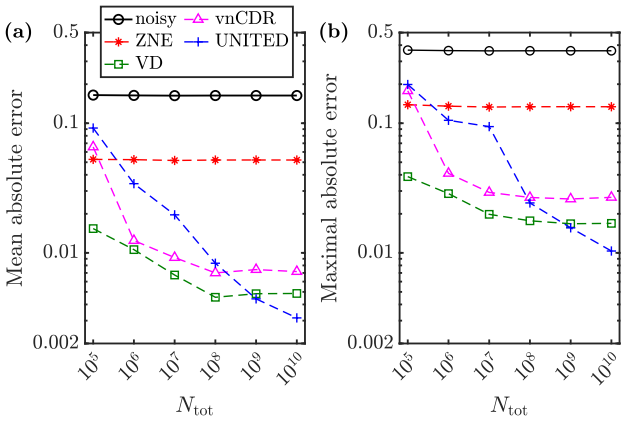

Figure 6: **Convergence of error mitigation methods applied to the Max-Cut QAOA for $Q = 8$ with increasing $N_{\rm tot}$.** Mean (a) and maximal (b) absolute errors of the expectation value of a random term of the Max-Cut Hamiltonian on 28 instances of Erdös-Renyi graphs. ZNE has already converged at $N_{\rm tot} = 10^5$, while VD is the best performer up to its convergence at $N_{\rm tot} = 10^8$, after which ($N_{\rm tot} = 10^9$) UNITED becomes the best performer. vnCDR converges at $N_{\rm tot} = 10^8$.

5.2 variable noise Clifford Data Regression

When constructing the noise-scaled circuits necessary for vnCDR we use the same method of noise scaling as used in the ZNE implementation. In our simulations above we use $c \in \{1, 2, 3\}$. For the case of random quantum circuits, we found that this choice of noise levels results in better results than using $c \in \{1, 2\}$. The method used to construct the near-Clifford training circuits is discussed in Section 5.5.

5.3 Virtual Distillation

We implement VD by computing $\frac{\mathrm{Tr}[\rho^M X]}{\mathrm{Tr}[\rho^M]}$ for the noisy state ρ and neglecting errors that occur during the derangement circuit execution. This assumption enables classical simulation of VD for a Q-qubit system while simulating Q noisy qubits instead of the $2Q+1$ that would otherwise be required to simulate the derangement. We analyze the performance of VD with derangement errors for $Q = 4$ systems in Appendix D. We present results of VD using 2 copies as we observed that this typically gave the best results. We explore the performance of VD for various numbers of copies in Appendix E.

5.4 Unified Data-driven Error Mitigation

To implement UNITED we choose the same near-Clifford circuits as used in vnCDR. The training set construction methods used to construct these circuits are discussed in Section 5.5. We choose $M_{\max} = 3$ and state preparation noise levels $c \in \{1, 2, 3\}$. We found that these parameters performed best across the regimes investigated for the RQC experiments.

5.5 Constructing the near-Clifford training circuits

To construct the near-Clifford training circuits for the random quantum circuits we use an algorithm from [9]. We replace XX and R_Y gates with their closest Clifford gates of the same type. We also replace a fraction of R_Z gates by their closest R_Z Clifford gates, leaving $N = 10$ R_Z gates intact.

In order to maximize the information contained within the training set while minimizing the number of shots used to evaluate their outputs on noisy hardware we inspect the distribution of the exact values produced by these circuits. Initially, we construct 100 near-Clifford circuits using the method described above. We find that with an increasing number of qubits, an increased fraction of their exact expectation values approaches 0. This clustering effect is a property of near-Clifford quantum circuits [11]. Shot efficiency of the learning-based error mitigation method can be improved by constructing a training set with significant variance in the exact expectation values, as explored in detail in [11]. Therefore, to construct the training set in this work, we post-select $N_t = 50$ near-Clifford circuits with the largest absolute value of the exact observable from the initial 100. We find that this strategy does not decrease the quality of the mitigation while decreasing the shot cost, see Appendix G for details.

However, the strategy outlined above is sub-optimal when applied to the Max-Cut QAOA problems, as for $Q = 10$ we see instances for which all 100 training circuits generated using this approach have the exact expectation value 0. Therefore, we use a more advanced training set construction strategy. This strategy is based on the method proposed in [11]. We use $N_t = 100$ training circuits with X^{exact} evenly distributed between -0.495 and 0.495, i.e. $X^{\text{exact}} \in \{-0.495 + 0.01n\}_{n=0}^{99}$. To construct these training circuits we start by building 10^5 near-Clifford circuits using the same gate replacement strategy as outlined above, including the same number of non-Clifford gates ($N = 10$). We then post-select these circuits in order to match the desired distribution of exact expectation values. We wish to reinforce the point that this process is scalable as near-Clifford circuits are classically simulable. While previous studies used Markov-Chain Monte Carlo sampling to generate such distribution [9], the simpler method of post-selection can also be used to find the desired distribution in certain cases.

5.6 Shot budgeting

Each technique described above requires a certain total number of circuits in order to mitigate the observable of interest. This, therefore, corresponds to a different number of observable evaluations. For VD, it is necessary to perform a measurement for 2 different circuits obtained with controlled derangement [26]. In the case of ZNE, one needs to measure the expectation value of $n + 1$ circuits built from one circuit of interest at $n + 1$ noise levels. For vnCDR, we additionally need to evaluate each training circuit for $n + 1$ noise levels. This means that in total vnCDR requires $(n + 1)(N_t + 1)$ circuit evaluations. Finally, for UNITED we need to mitigate the observable estimates using VD with $M_{\max} - 1$ different copies for the training circuits, the circuit of interest, and for each of $n + 1$ noise levels. Therefore, the number of circuit evaluations required by UNITED is $(n + 1)(N_t + 1)(2M_{\max} - 1)$.

In this work we consider several total shot budgets N_{tot}. For each method, we distribute the shots across measurements dividing N_{tot} by the number of measurements. We note that this strategy may not be optimal as it may be beneficial to distribute the shots unequally between noise levels or copy numbers. We note that one could imagine fine-tuning the values of n, N_t, $M_{\max}$ for a given N_{tot} but leave the systematic investigation of more advanced budgeting strategies for future work.

6 Discussion

The ability to efficiently and accurately mitigate hardware errors will be essential to achieving practical application of near-term quantum computers. A number of approaches have been proposed to fill this need, each bringing useful new ideas. In this work, we have discussed some state-of-the-art methods and examined how the approaches might best be combined. Specifically, we have considered Zero Noise Extrapolation (ZNE), variable noise Clifford Data Regression (vnCDR), and Virtual Distillation (VD) and proposed two new methods that represent unifications of the ideas behind these approaches. We applied these methods to the mitigation of errors in local observables measured from states prepared with random circuits and observables produced from applying QAOA to Max-Cut problems. As these methods have different advantages, disadvantages, and resource requirements, we evaluated their performance with $10^5 - 10^{10}$ total shots to mitigate the expectation value of an observable of interest with up to $Q = 10$ qubits with depth $L = Q$ as well as up to $Q = 8$ for depth $L = 16Q$. We used a realistic simulation of a noisy

trapped-ion device, which was chosen as the trapped-ion architecture enables all-to-all connectivity.

Combining the near-Clifford circuit data and the multi-copy data from VD, we formulated Clifford-guided VD (CGVD) which uses Clifford data to guide a combination of VD-mitigated observables to obtain an improved estimate. We then generalized this method to perform a Clifford-guided Richardson-like extrapolation on VD-mitigated data at variable noise levels. This can be thought of as first suppressing the effect of orthogonal errors, by using a combination of VD-mitigated observables, and then extrapolating to the zero noise limit. The new method, UNIfied Technique for Error mitigation with Data (UNITED), can be viewed as a unification of the CDR, ZNE, and VD methods. Furthermore, vnCDR is a special case of the new general method.

We find that UNITED is the technique that offers the best mitigation in both the RQC and QAOA numerical experiments at $N_{\text{tot}} = 10^{10}$. We also find that vnCDR is preferable at budgets $N_{\text{tot}} = 10^6, 10^8$ in the case of RQC. For $N_{\text{tot}} = 10^5$, ZNE is optimal for circuits with depths $L = Q$ while vnCDR is optimal for more noisy $L = 16Q$. When considering the QAOA results, VD is optimal for all shot budgets apart from the largest for which UNITED is the best method. This is because the noise floor is significantly lower in QAOA due to the fact that the solution of the algorithm is also an eigenstate of the measurement operator [27]. We observe that increasing the shot budget improves the error-mitigated expectation value until a performance ceiling is reached. Different techniques have different performance ceilings and we find that UNITED, being able to draw from more data than VD and ZNE, reaches better mitigated values when more resources are available.

We note it may be difficult to assess when the performance ceiling of a given method has been reached. This is an issue particularly for methods that lack analytical scaling guarantees, or when such guarantees only exist under restrictive noise assumptions. In this case, we propose using the convergence of the mitigated observable with the number of shots employed to judge if the performance ceiling has been reached.

As the available quantum devices continue to grow in qubit number and reduce the rate of hardware errors, practical applications of quantum computing grow ever closer. Due to the high number of qubits that are required for full error correction, it seems likely that error mitigation will play a pivotal part in those early applications. This work provides a useful reference for those who are seeking such applications to determine which combination of strategies is best for their use case.

A natural question one might ask is if there are other forms of data that would be useful to include in an error mitigation approach. Error mitigation for deep quantum circuits is challenging, but perhaps some other source of data (other than those we considered) would extend the depth for which error mitigation is practical. Further unification with other error mitigation techniques such as noise-resilient quantum compiling [7, 8, 25, 34, 39], quasi-probabilistic error decomposition [44, 45], verified phase estimation [35], truncated Neumann series [49] and specific approaches leveraging symmetries and/or post-selection [1, 4, 31, 37] is likely to be a promising research direction.

Another interesting direction for future work is to consider how best to determine when to measure new expectation values (i.e. more near-Clifford circuits, more noise levels, more copies of the state, etc.) versus expanding more shots per expectation value in order to get the best mitigation possible with fixed resources. Finally, we note that investigation of the effects of noisy controlled derangement on VD and UNITED performance including large circuits challenging for noisy classical simulation will be necessary for a full understanding of these methods' potential.

