Efficient quantum measurement of Pauli operators in the presence of finite sampling error

Ophelia Crawford[*1], Barnaby van Straaten[*1], Daochen Wang[*1,2], Thomas Parks[1], Earl Campbell[1,3], and Stephen Brierley[1]

[1]Riverlane, Cambridge, UK

[2]Joint Center for Quantum Information and Computer Science, University of Maryland, College Park, USA

[3]Department of Physics and Astronomy, University of Sheffield, Sheffield, UK

Estimating the expectation value of an operator corresponding to an observable is a fundamental task in quantum computation. It is often impossible to obtain such estimates directly, as the computer is restricted to measuring in a fixed computational basis. One common solution splits the operator into a weighted sum of Pauli operators and measures each separately, at the cost of many measurements. An improved version collects mutually commuting Pauli operators together before measuring all operators within a collection simultaneously. The effectiveness of doing this depends on two factors. Firstly, we must understand the improvement offered by a given arrangement of Paulis in collections. In our work, we propose two natural metrics for quantifying this, operating under the assumption that measurements are distributed optimally among collections so as to minimise the overall finite sampling error. Motivated by the mathematical form of these metrics, we introduce SORTED INSERTION, a collecting strategy that exploits the weighting of each Pauli operator in the overall sum. Secondly, to measure all Pauli operators within a collection simultaneously, a circuit is required to rotate them to the computational basis. In our work, we present two efficient circuit constructions that suitably rotate any collection of k independent commuting n-qubit Pauli operators using at most $kn - k(k+1)/2$ and $O(kn/\log k)$ two-qubit gates respectively. Our methods are numerically illustrated in the context of the Variational Quantum Eigensolver, where the operators in question are molecular Hamiltonians. As measured by our metrics, SORTED INSERTION outperforms four conventional greedy colouring algorithms that seek the minimum number of collections.

[*]These authors contributed equally to this work.

1 Introduction

Estimating the expectation value of an operator corresponding to an observable on a quantum state is a fundamental task in quantum mechanical experiments. Expectation estimation of a Hamiltonian features prominently as the quantum sub-routine of the Variational Quantum Eigensolver (VQE) algorithm [1], which has emerged as a leading candidate for exhibiting quantum advantage in the Noisy Intermediate Scale Quantum era [2]. VQE is a hybrid quantum-classical algorithm designed to find the ground state [1, 3–9], or energy spectrum [10–15], of a physical or chemical system. Often there is no natural way to obtain the expectation of an operator directly and some indirect protocol is required. In particular, this is true of current quantum computers, which can only obtain measurements in the computational basis defined, by convention, as eigenstates of the Pauli-Z operator on each qubit.

One naive method of proceeding, therefore, is to decompose the operator into a weighted sum of Pauli operators (or Paulis) and then measure each separately. The paper that introduced VQE [1] proposed measuring the Hamiltonian in this way. However, this can be inefficient. For example, a second-quantised chemical Hamiltonian on n qubits decomposes into a very large number of Paulis that scales as n^4. An improved method, therefore, is to arrange the Paulis in commuting collections. All Paulis in a collection can be measured at the same time, as any set of commuting Paulis can be simultaneously diagonalised by a single unitary. There are typically many arrangements of Paulis in mutually commuting collections, and the reduction in the number of measurements required will depend on the arrangement.

In the context of VQE, McClean et al. [3] proposed this improved protocol and argued using a toy example that the arrangement with the fewest collections might not result in the fewest measurements as splitting a single commuting collection into two might sometimes offer an improvement. However, Ref. [3] did not show how to obtain commuting collections nor how to construct the unitary rotation, U, that enables simultaneous measurement.

Recently, a series of papers [16–22][*] have appeared that make progress on both the collecting strategy and rotation circuit construction problems. Our paper is in the same arena and addresses *both* problems.

