Purification-based quantum error mitigation of pair-correlated electron simulations: Supplementary Information

T. E. O'Brien,[1,2,*] G. Anselmetti,[3] F. Gkritsis,[3] V. E. Elfving,[4] S. Polla,[1,2] W. J. Huggins,[1] O. Oumarou,[3] K. Kechedzhi,[1] D. Abanin,[1] R. Acharya,[1] I. Aleiner,[1] R. Allen,[1] T. I. Andersen,[1] K. Anderson,[1] M. Ansmann,[1] F. Arute,[1] K. Arya,[1] A. Asfaw,[1] J. Atalaya,[1] D. Bacon,[1] J. C. Bardin,[1,5] A. Bengtsson,[1] S. Boixo,[1] G. Bortoli,[1] A. Bourassa,[1] J. Bovaird,[1] L. Brill,[1] M. Broughton,[1] B. Buckley,[1] D. A. Buell,[1] T. Burger,[1] B. Burkett,[1] N. Bushnell,[1] J. Campero,[1] Y. Chen,[1] Z. Chen,[1] B. Chiaro,[1] D. Chik,[1] J. Cogan,[1] R. Collins,[1] P. Conner,[1] W. Courtney,[1] A. L. Crook,[1] B. Curtin,[1] D. M. Debroy,[1] S. Demura,[1] I. Drozdov,[1,6] A. Dunsworth,[1] C. Erickson,[1] L. Faoro,[1] E. Farhi,[1] R. Fatemi,[1] V. S. Ferreira,[1] L. Flores Burgos,[1] E. Forati,[1] A. G. Fowler,[1] B. Foxen,[1] W. Giang,[1] C. Gidney,[1] D. Gilboa,[1] M. Giustina,[1] R. Gosula,[1] A. Grajales Dau,[1] J. A. Gross,[1] S. Habegger,[1] M. C. Hamilton,[1,7] M. Hansen,[1] M. P. Harrigan,[1] S. D. Harrington,[1] P. Heu,[1] J. Hilton,[1] M. R. Hoffmann,[1] S. Hong,[1] T. Huang,[1] A. Huff,[1] L. B. Ioffe,[1] S. V. Isakov,[1] J. Iveland,[1] E. Jeffrey,[1] Z. Jiang,[1] C. Jones,[1] P. Juhas,[1] D. Kafri,[1] J. Kelly,[1] T. Khattar,[1] M. Khezri,[1] M. Kieferová,[1,8] S. Kim,[1] P. V. Klimov,[1] A. R. Klots,[1] A. N. Korotkov,[1,9] F. Kostritsa,[1] J. M. Kreikebaum,[1] D. Landhuis,[1] P. Laptev,[1] K. Lau,[1] L. Laws,[1] J. Lee,[1,10] K. Lee,[1] B. J. Lester,[1] A. T. Lill,[1] W. Liu,[1] W. P. Livingston,[1] A. Locharla,[1] E. Lucero,[1] F. D. Malone,[1] S. Mandra,[1] O. Martin,[1] S. Martin,[1] J. R. McClean,[1] T. McCourt,[1] M. McEwen,[1,11] A. Megrant,[1] X. Mi,[1] A. Mieszala,[1] K. C. Miao,[1] M. Mohseni,[1] S. Montazeri,[1] A. Morvan,[1] R. Movassagh,[1] W. Mruczkiewicz,[1] O. Naaman,[1] M. Neeley,[1] C. Neill,[1] A. Nersisyan,[1] H. Neven,[1] M. Newman,[1] J. H. Ng,[1] A. Nguyen,[1] M. Nguyen,[1] M. Y. Niu,[1] S. Omonije,[1] A. Opremcak,[1] A. Petukhov,[1] R. Potter,[1] L. P. Pryadko,[1,12] C. Quintana,[1] C. Rocque,[1] P. Roushan,[1] N. Saei,[1] D. Sank,[1] K. Sankaragomathi,[1] K. J. Satzinger,[1] H. F. Schurkus,[1] C. Schuster,[1] M. J. Shearn,[1] A. Shorter,[1] N. Shutty,[1] V. Shvarts,[1] J. Skruzny,[1] V. Smelyanskiy,[1] W. C. Smith,[1] R. Somma,[1] G. Sterling,[1] D. Strain,[1] M. Szalay,[1] D. Thor,[1] A. Torres,[1] G. Vidal,[1] B. Villalonga,[1] C. Vollgraff Heidweiller,[1] T. White,[1] B. W. K. Woo,[1] C. Xing,[1] Z. J. Yao,[1] P. Yeh,[1] J. Yoo,[1] G. Young,[1] A. Zalcman,[1] Y. Zhang,[1] N. Zhu,[1] N. Zobrist,[1] C. Gogolin,[3,†] R. Babbush,[1,‡] and N. C. Rubin[1,§]

[1] *Google Research, Mountain View, CA 94043, United States*
[2] *Instituut-Lorentz, Universiteit Leiden, 2300 RA Leiden, The Netherlands*
[3] *Covestro Deutschland AG, 51365 Leverkusen, Germany*
[4] *PASQAL SAS, 91300 Massy, France*
[5] *Department of Electrical and Computer Engineering, University of Massachusetts Amherst, Amherst MA, USA*
[6] *Department of Physics, University of Connecticut, Storrs, CT*
[7] *Department of Electrical and Computer Engineering, Auburn University, Auburn, AL*
[8] *QSI, Faculty of Engineering and Information Technology, University of Technology Sydney, NSW, Australia*
[9] *Department of Electrical and Computer Engineering, University of California, Riverside, CA USA*
[10] *Department of Chemistry, Columbia University, New York, NY, USA*
[11] *Department of Physics, University of California, Santa Barbara, CA, USA*
[12] *Department of Physics and Astronomy, University of California, Riverside, CA*

I. CALIBRATION OF THE PROCESSOR

All experiments were implemented on a subgrid of a 25-qubit superconducting processor with the Sycamore architecture. For all methods other than virtual distillation, a $2 \times N/2$ qubit grid was calibrated to within 0.008 XEB fidelity [1] and 0.008 speckle purity [1]. For virtual distillation, a $4 \times N/2$ qubit grid was calibrated to within 0.01 XEB fidelity and 0.01 speckle purity.

We were further required to calibrate the single-qubit Z-phases accumulated during a CZ gate. This is a well-documented issue [2, 3], but is complicated in our case by the addition of microwave gates. These are observed to bleed into the CZ gate, which made standard Floquet calibration techniques inaccurate. To solve this issue, we calibrate CZ gates *in-situ*. The Givens-SWAP gate was altered by, after each CZ between qubits i and j, inserting virtual rotations $\exp(iZ_i\beta_i^{(j)}), \exp(iZ_j\beta_j^{(i)})$ on qubit i and j respectively. The phases $\beta_i^{(j)}$ were calibrated by running two experiments in series. Firstly, a single $GS(0) = \text{SWAP}$ gate was implemented between qubits i and j (with virtual gates inserted); the qubits were prepared in the state $|0+\rangle$ measured in the ZX or ZY basis, or prepared in the state $|+0\rangle$ and measured in the XZ or YZ basis. Sweeping $\beta_i^{(j)}$ and $\beta_j^{(i)}$ gave four datasets that could be fitted to extract optimal phase offsets. The resulting gate was then benchmarked by estimating $\langle XI\rangle$ and $\langle YI\rangle$ on the state $[GS(0)]^{2k}|0+\rangle$ and $\langle IX\rangle$ and $\langle IY\rangle$ on $[GS(0)]^{2k}|+0\rangle$, and fitting this to an oscillatory de-

* teobrien@google.com
† christian.gogolin@covestro.com
‡ babbush@google.com
§ nickrubin@google.com

cay curve. Under this benchmark, the initial calibration typically reduced the accumulated phase per CZ to less than 30 milliradians. This benchmark was further used to calibrate, by sweeping $\beta_i + \beta_j$ on pairs i, j that are being acted on by the same GS gate to remove the remaining oscillations. We find in practice that a cubic fit to 11 datasets is a robust way to perform a final estimate of $\beta_i + \beta_j$, with the residual phase less than 5 milliradians when calibration was successful. If the estimated fidelity of the resulting GS gate underperformed ($> 1.5\%$ error per CZ gate), qubit or coupler frequencies were reoptimized before recalibrating. Calibration was performed in parallel on sets of CZ gates that were run in parallel during an experiment, to mimic the local environment and compensate for 2-qubit gate crosstalk.

II. FURTHER DETAILS OF THE UPCCD ANSATZ

A. BQP-completeness of nearest neighbor Givens-swap circuits

Here we substantiate the claim that the UpCCD circuits realized on hardware in this work are in general not efficiently classically simulable. We do so by constructing a universal quantum gate set on a reduced Hilbert space (dual-rail encoding) with an $O(1)$ depth overhead. This construction shows that any nearest neighbor depth-$O(N)$ circuit on a line of qubits can be mapped to a depth-$O(N)$ UpCCD ansatz (and circuits with arbitrary connectivity to a depth-$O(N^2)$ UpCCD ansatz), when allowing for the omission of gates (as the identity is not a GS gate). For this to hold it is pivotal that the GS gate family includes the SWAP gate and is thus not a match-gate.

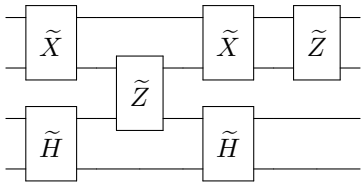
To demonstrate a universal gate set we use a dual-rail encoding of one logical qubit into two physical qubits (onto which the GS gates will act). We use tilde to denote logical states and operations and set

$$|\tilde{0}\rangle \coloneqq |01\rangle \tag{1}$$
$$|\tilde{1}\rangle \coloneqq |10\rangle. \tag{2}$$

It is then straightforward to verify by direct computation that a GS gate acting on two physical qubits belonging to the same logical qubit can be used to realize the following logical Hadamard, Pauli, and Pauli rotation gates:

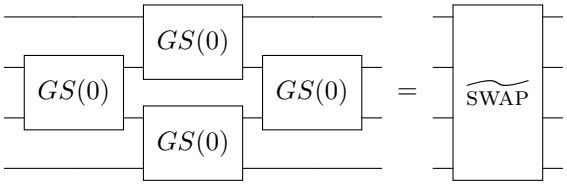
$$GS(\frac{\pi}{4}) = \widetilde{H} \tag{3}$$
$$GS(0) = \widetilde{X} \tag{4}$$
$$GS(\frac{\pi}{2}) = \widetilde{Z} \tag{5}$$
$$GS(0) \cdot GS(\theta) = \widetilde{RY}(\theta) \tag{6}$$

Logical two qubit entangling gates can be realized by acting with GS gates on qubits belonging to two different logical qubits. The $\widetilde{\text{CNOT}}$ gate can for instance be made by means of



(7)

and since $GS(0) = \text{SWAP}$ a $\widetilde{\text{SWAP}}$ is obtainable by means of the planar circuit



. (8)

Since CNOT, SWAP, and the above single qubit gates are universal for quantum computation, any circuit from a family recognizing a language in BQP can be represented as a planar GS gate circuit on twice as many qubits and with an at most polynomially larger depth.

