Low depth virtual distillation of quantum circuits by deterministic circuit decomposition

Akib Karim [1,*], Shaobo Zhang [2], and Muhammad Usman [1,2]

[1]*Data 61, CSIRO, Research Way Clayton 3168, Victoria, Australia*
[2]*School of Physics, University of Melbourne, Parkville 3010, Victoria, Australia*

Virtual distillation using measurements of multiple copies of a quantum circuit have recently been proposed as a method of noise mitigation of expectation values. Circuit decompositions known as B gates were found only for 1-local Hamiltonians, however practical problems for chemistry require n-local Hamiltonians, which cannot be corrected with B gates. We discover low depth circuit decompositions for expectation values for n-local Pauli strings by combining multiple projections to recover the correct measurement statistics or expectation values. Our method adds linear entangling gates with the number of qubits, but it requires extra measurements. Furthermore, in applications to find ground states such as the variational quantum eigensolver (VQE) algorithm, the variational principle is required which states that the energy cannot go below the ground-state energy. We find that the variational principle is violated when using B gates and is preserved if we use our low depth decomposition on all expectation values. We perform a demonstration on real devices, and we show that our decomposition can mitigate real experimental noise in VQE for the H_2 molecule with a two-qubit tapered mapping, H_3 with three qubits, and H_2 with four qubits. Our decomposition provides a way to perform duplicate circuit virtual distillation on real devices at significantly lower depth and for arbitrary observables.

I. INTRODUCTION

Quantum computing is a rapidly emerging field that is anticipated to run algorithms capable of solving certain computationally hard problems with improved scaling compared to its classical counterpart [1,2]. Unfortunately, current quantum devices are in the noisy intermediate scale quantum era (NISQ) because they suffer from significant noise, which reduces the accuracy of general algorithms or renders them useless for a quantum advantage. However, specific NISQ algorithms have been designed in which a quantum advantage is thought to have been found, e.g., boson sampling to calculate matrix permanents [3], increased information capacity for finite use of an amplitude damping channel [4], and the ability to generate statistics from random quantum circuits [5]. However, whether specific implementations show an advantage is a matter of debate [6–8]. For algorithms with more widespread applications, advantage is only reached with significantly more qubits and less noise. For example QAOA, which can calculate various graph theory problems like MAX-CUT, is thought to need 420 qubits for supremacy given some complexity assumptions [9]. Quantum error correction can guarantee that if noise is below a threshold, multiple physical qubits can be mapped to correctable logical qubits,

and the error rate can be bound arbitrarily [10]. These fault-tolerant devices will be able to show a quantum advantage without modifying the algorithms to consider noise. However, with our current NISQ devices, we require more specialized algorithms.

Some NISQ algorithms are considered tolerant to specific errors, such as the variational quantum eigensolver (VQE) for finding ground-state energies of Hamiltonians [11]. VQE uses a parametrized *Ansatz* circuit and measures the energy of the resultant state. The parameters are adjusted until the minimum is found, and this compensates for errors due to parameter setting or gate calibration. However, VQE cannot compensate for incoherent errors that result in mixed states, and there is a need to design noise mitigation techniques for these incoherent errors (see Fig. 1).

Many existing strategies of incoherent noise mitigation rely on a characterization of the device, such as Bayesian readout error mitigation [13], probabilistic error correction [14], random circuit sampling [15], or machine-learning-based methods [16]. Similarly, if the noise model is known analytically, there exist solutions to measure and undo its effect [15,17]. While these have shown success, these characterization methods become quickly outdated due to noise drift in real devices, and methods without this assumption are needed. Zero noise extrapolation has found success lately with the use of additional measurements on circuits of increasing noise to extrapolate back to a theoretical noise-free value [14]. However, this can lead to a biased estimator depending on the noise amplification method [18]. There also exist error detection and postselection methods that increase the number of measurements to select samples with less noise [19].
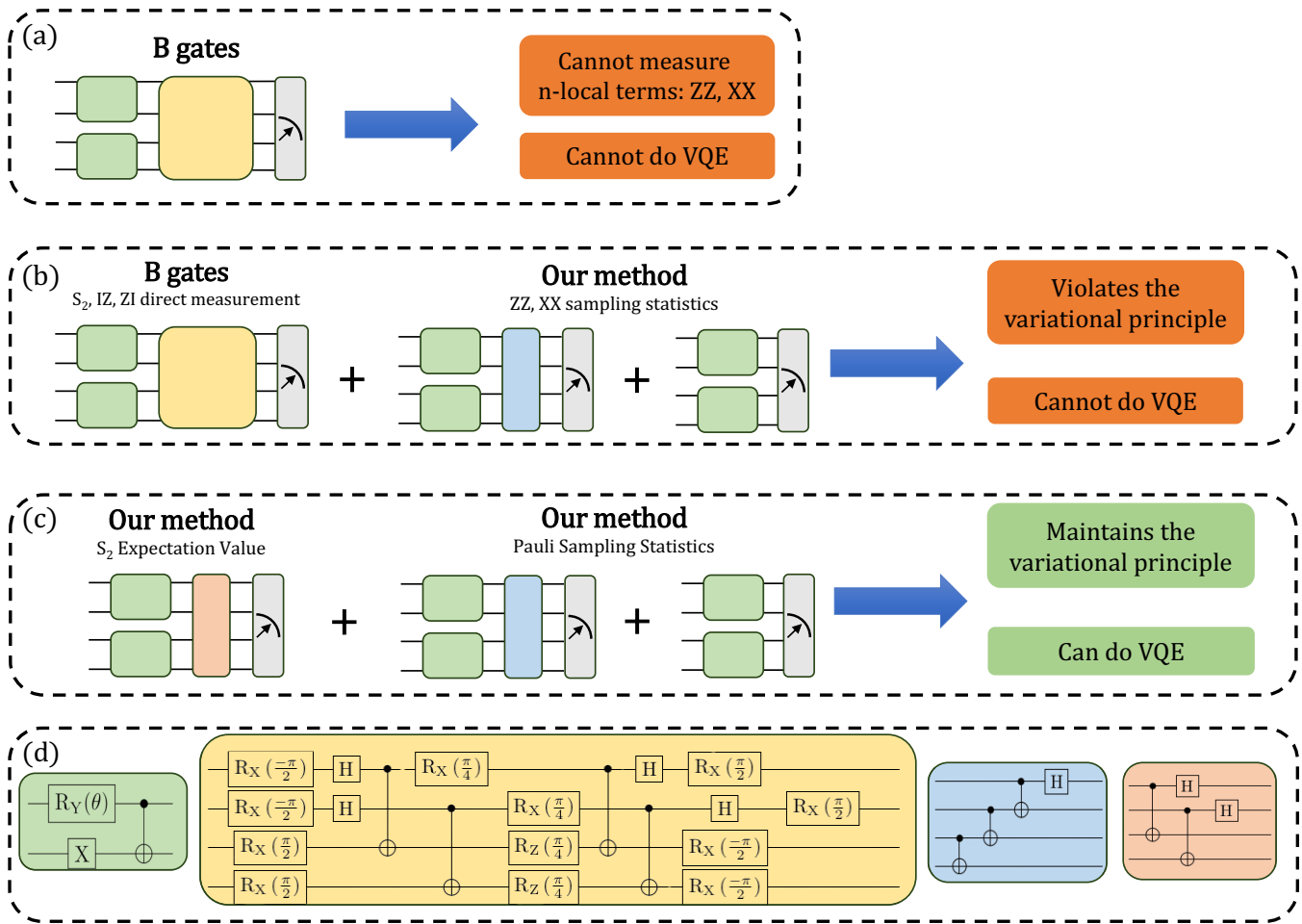
*Contact author: akib.karim@csiro.au

FIG. 1. Diagram of virtual distillation circuit implementations for a two-qubit H_2 VQE calculation. Part (a) is the original B gate proposal from Huggins *et al.* [12], which is only able to measure 1-local Hamiltonian terms and is unable to perform VQE calculation. In (b) for 1-local terms we use B gates, and for 2-local terms we use our method of decomposition, where the measurement statistics can be recreated with an entangling gate and computational measurement instead of direct measurement. This results in the variational principle being violated, and VQE is unable to be performed. (c) We replace the B gates with our method to recreate the expectation value of S_2 using an entangling projection and computational measurements. The Pauli terms are recreated with the same computational measurement and an entangling projection. Part (d) shows the circuits for the *Ansatz*, B gates, the entangling projection to recreate Pauli sampling statistics, and the entangling projection to recreate the S_2 expectation value.