7 Acknowledgments

This work was supported by the Quantum Science Center (QSC), a National Quantum Information Science Research Center of the U.S. Department of Energy (DOE). DB was supported by the U.S. DOE through a quantum computing program sponsored by the Los Alamos National Laboratory (LANL) Information Science & Technology Institute and by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 955479. M.H.G was supported by "la Caixa" Foundation (ID100010434), Grant No. LCF/BQ/DI19/11730056. Piotr C. and AA were supported by the Laboratory Directed Research and Development (LDRD) program of LANL under project numbers 20190659PRD4 (Piotr C.) and 20210116DR (AA and Piotr. C). MC was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Computational Partnerships program. PJC also acknowledges initial support from the LANL ASC Beyond Moore's Law project. LC was also initially supported by the U.S. DOE, Office of Science, Office of Advanced Scientific Computing Research, under the Quantum Computing Application Teams program.

References

[1] Xavi Bonet-Monroig, Ramiro Sagastizabal, M Singh, and TE O'Brien. Low-cost error mitigation by symmetry verification. *Physical Review A*, 98(6):062339, 2018. DOI: https://doi.org/10.1103/PhysRevA.98.062339.

[2] Sergey Bravyi, Sarah Sheldon, Abhinav Kandala,

David C Mckay, and Jay M Gambetta. Mitigating measurement errors in multiqubit experiments. *Physical Review A*, 103(4):042605, 2021. DOI: https://doi.org/10.1103/PhysRevA.103.042605.

[3] Zhenyu Cai. Multi-exponential error extrapolation and combining error mitigation techniques for NISQ applications. *npj Quantum Information*, 7(1):1–12, 2021. DOI: https://doi.org/10.1038/s41534-021-00404-3.

[4] Zhenyu Cai. Quantum error mitigation using symmetry expansion. *Quantum*, 5:548, 2021. DOI: https://doi.org/10.22331/q-2021-09-21-548.

[5] Zhenyu Cai. Resource-efficient purification-based quantum error mitigation. *arXiv preprint arXiv:2107.07279*, 2021. URL https://arxiv.org/abs/2107.07279.

[6] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(1):625–644, 2021. DOI: https://doi.org/10.1038/s42254-021-00348-9.

[7] Lukasz Cincio, Yiğit Subaşı, Andrew T Sornborger, and Patrick J Coles. Learning the quantum algorithm for state overlap. *New Journal of Physics*, 20(11):113022, nov 2018. DOI: https://doi.org/10.1088/1367-2630/aae94a.

[8] Lukasz Cincio, Kenneth Rudinger, Mohan Sarovar, and Patrick J. Coles. Machine learning of noise-resilient quantum circuits. *PRX Quantum*, 2:010324, Feb 2021. DOI: https://doi.org/10.1103/PRXQuantum.2.010324.

[9] Piotr Czarnik, Andrew Arrasmith, Lukasz Cincio, and Patrick J Coles. Qubit-efficient exponential suppression of errors. *arXiv preprint arXiv:2102.06056*, 2021. URL https://arxiv.org/abs/2102.06056.

[10] Piotr Czarnik, Andrew Arrasmith, Patrick J. Coles, and Lukasz Cincio. Error mitigation with Clifford quantum-circuit data. *Quantum*, 5:592, November 2021. ISSN 2521-327X. DOI: https://doi.org/10.22331/q-2021-11-26-592.

[11] Piotr Czarnik, Michael McKerns, Andrew T Sornborger, and Lukasz Cincio. Improving the efficiency of learning-based error mitigation. *arXiv preprint arXiv:2204.07109*, 2022. URL https://arxiv.org/abs/2204.07109.

[12] Eugene F Dumitrescu, Alex J McCaskey, Gaute Hagen, Gustav R Jansen, Titus D Morris, T Papenbrock, Raphael C Pooser, David Jarvis Dean, and Pavel Lougovski. Cloud quantum computing of an atomic nucleus. *Phys. Rev. Lett.*, 120(21):210501, 2018. DOI: https://doi.org/10.1103/PhysRevLett.120.210501.

[13] Suguru Endo, Simon C Benjamin, and Ying Li. Practical quantum error mitigation for near-future applications. *Physical Review X*, 8(3):031027, 2018. DOI: https://doi.org/10.1103/PhysRevX.8.031027.

[14] Suguru Endo, Zhenyu Cai, Simon C Benjamin, and Xiao Yuan. Hybrid quantum-classical algorithms and quantum error mitigation. *Journal of the Physical Society of Japan*, 90(3):032001, 2021. DOI: https://doi.org/10.7566/JPSJ.90.032001.

[15] P Erdös and A Rényi. On random graphs i. *Publ. math. debrecen*, 6(290-297):18, 1959. URL http://snap.stanford.edu/class/cs224w-readings/erdos59random.pdf.

[16] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014. URL https://arxiv.org/abs/1411.4028.

[17] Tudor Giurgica-Tiron, Yousef Hindy, Ryan LaRose, Andrea Mari, and William J Zeng. Digital zero noise extrapolation for quantum error mitigation. *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pages 306–316, 2020. DOI: https://doi.org/10.1109/QCE49297.2020.00045.

[18] Daniel Gottesman. The heisenberg representation of quantum computers, talk at. In *International Conference on Group Theoretic Methods in Physics*. Citeseer, 1998. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.252.9446.

[19] Stuart Hadfield, Zhihui Wang, Bryan O'Gorman, Eleanor G Rieffel, Davide Venturelli, and Rupak Biswas. From the quantum approximate optimization algorithm to a quantum alternating operator ansatz. *Algorithms*, 12(2):34, 2019. DOI: https://doi.org/10.3390/a12020034.

[20] Kathleen E Hamilton, Tyler Kharazi, Titus Morris, Alexander J McCaskey, Ryan S Bennink, and Raphael C Pooser. Scalable quantum processor noise characterization. In *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pages 430–440. IEEE, 2020. DOI: https://doi.org/10.1109/QCE49297.2020.00060.

[21] Andre He, Benjamin Nachman, Wibe A. de Jong, and Christian W. Bauer. Zero-noise extrapolation for quantum-gate error mitigation with identity insertions. *Physical Review A*, 102:012426, Jul 2020. DOI: https://doi.org/10.1103/PhysRevA.102.012426.

[22] William J Huggins, Sam McArdle, Thomas E O'Brien, Joonho Lee, Nicholas C Rubin, Sergio Boixo, K Birgitta Whaley, Ryan Babbush, and Jarrod R McClean. Virtual distillation for quantum error mitigation. *Physical Review X*, 11(4):041036, 2021. DOI: https://doi.org/10.1103/PhysRevX.11.041036.

[23] Mingxia Huo and Ying Li. Dual-state purification for practical quantum error mitigation.

Physical Review A, 105(2):022427, 2022. DOI: https://doi.org/10.1103/PhysRevA.105.022427.

[24] Abhinav Kandala, Kristan Temme, Antonio D. Córcoles, Antonio Mezzacapo, Jerry M. Chow, and Jay M. Gambetta. Error mitigation extends the computational reach of a noisy quantum processor. *Nature*, 567(7749): 491–495, Mar 2019. ISSN 1476-4687. DOI: https://doi.org/10.1038/s41586-019-1040-7.

[25] Sumeet Khatri, Ryan LaRose, Alexander Poremba, Lukasz Cincio, Andrew T Sornborger, and Patrick J Coles. Quantum-assisted quantum compiling. *Quantum*, 3:140, 2019. DOI: https://doi.org/10.22331/q-2019-05-13-140.

[26] Bálint Koczor. Exponential error suppression for near-term quantum devices. *Physical Review X*, 11(3):031057, 2021. DOI: https://doi.org/10.1103/PhysRevX.11.031057.

[27] Bálint Koczor. The dominant eigenvector of a noisy quantum state. *New Journal of Physics*, 23(12):123047, 2021. DOI: https://doi.org/10.1088/1367-2630/ac37ae.

[28] Angus Lowe, Max Hunter Gordon, Piotr Czarnik, Andrew Arrasmith, Patrick J. Coles, and Lukasz Cincio. Unified approach to data-driven quantum error mitigation. *Phys. Rev. Research*, 3:033098, Jul 2021. DOI: https://doi.org/10.1103/PhysRevResearch.3.033098.

[29] Andrea Mari, Nathan Shammah, and William J Zeng. Extending quantum probabilistic error cancellation by noise scaling. *Physical Review A*, 104(5):052607, 2021. DOI: https://doi.org/10.1103/PhysRevA.104.052607.

[30] Dmitri Maslov. Basic circuit compilation techniques for an ion-trap quantum machine. *New Journal of Physics*, 19(2):023035, 2017. DOI: https://doi.org/10.1088/1367-2630/aa5e47.

[31] Sam McArdle, Xiao Yuan, and Simon Benjamin. Error-mitigated digital quantum simulation. *Phys. Rev. Lett.*, 122:180501, May 2019. DOI: https://doi.org/10.1103/PhysRevLett.122.180501.

[32] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1):1–6, 2018. DOI: https://doi.org/10.1038/s41467-018-07090-4.

[33] Ashley Montanaro and Stasja Stanisic. Error mitigation by training with fermionic linear optics. *arXiv preprint arXiv:2102.02120*, 2021. URL https://arxiv.org/abs/2102.02120.

[34] Prakash Murali, Jonathan M. Baker, Ali Javadi-Abhari, Frederic T. Chong, and Margaret Martonosi. Noise-adaptive compiler mappings for noisy intermediate-scale quantum computers. ASPLOS '19, page 1015–1029, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362405. DOI: https://doi.org/10.1145/3297858.3304075.

[35] Thomas E. O'Brien, Stefano Polla, Nicholas C. Rubin, William J. Huggins, Sam McArdle, Sergio Boixo, Jarrod R. McClean, and Ryan Babbush. Error mitigation via verified phase estimation. *PRX Quantum*, 2:020317, May 2021. DOI: https://doi.org/10.1103/PRXQuantum.2.020317.

[36] Matthew Otten and Stephen K Gray. Recovering noise-free quantum observables. *Physical Review A*, 99(1):012338, 2019. DOI: https://doi.org/10.1103/PhysRevA.99.012338.

[37] Matthew Otten, Cristian L Cortes, and Stephen K Gray. Noise-resilient quantum dynamics using symmetry-preserving ansatzes. *arXiv preprint arXiv:1910.06284*, 2019. URL https://arxiv.org/abs/1910.06284.

[38] Lewis Fry Richardson and J. Arthur Gaunt. VIII. the deferred approach to the limit. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 226(636-646):299–361, January 1927. DOI: https://doi.org/10.1098/rsta.1927.0008.

[39] Kunal Sharma, Sumeet Khatri, M. Cerezo, and Patrick J Coles. Noise resilience of variational quantum compiling. *New Journal of Physics*, 22(4):043006, 2020. DOI: https://doi.org/10.1088/1367-2630/ab784c.