B. Gate decompositions

To implement the $GS(\theta)$ gate on superconducting hardware requires us to decompose it into native gates. In our case this is arbitrary single-qubit rotations and two-qubit number-conserving excitation gates [3]. To minimize calibration overhead, we limit ourselves to a fixed two-qubit gate; in all experiments performed this was a controlled-Z gate. We can write an arbitrary single-qubit rotation in phased-XZ form [1]

$$R(\alpha_x, \alpha_a, \alpha_z) = e^{i(\alpha_z+\alpha_a)Z} e^{i\alpha_x X} e^{-i\alpha_a Z} \tag{9}$$

A $GS(\theta)$ gate can be executed at arbitrary θ on qubits i and j using a combination of 3 CZ gates and single-qubit rotations

$$\begin{aligned} GS_{i,j}(\theta) =& R_i(\tfrac{\pi}{4}, \tfrac{\pi}{4}, 0) \cdot CZ_{i,j} \\ &\cdot R_i(\tfrac{\pi}{4}, \tfrac{-\pi}{4}, 0) \cdot R_j(\tfrac{\pi}{4} + \tfrac{\theta}{2}, \tfrac{-\pi}{4}, \tfrac{\pi}{2}) \cdot CZ_{i,j} \\ &\cdot R_i(\tfrac{\pi}{4}, \tfrac{\pi}{4}, 0) \cdot R_j(\tfrac{\pi}{4} + \tfrac{\theta}{2}, \tfrac{-\pi}{4}, \tfrac{\pi}{2}) \cdot CZ_{i,j} \\ &\cdot R_i(\tfrac{\pi}{4}, \tfrac{-\pi}{4}, 0). \end{aligned} \tag{10}$$

Alternatively, if one defines a $\sqrt{\text{SWAP}}$ gate to be

$$\sqrt{\text{SWAP}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{e^{-i\frac{\pi}{4}}}{\sqrt{2}} & \frac{e^{i\frac{\pi}{4}}}{\sqrt{2}} & 0 \\ 0 & \frac{e^{i\frac{\pi}{4}}}{\sqrt{2}} & \frac{e^{-i\frac{\pi}{4}}}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{11}$$

we can decompose a $GS(\theta)$ gate into two $\sqrt{\text{SWAP}}$ gates and single-qubit Z rotations

$$GS_{i,j}(\theta) = \sqrt{\text{SWAP}}_{i,j} e^{i\frac{\theta}{2}Z_i} e^{-i\frac{\theta}{2}Z_j} \sqrt{\text{SWAP}}_{i,j}. \tag{12}$$

This is similar to the decomposition of a Givens rotation gate into two $\sqrt{I\text{SWAP}}$ and z-rotation gates [4]: the $\sqrt{\text{SWAP}}$ and $\sqrt{I\text{SWAP}}$ gates differ by a $e^{i\frac{\pi}{4}ZZ}$ rotation, which doubles to yield the ZZ term which separates the Givens and GS rotation (up to a redefinition of angle and single-qubit Z rotations).

It is further possible to decompose a $GS(\theta)$ gate into three $\sqrt{I\text{SWAP}}$ gates and arbitrary single-qubit rotations, though the calculation is more involved. We start from a decomposition of a SWAP ($= GS(0)$) gate into three $\sqrt{I\text{SWAP}}$ gates

$$\begin{aligned}\text{SWAP} &= \left(e^{-i\frac{\pi}{4}(YI+IY)}\sqrt{\text{iSWAP}}e^{i\frac{\pi}{4}(YI+IY)}\right)\sqrt{\text{iSWAP}}\left(e^{i\frac{\pi}{4}(XI+IX)}\sqrt{\text{iSWAP}}e^{-i\frac{\pi}{4}(XI+IX)}\right) &(13)\\ &= \Big(\exp\left[i\tfrac{\pi}{8}(YY+ZZ)\right]\Big)\exp\left[i\tfrac{\pi}{8}(XX+YY)\right]\Big(\exp\left[i\tfrac{\pi}{8}(XX+ZZ)\right]\Big). &(14)\end{aligned}$$

If we perform a basis rotation on the bracketed terms by $\exp(i\alpha Z)$ on the appropriate qubit, we can generate a gate of the form

$$G(\alpha) = \exp\left[i\tfrac{\pi}{8}(\cos(\alpha)YY + \sin(\alpha)XY)\right]\exp\left[i\tfrac{\pi}{8}(XX+YY+2ZZ)\right]\exp\left[i\tfrac{\pi}{8}(\cos(\alpha)XX - \sin(\alpha)XY\right]. \quad (15)$$

This function can be expanded to give

$$G(\alpha) = g_1(\alpha)II + g_2(\alpha)(XX+YY) + g_3(\alpha)(IZ-ZI) + g_4(\alpha)ZZ, \quad (16)$$

with $g_i(\alpha)$ the following complex functions

$$\begin{aligned}g_1(\alpha) &= \frac{1}{\sqrt{2}}\Big\{\cos^2(\tfrac{\pi}{8})(\cos^2(\tfrac{\pi}{8}) + i\sin^2(\tfrac{\pi}{8})) + i\cos(\alpha)\cos(\tfrac{\pi}{8})\sin(\tfrac{\pi}{8})\frac{(1+i)}{\sqrt{2}}\\ &\quad + \cos^2(\alpha)\sin^2(\tfrac{\pi}{8})(i\cos^2(\tfrac{\pi}{8}) + \sin^2(\tfrac{\pi}{8})) + \sin^2(\alpha)\sin^2(\tfrac{\pi}{8})(\cos^2(\tfrac{\pi}{8}) + i\sin^2(\tfrac{\pi}{8}))\Big\} &(17)\\ g_2(\alpha) &= \frac{1+i}{4\sqrt{2}}(1+\cos(\alpha)) &(18)\\ g_3(\alpha) &= \frac{1}{\sqrt{2}}\Big\{\sin(\alpha)\sin(\tfrac{\pi}{8})\cos(\tfrac{\pi}{8})\frac{(1+i)}{\sqrt{2}} + i\sin(\alpha)\cos(\alpha)\sin^2(\tfrac{\pi}{8})(\cos^2(\tfrac{\pi}{8}) + i\sin^2(\tfrac{\pi}{8}))\\ &\quad -i\cos(\alpha)\sin(\alpha)\sin^2(\tfrac{\pi}{8})(i\cos^2(\tfrac{\pi}{8}) + \sin^2(\tfrac{\pi}{8}))\Big\} &(19)\\ g_4(\alpha) &= \frac{1}{\sqrt{2}}\Big\{\cos^2(\tfrac{\pi}{8})(i\cos^2(\tfrac{\pi}{8}) + \sin^2(\tfrac{\pi}{8})) - i\cos(\alpha)\cos(\tfrac{\pi}{8})\sin(\tfrac{\pi}{8})\frac{(1+i)}{\sqrt{2}}\\ &\quad + \cos^2(\alpha)\sin^2(\tfrac{\pi}{8})(\cos^2(\tfrac{\pi}{8}) + i\sin^2(\tfrac{\pi}{8})) + \sin^2(\alpha)\sin^2(\tfrac{\pi}{8})(i\cos^2(\tfrac{\pi}{8}) + \sin^2(\tfrac{\pi}{8}))\Big\}. &(20)\end{aligned}$$

This is of the correct form for our desired gate up to single-qubit Z rotations as long as the phase on $\langle 00|G(\alpha)|00\rangle$, $\langle 11|G(\alpha)|11\rangle$, $\langle 01|G(\alpha)|10\rangle$ and $\langle 10|G(\alpha)|01\rangle$ are equal. One can confirm that all have a phase of $e^{i\frac{\pi}{4}}$ for any angle of α. There remains a residual phase on the two on-diagonal elements of $G(\alpha)$,

$$\begin{aligned}\langle 01|G(\alpha)|01\rangle &= g_1(\alpha) - g_4(\alpha) + 2g_3(\alpha)\\ &= A(\alpha)e^{i\phi(\alpha)}e^{i\pi/4} &(21)\\ \langle 10|G(\alpha)|10\rangle &= g_1(\alpha) - g_4(\alpha) - 2g_3(\alpha)\\ &= -A(\alpha)e^{-i\phi(\alpha)}e^{i\pi/4}, &(22)\end{aligned}$$

which can be removed by shifting

$$G(\alpha) \to e^{i\phi/4(IZ-ZI)}G(\alpha)e^{i\phi/4(IZ-ZI)}. \quad (23)$$

The precise value of ϕ here is

$$\phi = -\arctan\left\{\frac{\sqrt{2}\left(\cos^2(\tfrac{\pi}{8}) - \sin^2(\tfrac{\pi}{8})\cos(2\alpha) - \frac{1}{\sqrt{2}}\cos(\alpha)\right)}{\sin(\alpha) + \sin(2\alpha)\sin^2(\tfrac{\pi}{8})}\right\}. \quad (24)$$

We notice that our formula for $g_2(\alpha)$ only takes positive values. To get the full range of $GS(\theta)$, one can finally send $G(\alpha) \to e^{i\pi/2ZI}e^{i\pi/2IZ}G(\alpha)e^{i\pi/2ZI}e^{i\pi/2IZ}$ (again at no extra cost), and solve for $\sin(\theta) = 2g_2(\alpha)$.