Recently, a method of virtual distillation was designed that uses a duplicate copy of the circuit to correct expectation values after measurement [12]. This found success with small systems that used only 1-local expectation values, but no general scheme was provided for larger systems that also require *n*-local expectation values (see Fig. 1). The extension of the B gate scheme would require a unique circuit for each expectation value with no bound on circuit depth. As the depth of the noise mitigation circuit increases, it can add more noise than is removed, rendering it useless.

Subsequently, additional papers that used virtual distillation only measured observables that can be rewritten as 1-local expectation values [20] by calculating a density matrix from maximum likelihood [21], or shadow tomography methods rather than direct measurement, which will require exponential measurements for arbitrary states [22]. Recently, a method of circuit cutting for the swap gates of the B gate circuit was found so that the correcting gates can be simulated on a classical computer, but this also requires probabilistic

reconstruction [23]. There is a need to devise new methods of deterministic circuit decomposition to enable duplicate circuit virtual distillation for use in practical calculations.

In this work, we propose a method of recovering the statistics of the proposed unitary projections in Ref. [12] by combining measurements from computational and specific entangling projections. We find methods to reconstruct the output sampling statistics or reconstruct the expectation value. Our method ensures linear scaling maximum additional depth. In particular, we find that only *n* additional two-qubit entangling gates are required for a circuit with *n* qubits before duplication. We devise a circuit decomposition that reconstructs the sampling statistics as well as one that reconstructs the expectation value. We show that a combination of these allows the variational principle to be retained while also having lower depth than the B gates. In the Appendix, we show alternative methods that are useful for chip topologies, such as measuring Pauli strings by reconstructing the expectation value, which increases the depth for fewer measurements, or

direct decomposition of observables to Pauli strings, which has no extra entangling gates but exponentially more measurements for systems where shots are less costly than depth.

Our method works for any algorithm that requires expectation values, therefore we demonstrate its effectiveness on the variational quantum eigensolver (VQE) [11] for two qubits (H_2), three qubits (H_3), and four qubits (H_2, no symmetry reduction). We show noise resilience using simulations as well as by running demonstrations on IBM quantum devices through cloud-based access.

II. THEORETICAL FRAMEWORK

For the trial state ρ and the Pauli string O, the expectation value is given by

$$\mathrm{tr}(O\rho). \tag{1}$$

Normally, we rotate ρ such that O is diagonalized, measure in the computational basis, and combine the counts based on the corresponding eigenvalues, i.e., if $O = UDU^\dagger$, where D is a diagonal matrix and U is a unitary matrix, then the measurement becomes

$$\mathrm{tr}(UDU^\dagger\rho) = \mathrm{tr}(DU^\dagger\rho U), \tag{2}$$

since the trace is invariant under cyclic permutations.

If there are errors, we can virtually purify the density matrix by replacing these measurements with

$$\frac{\mathrm{tr}(O\rho^2)}{\mathrm{tr}(\rho^2)}. \tag{3}$$

It has been shown in Ref. [12] that this quantity can be calculated using two copies of the circuit to create the density matrix $\rho^{\otimes 2}$ and measuring

$$\frac{\mathrm{tr}(OS_2\rho^{\otimes 2})}{\mathrm{tr}(S_2\rho^{\otimes 2})}, \tag{4}$$

where S_2 is the operator that swaps each qubit with their duplicate. It should be noted that $\mathrm{tr}(S_2\rho^{\otimes 2})$ is the definition of the purity of ρ.

For general observables, we construct the symmetrized operator:

$$O^{\mathrm{sym}} = \frac{O + S_2O}{2}. \tag{5}$$

This is the expectation value of O on the original circuit added to the expectation value on the duplicate circuit. This is desirable as the null space of the new operator is much larger as it consists of the states for which both the circuit and the duplicate circuit have zero eigenvalue, but also when they differ by a sign. The components of a state in this null space will not contribute to the expectation value, and they can be ignored when reconstructing the measurements.

Huggins *et al.* [12] found that 1-local observables, i.e., $O = ZII\ldots$, commute with S_2 and therefore they can construct one unitary [U from Eq. (2)] to project on to measure the entire quantity in Eq. (4). Their specific circuit decomposition is called the B gate.

Unfortunately, multiqubit operators do not commute with S_2 in general. We still need the B gate unitary to measure $\mathrm{tr}(S_2\rho^{\otimes 2})$, but now we need a new matrix that rotates the state

to measure the numerator. We can naively decompose this matrix into circuits using general decomposition algorithms, however this is incredibly deep and requires a new circuit for every set of commuting Pauli strings.

To enable use for arbitrary systems, we require a general decomposition scheme for all Pauli strings. We first investigate the eigenvectors of the observables. For S_2, the eigenvectors will be the states invariant to swapping. For example, consider a four-qubit state $|abcd\rangle$, $a, b, c, d \in \{0, 1\}$. The effect of S_2 is

$$S_2|abcd\rangle = |cdab\rangle. \tag{6}$$

The eigenvectors are therefore the following:
(i) $|abcd\rangle$ for $a = c; b = d$.
(ii) $\frac{1}{\sqrt{2}}(|abcd\rangle + |cdab\rangle)$ otherwise.

The computational basis states are exclusively either already eigenvectors or they can be rotated into the eigenbasis by superposition with their swapped basis. The technique from Huggins *et al.* [12] finds a single unitary that can project on the superposition states while leaving the rest in the computational basis. We show that we can recreate the same statistics by combining measurements on a computational basis and entangled bases with less depth.

We now show how to decompose this superposition basis into CNOT and Hadamard gates. Depending on the native entangling gates of the hardware, further optimizations can be made. We choose these as they are a subset of Clifford gates, which can be simulated efficiently classically and allow for noise mitigation schemes that characterize noise by comparing Clifford gate simulations to experiment.

A. Reconstructing sampling statistics for Pauli strings

For a general four-qubit state $|abcd\rangle$, $a, b, c, d \in \{0, 1\}$, the effect of applying a CNOT with control qubit 0 and target qubit 1 and then a Hadamard on qubit 0 is

$$H_0\,\mathrm{CNOT}_{0,1}|abcd\rangle = |0bcd\rangle + XXII|0bcd\rangle, \tag{7}$$

where the subscript labels the qubits. We therefore create the superposition of the first qubit in $|0\rangle$ and $|1\rangle$ and the second qubit identity or bit flipped. Using this, we can decompose the superpositions in Eq. (6) into equivalence classes of the form in Eq. (7). For a duplicate circuit of $2n$ qubits, the number of total equivalence classes is 2^n, but they only all need to be measured for $\langle S_2 \rangle$, which is one reason we find an alternative method for specifically S_2 below. For observables, only a subset need to be measured because we use the symmetrized form in Eq. (5). If the expectation values of O and S_2O differ by a sign, the value becomes 0. Furthermore, since Hamiltonians can be decomposed into multiple Pauli strings, they may share equivalence classes. Any technique that can remove the need to measure strings of one class will reduce measurements. For example, Pauli grouping to find a unitary to maximize simultaneous measurement of Pauli strings could be used to reduce the number of string classes and is compatible with our decomposition as we recreate sampling statistics.

The final procedure is then as follows: identify the set of entangling projections necessary for the specific Paulis required; measure those and the computational basis; directly reconstruct the measurement vectors where statistics

from states invariant to swap are taken from computational measurements and the rest are taken from their respective entangled measurement; normalize the measurement vector; and calculate relevant Pauli expectation values.

We note that our circuit for the 1-local measurements has significantly lower depth than the B gate proposed in Ref. [12].

B. Recreating the expectation value of S_2

Rather than recreate the sampling statistics, we can instead decompose the expectation value directly. Previously, we could normalize the sampling statistics, but this cannot be performed with only the expectation value. Physicality requires the expectation value to fall in the range $[-1, 1]$, and for $\langle S_2 \rangle$ this is not a problem as it is only close to 1 with no noise.