[40] John A. Smolin and David P. DiVincenzo. Five two-bit quantum gates are sufficient to implement the quantum fredkin gate. *Physical Review A*, 53:2855–2856, 1996. DOI: https://doi.org/10.1103/PhysRevA.53.2855.

[41] Alejandro Sopena, Max Hunter Gordon, German Sierra, and Esperanza López. Simulating quench dynamics on a digital quantum computer with data-driven error mitigation. *Quantum Science and Technology*, 2021. DOI: https://doi.org/10.1088/2058-9565/ac0e7a.

[42] Daniel Stilck França and Raul Garcia-Patron. Limitations of optimization algorithms on noisy quantum devices. *Nature Physics*, 17(11):1221–1227, 2021. DOI: https://doi.org/10.1038/s41567-021-01356-3.

[43] Armands Strikis, Dayue Qin, Yanzhu Chen, Simon C Benjamin, and Ying Li. Learning-based quantum error mitigation. *PRX Quantum*, 2(4):040330, 2021. DOI: https://doi.org/10.1103/PRXQuantum.2.040330.

[44] Ryuji Takagi. Optimal resource cost for error mitigation. *Phys. Rev. Res.*, 3:033178, Aug 2021. DOI: https://doi.org/10.1103/PhysRevResearch.3.033178.

[45] Kristan Temme, Sergey Bravyi, and Jay M. Gambetta. Error mitigation for short-depth quantum circuits. *Phys.*

Rev. Lett., 119:180509, Nov 2017. DOI: https://doi.org/10.1103/PhysRevLett.119.180509.

[46] Colin J Trout, Muyuan Li, Mauricio Gutiérrez, Yukai Wu, Sheng-Tao Wang, Luming Duan, and Kenneth R Brown. Simulating the performance of a distance-3 surface code in a linear ion trap. *New Journal of Physics*, 20(4):043038, 2018. DOI: https://doi.org/10.1088/1367-2630/aab341.

[47] Miroslav Urbanek, Benjamin Nachman, Vincent R Pascuzzi, Andre He, Christian W Bauer, and Wibe A de Jong. Mitigating depolarizing noise on quantum computers with noise-estimation circuits. *Phys. Rev. Lett.*, 127(27):270502, 2021. DOI: https://doi.org/10.1103/PhysRevLett.127.270502.

[48] Joseph Vovrosh, Kiran E Khosla, Sean Greenaway, Christopher Self, Myungshik S Kim, and Johannes Knolle. Simple mitigation of global depolarizing errors in quantum simulations. *Physical Review E*, 104(3):035309, 2021. DOI: 10.1103/PhysRevE.104.035309.

[49] Kun Wang, Yu-Ao Chen, and Xin Wang. Mitigating quantum errors via truncated neumann series. *arXiv preprint arXiv:2111.00691*, 2021. URL https://arxiv.org/abs/2111.00691.

[50] Samson Wang, Enrico Fontana, M. Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. Noise-induced barren plateaus in variational quantum algorithms. *Nature Communications*, 12(1):1–11, 2021. DOI: https://doi.org/10.1038/s41467-021-27045-6.

[51] Yifeng Xiong, Soon Xin Ng, and Lajos Hanzo. Quantum error mitigation relying on permutation filtering. *IEEE Transactions on Communications*, 70(3):1927–1942, 2022. DOI: https://doi.org/10.1109/TCOMM.2021.3132914.

[52] Nobuyuki Yoshioka, Hideaki Hakoshima, Yuichiro Matsuzaki, Yuuki Tokunaga, Yasunari Suzuki, and Suguru Endo. Generalized quantum subspace expansion. *Phys. Rev. Lett.*, 129:020502, Jul 2022. DOI: https://doi.org/10.1103/PhysRevLett.129.020502.

A Cancellation of errors using Richardson extrapolation

By using Richardson extrapolation one can cancel the leading orders of a Taylor expansion of a noisy observable expectation value $\hat{\mu}^{\mathrm{ZNE}}(\lambda)$ to suppress its error. The order to which one can suppress the error depends on how many noisy observables one has access to. Each observable used must be evaluated with a different noise strength, which should be known. Here λ is the noise strength. More precisely, Richardson extrapolation uses a combination of $n+1$ noisy expectation values of the mitigated observable $\hat{\mu}_0^{\mathrm{ZNE}}, \hat{\mu}_1^{\mathrm{ZNE}}, \ldots, \hat{\mu}_n^{\mathrm{ZNE}}$ with the noise strengths $\lambda_0 < \lambda_1 < \cdots < \lambda_n$ to suppress the error up to $\mathcal{O}(\lambda_0^{n+1})$. Writing the series expansion of the noisy observable as a function of λ we have

$$\hat{\mu}(\lambda) = \mu + \sum_{i=1}^{\infty} \alpha_i \lambda^i, \tag{23}$$

where μ is the exact expectation value and α_i are the expansion coefficients. We introduce the noise levels $c_0, c_1, c_2, \ldots, c_n$ such that $\lambda_0, \lambda_1, \lambda_2, ..., \lambda_n = c_0\lambda_0, c_1\lambda_0, c_2\lambda_0, ..., c_n\lambda_0$, which enables us to write

$$\hat{\mu}_j^{\mathrm{ZNE}} = \hat{\mu}(c_j\lambda_0) = \mu + \sum_{i=1}^{\infty} \alpha_i c_j^i \lambda_0^i, \tag{24}$$

To build intuition, we first consider the combination of just two noisy observables, $\hat{\mu}_0^{\mathrm{ZNE}}$ and $\hat{\mu}_1^{\mathrm{ZNE}}$, evaluated at noise levels c_0 and c_1, respectively. We show how by using Richardson extrapolation we can cancel the error in an estimate of μ up to order $\mathcal{O}(\lambda_0^2)$. To begin with, we write the series expansion for $\hat{\mu}_0^{\mathrm{ZNE}}$ and $\hat{\mu}_1^{\mathrm{ZNE}}$ as

$$\hat{\mu}_0^{\mathrm{ZNE}} = \mu + \alpha_0\lambda_0 + \mathcal{O}(\lambda_0^2), \tag{25}$$
$$\hat{\mu}_1^{\mathrm{ZNE}} = \mu + \alpha_0 c_1\lambda_0 + \mathcal{O}(\lambda_0^2). \tag{26}$$

Note that we use $c_0 = 1$ above. Therefore, combining $\hat{\mu}_0^{\mathrm{ZNE}}$ and $\hat{\mu}_1^{\mathrm{ZNE}}$ with coefficients γ_0 and γ_1 we write

$$\begin{aligned}\mu_1^{\mathrm{ZNE}} &= \gamma_0\hat{\mu}_0 + \gamma_1\hat{\mu}_1 \\ &= (\gamma_0+\gamma_1)\mu + \alpha_0\lambda_0(\gamma_0 + c_1\gamma_1) + \mathcal{O}(\lambda_0^2),\end{aligned} \tag{27}$$

Therefore, to cancel terms up to order $\mathcal{O}(\lambda_0^2)$ we require:

$$\gamma_0 + \gamma_1 = 1, \tag{28}$$
$$\gamma_0 + c_1\gamma_1 = 0. \tag{29}$$

One can generalize this procedure to larger numbers of noise levels. In such a case we take a combination of $n+1$ $\hat{\mu}_j^{\mathrm{ZNE}}$ observables

$$\mu_n^{\mathrm{ZNE}} = \sum_{j=0}^{n} \gamma_j \hat{\mu}_j^{\mathrm{ZNE}}, \tag{30}$$

with γ_j being the coefficients. Expanding this sum in powers of λ_j gives

$$\begin{aligned}\mu_n^{\mathrm{ZNE}} &= \sum_{j=0}^{n}\gamma_j\mu + \sum_{j=0}^{n}\gamma_j\sum_{i=1}^{\infty}\alpha_i\lambda_j^i \\ &= \sum_{j=0}^{n}\gamma_j\mu + \sum_{i=1}^{\infty}\alpha_i\sum_{j=0}^{n}\gamma_j\lambda_j^i.\end{aligned} \tag{31}$$

Using the noise levels we can write above as

$$\mu_n^{\mathrm{ZNE}} = \sum_{j=0}^{n}\gamma_j\mu + \sum_{i=1}^{\infty}\alpha_i\lambda_0^i\sum_{j=0}^{n}\gamma_j c_j^i. \tag{32}$$

Therefore, to obtain the error of order $\mathcal{O}(\lambda_0^{n+1})$ the Richardson extrapolation chooses γ_j coefficients to cancel the leading n orders of the error in (32). This leads to a system of linear equations for γ_j

$$\sum_{j=0}^{n}\gamma_j = 1, \quad \sum_{j=0}^{n}\gamma_j c_j^i = 0 \ \text{ for } i = 1,\dots,n. \tag{33}$$

To sum up, the Richardson extrapolation uses a combination of different noisy estimates with different noise levels to mitigate the error by canceling the leading error terms.

B Suppressing error through a combination of Virtual-Distillation-mitigated observables

Here we consider a combination of VD-mitigated observables $\hat{\mu}_M^{\mathrm{VD}}$ obtained for different numbers of VD copies $M = 1, 2, \dots, M_{\max}$. By analogy to the Richardson extrapolation, we look for a combination

$$\mu'_{M_{\max}} = \sum_{M=1}^{M_{\max}} b_M \hat{\mu}_M^{\mathrm{VD}} \tag{34}$$

with the error smaller than the error of each of the individual observables. As proved below, Virtual Distillation suppresses contributions to the observable from eigenstates of the noisy state other than the dominant one leading to

$$|\hat{\mu}_M^{\mathrm{VD}} - \mu - \epsilon| \in \mathcal{O}(\xi^M), \tag{35}$$

where μ is the exact observable, ϵ is the noise floor defined in (8) and repeated below for convenience

$$\epsilon = \langle\psi_0|X|\psi_0\rangle - \langle\psi_{\mathrm{exact}}|X|\psi_{\mathrm{exact}}\rangle, \tag{36}$$

and

$$\xi = \sum_{i=1}^{2^Q-1} p_i = 1 - p_0 \tag{37}$$

is a sum of eigenvalues of the mitigated state (6) other than the dominant one.