C. Scheduling details

A key advance in this work was the development of a mapping of our UpCCD ansatz to a 2D grid with local connectivity, such that a) the entire ansatz could be implemented in depth $N/2$ GS gates, and b) all $X_iX_j+Y_iY_j$ terms could be estimated using only N unique mappings. In this section, we explain this mapping in more detail and prove both a) and b) true.

Let us first expand the discussion of the implementation of the UpCCD ansatz. The standard UpCCD ansatz takes the form

$$U(\theta) = \exp\left\{ \sum_{p\in\text{unocc},q\in\text{occ}} \theta_{pq} a^\dagger_{p\alpha} a^\dagger_{p\beta} a_{q\alpha} a_{q\beta} - \text{h.c.} \right\}, \tag{25}$$

which when mapped to qubits in the S_0 approximation, becomes

$$U(\theta) = \exp\left\{ \sum_{p\in\text{unocc},q\in\text{occ}} \theta_{pq} X_p Y_q - \text{h.c.} \right\}. \tag{26}$$

This in turn can be Trotterized to a product of coherent pair excitations $\exp(\theta_{pq}X_pY_q - \text{h.c.})$. As mentioned in the main text, the effect of a single GS gate in the fermionic picture is to implement a single coherent pair excitation between the spatial orbitals assigned to qubits i and j, and then to SWAP the orbitals. This means a given orbital is not assigned to a fixed qubit throughout the experiment. For our implementation of the UpCCD ansatz (see e.g. Fig. 1 of the methods and materials, bottom) GS gates are applied in layers: between qubits $i,j = 2n, 2n+1$ during an even-numbered layer, and between qubits $i,j = 2n+1, (2n+2)\%N$ during an odd-numbered layer (for $n = 0,\ldots,N/2-1$). We claim that, at half-filling, any initial assignment of occupied orbitals to odd-numbered qubits and unoccupied orbitals to even-numbered qubits will cause $N/2$ such layers to implement a Trotterized form of Eq. (26). To see this, note that during an even-numbered layer, orbitals sitting on even-numbered (odd-numbered) qubits shift to the right (left) around the ring of qubits, and vice-versa during an odd-numbered layer. This in turn implies that the empty orbitals, that are initially assigned to an even-numbered qubit, will only ever move to the right around this ring as the ansatz proceeds, as they will be assigned to an odd-numbered qubit on odd-numbered rounds. Likewise, the filled orbitals will only ever move to the left around this ring. As an occupied orbital must cross (and thus interact with) every unoccupied orbital before it encounters the same one twice, this implies that the first $N/2$ layers of our UpCCD ansatz will yield precisely the $(N/2)^2$ coherent pair excitations between each unoccupied and each occupied orbital, as required. (This is demonstrated for 10 qubits in Fig. 1[top].)

Let us now consider the measurement of non-local $X_iX_j + Y_iY_j$ terms as performed in this work. As mentioned in the main text, these are diagonalized on a pair of orbitals i,j via a $GS(\pi/4)$ rotation, which necessitates the orbitals be on neighbouring qubits. The $GS(\pi/4)$ rotation maps the operators $Z_i, Z_j \to D^+_{ij}, D^-_{ij}$, where we define

$$D^\pm_{ij} = \frac{1}{2}\Big[Z_i + Z_j \pm \big(X_iX_j + Y_iY_j\big)\Big] \tag{27}$$

As we implement our ansatz on a $2\times N/2$ grid, qubit i is not only connected to qubit $(i\pm1)\%N$; the cross-links in the grid connect qubit i and qubit $N-i-1$. Moreover, note that our ansatz remains unchanged if we perform the cyclic permutation $i \to (i+k)\%N$; that is, we shift our orbital assignments, and all ansatz gates and parameters, around the ring of qubits. (Note that this technically SWAPs the definition of even and odd layers when k is odd: after this permutation the first layer of GS gates will be between qubits $i,j = 2n+1, (2n+2)\%N$.) Following this permutation, cross-links will connect the orbital that was on qubit i to that which was on qubit $[N-(i+2k)-1]\%N$. One can confirm that by running over $k = 0,\ldots,N/2-1$ we find $N/2$ cyclic permutations, such that each occupied orbital is coupled to each unoccupied orbital by a cross-link for exactly one permutation. With just the cyclic permutation operation and no additional SWAP gates, this gives $N/2$ unique circuits that allow for measurement of all $D^\pm_{ij}$ where i is occupied and j unoccupied. To obtain the circuits that couple occupied orbitals to occupied orbitals (and unoccupied to unoccupied), we require an additional layer of SWAP gates. (This is unavoidable given our initial ansatz ordering: all occupied qubits are connected by an even number of couplings, so a further direct coupling would yield an odd-order cycle, which does not exist on a square lattice.) After each permutation k above, we perform SWAPs between qubits $l+(k\%2), l+(k\%2)+1$ for $0 \le l < N/4$. One can confirm that this yields $N/2$ circuits such that a coupling between any pair of occupied orbitals can be achieved in one such circuit. (In this second set of circuits, some qubits are not coupled, and were not used to estimate expectation values in this experiment.) This can be confirmed by a visual inspection of Fig. 1[bottom]. Code that implements this scheduling has also been uploaded to ReCirq [5].

D. Scheduling of EV circuits

In this section we outline the additional experimental details required to implement EV in this work. We implemented control-free EV for $O = Z_i$, or Z_iZ_j, or D^+_{ij}(Eq. (27)), using a vacuum reference state $|\phi\rangle = |00\ldots\rangle$. We prepare the superposition $\frac{1}{\sqrt{2}}(|\psi\rangle + |\phi\rangle)$ by acting the UpCCD ansatz circuit on the cat state $\frac{1}{\sqrt{2}}(|0000\ldots\rangle + |0101\ldots\rangle)$. Then, all operators O were implemented by compiling them to a virtual Z rotation on a single qubit. Finally, to measure $\langle\Psi|\phi\rangle\langle\psi|\Psi\rangle$, we inverted the UpCCD circuit and cat state preparation. This maps the desired matrix element $|\phi\rangle\langle\psi| \to |0\rangle\langle 1| \otimes$

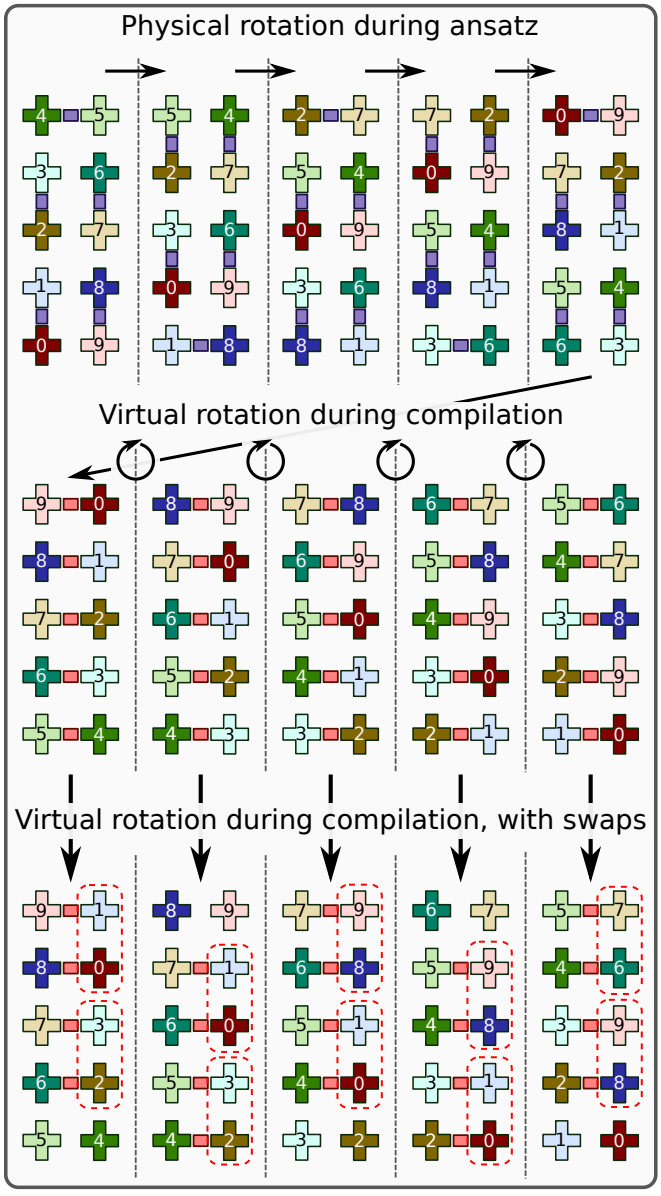


FIG. 1. Detail of the scheduling of pair excitation interactions during the UpCCD ansatz, and measurement scheduling. (top) The physical rotations during the execution of SWAP layers of the UpCCD ansatz, showing that all occupied (even index) and all unoccupied (odd index) orbitals are coupled (purple squares) at some point during the ansatz. (middle) After executing the UpCCD ansatz we have not used the cross-couplers (red squares), allowing us to virtually rotate the entire grid during compilation for the purposes of measurement. This virtual rotation allows all occupied and unoccupied orbitals to be coupled at some step, without any increase in circuit depth. (bottom) The virtual rotation cannot however, bring two occupied or two unoccupied orbitals to nearest-neighbour qubits so that they may be coupled. To achieve this, we require an extra round of SWAPs (red dashed boxes). One can confirm that across the middle and bottom layers, all pairs of qubits are coupled in at least one configuration.

$|00\ldots\rangle\langle 00\ldots|$, allowing its measurement via single-qubit rotation on a single 'measurement' qubit, and readout of all qubits in the computational basis. (Reading out all qubits is essential, as we record an estimate of 0 for the measurement of $\langle\Psi|\phi\rangle\langle\psi|\Psi\rangle$ unless all qubits other than the measurement qubit read out 0.)

Our implementation of O as a virtual Z rotation allows us to replace $O \to O^\alpha = \cos(\alpha) + i\sin(\alpha)O$ to remove our susceptibility to a uniform background magnetic field $e^{ih\sum_j Z_j}$. Such a field transforms Eq. 9 of the methods and materials to $\frac{1}{2}\langle\psi|O^\alpha|\psi\rangle e^{i(\phi+hN/2)}$; fitting this to three points of α allows us to simultaneously estimate h, the fidelity F, and $\langle O\rangle$. (This is preferable to independent calibration as h fluctuates with a $1/f$ spectrum.)