Consider $|\psi\rangle = a_0|00\rangle + a_1|01\rangle + a_2|10\rangle + a_3|11\rangle$. We can expand the expectation value $\langle \psi|S_2|\psi \rangle$ as

$$a_0 a_0^\dagger \langle 00|S_2|00\rangle + a_0 a_1^\dagger \langle 00|S_2|01\rangle + \cdots . \qquad (8)$$

We can group these into our two sets of basis states from Eq. (6). If the basis state $|i\rangle$ is invariant, then they will contribute terms of the form $|a_i|^2$, whereas if a pair of basis sets, $|i\rangle$ and $|j\rangle$, are mapped to each other by S, then they contribute $a_i a_j^\dagger + a_j a_i^\dagger$.

To calculate the full expectation value, we consider a circuit where we apply CNOTs on pairs of qubit copies (qubits labeled by subscript) and Hadamards on all qubits of one copy. For a two-qubit circuit, this is

$$(H \otimes H \otimes I \otimes I)\text{CNOT}_{0,2}\text{CNOT}_{1,3}. \qquad (9)$$

When applied to a computational basis state $|abcd\rangle$, this creates a superposition over four states,

$$IIII|00cd\rangle \pm IIIX|01cd\rangle \pm IIXI|10cd\rangle \pm IIXX|11cd\rangle, \qquad (10)$$

where the sign depends on the initial state. Importantly, for any number of qubits, it will contain the original computational state and the generator Pauli strings calculated above; however, it will also contain the superposition of additional states. Consider an arbitrary state $\sum_{abcd} A_{abcd}|abcd\rangle$. Applying the circuit and then measuring in the computational basis will give outcomes that are linear combinations of all four states in Eq. (10). It is always possible to find a linear combination of outcomes that recreate the expectation value for S_2 because it is simply the addition of terms where the states are swapped. Appendix A 1 details how to find the linear combination of outcomes with examples for one and two qubits. Appendix A 2 then describes how to use this method for measuring observables, which requires multicontrol Z gates.

The full procedure is as follows: first calculate the linear combination of outcomes that eliminates the additional cross terms; then perform measurements in both the computational basis and the pairwise entangled basis projection; the expectation value of S_2 is the addition of the computational measurements for states invariant to S_2 with the linear combination of outcomes from the entangled projection. We note that the pairwise entanglement is identical to the B gate and requires only one entangling gate per pair compared to the B

gate, which needs two. It also requires no rotations and uses only Clifford gates.

III. DEMONSTRATIONS ON IBM DEVICES

To benchmark the performance of our methodologies, we run our circuits on IBM superconducting devices. It should be noted that in this paper, we report all states in conventional (big endian) format, which is reversed from IBM measurement output.

For two qubits, we generate a Hamiltonian for the H-H molecule at a fixed distance using PySCF [24] with a STO-3G basis [25]. This is mapped to a two-qubit basis using the parity mapping with Z_2 symmetry reduction. We then use a manually simplified UCCSD *Ansatz* with one parameter, θ, given by preparation in the computational basis, $Ry(\theta) \otimes X$ CNOT, and measurement in the computational basis. As there is only one parameter, we sweep from $[-\pi, \pi]$ with 51 steps. This is repeated with auxiliary error correcting circuits. For each Hamiltonian, the angle that minimizes the energy for the raw data was found and the corrected values were calculated for that angle. Demonstrations were performed on *ibm_hanoi* using a mapping to qubits 1,4,7,6 as indicated in Appendix A 6.

Similarly, for three qubits, we used a linear chain H-H-H with equal distances between hydrogens. We used STO-3G with the Jordan-Wigner mapping and qubit tapering [26] to get a three-qubit basis. A hardware-efficient *Ansatz* with nine parameters was used given by two layers of Ry, CNOT, Ry, CNOT, Ry. Calculations confirming the ability of the *Ansatz* to find the ground state can be found in Appendix A 4. Optimal values were found with simulated data using a provided noise model as performed by Li *et al.* [23]. The optimal circuit was then measured in a computational basis and on entangled corrective projections. Demonstrations were performed on *ibm_hanoi* using a mapping to qubits 5,3,2,1,4,7 as indicated in Appendix A 6.

Finally, for four qubits, we used the H-H molecule with STO-3G and parity mapping but without any symmetry reduction. We used a hardware-efficient circuit with Ry, Rz, CNOT, Ry, Rz, CNOT, Ry, Rz layers across all qubits. This *Ansatz* is able to find the ground state for the Hamiltonian with parity mapping but not Jordan-Wigner mapping as detailed in Appendix A 4. Optimal values were found using a provided noise model using simulated data. The optimal circuit was then measured in the computational basis and on entangled corrective projections. Experiments were performed on *ibm_hanoi* with mapping to qubits 1,2,3,6,4,7,5,8 as indicated in Appendix A 6.

The entangled projections are determined to recreate sampling statistics for each Pauli string and pairwise entangling gates to recreate the expectation value of S_2.

IV. RESULTS AND DISCUSSION

We first perform a calculation with B gates from Ref. [12] as a comparison to previous work. We will use the two-qubit H_2 molecule calculation as an example, which has a Hamiltonian that decomposes to IZ, ZI, ZZ, XX strings. Since the B gate cannot calculate multiqubit measurements, we must use the method developed in this work for reconstructing the sampling statistics for ZZ and XX Paulis.
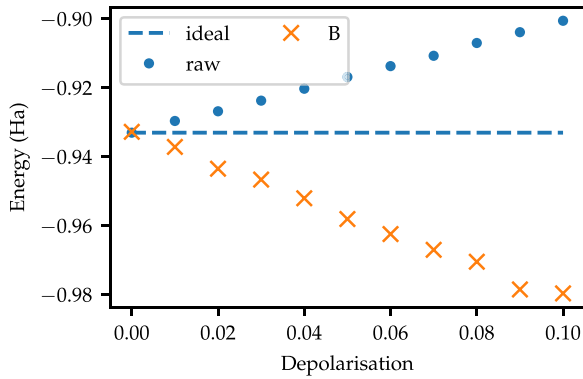
FIG. 2. H_2 energy at a distance of 2 Å as depolarization noise on CNOT gates increases. Ideal refers to the noise-free value at 0 depolarization. Raw refers to directly measuring the *Ansatz*; B refers to using B gates [12] for VD and violates the variational principle by being below ideal.

Figure 2 gives the measured ground-state energy of H_2 at 2 distance as depolarization on each two-qubit gate increases. The depolarization parameter ranges between 0 at no noise and 1 where the qubit is in the maximally mixed state. The dotted line refers to the ideal noiseless measurement; the blue dots give the raw measurement without any correction; the yellow crosses give the B gate corrected values.

It is obvious that depolarization noise causes the B gates to immediately violate the variational principle (see Fig. 1). This is because it has four CNOTs to calculate S_2, whereas our work only uses two at most. This means S_2 sees four depolarization channels and underestimates the purity.

O'Brien *et al.* [20] also noted violation of the variational principle with virtual distillation, however they bound expectation values between $[-1, 1]$, which they state is clearly qualitatively incorrect. They are correct in that assessment: it is obvious that if S_2 measurements have more error than Pauli measurements, observables will appear to take their maximum value when the noiseless value may be nowhere near it. We will now provide results for the method developed in our work, which takes this into account and ensures we can retain the variational principle.

Figure 3 shows the same two-qubit CNOT gate depolarization noise as in Fig. 2. Figure 3(a) shows the same two-qubit H_2 system, and the raw measurement given by blue dots is the same as in Fig. 2. Figure 3(b) shows the three-qubit H_3 linear chain molecule, and Fig. 3(c) shows the four-qubit H_2 molecule with no symmetry reduction.

First, our method does not violate the variational principle, as the computed energy with noise is always higher than the noiseless ideal dotted blue line (see Fig. 1). Second, the corrected values are always a lower energy than the raw values, which means our method is successfully mitigating noise. Specifically, it means the additional entangling gates do not add more noise than they cancel out. We note that while the uncorrected raw energies increase linearly, in this noise range the corrected values increase much slower, especially for four qubits in Fig. 3(c).