B.1 Proof of VD error scaling with the number of state copies

To prove (35) we use the eigendecomposition of a noisy state ρ given by Eq. (6). Then we have

$$\hat{\mu}_M^{\mathrm{VD}} = \frac{\mathrm{Tr}[\rho^M X]}{\mathrm{Tr}[\rho^M]} = \frac{\sum_{i=0}^{2^Q-1} p_i^M X_i}{\sum_{i=0}^{2^Q-1} p_i^M}, \tag{38}$$

where $X_i = \langle\psi_i|X|\psi_i\rangle$. Transforming it further, we obtain

$$\begin{aligned}\hat{\mu}_M^{\mathrm{VD}} &= \frac{p_0^M X_0 + \sum_{i=1}^{2^Q-1} p_i^M X_i}{p_0^M + \sum_{i=1}^{2^Q-1} p_i^M} \\ &= X_0 + \frac{\sum_{i=1}^{2^Q-1} p_i^M (X_i - X_0)}{p_0^M + \sum_{i=1}^{2^Q-1} p_i^M}.\end{aligned} \tag{39}$$

Therefore, we obtain

$$|\hat{\mu}_M^{\mathrm{VD}} - X_0| = \frac{\sum_{i=1}^{2^Q-1} p_i^M |X_i - X_0|}{p_0^M + \sum_{i=1}^{2^Q-1} p_i^M} \le \frac{2\|X\|_\infty \sum_{i=1}^{2^Q-1} p_i^M}{p_0^M + \sum_{i=1}^{2^Q-1} p_i^M}, \tag{40}$$

where we use that $p_i \ge 0$, and $\|X\|_\infty$ is an infinity norm of X. But,

$$\frac{\sum_{i=1}^{2^Q-1} p_i^M}{p_0^M + \sum_{i=1}^{2^Q-1} p_i^M} \le \frac{\sum_{i=1}^{2^Q-1} p_i^M}{p_0^M} \le \frac{(\sum_{i=1}^{2^Q-1} p_i)^M}{p_0^M}, \tag{41}$$

where we again use that $p_i \ge 0$. Using (40), (41), and (37), this leads to

$$|\hat{\mu}_M^{\mathrm{VD}} - X_0| \le \frac{2\|X\|_\infty \xi^M}{(1-\xi)^M}. \tag{42}$$

Therefore, if $\|X\|_\infty$ is finite, like for a Pauli string, we obtain

$$|\hat{\mu}_M^{\mathrm{VD}} - X_0| = |\hat{\mu}_M^{\mathrm{VD}} - \mu - \epsilon| \in \mathcal{O}(\xi^M), \tag{43}$$

that was to be proved.

B.2 Error suppression with a combination of VD terms

Here, to motivate the CGVD ansatz (16) - which is a linear combination of VD-mitigated observables with up to $M_{\max}$ copies - we find a combination of VD-mitigated observables that estimates the noise-free observable of interest with an error smaller than its parts. We start with $M_{\max} = 2$ and then generalize this approach to more copies.

The noise model we consider is a simple model where for each gate an error occurs with probability λ, and every two different sequences of errors result in orthogonal states. Additionally, at the end of the circuit we place a coherent error given by a unitary $U(\theta) = e^{-i\theta E}$, where E is a Hermitian operator. Such an error produces a non-zero noise floor for $\theta \neq 0$, while preserving the orthogonality of states corresponding to different sequences of the gate errors. For a circuit with G gates the noisy output ρ can be written as

$$\begin{aligned}\rho = (1-\lambda)^G \rho_0 + (1-\lambda)^{G-1}\lambda \sum_{j=1}^{G} \rho_j \\ + (1-\lambda)^{G-2}\lambda^2 \sum_{j_1 \neq j_2} \rho_{j_1,j_2} + \cdots,\end{aligned} \tag{44}$$

where every state is orthogonal. For small enough error rate, ρ_0 is the dominant eigenvector, i.e. $\rho_0 = |\psi_0\rangle\langle\psi_0|$. We consider taking linear combinations of observables mitigated with VD using different numbers of copies up to $M = M_{\max}$. That is,

$$\mu'_{M_{\max}} = \sum_{M=1}^{M_{\max}} b_M \hat{\mu}_M^{\mathrm{VD}} \,, \tag{45}$$

where b_M are some coefficients to be found. We show that this ansatz contains solutions that lead to an estimate of the observable of interest with a smaller error than $\hat{\mu}_{M_{\max}}^{\mathrm{VD}}$.

B.2.1 Combining two VD-mitigated observables

To build intuition we first consider the combination of VD-mitigated observables with $M = 1$ and $M = 2$ copies. We can write

$$\begin{aligned}
\hat{\mu}_1^{\mathrm{VD}} &= \frac{\mathrm{Tr}[\rho X]}{\mathrm{Tr}[\rho]} \\
&= (1-\lambda)^G X_0 + (1-\lambda)^{G-1}\lambda \sum_{j=1}^{G} \mathrm{Tr}[\rho_j X] + \cdots \\
&= \big[1 - G\lambda + 1/2(G-1)G\lambda^2 + \cdots\big] X_0 \\
&\quad + \big[1 - (G-1)\lambda + \\
&\quad + 1/2(G-2)(G-1)\lambda^2 + \cdots\big]\lambda \sum_{j=1}^{G} \mathrm{Tr}[\rho_j X] \\
&\quad + \cdots .
\end{aligned} \tag{46}$$

By grouping terms according to powers of λ we rewrite this as

$$\hat{\mu}_1^{\mathrm{VD}} = X_0 + \lambda\alpha_1^{(1)} + \lambda^2\alpha_2^{(1)} + \lambda^3\alpha_3^{(1)} + \cdots \,, \tag{47}$$

where we have defined $\alpha_1^{(1)} = \sum_{j=1}^{G} \mathrm{Tr}[\rho_j X] - GX_0$, and $\alpha_2^{(1)} = 1/2(G-1)GX_0 - (G-1)\sum_{j=1}^{G}\mathrm{Tr}[\rho_j X] + \sum_{j_1\neq j_2}\mathrm{Tr}[\rho_{j_1,j_2}X]$. In the following we assume that $\alpha_1^{(1)} \neq 0$ so that error is of order $\mathcal{O}(\lambda)$. Taking into account that

$$\mathrm{Tr}[\rho^2] = \big[(1-\lambda)^2 + \lambda^2\big]^G \tag{48}$$

we arrive at a similar expression for $\hat{\mu}_2^{\mathrm{VD}}$,

$$\begin{aligned}
\hat{\mu}_2^{\mathrm{VD}} &= \frac{(1-\lambda)^{2G}X_0 + (1-\lambda)^{2(G-1)}\lambda^2\sum_{j=1}^{G}\mathrm{Tr}[\rho_j X] + \cdots}{\big[(1-\lambda)^2 + \lambda^2\big]^G} \\
&= X_0 + \lambda^2\alpha_1^{(2)} + \lambda^3\alpha_2^{(2)} + \cdots \,,
\end{aligned} \tag{49}$$

where the coefficients $\alpha_k^{(2)}$ are defined similarly to the $\alpha_k^{(1)}$ ones. In particular, one can find $\alpha_1^{(2)} = \alpha_1^{(1)}$. As expected, when computing the expectation value with two copies of the state, Virtual Distillation suppresses the error contribution from the orthogonal noise of order $\mathcal{O}(\lambda)$.

Let us now show how one can take a linear combination of $\hat{\mu}_1^{\mathrm{VD}}$ and $\hat{\mu}_2^{\mathrm{VD}}$ to further suppress the error. Consider the combination

$$\mu'_2 = b_1\hat{\mu}_1^{\mathrm{VD}} + b_2\hat{\mu}_2^{\mathrm{VD}}, \tag{50}$$

where b_1 and b_2 are coefficients to be found. Expanding this out we obtain

$$\begin{aligned}
\mu'_2 = X_0\big(b_1 + b_2\big) &+ \big(\lambda\alpha_1^{(1)}b_1 + \lambda^2\alpha_1^{(2)}b_2\big) \\
&+ \lambda^2\alpha_2^{(1)}b_1 + \cdots .
\end{aligned} \tag{51}$$

Consider taking

$$b_1 + b_2 = 1 \tag{52}$$

$$\lambda\alpha_1^{(1)}b_1 + \lambda^2\alpha_1^{(2)}b_2 = 0. \tag{53}$$

Solving for b_1 and b_2, and recalling that $\alpha_1^{(2)} = \alpha_1^{(1)}$ gives

$$b_1 = \frac{-\lambda\alpha_1^{(2)}}{\alpha_1^{(1)} - \lambda\alpha_1^{(2)}} = \frac{-\lambda}{(1-\lambda)} \,, \tag{54}$$

$$b_2 = \frac{\alpha_1^{(1)}}{\alpha_1^{(1)} - \lambda\alpha_1^{(2)}} = \frac{1}{(1-\lambda)} \,. \tag{55}$$

Putting these coefficients into Eq. (51) leads to

$$\mu'_2 = X_0 + \lambda^2\alpha_2^{(1)}b_1 + \cdots . \tag{56}$$

Considering the term $\lambda^2\alpha_2^{(1)}b_1$ we obtain

$$\lambda^2\alpha_2^{(1)}b_1 = -\lambda^3\frac{\alpha_2^{(1)}}{1-\lambda} \in \mathcal{O}(\lambda^3) \,. \tag{57}$$

The previous then implies

$$|\mu'_2 - X_0| = |\mu'_2 - \mu - \epsilon| \in \mathcal{O}(\lambda^3). \tag{58}$$

Therefore, we have shown that taking a linear combination of VD-mitigated observables further reduces the error on the estimate of the noiseless observable. Having seen how the errors can be canceled for $M_{\max} = 2$ let's generalize this same error cancellation to the many copy case.