Some further experimental optimization was made for the EV circuit that was not available for the VQE or VD circuits. Many of the gates in the final EV circuit [Fig. 1 of the methods and materials, bottom] cancel to the identity, as the second half of the circuit is the inversion of the first half. We identify and prune these to increase the overall circuit fidelity, and insert echo pulses into the resulting empty space. We further compile an echo pulse for the entire second half of the EV circuit [this can be done as $XX \cdot GS(\theta) \cdot XX = GS(\theta)$]. To unbias readout, we measure the single measurement qubit in the $\pm X$ and $\pm Y$ bases. This, combined with the additional circuits to remove a background magnetic field, raises the total number of circuits to $12N^2$ $(6N^2 + 6N)$ to estimate the expectation value of our chemistry (RG) Hamiltonian. Shots were distributed across these circuits following the term weight in the Hamiltonian with some additional restrictions imposed by classical readout hardware (see App. IV C).

III. DATA PROCESSING

A. Optimal linear combinations of non-independent expectation values

Once we have used our variational ansatz to prepare an approximation $\rho(\boldsymbol{\theta}) \sim |\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|$ to the ground state of our target problem, it remains to measure the quantum device to estimate the energy (or other properties of the state). As our devices are heavily coherence limited, rather than attempting to perform this estimation in a single shot, we write our Hamitonian as a sum of simpler terms

$$H = \sum_i c_i Q_i, \tag{28}$$

and estimate the expectation value of each such term

$$E = \mathrm{Trace}[H\rho] = \sum_i c_i \mathrm{Trace}[Q_i\rho]. \tag{29}$$

(Note that we are not placing any restrictions on Q_i at this point.) The method in which we estimate $\mathrm{Trace}[Q_i\rho]$

will depend on which error mitigation methods are being implemented. However, all schemes will return a set of estimates of $\mathrm{Trace}[Q_i\rho]$ with a covariance matrix

$$\Sigma_{i,j} = \mathrm{Covar}\Big[\mathrm{Trace}[Q_i\rho], \mathrm{Trace}[Q_j\rho]\Big]. \tag{30}$$

In this experiment, it turns out that our choice of $\{Q_i\}$ will not be linearly independent. The reason for this is post-selection: we desire our choice of $\{Q_i\}$ to allow us to measure $S_z = \sum_i Z_i$ for each experiment. To achieve this, we measure the operators Z_j, Z_jZ_k, and $D^{\pm}_{jk}$ (Eq. (27)), but $D^{+}_{jk} + D^{-}_{jk} = Z_j + Z_k$. This leaves us with a degree of freedom in our choice of c_i that we may optimize upon, once a dataset is taken.

In order to choose c_i, we perform a constrained Lagrangian minimization. Our target cost function is the variance on the estimate in Eq. (29)

$$\mathrm{Var}(E) = \sum_{i,j} c_i\Sigma_{i,j}c_j, \tag{31}$$

subject to Eq. (28) as a constraint. Let us fix some linearly independent basis of operators $\{P_j\}$ (e.g. Pauli operators), and we can write $H = \sum_j h_jP_j$, and $Q_i = \sum_j q_{i,j}P_j$. (As P_j is a basis, we have no freedom in our choice of h_j or $q_{i,j}$.) Our constraints then take the form

$$\sum_i c_iq_{i,j} = h_j. \tag{32}$$

This can be written as a Lagrange multiplier, yielding a Lagrangian

$$\mathcal{L} = \sum_{i,j} c_i\Sigma_{i,j}c_j - \sum_j \Big(\sum_i c_iq_{i,j} - h_j\Big)\lambda_j. \tag{33}$$

Differentiating with respect to c_i and setting equal to 0 yields (using the fact that $\Sigma_{i,j}$ is a positive matrix)

$$2\sum_j \Sigma_{i,j}c_j - \sum_j q_{i,j}\lambda_j = 0 \to c_j. \tag{34}$$

Recognising this as a vector equation, as the matrix Σ is invertible we have

$$\mathbf{c} = \frac{1}{2}\Sigma^{-1}q\boldsymbol{\lambda}. \tag{35}$$

Here, $\boldsymbol{\lambda}, \mathbf{c}$ are vectors containing the λ_j and c_j components, and q and Σ are matrices containing the $q_{i,j}$ and $\Sigma_{i,j}$ components respectively. Substituting this into our Lagrangian yields

$$\mathcal{L} = -\frac{1}{4}\boldsymbol{\lambda}^Tq^T\Sigma^{-1}q\boldsymbol{\lambda} + \boldsymbol{\lambda}^T\mathbf{h}. \tag{36}$$

Then, differentiating with respect to λ (and using the fact that $q^T\Sigma^{-1}q$ is Hermitian), we have

$$q^T\Sigma^{-1}q\boldsymbol{\lambda} = 2\mathbf{h} \to \boldsymbol{\lambda} = 2(q^T\Sigma^{-1}q)^*\mathbf{h} \tag{37}$$

$$\to \mathbf{c} = \Sigma^{-1}q(q^T\Sigma^{-1}q)^*\mathbf{h}, \tag{38}$$

where here, $*$ denotes the Moore-Penrose pseudoinverse of a matrix.

In practice, though we find that this produces low-variance estimates, it is unstable to uncertainty in our estimate of Σ. As such, we set $\Sigma = I$ for the purposes of determining $\mathbf{c}$ (which corresponds to assuming that all uncertainties are equal and all covariances are 0). This yields

$$\mathbf{c} = q(q^Tq)^*\mathbf{h}. \tag{39}$$

B. Error propagation and bootstrapping

The exact form of the error for raw and post-selected VQE is well-known. The covariance between estimates of the expectation value of two reflection operators P_i and P_j, estimated simultaneously from M repeated preparations and measurements on a target state, is given by

$$\mathrm{Covar}[\langle P_i\rangle\langle P_j\rangle] = \frac{\langle P_iP_j\rangle - \langle P_i\rangle\langle P_j\rangle}{M}. \tag{40}$$

The resulting number can be substituted into the propagation of variance formula described in App. III A above to get an estimate of the variance in energy, whilst errors in order parameters can be obtained by propagation of variance through ($\Delta = \frac{1}{N}\sum_{j,\sigma}\sqrt{\langle n_{j\sigma}^2\rangle - \langle n_{j\sigma}\rangle^2}$)

$$\frac{\partial\Delta}{\partial\langle n_{j\sigma}\rangle} = \frac{1}{2N}\frac{1-2\langle n_{j\sigma}\rangle}{\sqrt{\langle n_{j\sigma}\rangle - \langle n_{j\sigma}\rangle^2}} \tag{41}$$

$$\mathrm{Var}[\Delta] = \sum_{j\sigma}\frac{1}{2N}\frac{(1-2\langle n_{j\sigma}\rangle)^2}{\langle n_{j\sigma}\rangle - \langle n_{j\sigma}\rangle^2}\mathrm{Var}[\langle n_{j\sigma}\rangle]. \tag{42}$$

We note that this diverges when $\langle n_{j,\sigma}\rangle \to 0, 1$, which is what happens at $g = 0$ when $\Delta \to 0$. We suggest that this explains the peak in the observed experimental error in the order parameter around $g = 0$ somewhat.

The experimental covariances (Eq. (40) for raw and postselected VQE when $i \neq j$ were corrupted during data taking; these terms were set to 0 when generating error bars for Fig. 1 and Fig. 2 of the main text. As said error bars are negligible due to the large number of samples used in this experiment, more complex recovery procedures will not noticeably change the figures.

For echo verification and virtual distillation, the formulas for variance are more complicated (the form derived in Ref. [6] is not appropriate here due to our fitting to remove a background magnetic field). Instead, error bars were determined by bootstrapping the raw data (resampling with replacement and taking the sample standard deviation) over 100 and 25 samples respectively. This was made complicated due to data loss during the the EV experiment. Arrays were stored of the verified expectation value of X and Y on the measurement qubit; this is insufficient to recover the shot distribution, as the EV measurement can return three values; $+1$, -1, and 0 [7]. Counts M_+, M_-, and M_0 of these values were

approximated from these expectation valus using the estimated EV fidelity F_{EV} and the fact that in the absence of error

$$\frac{M_0}{M} = 1 - |\langle U\rangle|^2, \tag{43}$$

where U is the operator to be estimated via EV, and $M = M_+ + M + - + M_0$. Assuming the fraction $1 - F_{\mathrm{EV}}$ fails verification, we have $M_0 = M(1 - F_{\mathrm{EV}}|\langle U\rangle|^2)$, and M_+ and M_- may be then distributed so that $(M_+ - M_-)/M$ is the observed expectation value on the measurement qubit. (This can result in $M_+ < 0$ or $M_- < 0$, in which case we reduced $M_0 \to M_0 - 2\min(M_+, M_-)$, $M_+, M_- \to M_+ + \min(M_+, M_-), M_- + \min(M_+, M_-)$.)

IV. QUANTUM RUN TIME ESTIMATION

In this section, we estimate the cost of running our quantum experiments in terms of number of experiments (shots) and the wall-clock time. This is done in terms of established theoretical cost estimates [8, 9], which allows us to compare between this and what was implemented in the actual experiment in Fig. 3 of the main text (top). This is further necessary for the tuning of hyperparameters of the variational optimizers we will present in App. V.