Finally, it is observed that the larger systems are more sensitive to small noise. Around 0.01 depolarization, the two-
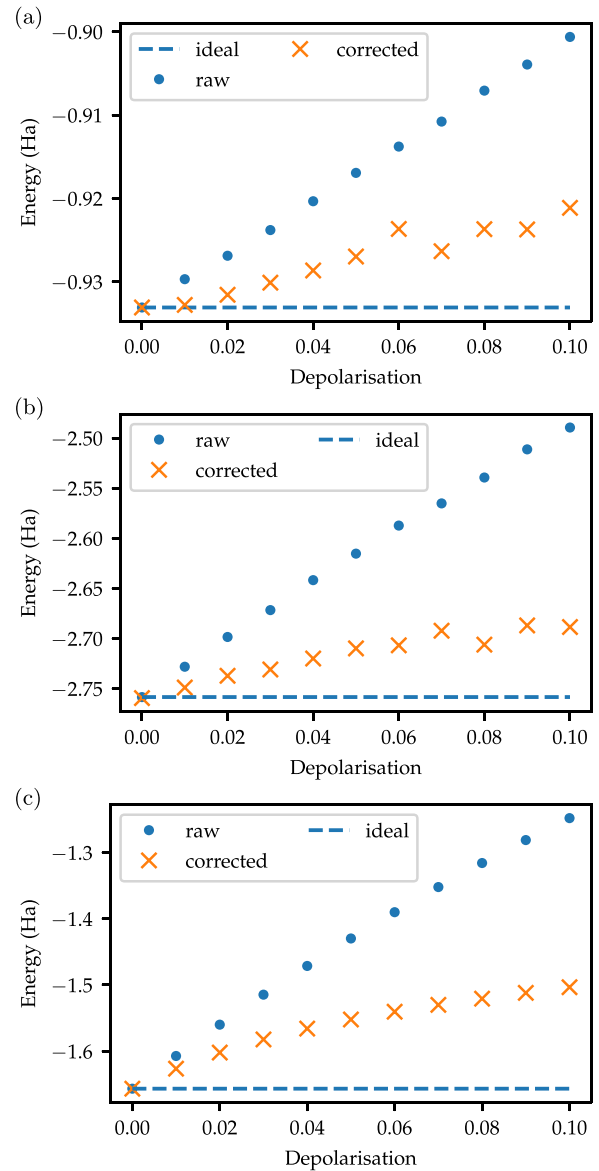

FIG. 3. Effect of depolarization on ground-state energy for two-qubit (H_2), three-qubit (H_3), and four-qubit (H_2, no symmetry reduction) systems at a distance of 2 Å. Simulated data from depolarization channel on each CNOT gate. The dotted blue line gives the ideal noiseless value. Blue dots show raw measured energy. Yellow crosses give corrected energy.

qubit system is able to be corrected to almost ideal, whereas the four-qubit system cannot. In comparison, due to the slow increase of the corrected values to noise, at high noise values, larger systems perform well. VD by itself is therefore optimal at relatively higher noise (up to reasonable noise for a current device). However, since it always corrects, there are other techniques that can be used in conjunction that could mitigate VD at small noise and allow it to reach the ideal case.

After a basic test of our method in noisy simulation environments, we can now implement it on a noisy IBM quantum device. We calculate the full dissociation curve for the molecules in question as distance increases. The results are given in Fig. 4. The ideal curve is given as a dotted blue
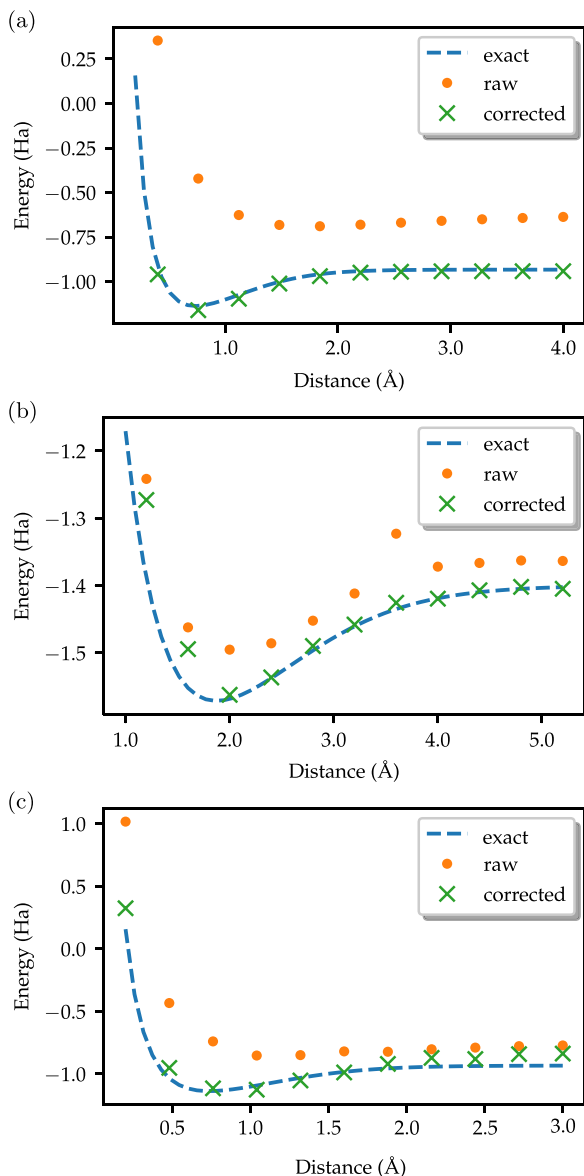
FIG. 4. Dissociation curves for two-qubit, three-qubit, and four-qubit systems on *ibmq_hanoi*. Exact refers to direct diagonalization of the Hamiltonian after qubit mapping. Raw refers to directly measuring the *Ansatz*. Corrected refers to using VD with our circuit decomposition. Part (a) shows H_2 molecule with UCC *Ansatz* and parity reduction. Part (b) shows H_3 linear chain molecule with hardware-efficient *Ansatz*. Part (c) shows H_2 molecule with hardware-efficient *Ansatz*.
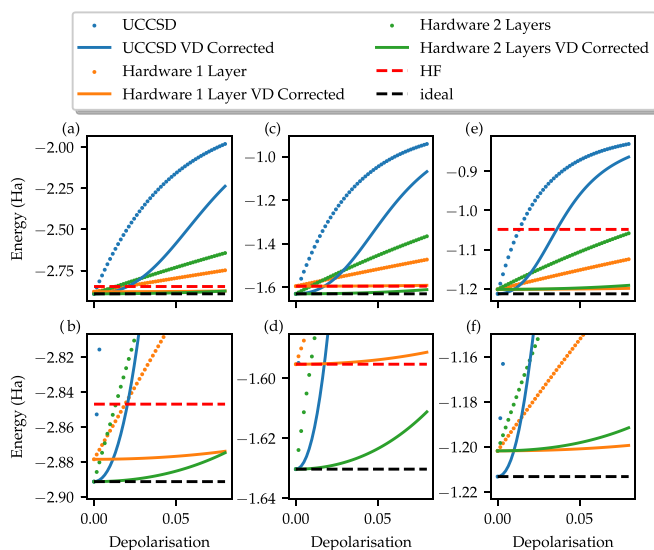

FIG. 5. Simulated *Ansatz* minimum energies raw and corrected with virtual distillation as depolarisation noise increases for UCCSD and hardware-efficient *Ansatz* with one and two layers. Part (a) shows the three-qubit H_3 in Jordan-Wigner mapping and Z_2 symmetry reduction with an expanded section in (b). Part (c) shows the four-qubit H_4 with parity mapping with an expanded section in (d). Part (e) shows the four-qubit H_4 with Jordan-Wigner mapping with an expanded section in (f). HF refers to Hartree-Fock initial guess and is the lowest depth circuit as it contains no entangling gates. Nuclear repulsion energy is not added.

line, the raw values are yellow dots, and the corrected values are green crosses. As above, Fig. 4(a) is the two-qubit H_2 calculation, (b) is the three-qubit H_3 linear chain, and (c) is the four-qubit H_2 calculation with no symmetry reduction.