B.2.2 Generalizing to the many copy case

Analogously to the case of $M_{\max} = 2$ one can use combination (45) to suppress the error also for larger $M_{\max}$. For M copies and our simple noise model we have

$$\begin{aligned}
\hat{\mu}_M^{\mathrm{VD}} &= \frac{\mathrm{Tr}[\rho^M X]}{\mathrm{Tr}[\rho^M]} = \\
&\frac{(1-\lambda)^{MG}X_0 + (1-\lambda)^{M(G-1)}\lambda^M\sum_{j=1}^{G}\mathrm{Tr}[\rho_j X] + \cdots}{\big[(1-\lambda)^M + \lambda^M\big]^G} \\
&= X_0 + \lambda^M\alpha_1^{(M)} + \lambda^{M+1}\alpha_2^{(M)} + \cdots \,,
\end{aligned} \tag{59}$$

where $\alpha_1^{(M)}$, $\alpha_2^{(M)}$ are coefficients of $\hat{\mu}_M^{\mathrm{VD}}$ series expansion in λ. Using that and grouping terms based on the index j in $\alpha_j^{(M)}$ we obtain

$$\begin{aligned}\mu'_{M_{\max}} &= \sum_{M=1}^{M_{\max}} b_M \hat{\mu}_M^{\mathrm{VD}} \\ &= X_0 \sum_{M=1}^{M_{\max}} b_M + \sum_{M=1}^{M_{\max}} b_M \sum_{j=1}^{\infty} \alpha_j^{(M)} \lambda^{j+M-1} = \\ &= X_0 \sum_{M=1}^{M_{\max}} b_M + \sum_{j=1}^{\infty} \sum_{M=1}^{M_{\max}} b_M \alpha_j^{(M)} \lambda^{j+M-1} = \\ &= X_0 \sum_{M=1}^{M_{\max}} b_M + \sum_{j=1}^{M_{\max}-1} \sum_{M=1}^{M_{\max}} b_M \alpha_j^{(M)} \lambda^{j+M-1} + \\ &+ b_1 \alpha_{M_{\max}}^{(1)} \lambda^{M_{\max}} + \mathcal{O}(\lambda^{M_{\max}+1}). \end{aligned} \tag{60}$$

Taking

$$\sum_{M=1}^{M_{\max}} b_M = 1, \tag{61}$$

$$\sum_{M=1}^{M_{\max}} b_M \alpha_j^{(M)} \lambda^{j+M-1} = 0\,, \quad \forall j = 1, ..., M_{\max} - 1\,, \tag{62}$$

we get

$$\mu'_{M_{\max}} = X_0 + b_1 \alpha_{M_{\max}}^{(1)} \lambda^{M_{\max}} + \mathcal{O}(\lambda^{M_{\max}+1})\,. \tag{63}$$

To obtain (62) in the limit of small λ while taking into account that the $\alpha_j^{(M)}$ coefficients do not depend on λ we need terms b_1, $b_2\lambda$, $b_3\lambda^2$,...,$b_{M_{\max}}\lambda^{M_{\max}-1}$ to be of the same order. Otherwise, a system of equations (61, 62) could be reduced to systems of overdetermined linear equations which usually do not have a solution. Furthermore, Eqs. 61, 62 imply that $b_{M_{\max}}$ is of order 1, which leads to $b_1 \in \mathcal{O}(\lambda^{M_{\max}-1})$. Assuming that the system of equations (61, 62) has a solution, which is only false in the special case of a singular system of linear equations, we obtain

$$|\mu'_{M_{\max}} - X_0| = |\mu'_{M_{\max}} - \mu - \epsilon| \in \mathcal{O}(\lambda^{M_{\max}+1}). \tag{64}$$

Therefore, the combination improves the VD error suppression for the largest used $M = M_{\max}$.

This shows that in general, we can combine many VD-mitigated observables with different numbers of copies to achieve a better quality of error mitigation. Consequently, it motivates the linear combinations of VD-mitigated observables we take in the CGVD (16) and UNITED (12) Ansätze proving that such combinations can be advantageous. We note that the coefficients b_M given by the system of equations (61, 62) are not guaranteed to be a unique choice that leads to an improvement of the error of combination (45) with respect to its constituent parts.

B.3 Combining VD-mitigated observables obtained at different noise levels

Now we motivate the ansatz used in the UNITED method, which takes the form of a linear combination of VD-mitigated observables obtained with different numbers of copies and different noise levels. We consider the same noise model as introduced in the previous section (Appendix B.2), where we showed that the orthogonal error can be suppressed by taking a combination of the VD-mitigated observables with different numbers of copies. Such a combination in itself cannot mitigate a noise floor that results from a coherent error of the form $U(\theta) = e^{-i\theta E}$. To mitigate the noise floor we consider taking a combination of VD-mitigated observables obtained at various noise levels.

We start by recalling from the previous section that

$$\hat{\mu}_M^{\mathrm{VD}} = X_0 + \lambda^M \alpha_1^{(M)} + \lambda^{M+1} \alpha_2^{(M)} + \cdots \tag{65}$$

and

$$X_0 = \langle \psi_0 | X | \psi_0 \rangle\,, \tag{66}$$

where $|\psi_0\rangle\langle\psi_0|$ is the dominant eigenvector of the noisy state ρ. For our noise model we have

$$X_0 = \mathrm{Tr}[e^{-i\theta E} |\psi_{\mathrm{exact}}\rangle\langle\psi_{\mathrm{exact}}| e^{i\theta E} X]\,, \tag{67}$$

where we assume $[X, E] \neq 0$. This leads to

$$\begin{aligned}\hat{\mu}_M^{\mathrm{VD}}(\theta, \lambda) = \mathrm{Tr}[e^{-i\theta E} |\psi_{\mathrm{exact}}\rangle\langle\psi_{\mathrm{exact}}| e^{i\theta E} X] + \lambda^M \alpha_1^{(M)} \\ + \lambda^{M+1} \alpha_2^{(M)} + \cdots\,. \end{aligned} \tag{68}$$

As shown in the previous section, combining VD-mitigated observables as

$$\mu'_{M_{\max}}(\theta, \lambda) = \sum_{M=1}^{M_{\max}} b_M \hat{\mu}_M^{\mathrm{VD}}(\theta, \lambda)\,, \tag{69}$$

with coefficients (61, 62) leads to

$$\begin{aligned}\mu'_{M_{\max}}(\theta, \lambda) = \mathrm{Tr}[e^{-i\theta E} |\psi_{\mathrm{exact}}\rangle\langle\psi_{\mathrm{exact}}| e^{i\theta E} X] \\ + \mathcal{O}(\lambda^{M_{\max}+1}). \end{aligned} \tag{70}$$

For large enough $M_{\max}$ the dominant part of the error comes from the noise floor. Therefore, in this case we can neglect dependence of $\mu'_{M_{\max}}$ on λ

$$\mu'_{M_{\max}}(\theta, \lambda) \approx \mathrm{Tr}[e^{-i\theta E} |\psi_{\mathrm{exact}}\rangle\langle\psi_{\mathrm{exact}}| e^{i\theta E} X]. \tag{71}$$

Performing a Taylor expansion of this approximation around $\theta = 0$, we obtain

$$\mu'_{M_{\max}}(\theta) = \mu + \beta_1 \theta + \beta_2 \theta^2 + \cdots\,, \tag{72}$$

where $\mu = \mathrm{Tr}[|\psi_{\mathrm{exact}}\rangle\langle\psi_{\mathrm{exact}}| X]$, and $\beta_1, \beta_2, \ldots$ are the expansion coefficients. By considering $n+1$ values of θ, i.e. $c_0\theta_0 = \theta_0$, $c_1\theta_0$, ..., $c_n\theta_0$, we perform a

Richardson extrapolation obtaining a mitigated estimate of the observable

$$\mu''_{M_{\max},n} = \sum_{j=0}^{n} \gamma_j \mu'_{M_{\max}}(c_j\theta_0) \ , \tag{73}$$

where

$$\sum_{j=0}^{n} \gamma_j = 1, \quad \sum_{j=0}^{n} \gamma_j c_j^k = 0 \ \text{ for } k = 1, \ldots, n. \tag{74}$$

Recalling (69), (72) we finally obtain

$$\mu''_{M_{\max},n} = \sum_{j=0}^{n} \sum_{M=1}^{M_{\max}} \gamma_j b_M \hat{\mu}_M^{\mathrm{VD}}(c_j\theta_0, \lambda) \ , \tag{75}$$

which is a linear combination of VD-mitigated observables obtained with different numbers of copies at different coherent noise levels. This combination has exactly the same form as the UNITED ansatz (12). As previously discussed in Appendix A this application of the Richardson extrapolation results in an error

$$\mu''_{M_{\max},n} - \mu = \mathcal{O}(\theta_0^{n+1}), \tag{76}$$

assuming that the orthogonal errors are small enough. We note that as long as the orthogonal error rates are not too large, one can scale both θ an λ, e.g. choosing $\theta_j = c_j\theta_0$ and $\lambda_j = c_j\lambda_0$, which results in

$$\mu''_{M_{\max},n} = \sum_{j=0}^{n} \sum_{M=1}^{M_{\max}} \gamma_j b_M \hat{\mu}_M^{\mathrm{VD}}(c_j\theta_0, c_j\lambda_0), \tag{77}$$

and obtaining the same error scaling as in (76), since for large enough $M_{\max}$ the error will always be dominated by the coherent mismatch.

C Correcting for global depolarizing noise with CGVD or UNITED

Here we motivate the UNITED ansatz by considering its performance in mitigating the action of a global depolarizing noise channel on some observable of interest. The ansatzes used in CDR and vnCDR have been shown to perfectly correct global depolarizing noise and this fact is used to motivate the form of the CGVD ansatz. We note in this setting there is no limit set by coherent mismatch when using VD, so one expects a combination of VD and CDR to perform well. First, we consider CGVD since the treatment is the same as in the UNITED case and can be generalized very simply.

Consider a global depolarizing noise channel that acts on all Q qubits. For some state $\rho = |\psi\rangle\langle\psi|$ produced by quantum circuit U, this channel takes the form:

$$\rho \longrightarrow (1-\lambda)\rho + \lambda\frac{\mathbb{1}}{2^Q}. \tag{78}$$

Consider the effect of the above channel on some observable of interest X

$$\hat{\mu} = (1-\lambda)\mu + \lambda\frac{\mathrm{Tr}(X)}{2^Q}, \tag{79}$$

where $\mu = \mathrm{Tr}(\rho X)$. We consider the CGVD ansatz that combines observables evaluated at several numbers of copies in the VD circuit structure. For simplicity, we omit the small changes in the rate of errors for different numbers of copies due to the swap noise. We can rewrite (16) as follows

$$\mu'_{M_{\max}} = \sum_{M=1}^{M_{\max}} b_M \hat{\mu}_M^{\mathrm{VD}}, \tag{80}$$

where $\mu'_{M_{\max}}$ is the estimate of the observable of interest, a_M are the parameters fitted using least squares regression on the data generated by the near-Clifford training circuits mitigated using VD with up to M copies.