A. Wall-clock model

The cost of running a set of experiments on real-world hardware in terms of the actual time spent or *wall-clock time* is not a linear function of the number of samples used. We can estimate the wall-clock time as a function of three parameters. First, the number a of calls to the device from the computed executing the Cirq code. Second, the number b of distinct circuits the device needs to execute. Third, the total number of shots c spend over all calls and distinct circuits. We found empirically that for the device used the time for a call to the device from Cirq is about $1s$. In each call, a batch of distinct circuits can be sent to the device for execution but re-programming the device to execute a different circuit takes $0.042s$. When estimating energy expectation values or order parameters, all circuits that need to be executed are known in advance and thus can be sent in a single batch. However, during variational optimization, depending on the optimizer used, at least one call to the device per epoch is needed, because the quantum circuits to be executed next depend on the step taken by the optimizer, which in turn depends on the measurement results of the first batch. This needs to be taken into account when comparing the performance of different optimizers. The time per shot was found to be $1 * 10^{-5}$, essentially independently of the circuit depth (probably because it is dominated by readout, reset, and time needed by the control electronics). The number of shots c of a circuit can currently only be set for a batch as a whole, which in practice limits our ability to distribute shots in a way that would minimize the variance of the resulting estimator. The total wall-clock time this takes the form

$$t_{\mathrm{wall}} = a * 1s + b * 0.042s + c * 5 * 10^{-5}s \tag{44}$$

B. Hamiltonian decomposition schemes

In this section we define the different options chosen to decompose a Hamiltonian into terms Q_i for measurement. We are free to group the measurement of all Q_i terms together, as long as they commute and an appropriate diagonalization circuit is found. For our numerics, we study the following measurement strategies:

- *Termwise* — we take individual Pauli operators $Q_i \in \{Z_j, X_iX_j, Y_iY_j, Z_iZ_j\}$, and measure each Q_i using an independent measurement circuit.
- $XX + YY$ — we first measure all $Q_i \in \{Z_j, Z_iZ_j\}$ in a single-shot measurement. Next, we measure $Q_i \in \{X_jX_k + Y_jY_k\}$, grouping disjoint pairs j, k together into N sets following the scheduling outlined in Sec. II C. This mostly matches the scheme used for the estimation of expectation values Raw VQE, PS VQE and PS VD in Fig. 1 of the main text, and the scheme used for the PS VQE estimates in Fig. 2 of the main text.
- $XX + YY + IZ + ZI$ — here, we draw $Q_i \in \{Z_i, Z_iZ_j, D^+_{ij}\}$ ($D^{\pm}_{ij}$ is defined in Eq. (27)), and measure each term separately. This matches the measurement scheme used for EV. As the operators chosen are Hermitian and unitary, the EV variance is equivalent to the standard tomography variance [6]. By choosing only D^+_{ij} and not adding D^-_{ij} to our set of measurements, the set $\{Q_i\}$ becomes linearly independent, and so the relative coefficients c_i (Eq.(29)) are fixed.

Some notable differences occur between the numerical estimates made here and those implemented on hardware. (Ultimately the number of shots in the experiment was chosen to be low enough that the experimental error mostly dominated, rather than being optimized based on preliminary calculations.) In the experiment, additional estimates of $Z_j + Z_k$ were extracted alongside each measurement of $X_jX_k + Y_jY_k$ and combined using the techniques outlined in App. III A. Each circuit was repeated with and without an additional π pulse on all qubits to unbias readout noise. In the experiment, the shot distribution was chosen to be uniform for VQE (40,000 per circuit) and VD (100,000 per circuit). (One can observe in Fig. 3 of the main text that an excess of shots was taken in both cases.) For EV, shots were distributed according to the relative weight of Q_i in the Hamiltonian. However, this was made more complicated by a technical requirement that all shots executed in a single batch must

have the same number of repetitions. To accommodate this, the number of shots used to estimate a single Q_i was rounded up or down to be an integer multiple of a fixed base (40000). Furthermore, as mentioned in App. II D, when performing EV to estimate each Q_i, 12 unique circuits were run to cancel out a background magnetic field and depolarize readout noise.

C. Shot distribution

Once the decomposition of the Hamiltonian is decided, the variance of the resulting energy estimator further depends on how the available shot budget is distributed over the Q_i (or the jointly measurable groups of Q_i). In principle, estimates of the (co-)variances from a small number of shots, from previous measurements at close by VQE parameter values according to (40), or in adaptive schemes can be used to distribute shots in an asymptotically optimal way to reduce the variance. In practice, one is limited by the overhead of calling the device and limitations in setting the shots for individual circuits inside a batch (see Section IV A).

In our estimates we assume the ability to allocate shots per distinct circuit (and not only on a per-batch basis) but only consider the two non-adaptive shot distribution schemes that need no input from the quantum computer:

- *Uniform Distribution* — distribute the total shot budget per expectation value uniformly over the term groups, i.e., spend the same number of shots on each group of Q_i that are measured jointly according to the decomposition scheme.
- *According to term group weights* — the total shot budget is distributed proportional to the coefficients $|c_i|$ in Eq. 29. In the case that multiple Q_i are measured, shots are distributed according to $[\sum_i |c_i|^2]^{1/2}$. This is optimal in case the variances of all Q_i are equal and covariances vanish [10]. (Note that as the three measurement strategies measure only linearly-independent operators, the $|c_i|$ can be fixed prior to measurement.)

D. Wall-clock time extrapolation to large system sizes

Using the protocol described in the previous sections, we are able to estimate the cost of measuring the RG model at $g = -0.9$ to within 0.1 a.u. as we enlarge the system size, while optimizing our shot distribution (App. IV C) for the wall-clock time (App. IV A). In Fig. 2, we plot the cost in terms of the number of shots (top) and in wall-clock time for superconducting hardware (bottom). Estimates are performed for $N = 6, 8, 10, 12$; given good fit to a line on a log-log plot and that a powerlaw scaling is expected, we are able to extrapolate this to larger system sizes. This data was combined with fidelity estimates for a 10-qubit system (Fig. 3[bottom-right] of the main text) to yield the curves in Fig. 3[top] of the main text.

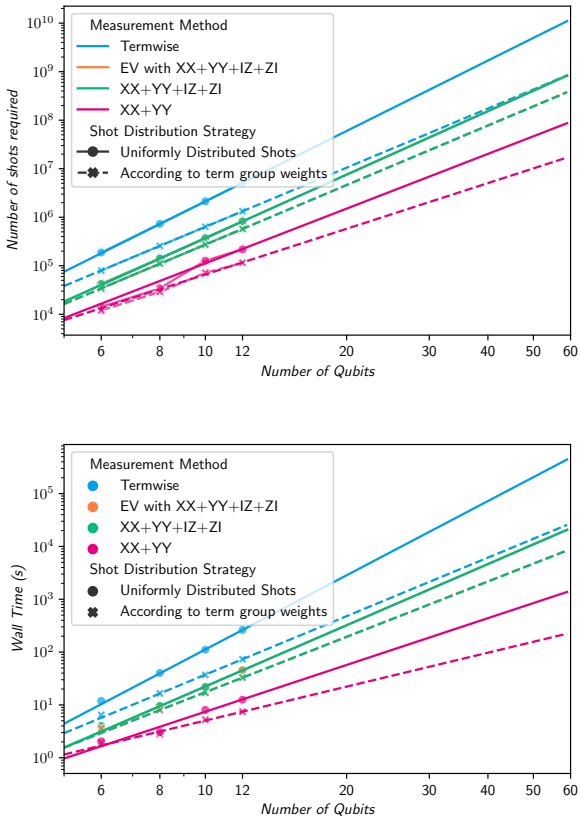


FIG. 2. Number of shots (top) and wall-clock time (bottom) required to estimate the ground-state energy of the RG Hamiltonian at $g = -0.9$ to within 0.1a.u.. Estimates were obtained by standard Lagrangian optimization [9] using term distributions defined in App. IV C and measurement strategies defined in App. IV B. Wall-clock time model is described in App. IV A. The orange and green data points for the $XX + YY + IZ + ZI$ scheme with and without EV coincide.

V. VARIATIONAL OPTIMIZATION OF A 6-QUBIT EXPERIMENT

A. The Conjugate Model Gradient Descent optimizer

The large number of evaluations needed for ansatz parameter optimization on quantum hardware is a major impediment towards keeping the cost of variational algorithms (in terms of wall-clock time) manageable. To mitigate the overhead incurred by sending jobs to and receiving results from a device (see App. IV A), it is beneficial if the optimizer can request a batch of cost-function

evaluations at once before making a step. In [11], surrogate model based optimizers were found to have good performance under this constraint. Here, a quadratic model function was fitted to present and past expectation value estimates in the vicinity of the current ansatz parameter vector. Then, after making a step, a batch of circuits corresponding to points in the vicinity of the new parameter vector were evaluated and the stepping procedure repeated. In this appendix we develop a natural extension of this procedure, by combining it with the conjugate descent algorithm.

The Conjugate Model Gradient Descent optimizer is a surrogate model-based optimization algorithm, with the additional improvement that the gradient which is extrapolated from fitting a quadratic model to the cost function, is used in the framework of conjugate gradient descent to make a step in the parameter space. The Conjugate Gradient Descent method was developed by Hestenes and Stiefel in [12]. For the special case of quadratic cost functions over an n dimensional space, conjugate gradient methods can be proven to converge in n iterations [13]. In practice the conjugate gradient method is found to work well for cost functions that are locally approximately quadratic. This is an assumption one anyway needs to make when using quadratic model based optimizers and thus makes it natural to combine conjugate gradient descent and quadratic surrogate model based optimizers with quadratic model functions. In conjugate gradient descent the steps are taken in the direction of the so-called conjugate gradient $s_n = g_n + \beta_n s_{n-1}$, where g_n is the gradient (in our case the one of the quadratic model fitted to the samples around the current position) and β_n is a scaling factor. The formula for the n-th iteration scaling factor is given in [14] as: $\beta_n^{FR} = \frac{g_n^T g_n}{g_{n-1}^T g_{n-1}}$. With $s_0 = g_0$ fixed for the first step of the algorithm.

In Alg. 1, we present our Conjugate Model Gradient Descent algorithm. We assume in this algorithm access to a (noisy) oracle to the target function f to be optimized. In practice, this is given by a call to a quantum device with a target error, that must be made small enough to enable convergence.

Algorithm 1 Conjugate Model Gradient Descent

Input: Initial point x_0, learning rate γ, sample radius δ, maximum iterations n, number of evaluations per iteration k, rate decay exponent α, stability constant A, sample radius decay exponent ξ, tolerance ε, oracle for function f.