The results follow similar trends to the simulated data. For the smallest system, in Fig. 4(a) we note that the ideal energy can be recreated up to shot noise, despite large errors on the raw data. For the largest system in Fig. 4(c), we can note a similar correction close to ideal, however the curve as the hydrogen atoms are separated is not able to recreate the exact energies. Furthermore, from the comparison of raw and exact

data, it appears that the noise in Fig. 4(c) is much less than in Fig. 4(a).

Similarly, the effect of shot noise was considered. The demonstrations from Fig. 3 were simulated with 8196 shots and no other noise. This was repeated 100 times and the standard deviations in the energy recorded. The standard deviations for the corrected energies are 1.296×10^{-3} Ha for two qubits, 1.386×10^{-3} Ha for three qubits, and 2.038×10^{-3} Ha for four qubits. These are significantly smaller than the discrepancy between all measured and exact energies, which shows that incoherent noise must be responsible rather than shot noise. However, shot noise is of the order of chemical accuracy of approximately 1.594×10^{-3} Ha, and this means significantly more shots are required in applications to reach chemical accuracy even if incoherent noise can be mitigated.

While this method is shown to work, care must be taken to know the exact hardware implementation to ensure that the S_2 measurements do not have significantly more error than the Pauli measurements. In particular, additional swap gates or dynamical decoupling steps may be introduced by a compiler that can violate this. In the Appendix, we detail
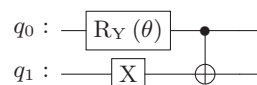

FIG. 6. Manually reduced *Ansatz* for Z_2 symmetry reduced UCC H_2 two-qubit Hamiltonian. The Hartree-Fock initial guess is mapped to $|01\rangle$. The excitations were found to map to the space spanned by $|01\rangle$ and $|10\rangle$ with real coefficients.
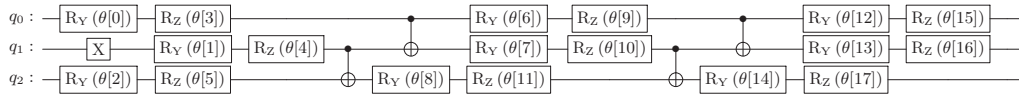
FIG. 7. Three-qubit two-layer hardware-efficient *Ansatz* for the Jordan-Wigner H$_3$ Hamiltonian. The initial Hartree-Fock guess is mapped to $|010\rangle$.

several alternative methods that can be used to tailor this method to hardware requirements. We show how to calculate Paulis by recreating their expectation value (as we did with S_2 in the main text), however this adds a multi-controlled-Z gate, which can be prohibitive, and if three-qubit gates can be implemented, the ancilla-based Fredkin approach may be more desirable. Similarly, we show how to decompose all expectation values into Pauli strings, which requires no extra entangling gates and has the least noise, but will require the most measurements. At worst, it will need to measure all Paulis for all doubled qubits, at which point it will take fewer measurements to do a tomography on the original circuit and analytically calculate powers of the density matrix. In that sense, for specific problems and hardware limitations, these alternate methods could be useful, however in the specific case of chemical calculations on IBM architecture our method of recreating Pauli sampling statistics and S_2 expectation values was found to be optimal.

In conclusion, we have demonstrated that the B gate circuit decomposition for VD only works for observables that commute with S_2 and can violate the variational principle otherwise. We found a method for circuit decomposition that maintains the variational principle which enables VQE applications. Our method increases the number of measurements but reduces the number of entangling gates compared to the B gate method. We found that it performs well at different noise scales and on real devices. Future work can investigate combining VD with other noise mitigation strategies to see if chemical accuracy can be reached.

ACKNOWLEDGMENT

APPENDIX

1. Linear combination of measurements for $\langle S_2 \rangle$

Recall that the goal is to recreate the expectation value of the swap operator. Let us say we have $|\psi\rangle = A|00\rangle + B|01\rangle + C|10\rangle$, then

$$\langle\psi|S_2|\psi\rangle = |A|^2 + \overline{B}C + \overline{C}B. \tag{A1}$$

The squared terms are probabilities measured from computational measurement. The goal is thus to measure a quantity that corresponds to the cross terms added together with their complex conjugates. This will be accomplished by the CNOT H circuit (written in circuit order). We want to know what probabilities occur after an arbitrary state $|\psi\rangle$ is projected by the circuit CNOT H on each pair. Ultimately we will be taking probabilities from measuring in the computational basis and taking linear combinations of them. Therefore, the first step is to see what the projection of CNOT H looks like on an arbitrary state $|\psi\rangle$.

To make the analysis easier, we reorder qubits such that each qubit is adjacent with its corresponding qubit from the copied system, i.e., for a two-qubit system $|abcd\rangle \to |acbd\rangle$. This now makes the swap operator S_2 act as

$$S_2|abcd\cdots\rangle = |badc\cdots\rangle. \tag{A2}$$

In this qubit ordering, the projection circuit is now CNOT H on each pair in order, so we can consider the action only on separable pairs of qubits, specifically now only the top two qubits. The operation of CNOT H has the form

$$|+\rangle|0\rangle \otimes I + |-\rangle|1\rangle \otimes X. \tag{A3}$$

We want to know what the outcome of measuring in the computational basis looks like after this. If we project onto state $\langle01|$, we get

$$\langle01|H\text{CNOT} = \frac{1}{\sqrt{2}}\langle01| + \langle10|, \tag{A4}$$

which means that if we start with the state $|\psi\rangle$ and project using the circuit, we will measure the outcome 01 with probability

$$\frac{1}{2}(\langle\psi|01\rangle + \langle\psi|10\rangle)(\langle01|\psi\rangle + \langle10|\psi\rangle)$$
$$= \frac{1}{2}(|B|^2 + |C|^2 + \overline{B}C + \overline{C}B). \tag{A5}$$

Similarly, the outcome 10 occurs with probability

$$\frac{1}{2}(\langle\psi|01\rangle - \langle\psi|10\rangle)(\langle01|\psi\rangle - \langle10|\psi\rangle)$$
$$= \frac{1}{2}(|B|^2 + |C|^2 - \overline{B}C - \overline{C}B). \tag{A6}$$

Therefore, if we subtract the probability of measuring outcome 10 from 01, we get $\overline{B}C + \overline{C}B$, which matches the quantity required in Eq. (A1).
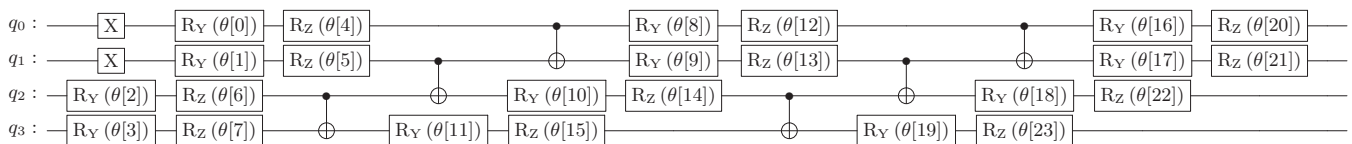

FIG. 8. Four-qubit two-layer hardware-efficient *Ansatz* for H$_2$ molecule with parity mapping but no symmetry reduction. Hartree-Fock state is mapped to $|1100\rangle$.
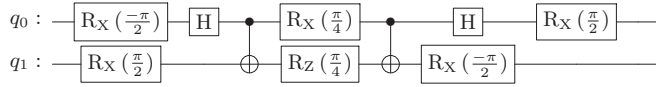
FIG. 9. B gate over two qubits. For a larger number of qubits, we require B gates over every pair of qubits.