We can express an observable mitigated with VD using M copies as

$$\hat{\mu}_M^{\mathrm{VD}} = \frac{\mathrm{Tr}\left[\rho^M X\right]}{\mathrm{Tr}\left[\rho^M\right]}. \tag{81}$$

We can evaluate $\hat{\mu}_M^{\mathrm{VD}}$, assuming ρ has been acted upon by a global depolarizing channel as defined in (78). First focusing on the denominator,

$$\begin{aligned}\mathrm{Tr}\left[\rho^M\right] &= \mathrm{Tr}\left[\sum_{k=0}^{M}\binom{M}{k}\left((1-\lambda)\rho\right)^{(M-k)}\left(\frac{\lambda}{2^Q}\right)^k\right] \\ &= \sum_{k=0}^{M}\binom{M}{k}(1-\lambda)^{(M-k)}\mathrm{Tr}(\rho^{M-k})\left(\frac{\lambda}{2^Q}\right)^k.\end{aligned} \tag{82}$$

Noting that when $M = k$ we are tracing over the identity, we obtain:

$$\mathrm{Tr}\left[\rho^M\right] = \left(\frac{1+(2^Q-1)(1-\lambda)}{2^Q}\right)^M + (2^Q-1)\left(\frac{\lambda}{2^Q}\right)^M. \tag{83}$$

Now evaluating the numerator of (81) we obtain

$$\begin{aligned}\mathrm{Tr}\left[\rho^M X\right] = \mathrm{Tr}[\rho X]&\left(\frac{1+(2^Q-1)(1-\lambda)}{2^Q}\right)^M \\ &+\left(\frac{\lambda}{2^Q}\right)^M\Big(\mathrm{Tr}[X] - \mathrm{Tr}[\rho X]\Big).\end{aligned} \tag{84}$$

Combining the above we arrive at the following expression,

$$\hat{\mu}_M^{\mathrm{VD}} = f_M \mathrm{Tr}[\rho X] + \frac{1-f_M}{2^Q}\mathrm{Tr}[X] \ , \tag{85}$$

where

$$f_M = 1 - \frac{2^Q\lambda^M}{(2^Q-1)\lambda^M + (2^Q - \lambda(2^Q-1))^M} \tag{86}$$

Taking $M = 1$ leads to $f_1 = 1 - \lambda$ which gives the expected result. Using (80) we can determine the criteria for this ansatz to give perfect mitigation,

$$\mu'_{M_{\max}} = \sum_{M=1}^{M_{\max}} b_M \big(f_M \mathrm{Tr}[\rho X] + \frac{1 - f_M}{2^Q} \mathrm{Tr}[X]\big). \quad (87)$$

Therefore, we require

$$\sum_{M=1}^{M_{\max}} b_M f_M = 1, \quad \mathrm{Tr}[X] = 0. \quad (88)$$

A global depolarizing noise channel will have the same effect on the training circuits as it will on the circuit of interest. As we have access to the exact observable for these circuits, the coefficients required to lead to perfect mitigation can be learned. As such, the ansatz used in CGVD can be used to perfectly mitigate global depolarizing noise using a near-Clifford training set.

It is straightforward to extend this analysis to involve the ansatz including multiple noise levels. Notice that under the action of global depolarizing noise defined in (78), if this channel acts j times throughout the dynamics we can write

$$\rho_j \longrightarrow (1 - \lambda^j)\rho + \lambda^j \frac{\mathbb{1}}{2^Q}. \quad (89)$$

Now let us consider UNITED ansatz that combines observables mitigated with VD using various numbers of copies at various noise levels. This ansatz can be written as follows

$$\mu''_{M_{\max},n} = \sum_{M=1}^{M_{\max}} \sum_{j=1}^{n} d_{M,j} \hat{\mu}^{\mathrm{VD}}_{M,j}, \quad (90)$$

where $\mu''_{M_{\max},n}$ is the mitigated observable, $d_{M,j}$ are the parameters fitted using least squares regression on the data generated by the near-Clifford VD training circuits with up to $M_{\max}$ copies, evaluated with $n+1$ different noise levels.

Working through the same analysis with this ansatz, we find that, for the effects of global depolarizing noise to be perfectly mitigated we require

$$\sum_{M=1}^{M_{\max}} \sum_{j=1}^{n} d_{M,j} f_{M,j} = 1, \quad \mathrm{Tr}[X] = 0\ , \quad (91)$$

where

$$f_{M,j} = 1 - \frac{2^Q \lambda^{Mj}}{(2^Q - 1)\lambda^{Mj} + (2^Q - \lambda^j(2^Q - 1))^M}. \quad (92)$$

Therefore, we can conclude that both CGVD and UNITED can perfectly mitigate the effect of global depolarizing noise on some observable of interest.

The assumption $\mathrm{Tr}[X] = 0$ is not restrictive due to the common decomposition of operators into Pauli terms (which have zero trace) and the identity. However, we can consider an ansatz which includes an additional constant term to be fitted. In which case (80) becomes

$$\mu'_{M_{\max}} = \sum_{M=1}^{M_{\max}} b_M \hat{\mu}^{\mathrm{VD}}_M + b_0, \quad (93)$$

Thus, (87) becomes

$$\mu'_{M_{\max}} = \sum_{M=1}^{M_{\max}} b_M \big(f_M \mathrm{Tr}[\rho X] + \frac{1 - f_M}{2^Q} \mathrm{Tr}[X]\big) + b_0. \quad (94)$$

Therefore, including a constant term in the ansatz, to perfectly mitigate the effect of depolarizing noise we require

$$\sum_{M=1}^{M_{\max}} b_M f_M = 1, \quad \sum_{M=1}^{M_{\max}} \frac{1 - b_M}{2^Q} \mathrm{Tr}[X] = b_0\ . \quad (95)$$

Including a constant term in the UNITED ansatz leads to a similar conclusion. Therefore, one can see that global depolarizing noise can be perfectly mitigated using near-Clifford circuits to train the parameters of the CGVD and UNITED ansatzes.

D Effects of the controlled derangement noise.

VD implementation on real hardware requires performing a controlled derangement which can be decomposed into controlled swap operations [26]. In the main text we neglect the effects of noise in controlled swap implementations as the cost of classical simulation of the noisy controlled derangement scales as 2^{4Q+2} instead of the 2^{2Q} for the noiseless derangement simulation [9]. Here we examine the effects of the noisy derangement for $Q = 4$ random quantum circuits. We perform the simulation by decomposing a controlled swap gate into two CNOTs and a Toffoli gate as described in [40]. Subsequently, we decompose Toffoli gates and CNOTs to the native gates using decomposition from [30].

In Figure 7 we compare results for $Q = 4$, $L = 5$ random quantum circuits simulated with and without the derangement noise. We analyze error $\langle \sigma_Z^1 \rangle$ average over 30 instances. We find that the derangement noise reduces the performance ceiling of VD mitigation and the performance of UNITED for large $N_{\mathrm{tot}} = 10^9, 10^{10}$. Nevertheless, we still find a factor of 2.5 VD improvement over the noisy simulations and a factor of 1.2 UNITED improvement over vnCDR for $N_{\mathrm{tot}} = 10^{10}$. These results indicate that UNITED can be useful in the presence of the derangement noise even without derangement noise error mitigation.

Further improvements of UNITED and VD performance can be obtained by optimizing the decomposition of the controlled derangement circuit to native

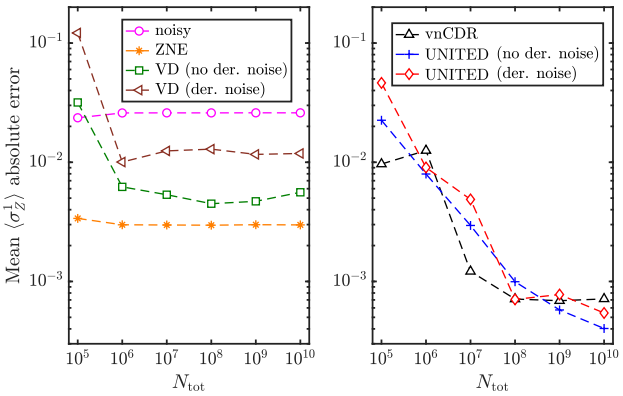

Figure 7: **Comparing VD and UNITED error mitigation with (the brown and the red curves) and without controlled derangement noise (the green and the blue curves).** We perform the comparison for the random quantum circuits with $Q = 4$ and $L = 5$. We plot an error of mitigated and noisy $\langle\sigma_Z^1\rangle$ averaged over 30 instances of the random quantum circuits. The derangement noise affects VD and UNITED implementations decreasing the performance of VD and UNITED for large $N_{\rm tot}$. Nevertheless, even with the derangement noise, we obtain a factor of 2.5 improvement for VD with $N_{\rm tot} \geq 10^6$ and factor 1.2 improvement of UNITED over vnCDR for the largest considered $N_{\rm tot} = 10^{10}$. For reference, we show also the noisy, ZNE, and vnCDR results. Here, ZNE is performed using linear extrapolation with noise levels $c = \{1, 2\}$. VD uses $M = 3$ copies. We obtain the vnCDR correction using noise levels $c = \{1, 2, 3\}$ and training circuits with $N = 30$ R_Z non-Clifford gates left intact. In the case of UNITED we have $c = \{1, 2, 3\}$, $M_{\max} = 3$, and $N = 30$.

gates [26] and by using further error mitigation of controlled derangement noise [26]. Another interesting direction is VD implementations based on state verification instead of controlled derangement [5, 23].

E Optimal parameters for Virtual Distillation

Here we systematically explore the convergence of VD estimates with an increasing number of copies M. In Figure 8 we plot the mean and maximal errors of VD $\langle\sigma_Z^1\rangle$ estimate for 30 random quantum circuits and $M = 2 - 6$. In this figure, we show results for the smallest $N_{\rm tot} = 10^5$ and the largest $N_{\rm tot} = 10^{10}$ considered here. We find that $Q = 4 - 10$ and $g = 1$, $M = 2$ typically gives optimal results for both shot budgets. We note that for $Q = 4 - 10$, $g = 1$ and $N_{\rm tot} = 10^{10}$, $M = 2 - 5$ VD gives very similar results indicating that $M = 2$ is large enough to reach the noise floor. Our findings agree with [27], which argues that, with the exception of highly noisy circuits, $M \leq 3$ is sufficient to reach the noise floor. For noisier $Q = 4, g = 16$ circuits we find that the optimal M depends strongly on $N_{\rm tot}$. For $N_{\rm tot} = 10^5$, $M = 2$ is optimal but it gives on average a small factor of 1.2 improvement over the noisy results. For $N_{\rm tot} = 10^{10}$, $M = 4$ is optimal giving a factor of 2.5 improvement. Based on the obtained results, we use $M = 2$ in the main text as it is optimal for most circuits. We note that although with $M = 4$ VD performance can be improved for $Q = 4, g = 16$ and large $N_{\rm tot}$, this improvement is not large enough to outperform vnCDR and UNITED.