Initialize lists L, L'
Initialize a list G
Initialize a list H
Let $x \leftarrow x_0$
for m in 0...n **do**
  Let $\delta' \leftarrow \delta/(m+1)^\xi$
  Sample k points uniformly at random from the δ'-neighborhood of x to generate a set S
  for each x' in $S \cup \{x\}$ **do**
    Add $(x', f(x'))$ to L
  end for
  Clear list L'
  for each tuple (x', y') in L **do**
    if $|x' - x| < \delta$ **then**
      Add (x', y') in L'
    end if
  end for
  Fit $f(x) = x^T A x + bx + c \sim y$ to the points (x, y) in L' using least squares linear regression.
  Let g_m be the gradient of f at x (i.e. $g_m = b$).
  if $|g_m| < \varepsilon$ **then**
    return x
  end if
  if $m = 0$ **then**
    Let $h_0 \leftarrow g_0$
  else
    $\beta_m \leftarrow g_m^T g_m / g_{m-1}^T g_{m-1}$
    $h_m \leftarrow g_m + \beta_m h_{m-1}$
  end if
  $\gamma' = \gamma/(m + 1 + A)^\alpha$
  Add g_m to the list G
  Add cg_m to the list H
  Let $x \leftarrow x - \gamma' \cdot h_m$
  Let $m \leftarrow m + 1$
end for
return x

B. Hyperparameter tuning

For all experiments shown in the main text, parameters were obtained by a noiseless simulation using L-BFGS-B, starting from $\boldsymbol{\theta} = 0$. We find that in the absence of experimental noise, gradient-based optimizers such as L-BFGS-B converge well to an optimal solution. However, this is not the case in the presence of experimental or sampling noise, which can make many estimators unstable. This has lead previous efforts towards stochastic-based [15] or model-based [4] optimizers, including the Model Gradient Descent optimizer that we have based our conjugated Model Gradient Descent algorithm upon in the previous section.

The hyperparameters used in the Conjugate Model Gradient Descent algorithm were either chosen through an intuitive approach due to their physical meaning, such as the sample radius due to the parameters being per-

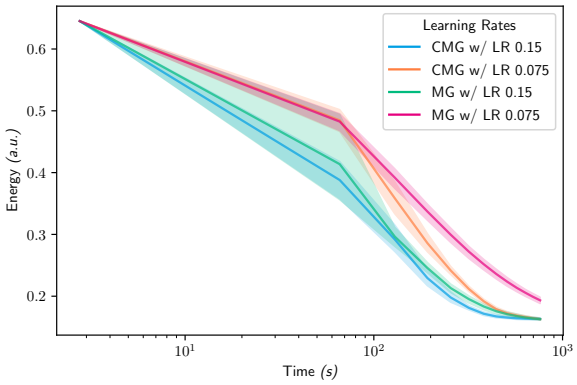


FIG. 3. In this figure, the distance between the energy that was calculated in the noiseless regime by the L-BFGS-B optimizer is plotted over the wall-time required to reach such resolution. The system displayed is a RG Hamiltonian for 6 qubits, with a coupling constant of $g = -0.9$, consisting of 10 runs and the lines being the average of those runs for the Conjugate Model Gradient Descent (CMG) and the Model Gradient Descent (MG) found in [11].

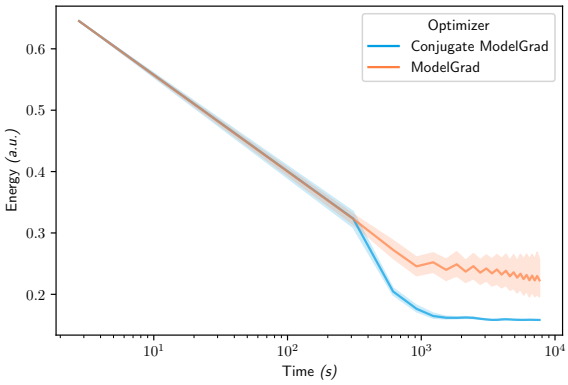


FIG. 4. Comparison of the two variants of the Model Gradient Descent optimizer, where in this case the number of samples used to construct the quadratic model one uses is higher, namely $k = 4(N+1)(N+2)$, allowing for the performance difference of the two optimizers to become more pronounced, with Conjugate Model Gradient Descent managing to consistently converge at better parameters, while Model Gradient Descent gets stuck in a local minimum, averaged over 10 runs for each optimizer, for 25 maximum iterations.

turbed by a random variable from [-0.25,0.25] and the maximum number of iterations to stay within reasonable wall-time, or via grid search for the rest of the hyperparameters, ensuring they are good enough to work for the variety of cases examined while avoiding overfitting. Explicitly, the hyperparameters used are included in Table I:

Hyperparameter	Values Used
Sample radius δ	1.0
Learning rate γ	0.15
Stability constant A	0
Sample number k	$0.409(N+1)(N+2)$
Sample radius decay exponent ξ	0
Rate decay exponent α	0.2
Maximum evaluations n	12

TABLE I. Hyperparameters for the Conjugate Model Gradient Descent optimizer during the experiment, where N is the number of parameters in the optimization. These were also used in the later comparisons with Model Gradient Descent, to gauge performance in equal footing.

The hyperparameters we chose work equally well for Model Gradient Descent, as well our Conjugate variant. We found that the conjugate method can sometimes speed up convergence or help prevent getting stuck in local minima. We illustrate this with two examples: In Fig 3 how Conjugate Model Gradient Descent has the ability to speed up in case the initial learning rate was chosen to be too small. In Fig. 4 we show an example of hyper parameters for which Conjugate Model Gradient Descent is able to converge to a lower minimum than plain Model Gradient Descent.

C. Experimental results

We test the hyperparameter-tuned Conjugate Model Gradient Descent algorithm on the problem of optimizing the UpCCD ansatz for the ground state of the RG Hamiltonian at a coupling $g = -0.9$. In order to demonstrate convergence, we perturb our ansatz parameters from values optimized on a noiseless classical simulation (using the L-BFGS-B optimizer as described above), by a random variable drawn from $[-0.25, 0.25]$. From this point, Conjugate Model Gradient Descent converged in 9 iterations with each iteration requiring 46 calls to the cost function (414 calls total). We observe that the optimizer successfully finds an energy below the initial point ($0.07 a.u.$), which demonstrates the well-known VQE ability to mitigate coherent noise [16, 17]. This improvement is reflected in the results mitigated with echo verification as well, which demonstrated a $0.04 a.u.$ reduction in error. However, this is a relatively small reduction compared to the overall error observed. We note that this is a significantly higher error than that observed in Fig. 8, which we attribute to poor calibration on the day. If the absolute gain in energy was replicated in Fig. 8, this would account for $\sim 40\%$ of the overall error. However, the relative gain in error in this case was only $\sim 10\%$. Ultimately, as the cost of optimization was already significant enough for this 6-qubit example, and as the number of calls for Conjugate Model Gradient Descent scales as $O(\#\text{parameters}) \sim O(N^2)$, we did not see it practical to continue this line of research in this work. With current qubit counts and shot repetition rates the optimization of variational algorithms of meaningful size remains a major

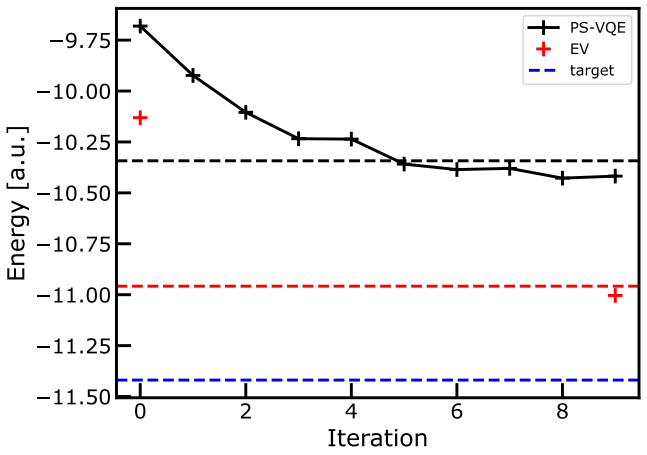


FIG. 5. Optimization trace of a 6 qubit RG simulation at $g = -0.9$, targeting the blue dashed line (ground state energy in the seniority-zero space). After an energy estimation using postselection (black dashed line) and echo verification (red dashed line) at optimized parameters from a noiseless simulation, all parameters were perturbed by a random variable drawn from $[-0.25, 0.25]$, and reoptimized over 9 iterations of Conjugate Model Gradient Descent, using the post-selected experimental data as a cost function. After perturbation and after convergence, a single call was made with the converged parameters to an echo verified energy estimation (red cross).

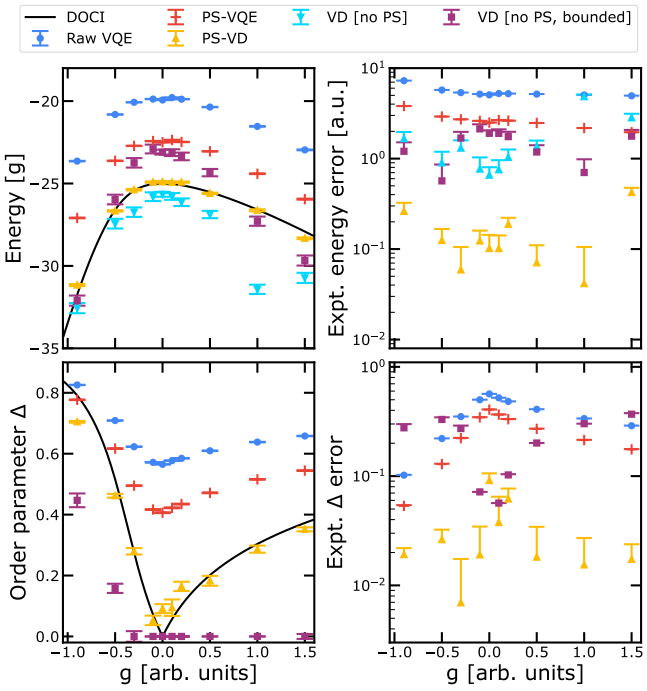


FIG. 6. Comparison of virtual distillation with postselection (yellow), without postselection (light blue), and without postselection but while bounding expectation values of Z_j and $D^{\pm}_{ij}$ in $[-1, 1]$ (purple), for a 10-qubit simulation of ground states of the RG Hamiltonian across a range of g values. Raw VQE (blue), postselected VQE (red), and post-selected VD [PS VD] data same as in Fig. 1 of the main text. VQE error bars obtained by error propagation (1 standard deviation, see App. III B), VD error bars obtained by bootstrapping (1 standard deviation).

obstacle.