Explicitly, for arbitrary numbers of qubits, the action of CNOT H on the top two qubits is to map

$$\langle 00| \otimes \langle abc\cdots| \rightarrow (\langle 00| + \langle 11|) \otimes \langle abc\cdots|,$$
$$\langle 01| \otimes \langle abc\cdots| \rightarrow (\langle 01| + \langle 10|) \otimes \langle abc\cdots|,$$
$$\langle 10| \otimes \langle abc\cdots| \rightarrow (\langle 00| - \langle 11|) \otimes \langle abc\cdots|,$$
$$\langle 11| \otimes \langle abc\cdots| \rightarrow (\langle 01| - \langle 10|) \otimes \langle abc\cdots|. \quad \text{(A7)}$$

If we have CNOT H gates on every pair, we end up with all qubit pairs having entangled projections as above. Let us label

$$\langle S+| = \frac{1}{\sqrt{2}}(\langle 00| + \langle 11|),$$

$$\langle A+| = \frac{1}{\sqrt{2}}(\langle 01| + \langle 10|),$$

$$\langle S-| = \frac{1}{\sqrt{2}}(\langle 00| - \langle 11|),$$

$$\langle A-| = \frac{1}{\sqrt{2}}(\langle 01| - \langle 10|). \quad \text{(A8)}$$

The action of CNOT H creates several orbits, i.e., it will create entanglement between a few select states with each other. Specifically, for some fixed number of S and A states, the orbit will contain every combination of $+$ and $-$. This is important because fixing the S and A nature of each qubit pair fixes the projected states, whereas the $+$, $-$ nature only affects the relative sign.

We can write the linear combination of measurements in an example orbit of just S as

$$\langle \psi|(a|S+\rangle + b|S-\rangle)|\psi\rangle. \quad \text{(A9)}$$

We want this to equal $\langle S_2 \rangle$. We can find coefficients by projecting onto the operator, for example,

$$a = \text{tr}(S_2|S+\rangle\langle S+|). \quad \text{(A10)}$$

Specifically, we see that

$$S_2|S+\rangle = |S+\rangle,$$
$$S_2|S-\rangle = |S-\rangle,$$
$$S_2|A+\rangle = |A+\rangle,$$
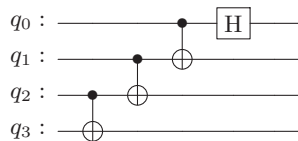$$S_2|A-\rangle = -|A-\rangle. \quad \text{(A11)}$$

FIG. 10. Entangling projection required to recreate Pauli sampling probabilities for H_2.
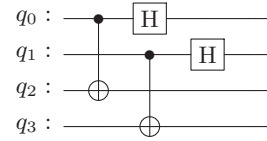

FIG. 11. Entangling projection required to recreate expectation value of S_2 for H_2.

This means the coefficients are either 1 or -1, and once they are solved, the linear combination of measurements can be made to estimate $\langle S_2 \rangle$.

As an example, for H_2, the orbit of $|0001\rangle$ is $|0001\rangle$, $|0100\rangle$, $|1011\rangle$, $|1110\rangle$. We can reorder the qubits and rewrite this as $|0001\rangle$, $|0010\rangle$, $|1101\rangle$, $|1110\rangle$. This can be decomposed into

$$|S+\rangle|A+\rangle, |S+\rangle|A-\rangle, |S-\rangle|A+\rangle, |S-\rangle|A-\rangle. \quad \text{(A12)}$$

Therefore, only the terms with $|A-\rangle$ will pick up a negative sign, and we can write the expectation value as

$$\langle \psi|(|S+\rangle|A+\rangle\langle S+|\langle A+| - |S+\rangle|A-\rangle\langle S+|\langle A-| \\ + |S-\rangle|A+\rangle\langle S-|\langle A+| - |S-\rangle|A-\rangle\langle S-|\langle A-|)|\psi\rangle. \quad \text{(A13)}$$

We can also identify which measurements these correspond to by looking at the action of CNOT H on the states to be projected. Let P_0 be the probability of measuring outcome 0. We can now write the final expectation value as

$$P_{0001} - P_{0101} + P_{1001} - P_{1101}. \quad \text{(A14)}$$

This process is then repeated for each orbit. This can be used for arbitrary numbers of qubits, and it serves as a low depth method of measuring the purity with two copies of a quantum circuit.

2. Recreating Pauli expectation values

While $\langle S_2 \rangle$ only requires pairwise CNOT gates and Hadamards on one copy of the *Ansatz*, additional gates are needed to reconstruct the expectation value of arbitrary Pauli strings. First, any X and Y observables can be mapped to Z by applying a Hadamard or Hadamard and $-\frac{\pi}{2}$ phase gate, respectively, so we only consider Pauli observables O that contain I or Z. Thus, $\langle S_2 O \rangle$, is only modified by adding a -1 phase to states where $O|\psi\rangle = -|\psi\rangle$. The additional gates we require must add this phase to those specific states.

Unfortunately, in the general case, this will require multi-control-Z gates with control on arbitrary many qubits. Multicontrol gates decompose into deep two-qubit entangling gates, and if even three-qubit entangling gates are efficient, it may be better to simply measure expectation values with an ancilla and Hadamard test. However, the advantage depends on the topology, and entangling to one ancilla may cause more swap gates than multicontrol Z.

3. Decomposing observables to Pauli strings

Direct decomposition of the observables into Pauli strings is possible and requires no additional entangling gates.
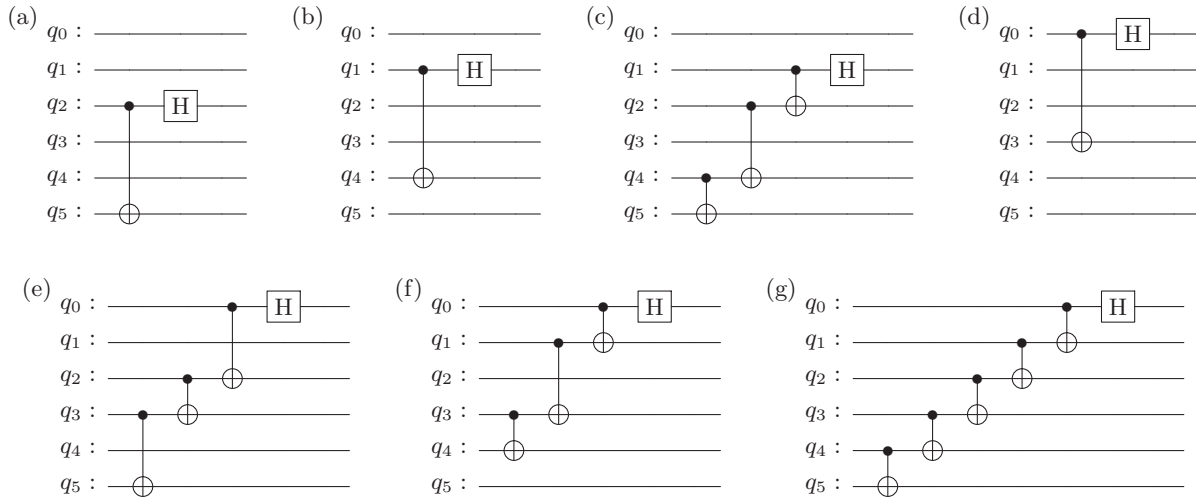
FIG. 12. Entangling projections required to recreate Pauli sampling probabilities for H_3. Seven are required since the Pauli decomposition of the Hamiltonian includes all seven strings of the form $IIP, IPI, PII, IPP, PIP, PPI, PPP$, where $P \in X, Y, Z$. Pauli grouping or rotation can reduce the required number of circuits.

An observable, O, can be decomposed into Pauli strings by calculating the projection along each:

$$a_i = \frac{1}{n}\text{tr}(P_i O). \tag{A15}$$

We can then write

$$O = \sum_i a_i P_i. \tag{A16}$$

As above, we can map measurements of Pauli strings with X and Y to Z by using a Hadamard or Hadamard and $-\frac{\pi}{2}$ phase gate on the corresponding qubit. Then the measurement sampling statistics can be combined with the relevant eigenvalues to get expectation values for every Pauli string with I or Z, since they commute.

This procedure can be performed on the expectation value from Eq. (4). However, since these observables act on twice the qubit number, the corresponding Pauli decomposition will also act on twice the qubits. This means that in the worst case scenario, it may take all 4^{2n} Pauli string measurements to measure Eq. (4).

As mentioned in the main text, a full tomography of one copy of the circuit is sufficient to get the density matrix and analytically perform virtual distillation up to arbitrarily high copies. While this technique adds no extra entangling gates, the exponential scaling in measurements means it can only be used if the Pauli decomposition is guaranteed to be small.
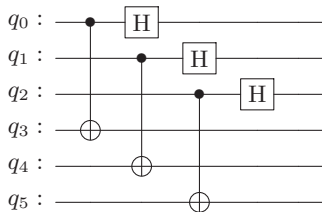

FIG. 13. Entangling projection required to recreate expectation value of S_2 for H_3.