F Comparison of ZNE implementations

For ZNE we compare the mitigation performance when using two or three noise levels, $c \in \{1, 2, 3\}$ and $c \in \{1, 2\}$ respectively. For three noise levels, we apply both linear and exponential extrapolations, while the data coming from two noise levels only allows us to perform a linear extrapolation. We benchmark the extrapolation approaches when mitigating observables produced by RQCs and taking $N_{\rm tot} = 10^5, 10^6, \ldots, 10^{10}$. For $L = Q$ and $Q = 4, 6, 8, 10$ we find that with the exception of $Q = 4, N_{\rm tot} = 10^5$ using two noise levels leads to better mitigation quality. The exponential extrapolation results in an order of magnitude worse results than the linear one. The results for $Q = 4, 10$ are shown in Figure 9. For $L = 16Q$ and $Q = 4, 6, 8$, we find that both variants of the linear extrapolation ZNE lead to similar mitigation quality. The exponential extrapolation again produces an order of magnitude worse results than the linear extrapolation.

G Clustering of the near-Clifford circuits

It is a well-known property of near-Clifford circuits that exact expectation values of Pauli strings cluster around 0. This clustering negatively impacts the shot-efficiency of learning-based error mitigation methods using near-Clifford circuits as training circuits [11]. Here we demonstrate that the clustering does indeed occur for the training sets used to mitigate the random quantum circuit expectation values in Section 4.2.

In Figure 10 we show relative frequencies of exact expectation values of $\langle\sigma_Z^1\rangle$ for a sample of 30 $Q = 4, L = 4$ random quantum circuits and 3000 near-Clifford circuits obtained by random substitutions of non-Clifford gates in these random quantum circuits. For each circuit of interest, 100 near-Clifford circuits were generated. We see strong clustering of the expectation values for the near-Clifford circuits despite the lack of this effect in the random quantum circuits used to generate them. The training circuits used in Section 4.2 were then chosen by post-selecting on these near-Clifford circuits.

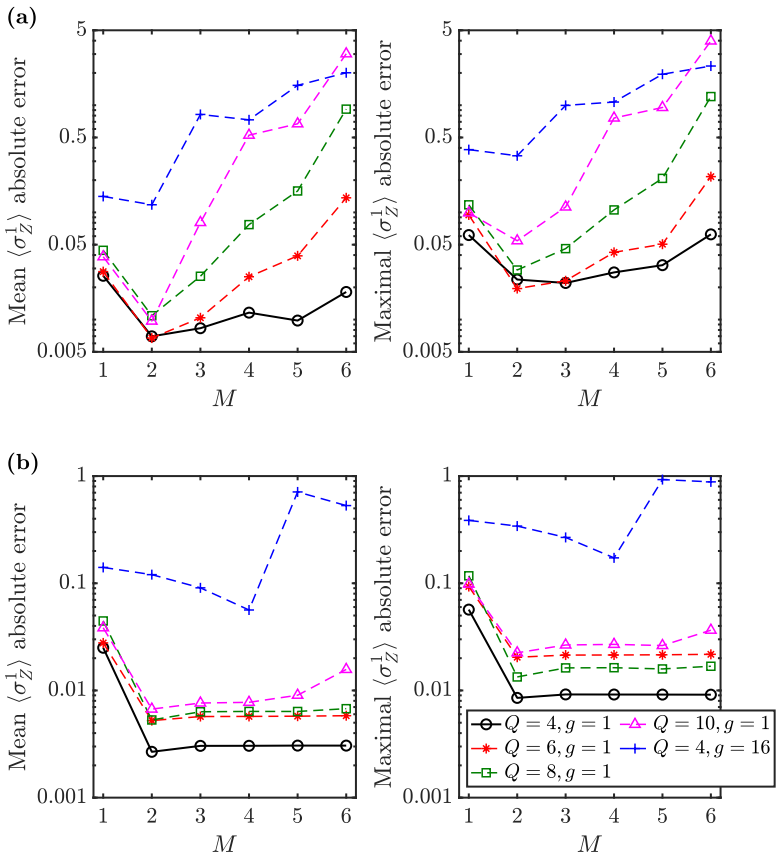

Figure 8: **Convergence of VD estimates with increasing M.** In (a) results for $N_{\text{tot}} = 10^5$ and in (b) results for $N_{\text{tot}} = 10^{10}$ plotted versus number of VD copies M. For reference purposes, we include single-copy noisy results as $M = 1$. The plots show that for $g = 1$ $M = 2$ is optimal while for $Q = 4$, $g = 16$ the optimal M increases with increasing N_{tot}. The left plots show the mean values while the right plots show the maximal values.

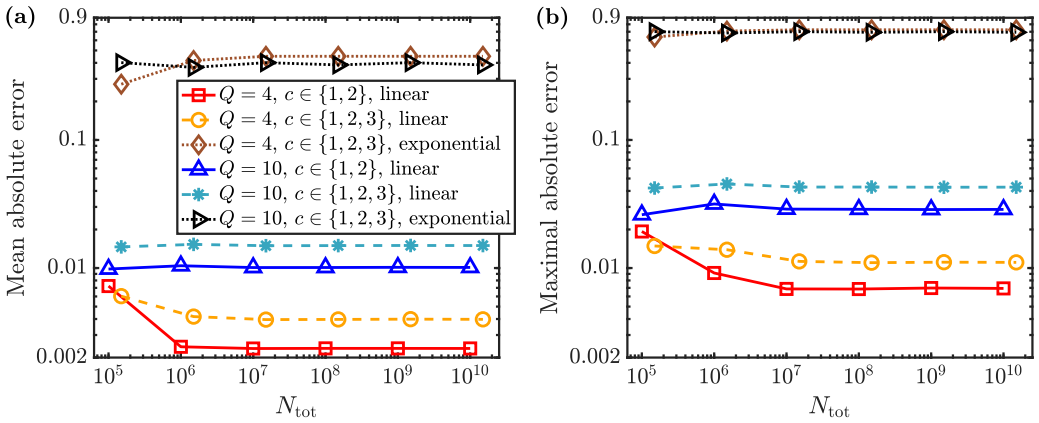

Figure 9: **Convergence of different ZNE variants applied to RQCs.** Mean and maximal absolute errors of the expectation value of σ_Z^1 plotted versus the shot budget. We use 30 instances of random quantum circuits with $Q = 4, 10$ and $L = Q$. The red and blue curves are obtained with two noise levels $c \in \{1, 2\}$ and a linear extrapolation while the orange and magenta curves are obtained with the linear extrapolation and three noise levels $c \in \{1, 2, 3\}$. The brown and the black curves are obtained with an exponential extrapolation and $c \in \{1, 2, 3\}$.

In Table 2 we further analyze the clustering of the near-Clifford circuits. The table shows the median values $|\langle\sigma_Z^1\rangle|_2$, and third quartiles $|\langle\sigma_Z^1\rangle|_3$ of the absolute value of the observable of interest, obtained for all the random quantum circuits mitigated in Section 4.2 and all near-Clifford circuits generated before post-selection. We see that $|\langle\sigma_Z^1\rangle|_2$ and $|\langle\sigma_Z^1\rangle|_3$ are decreasing slowly with increasing Q for the random quantum circuits. For example, $Q = 4, L = 4$ the RQCs have $|\langle\sigma_Z^1\rangle|_2 = 0.270$ and $|\langle\sigma_Z^1\rangle|_2 = 0.438$, while for the largest $Q = 10$ ($L = 10$) we obtain $|\langle\sigma_Z^1\rangle|_2 = 0.184$ and $|\langle\sigma_Z^1\rangle|_2 = 0.265$. Unlike the

Q	L	$\lvert\langle\sigma_Z^1\rangle\rvert_2$ RQC	$\lvert\langle\sigma_Z^1\rangle\rvert_3$ RQC	$\lvert\langle\sigma_Z^1\rangle\rvert_2$ near-Clifford circuits	$\lvert\langle\sigma_Z^1\rangle\rvert_3$ near-Clifford circuits	95th $\lvert\langle\sigma_Z^1\rangle\rvert$ percentile near-Clifford circuits
4	4	0.270	0.438	0	0.319	0.714
4	64	0.125	0.233	0.122	0.262	0.390
6	6	0.185	0.300	0	0.081	0.584
6	96	0.108	0.146	0	0.074	0.212
8	8	0.231	0.378	0	0	0.527
8	128	0.030	0.063	0	0	0.067
10	10	0.184	0.265	0	0	0.363

Table 2: **Clustering of the near-Clifford circuits.** Here we gather medians $\lvert\langle\sigma_Z^1\rangle\rvert_2$ and third quartiles $\lvert\langle\sigma_Z^1\rangle\rvert_3$ of $\lvert\langle\sigma_Z^1\rangle\rvert$, obtained for samples of 30 random quantum circuits with various Q and L values mitigated in Section 4.2. We also show $\lvert\langle\sigma_Z^1\rangle\rvert_2$, $\lvert\langle\sigma_Z^1\rangle\rvert_3$ and 95$^{\text{th}}$ percentile of $\lvert\langle\sigma_Z^1\rangle\rvert$ for 3000 near-Clifford circuits obtained by substituting non-Clifford gates in the random quantum circuits. These circuits are then post-selected, choosing a subset that is used as the training circuits in UNITED and vnCDR in Section 4.2. When the obtained quartiles are of the order of numerical precision (10^{-15}) or smaller we simply denoted it as 0 in the table.

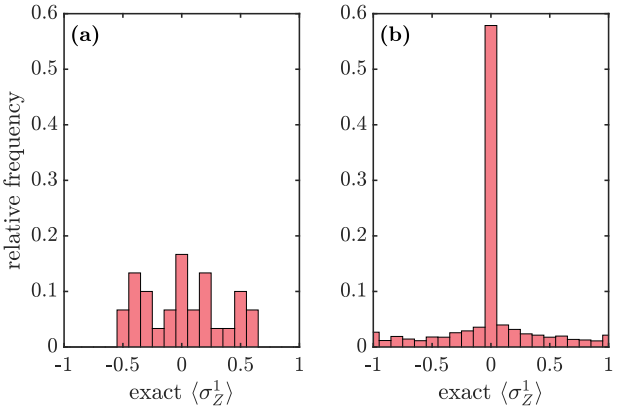

Figure 10: **Clustering of the near-Clifford circuits.** In (a) we plot relative frequencies of different exact expectation values of the mitigated observable in a sample of 30 random quantum circuits with $Q = 4, L = 4$. In (b) we show the same for 3000 near-Clifford circuits obtained by substitution of non-Clifford gates in the circuits analyzed in (a). For each random training circuit, 100 near-Clifford circuits are generated.