VI. ADDITIONAL RESULTS AND ANALYSIS

A. Virtual distillation with and without postselection

In this appendix, we show the effect that postselection has on virtual distillation. In Fig. 6, we repeat the plot of Fig. 1 of the main text, but with data from VD without postselection on top. (Other lines are removed to make the plot clearer.) We see that by itself, VD [light blue curve] outperforms postselection in terms of energy estimation across most points of the energy curve, but it is not variationally bound. This is because without postselection, VD returns unphysical estimates $|\langle Z_i \rangle| \geq 1$ and $|\langle D^{\pm}_{ij} \rangle| \geq 1$. (This is a consequence of the two copies of the state not necessarily being subject to the same noise.) As a consequence of this, our estimate of the order parameter is complex and non-physical, and therefore not plotted. This issue can be rectified somewhat by bounding the estimates of $\langle Z_i \rangle$ and $\langle D^{\pm}_{ij} \rangle$ to lie in $[-1, 1]$. Performing this (Fig. 6, purple curve) allows for an estimate of the order parameter to be made, however it is clearly qualitatively incorrect. Moreover, although the energies are shifted up, some are still not variationally bounded, and those that are, often overshoot the energy, making the absolute error worse. (The variational bound could be rectified by enforcing positivity conditions on the set of generated estimates [9].) In summary, the combination of postselection and virtual distillation is seen here to have a greater effect than the sum of its parts for energy error.

We attribute the gain from postselection in virtual distillation to two sources. Firstly, postselection removes final readout noise, which VD does not naturally correct against. (As the estimation of $\mathrm{Tr}[\rho^2]$ involves a highly correlated measurement, this cannot be easily corrected by most readout correction schemes designed to estimate local Pauli expectation values.) Secondly, though we are in principle only post-selecting on the sum of the number of excitations in $\rho^{(1)}$ and the number of excitations in $\rho^{(2)}$ being equal to N, when combined with virtual distillation this projects out all noise such as T1 decay and suppresses any particle-non-conserving noise to second order. In short, this is because breaking number conservation separately on $\rho^{(1)}$ and $\rho^{(2)}$ yields states which do not overlap (as they must have different particle number), and these states cancel out when taking the product $\rho^{(1)}\rho^{(2)}$. Let us study this in more detail. Consider a post-selected estimate of $\mathrm{Tr}[\rho^2 O]$ using VD for $O = Z_j$ (as described in the main text). After postselecting on

$\sum_j (1 \otimes Z_j + Z_j \otimes 1) = N$, we can write our global state as

$$\rho_2 = \sum_{p,q,r,s} c_{p,q,r,s} |p\rangle\langle q| \otimes |r\rangle\langle s|, \tag{45}$$

where p, q, r, s index the basis states of both systems. (Following projection, ρ_2 will no longer be a product state.) Let us write $n_p = \langle p| \sum_j Z_j |p\rangle$ etc as the number of excitations in these basis states (i.e. the Hamming weight of the index), and our projection requires $c_{p,q,r,s} = 0$ unless $n_p + n_r = n_q + n_s = N$. (This assumes WLOG we are at half-filling, and ideally $n_p = n_r = n_q = n_s = N/2$.) Then, as O_s preserves particle number for $O = Z_j$, the only contribution to

$$\begin{aligned} &\mathrm{Tr}\left[\rho_2 S^{(2)} (Z_j \otimes I + I \otimes Z_j)\right] \\ &= \sum_{p,q,r,s} c_{p,q,r,s} \delta_{s,p} \delta_{r,q} \Big(\langle s|Z_j|s\rangle + \langle r|Z_j|r\rangle \Big), \end{aligned} \tag{46}$$

comes from those terms $c_{p,q,p,q}$. (The same is true for our estimate of Trace$[\rho_2 S^{(2)}]$.) This implies that a non-zero contribution to Trace$[\rho_2 S^{(2)}]$ comes only from matrix elements $|p\rangle\langle q| \otimes |p\rangle\langle q|$ where $n_p = \frac{N}{2} + \delta$, $n_q = \frac{N}{2} - \delta$. When $\delta = 0$, our matrix element $|p\rangle\langle q| \otimes |p\rangle\langle q|$ is in the number-conserving sector. When $\delta \neq 0$, our matrix element corresponds to a product of coherent superpositions between the $N/2+1$ and $N/2-1$ sector on both qubits. This implies that noise channels such as the T1 channel will be completely mitigated as these off-diagonal elements do not exist. Coherent noise may cause some of these off-diagonal elements to appear (consider e.g. a single-qubit X rotation on the state $\frac{1}{\sqrt{2}}(|01\rangle + |10\rangle)$.), however this is required to happen on both states to contribute, which suppresses it to second order. (This is in contrast to coherent noise that preserves number, to which we can be first-order sensitive.) The above analysis has assumed that the postselection is perfect; readout noise would complicate the above.

B. Distribution of errors in Pauli expectation value estimation

In Fig. 7, we plot a histogram of the error in estimating the expectation values of Pauli operators, taking data from Fig. 1 of the main text, Fig. 8, and Fig. 2 of the main text. We observe that the mean error in both cases for cyclobutene is slightly worse than for the RG model, but only by a small percentage. (This justifies our claim in the main text that a 0.1 a.u. error in the RG Hamiltonian is approximately $1-2$ orders of magnitude larger than chemical accuracy for a 10-qubit system.) As the difference between systems here is minimal (changing only the value of some virtual Z rotations), we can only attribute this difference to the performance of the device whilst taking these datasets. Going from 6 to 10 qubits increases the mean error in all experiments by a factor

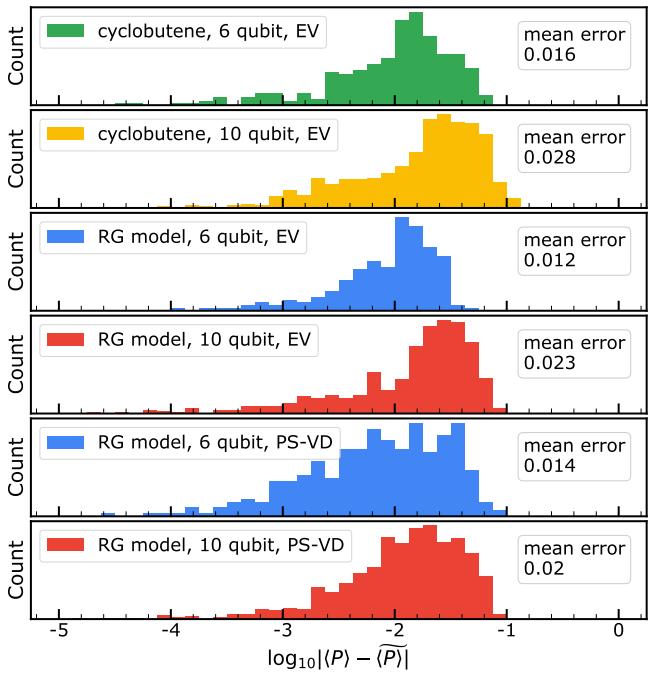


FIG. 7. Histogram of the expectation value error in Pauli operators $P = Z_j$, $P = Z_i Z_j$ or $P = D_{ij}^{\pm}$ (Eq. (27)), across all data taken for the experiment mentioned. Mean error across the dataset is given.

$1.5-2\times$. As the Hamiltonian for both the RG model and cyclobutene has a number of terms scaling as $O(N^2)$; assuming that the errors follow a roughly Gaussian distribution would predict the energy error scales $O(N)\times$ the error per each individual operator. This then predicts a gain in energy of $2.5-3.3\times$, which lies in between the observation of Fig. 2 of the main text and Fig. 3 of the main text. This suggests that the RG model energy estimation at large N may have had some benevolent cancellation of noise, and likewise for the cyclobutene energy estimation at small N.

C. Smaller studies of the RG Hamiltonian

In Fig. 8, we present experimental simulations of the ground state of the RG Hamiltonian for 4, 6, and 8 qubits. The method to produce these figures is identical to that used in the production of Fig. 1 of the main text, save for the number of qubits and shots used. Aggregated data from these figures was used to generate the scaling plots of Fig. 3 of the main text.