4. *Ansatz* evaluation

For three- and four-qubit calculations, we simulated UCCSD and the hardware-efficient *Ansatz* of increasing depth to determine the lowest depth *Ansatz* of appropriate accuracy to run on real devices. VQE was performed on noise-free simulation with the basin-hopping algorithm to find the global minimum. Once the final parameters were found, the circuit was simulated with increasing depolarization noise up to 0.1 along with the full virtual distillation calculation, and the results for all *Ansätze* are shown in Fig. 5.

UCCSD is able to find the ground state in all cases. However, due to the circuit depth, depolarization beyond 0.01 causes it to be worse than the Hartree-Fock initial guess, even with VD corrections. A hardware-efficient *Ansatz* with one layer is not expressive enough to recover the ground-state energy in any cases. It is also interesting to note that it is worse than the Hartree-Fock initial guess for the four-qubit parity Hamiltonian shown in (d). The hardware-efficient *Ansatz* with two layers is able to recover the ground state and does not exceed the Hartree-Fock limit at 0.1 depolarization for the three-qubit Hamiltonian in (b) and the four-qubit parity Hamiltonian in (d), however it is not expressive enough to find the ground state for the four-qubit Jordan-Wigner Hamiltonian in (f). As such, the obvious choice of *Ansatz* is the hardware-efficient *Ansatz* with two layers for three-qubit H_3 calculation and four-qubit H_2 calculation with the parity mapping.

5. Circuit diagrams

In this section, we will show that the circuits run on real devices in this work. First we show the *Ansatz* used. For H_2 (two qubit) shown in Fig. 6 we design our own *Ansatz* based on the symmetry of the ground state in the parity mapping with Z_2 symmetry reduction. For H_3 (three qubit) and H_2 (four qubit) we use a basic SU2 hardware-efficient circuit with two repetitions in Figs. 7 and 8. Next we show that the B gates directly measure the S_2 expectation value in Fig. 9. This is applied to every pair of qubits. For this work, we replace
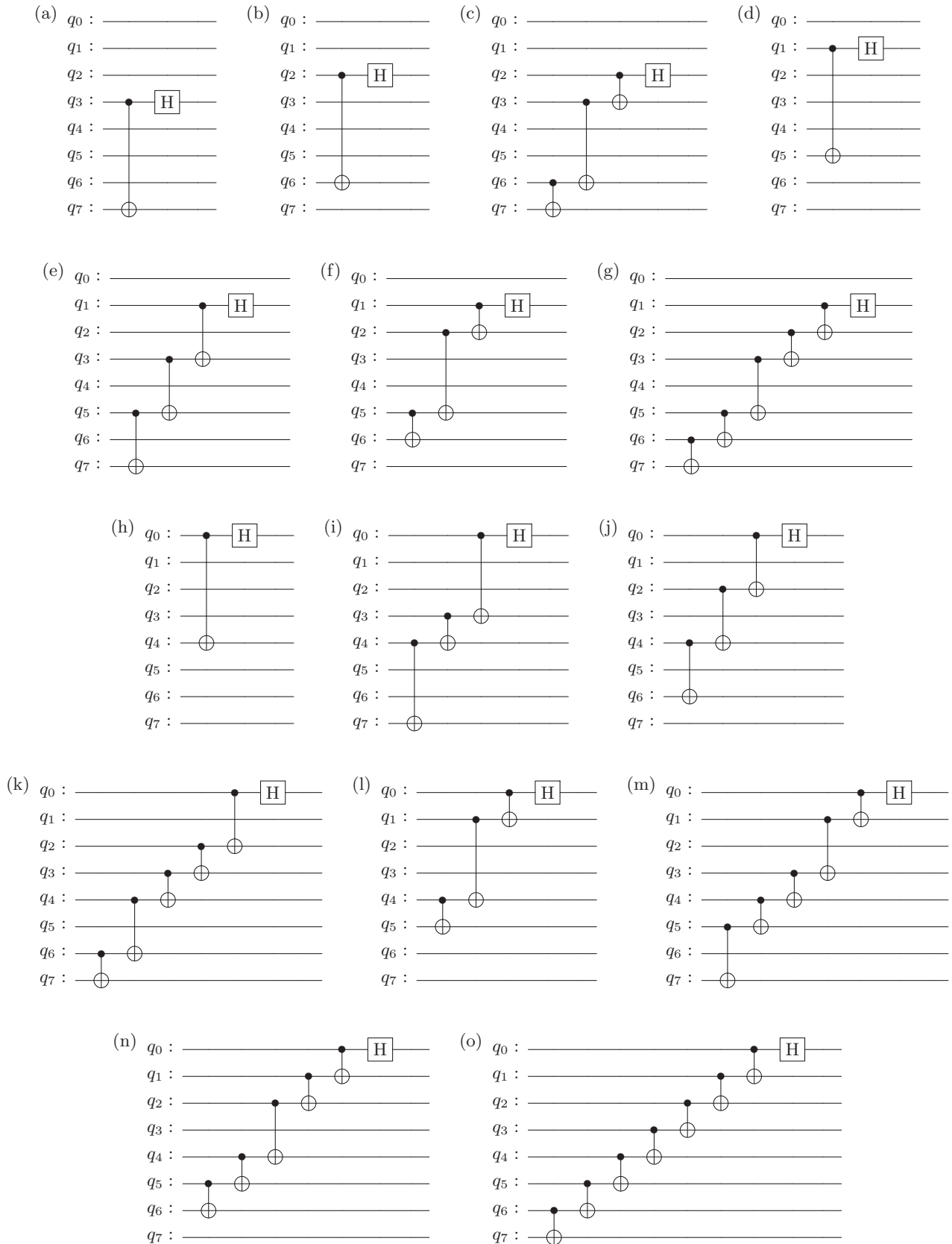
FIG. 14. Entangling projections required to recreate Pauli sampling probabilities for H_2 mapped to four qubits. All 15 are required due to the Pauli decomposition of the Hamiltonian. Pauli grouping or rotation can reduce the required number of circuits.

TABLE I. Table of *ibm_hanoi* device parameters. The frequency, T_1, T_2, and readout noise are given for each qubit, and the CX error is given for pairs of qubits.

|	q_0	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}	q_{11}	q_{12}	q_{13}	q_{14}	q_{15}	q_{16}	q_{17}	q_{18}	q_{19}	q_{20}	q_{21}	q_{22}	q_{23}	q_{24}	q_{25}	q_{26}
Frequency (GHz)	5.035	5.155	5.256	5.097	5.072	5.206	5.020	4.919	5.031	4.874	4.821	5.162	4.718	4.963	5.046	4.923	4.883	5.222	4.967	5.002	5.095	4.839	4.919	4.920	4.991	4.812	5.019
T_1 (μs)	174.9	224.6	127.8	103.4	134.0	179.8	132.2	170.0	103.7	135.1	114.3	119.3	183.7	249.7	104.3	164.7	167.8	119.4	123.0	145.7	66.70	158.4	80.58	166.4	136.9	175.7	83.87
T_2 (μs)	210.5	159.4	183.0	32.45	14.13	231.8	278.4	125.8	185.2	114.7	102.6	243.0	160.8	114.3	23.30	37.32	228.6	67.57	162.9	126.8	93.77	36.04	88.66	241.7	27.89	108.9	31.50
Readout error	0.0082	0.0098	0.0121	0.0083	0.0078	0.0087	0.046	0.0105	0.0106	0.0072	0.0121	0.088	0.146	0.068	0.0103	0.0521	0.0119	0.0178	0.0154	0.0073	0.0100	0.0271	0.0125	0.0127	0.0116	0.0080	0.0080
CX error q_0		0.52																									
CX error q_1	0.52		0.30		0.38																						
CX error q_2		0.30		0.95																							
CX error q_3			0.95			0.37																					
CX error q_4		0.38						1.40																			
CX error q_5				0.37					100																		
CX error q_6								0.71																			
CX error q_7					1.40		0.71				0.67																
CX error q_8						100				1.03		0.55															
CX error q_9									1.03																		
CX error q_{10}								0.67					0.60														
CX error q_{11}									0.55						0.72												
CX error q_{12}											0.60			0.53		0.88											
CX error q_{13}													0.53		0.55												
CX error q_{14}												0.72		0.55			1.32										
CX error q_{15}													0.88						1.22								
CX error q_{16}															1.32					0.38							
CX error q_{17}																			0.57								
CX error q_{18}																1.22		0.57				0.40					
CX error q_{19}																	0.38				100		0.69				
CX error q_{20}																				100							
CX error q_{21}																			0.40					0.66			
CX error q_{22}																				0.69						0.83	
CX error q_{23}																						0.66			1.54		
CX error q_{24}																								1.54		2.54	
CX error q_{25}																							0.83		2.54		0.68
CX error q_{26}																										0.68	
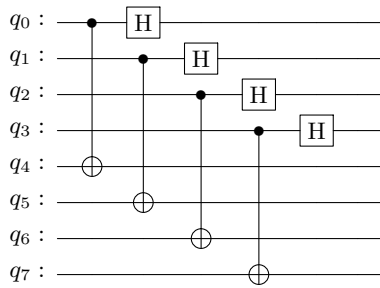
FIG. 15. Entangling projection required to recreate expectation value of S_2 for H_2 mapped to four qubits.