RQCs, the near-Clifford circuits are all strongly clustered around zero. Across all Q and L values mitigated in Section 4.2, only $Q = 4, L = 64$ non-Clifford circuits have non-zero $\lvert\langle\sigma_Z^1\rangle\rvert_2$. Furthermore, $\lvert\langle\sigma_Z^1\rangle\rvert_3 = 0$ for all $Q = 8$ and $Q = 10$ near-Clifford circuits. Finally, we gather in Table 2 95$^{\text{th}}$ percentile of $\lvert\langle\sigma_Z^1\rangle\rvert$ for the near-Clifford circuits to give information about tails of the near-Clifford circuits distribution. We find that it decreases with decreasing Q. Unlike in the case of $\lvert\langle\sigma_Z^1\rangle\rvert_2$ and $\lvert\langle\sigma_Z^1\rangle\rvert_3$, its decay is much faster for deep $L = 16Q$ circuits. For the deepest considered $Q = 8, L = 128$, it equals 0.067, which indicates that nearly all generated near-Clifford circuits cluster around 0.

These results show that the clustering of observables produced by near-Clifford circuits becomes stronger with increasing Q much faster than in the case of RQCs. Consequently, they justify post-selection on the exact expectation values of the training circuits employed while constructing the training circuits, which is described in detail in Section 5.5. This technique improves the shot-efficiency of the error mitigation scheme as explored in more detail in [11]. In Figure 11 we compare the mean and maximal absolute errors of the mitigated observable obtained with and without post-selection on the training circuits. We show that indeed, the usage of the post-selected training set improves the quality of the error mitigation for a finite number of shots, $N_{\text{tot}} = 10^{10}$, in both UNITED and vnCDR. We also compare the performance of both training set construction methods in the limit of an infinite number of shots. In that limit, both training sets with and without post-selection lead to a similar quality of error mitigation. The improvement of the mitigation quality for $N_{\text{tot}} = 10^{10}$ is much more pronounced for UNITED which converges slower with N_{tot} than vnCDR. This result further demonstrates that the post-selection is of particular importance when the quality of the mitigation is limited by the shot budget N_{tot}.

H Details of the noise model

Here we use a noise model based on Trout *et al.* [46] and developed in [8]. The noisy quantum gates R_X, R_Y, R_Z and XX are defined with noise channels $\mathcal{S}_\Xi$, $\mathcal{S}_{XX}$, respectively, where Ξ is one of X, Y or Z:

$$\begin{aligned}\mathcal{S}_\Xi &= \mathcal{W}^{p_d} \circ \mathcal{D}^{p_{\text{dep}}} \circ \mathcal{R}_\Xi^{p_\alpha}\,,\\ \mathcal{S}_{XX} &= \left[\mathcal{W}_1^{p_{d_1}} \otimes \mathcal{W}_2^{p_{d_2}}\right] \circ \left[\mathcal{D}_1^{p_{\text{dep}}} \otimes \mathcal{D}_2^{p_{\text{dep}}}\right] \circ \mathcal{H}^{p_{\text{xx}}} \circ \mathcal{H}^{p_h}.\end{aligned} \tag{96}$$

Here, $\mathcal{R}_\Xi^{p_\alpha}(\rho) = (1-p_\alpha)\rho + p_\alpha \Xi\rho\Xi$, with Ξ being one of σ_X, σ_Y or σ_Z, which represent imprecision in the angle of rotation about the X, Y, Z axes, respectively. $\mathcal{D}^{p_{\text{dep}}}$ is the local depolarizing channel with the error rate p^{dep}, defined as $\mathcal{D}^p : \rho \mapsto (1-p)\rho + \frac{p}{3}(\sigma_X\rho\sigma_X + \sigma_Y\rho\sigma_Y + \sigma_Z\rho\sigma_Z)$. $\mathcal{W}^{p_d} : \rho \mapsto (1-p_d)\rho + \sigma_Z\rho\sigma_Z$ is the dephasing channel, and $\mathcal{H}^p : \rho \mapsto (1-p)\rho + p(\sigma_X \otimes \sigma_X)\rho(\sigma_X \otimes \sigma_X)$ is a two-qubit channel that represents both an imprecise rotation when $p = p_{\text{xx}}$

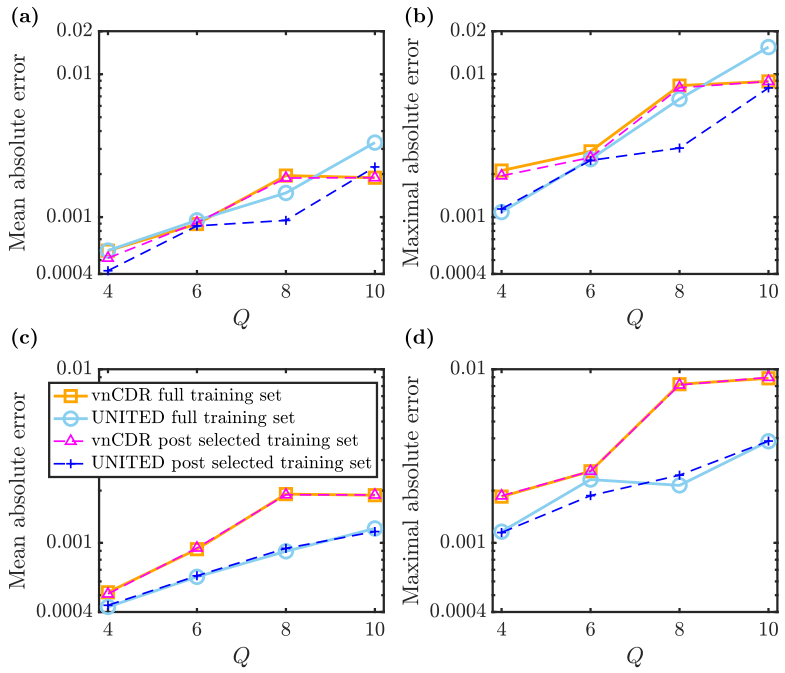

Figure 11: **Effects of training circuits post-selection on learning-based RQC error mitigation (UNITED and vnCDR).** Here we compare the mean and maximal absolute errors of mitigated $\langle \sigma_Z^1 \rangle$ for two methods of training circuits choice. First, we the $N_t = 100$ near-Clifford circuits obtained by random substitution of non-Clifford gates as the training circuits (the solid lines). Second, we post-select those circuits choosing the $N_t = 50$ with the largest exact $\langle \sigma_Z^1 \rangle$ as the training circuits (the dashed lines). The second strategy is used to obtain results in the main text. Panels **(a)** and **(b)** show the results for the total error mitigation shot cost $N_{\text{tot}} = 10^{10}$. The results in the infinite N_{tot} limit are shown in **(c)** and **(d)**. Here we mitigate 30 instances of $L = Q$ RQCs analyzed in Section 4.2.

and ion heating when $p = p_h$. Furthermore, we apply $\mathcal{D}^{p_{\text{idle}}}$ to simulate idling noise. The error rates were chosen as

$$\begin{aligned} p_d &= 1.5 \cdot 10^{-4}, \quad p_{\text{dep}} = 8 \cdot 10^{-4}, \\ p_{d_1} &= p_{d_2} = 7.5 \cdot 10^{-4}, \quad p_\alpha = 1 \cdot 10^{-4}, \\ p_{\text{xx}} &= 1 \cdot 10^{-3}, \quad p_h = 1.25 \cdot 10^{-3}, \\ p_{\text{idle}} &= 8 \cdot 10^{-4}. \end{aligned} \tag{97}$$

We neglect measurement error in our simulations as it can be mitigated by specialized techniques [2, 20].

I Compilation of the QAOA circuits to the native trapped ion gate set

In Section 4.4 we mitigate expectation values of terms of the Max-Cut Hamiltonian (21) for the QAOA ansatz (22) $\langle \Psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) | \sigma_Z^i \sigma_Z^j | \Psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) \rangle$. To evaluate them with the native trapped ion gate set we notice that

$$\langle \Psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) | \sigma_Z^i \sigma_Z^j | \Psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) \rangle = \langle \Psi'(\boldsymbol{\alpha}, \boldsymbol{\beta}) | \sigma_X^i \sigma_X^j | \Psi'(\boldsymbol{\alpha}, \boldsymbol{\beta}) \rangle, \tag{98}$$

where,

$$\begin{aligned} |\Psi'(\boldsymbol{\alpha}, \boldsymbol{\beta})\rangle = e^{-i\beta_L H'_M} e^{-i\alpha_L H'_P} \dots \\ \dots e^{-i\beta_2 H'_M} e^{-i\alpha_2 H'_P} e^{-i\beta_1 H'_M} e^{-i\alpha_1 H'_P} |\mathbf{0}\rangle, \end{aligned} \tag{99}$$

and $H'_P = \sum_{(i,j)\in E} \sigma_X^i \sigma_X^j$, $H'_M = \sum_{i=1}^Q \sigma_Z^i$. We have $e^{-i\alpha H'_P} = \prod_{(i,j)\in E} e^{-i\alpha \sigma_X^i \sigma_X^j} = \prod_{(i,j)\in E} XX(\alpha)$ and $e^{-i\beta H'_M} = \prod_{i=1,\dots,Q} e^{-i\beta\sigma_Z^i} = \prod_{i=1,\dots,Q} R_Z(\beta)$. That gives us the decomposition of $|\Psi'(\boldsymbol{\alpha}, \boldsymbol{\beta})\rangle$ to the native gate set of a trapped-ion quantum computer.

We note that H'_P and H'_M can be obtained from the original problem Hamiltonian H_P and mixer Hamiltonian H_M, respectively, by a rotation of the computational basis $U = \bigotimes_{i=1,\dots,Q} e^{-i(\pi/4)\sigma_Y^i}$, $H'_P = U H_P U^\dagger$, $H'_M = U H_M U^\dagger$. Therefore, $|\Psi(\boldsymbol{\alpha}, \boldsymbol{\beta})\rangle$ and $|\Psi'(\boldsymbol{\alpha}, \boldsymbol{\beta})\rangle$ are also related by U, $|\Psi'(\boldsymbol{\alpha}, \boldsymbol{\beta})\rangle = U|\Psi(\boldsymbol{\alpha}, \boldsymbol{\beta})\rangle$.