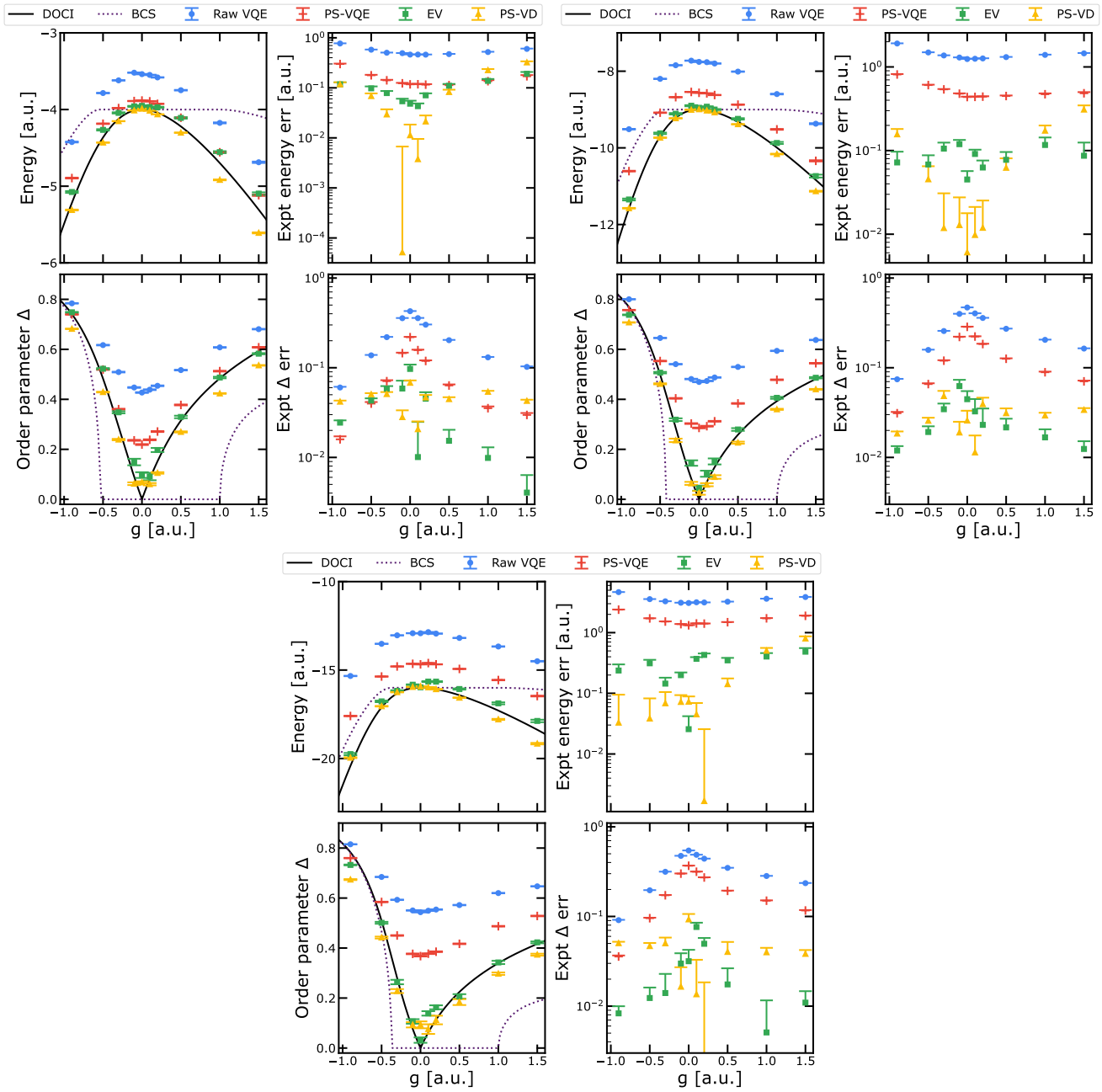


FIG. 8. Identical experiment to Fig. 1 of the main text, but for 4 (top-left), 6(top-right), and 8 (bottom) qubits instead of 10. Aggregate data is used in Fig. 3 of the main text. (top-left) Energy plot for the RG system as a function of the coupling parameter g, for an unmitigated state preparation [blue circles], and state preparation mitigated by postselection [red crosses], echo verification [yellow triangles], and postselected virtual distillation [green squares]. This is compared to the exact DOCI result [black solid line], and BCS [purple dashed line]. (top-right) Log plot of experimental energy error (ignoring the model error from the UpCCD approximation). (bottom-left) Superconducting order parameter for the RG Hamiltonian ($\Delta = \frac{1}{N}\sum_{j,\sigma}\sqrt{\langle n_{j\sigma}^2\rangle - \langle n_{j\sigma}\rangle^2}$), again compared to classical models. (bottom-right) Experimental error in estimating the superconducting order parameter vs the target state within the UpCCD approximation (again ignoring model error). 1 std. dev. error bars estimated by either propagating variance (Raw VQE, PS VQE) or bootstrapping (EV, PS VD).

[1] 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, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, T. S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven, and J. M. Martinis, Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019).

[2] C. Neill, T. McCourt, X. Mi, Z. Jiang, M. Y. Niu, W. Mruczkiewicz, I. Aleiner, F. Arute, K. Arya, J. Atalaya, R. Babbush, J. C. Bardin, R. Barends, A. Bengtsson, A. Bourassa, M. Broughton, B. B. Buckley, D. A. Buell, B. Burkett, N. Bushnell, J. Campero, Z. Chen, B. Chiaro, R. Collins, W. Courtney, S. Demura, A. R. Derk, A. Dunsworth, D. Eppens, C. Erickson, E. Farhi, A. G. Fowler, B. Foxen, C. Gidney, M. Giustina, J. A. Gross, M. P. Harrigan, S. D. Harrington, J. Hilton, A. Ho, S. Hong, T. Huang, W. J. Huggins, S. V. Isakov, M. Jacob-Mitos, E. Jeffrey, C. Jones, D. Kafri, K. Kechedzhi, J. Kelly, S. Kim, P. V. Klimov, A. N. Korotkov, F. Kostritsa, D. Landhuis, P. Laptev, E. Lucero, O. Martin, J. R. McClean, M. McEwen, A. Megrant, K. C. Miao, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, M. Newman, T. E. O'Brien, A. Opremcak, E. Ostby, B. Pató, A. Petukhov, C. Quintana, N. Redd, N. C. Rubin, D. Sank, K. J. Satzinger, V. Shvarts, D. Strain, M. Szalay, M. D. Trevithick, B. Villalonga, T. C. White, Z. Yao, P. Yeh, A. Zalcman, H. Neven, S. Boixo, L. B. Ioffe, P. Roushan, Y. Chen, and V. Smelyanskiy, Accurately computing the electronic properties of a quantum ring, Nature **594**, 508 (2021).

[3] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, A. Bengtsson, S. Boixo, M. Broughton, B. B. Buckley, D. A. Buell, B. Burkett, N. Bushnell, Y. Chen, Z. Chen, Y.-A. Chen, B. Chiaro, R. Collins, S. J. Cotton, W. Courtney, S. Demura, A. Derk, A. Dunsworth, D. Eppens, T. Eckl, C. Erickson, E. Farhi, A. Fowler, B. Foxen, C. Gidney, M. Giustina, R. Graff, J. A. Gross, S. Habegger, M. P. Harrigan, A. Ho, S. Hong, T. Huang, W. Huggins, L. B. Ioffe, S. V. Isakov, E. Jeffrey, Z. Jiang, C. Jones, D. Kafri, K. Kechedzhi, J. Kelly, S. Kim, P. V. Klimov, A. N. Korotkov, F. Kostritsa, D. Landhuis, P. Laptev, M. Lindmark, E. Lucero, M. Marthaler, O. Martin, J. M. Martinis, A. Marusczyk, S. McArdle, J. R. McClean, T. McCourt, M. McEwen, A. Megrant, C. Mejuto-Zaera, X. Mi, M. Mohseni, W. Mruczkiewicz, J. Mutus, O. Naaman, M. Neeley, C. Neill, H. Neven, M. Newman, M. Y. Niu, T. E. O'Brien, E. Ostby, B. Pató, A. Petukhov, H. Putterman, C. Quintana, J.-M. Reiner, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, D. Strain, K. J. Sung, P. Schmitteckert, M. Szalay, N. M. Tubman, A. Vainsencher, T. White, N. Vogt, Z. J. Yao, P. Yeh, A. Zalcman, and S. Zanker, Observation of separated dynamics of charge and spin in the fermi-hubbard model, ArXiv:2010.07965 (2020).

[4] F. Arute, K. Arya, R. Babbush, D. Bacon, J. Bardin, R. Barends, S. Boixo, M. Broughton, B. Buckley, D. Buell, B. Burkett, N. Bushnell, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, S. Demura, A. Dunsworth, E. Farhi, A. Fowler, B. Foxen, C. Gidney, M. Giustina, R. Graff, S. Habegger, M. Harrigan, A. Ho, S. Hong, T. Huang, W. Huggins, L. Ioffe, S. Isakov, E. Jeffrey, Z. Jiang, C. Jones, D. Kafri, K. Kechedzhi, J. Kelly, S. Kim, P. Klimov, A. Korotkov, F. Kostritsa, D. Landhuis, P. Laptev, M. Lindmark, E. Lucero, O. Martin, J. Martinis, J. McClean, M. McEwen, A. Megrant, X. Mi, M. Mohseni, W. Mruczkiewicz, J. Mutus, O. Naaman, M. Neeley, C. Neill, H. Neven, M. Niu, T. O'Brien, E. Ostby, A. Petukhov, H. Putterman, C. Quintana, P. Roushan, N. Rubin, D. Sank, K. Satzinger, V. Smelyanskiy, D. Strain, K. Sung, M. Szalay, T. Takeshita, A. Vainsencher, T. White, N. Wiebe, Z. Jamie Yao, P. Yeh, and A. Zalcman, Hartree-fock on a superconducting qubit quantum computer, Science **369**, 1084 (2020).

[5] Quantum AI team and collaborators, Recirq (2020).

[6] S. Polla, G.-L. R. Anselmetti, and T. E. O'Brien, Optimizing the information extracted by a single qubit measurement, ArXiv:2207.0947 (2022).

[7] T. E. O'Brien, S. Polla, N. C. Rubin, W. J. Huggins, S. McArdle, S. Boixo, J. R. McClean, and R. Babbush, Error mitigation via verified phase estimation, PRX Quantum **2**, 020317 (2021).

[8] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, Phys. Rev. A **92**, 042303 (2015).

[9] N. C. Rubin, R. Babbush, and J. McClean, Application of fermionic marginal constraints to hybrid quantum algorithms, New J. Phys. **20**, 053020 (2018).

[10] W. J. Huggins, J. McClean, N. Rubin, Z. Jiang, N. Wiebe, K. B. Whaley, and R. Babbush, Efficient and Noise Resilient Measurements for Quantum Chemistry on Near-Term Quantum Computers, npj Quantum Information volume **7** (2021).

[11] K. J. Sung, J. Yao, M. P. Harrigan, N. C. Rubin, Z. Jiang, L. Lin, R. Babbush, and J. R. McClean, Using models to improve optimizers for variational quantum algorithms, Quantum Science and Technology **5**, 044008 (2020).

[12] M. R. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, Journal of Research of the National Bureau of Standards **49**, 409 (1952).

[13] J. W. Daniel, Convergence of the conjugate gradient method with computationally convenient modifications, Numerische Mathematik **10**, 125 (1967).

[14] R. Fletcher and C. M. Reeves, Function minimization by conjugate gradients, The Computer Journal **7**, 149 (1964), https://academic.oup.com/comjnl/article-pdf/7/2/149/959725/070149.pdf.

[15] S. Stanisic, J. L. Bosse, F. M. Gambetta, R. A. Santos, W. Mruczkiewicz, T. E. O'Brien, E. Ostby, and A. Montanaro, Observing ground-state properties of the fermi-hubbard model using a scalable algorithm on a quantum computer, ArXiv:2112.02025 (2021).

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

[17] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The Theory of Variational Hybrid Quantum-Classical Algorithms, New Journal of Physics **18**, 23023 (2016).