the B gates with combinations of computational and entangled measurements. The entangled gates for Pauli sampling reconstructing of H_2 (two qubit) are given in Fig. 10. The entangling circuit for reconstructing the S_2 expectation value is given in Fig. 11. Similarly, the entangled gates for H_3 (three qubits) Paulis are given in Fig. 12 and S_2 in Fig. 13. The entangled gates for H_2 (four qubits) are given in Fig. 14 and S_2 in Fig. 15.

6. IBM device characteristics

In this section, we show the layout of the *ibm_hanoi* chip used for demonstrations in this work in Fig. 16 along with
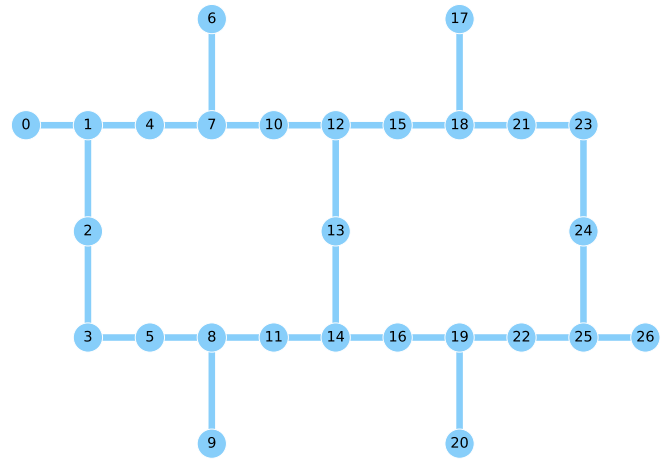

FIG. 16. Layout of the *ibm_hanoi* device. Two-qubit calculation uses qubits 1,4,7,8. Three-qubit calculation uses 5,3,2,1,4,7. Four-qubit calculation uses 1,2,3,6,4,7,5,8.

noise spectroscopy collected around the time of the simulations in Table I. It should be noted that other sources of noise exist in real devices, and this only represents the sources of noise that were measured.

[1] A. Montanaro, Quantum algorithms: An overview, npj Quantum Inf. **2**, 15023 (2016).

[2] P. Shor, Algorithms for quantum computation: Discrete logarithms and factoring, in *Proceedings of the 35th Annual Symposium on Foundations of Computer Science, Santa Fe, NM, USA* (IEEE, New York, 1994), pp. 124–134.

[3] M. A. Broome, A. Fedrizzi, S. Rahimi-Keshari, J. Dove, S. Aaronson, T. C. Ralph, and A. G. White, Photonic boson sampling in a tunable circuit, Science **339**, 794 (2013).

[4] R. J. Chapman, A. Karim, Z. Huang, S. T. Flammia, M. Tomamichel, and A. Peruzzo, Beating the classical limits of information transmission using a quantum decoder, Phys. Rev. A **97**, 012315 (2018).

[5] M. J. Bremner, A. Montanaro, and D. J. Shepherd, Achieving quantum supremacy with sparse and noisy commuting quantum computations, Quantum **1**, 8 (2017).

[6] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler *et al.*, Quantum supremacy using a programmable superconducting processor, Nature (London) **574**, 505 (2019).

[7] F. Pan, K. Chen, and P. Zhang, Solving the sampling problem of the sycamore quantum circuits, Phys. Rev. Lett. **129**, 090502 (2022).

[8] M. J. Bremner, B. Cheng, and Z. Ji, IQP sampling and verifiable quantum advantage: Stabilizer scheme and classical security, arXiv:2308.07152.

[9] A. M. Dalzell, A. W. Harrow, D. E. Koh, and R. L. La Placa, How many qubits are needed for quantum computational supremacy? Quantum **4**, 264 (2020).

[10] P. Shor, Fault-tolerant quantum computation, in *Proceedings of the 37th Conference on Foundations of Computer Science, Burlington, VT, USA* (IEEE, New York, 1996), pp. 56–65.

[11] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. OBrien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[12] W. J. Huggins, S. McArdle, T. E. OBrien, J. Lee, N. C. Rubin, S. Boixo, K. B. Whaley, R. Babbush, and J. R. McClean, Virtual distillation for quantum error mitigation, Phys. Rev. X **11**, 041036 (2021).

[13] B. Nachman, M. Urbanek, W. A. de Jong, and C. W. Bauer, Unfolding quantum computer readout noise, npj Quantum Inf. **6**, 84 (2020).

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

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

[16] S. H. Sack and D. J. Egger, Large-scale quantum approximate optimization on nonplanar graphs with machine learning noise mitigation, arXiv:2307.14427.

[17] M. Urbanek, B. Nachman, V. R. Pascuzzi, A. He, C. W. Bauer, and W. A. de Jong, Mitigating depolarizing noise on quantum computers with noise-estimation circuits, Phys. Rev. Lett. **127**, 270502 (2021).

[18] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. van den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme, and A. Kandala, Evidence for the utility of quantum computing before fault tolerance, Nature (London) **618**, 500 (2023).

[19] R. Mezher, J. Mills, and E. Kashefi, Mitigating errors by quantum verification and postselection, Phys. Rev. A **105**, 052608 (2022).

[20] T. E. O'Brien, G. Anselmetti, F. Gkritsis, V. E. Elfving, S. Polla, W. J. Huggins, O. Oumarou, K. Kechedzhi, D. Abanin, R. Acharya, I. Aleiner, R. Allen, T. I. Andersen, K. Anderson, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. Atalaya, J. C. Bardin *et al.*, Purification-based quantum error mitigation of pair-correlated electron simulations, Nat. Phys. **19**, 1787 (2023).

[21] C. Cao, Y. Yu, Z. Wu, N. Shannon, B. Zeng, and R. Joynt, Mitigating algorithmic errors in quantum optimization through energy extrapolation, Quantum Sci. Technol. **8**, 015004 (2023).

[22] A. Seif, Z.-P. Cian, S. Zhou, S. Chen, and L. Jiang, Shadow distillation: Quantum error mitigation with classical shadows

[23] for near-term quantum processors, PRX Quantum **4**, 010303 (2023).

[23] P. Li, J. Liu, H. P. Patil, P. Hovland, and H. Zhou, Enhancing virtual distillation with circuit cutting for quantum error mitigation, arXiv:2310.04708.

[24] Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, S. Wouters, and G. K.-L. Chan, PySCF: The Python-based simulations of chemistry framework, WIREs Comput. Molec. Sci. **8**, e1340 (2018).

[25] W. J. Hehre, R. F. Stewart, and J. A. Pople, Self-consistent molecular-orbital methods. I. use of Gaussian expansions of Slater-type atomic orbitals, J. Chem. Phys. **51**, 2657 (1969).

[26] S. Bravyi, J. M. Gambetta, A. Mezzacapo, and K. Temme, Tapering off qubits to simulate fermionic Hamiltonians, arXiv:1701.08213.