First, we define two natural metrics, R and $\hat{R}$, that quantify the performance of any given arrangement. R and $\hat{R}$ measure the ratio of the number of measurements required in the uncollected case to the collected case to attain a fixed level of accuracy. The key feature of these two metrics is that they assume measurements are distributed optimally between the collections to minimise the finite sampling error, following Refs [8, 25, 26]. The difference between them is that R is state-dependent but $\hat{R}$ is designed to approximate $\mathbb{E}[R]$ over the uniform spherical measure. Therefore, R is more suitable for use given knowledge of the underlying quantum state, while $\hat{R}$ is more suitable otherwise.

With R and $\hat{R}$ defined, we prove a first result that breaking a single commuting collection into two is never advantageous, as the number of measurements required to obtain an expectation estimate to a given accuracy is never reduced. This result already contradicts the conclusion of the aforementioned toy example used by McClean et al. [3], and analysed in full in Ref. [20, Sec. 10.1], that breaking a collection can be advantageous. The reason for the discrepancy is that we distribute measurements optimally among the collections so as to minimise the overall error due to finite sampling for a given number of measurements. In contrast, in the previous works, measurements are distributed uniformly between the collections.

We further propose an intrinsically new collecting strategy, which we call SORTED INSERTION, that is designed to maximise R and $\hat{R}$. Unlike all strategies used previously that seek the minimum number of collections [16, 17, 20, 21], SORTED INSERTION seeks to maximise $\hat{R}$ by explicitly exploiting the coefficients of the Paulis in its assignment of collections. It is important to stress that maximising $\hat{R}$ is not the same as minimising the number of collections, as we show by a toy example. Perhaps counter-intuitively, this does not contradict the above conclusion, which showed only that it is never better to break a single collection into two.

When considering the rotation circuit construction problem, we contribute two new methods for constructing Clifford rotation circuits that enable simultaneous measurement of a collection containing arbitrary commuting Paulis. Like Ref. [20], we approach the problem via the stabiliser formalism but further

consider the case where the number of independent operators, k, in a collection can be less than the number of qubits, n. We show that the number of two-qubit gates in the rotation circuit, U, can be reduced in a way that scales with k. This is important because it is atypical for actual collections to have exactly $k = n$ independent Paulis and reducing the number of two-qubit gates is important, especially in the near-term [27–41]. As far as we are aware, ours is the first paper to consider the $k < n$ case explicitly. Furthermore, we allow classical post-processing, which can save quantum resources.

More specifically, we introduce constructions "CZ" and "CNOT". The CZ-construction builds on work by Van den Nest, Dehaene, and De Moor [42] in the graph-state literature to yield U with a number of two-qubit gates at most

$$u_{\mathrm{cz}}(k, n) = kn - k(k+1)/2. \qquad (1)$$

The CNOT-construction builds on our CZ-construction, and work by Aaronson and Gottesman [43] and Patel, Markov, and Hayes [44] to yield U with a number of two-qubit gates at most

$$u_{\mathrm{cnot}}(k, n) = O(kn/\log k). \qquad (2)$$

We stress that u_{cnot} and u_{cz} are worst-case upper bounds. In practice, numerical simulations are needed to determine whether the CZ- or CNOT-construction is actually more efficient. We note that, in the case of $k = n$, our constructions have two-qubit gate-counts scaling no worse than the previous best of $O(n^2)$ [20]. Other works, such as Ref. [17, Appendix A], do not provide gate-counts, or demonstrate a scaling that is worse than $O(n^2)$ [18, Appendix B], or present a method that only works for the second-quantised fermionic Hamiltonian [19, 22].

We end our paper with a series of numerical results on molecular Hamiltonians that serve to illustrate and validate our theoretical work. In doing so, we first quantify the performance of SORTED INSERTION using the metric $\hat{R}$ for molecules ranging in size from hydrogen H_2, which requires two qubits, to hydrogen selenide H_2Se, which requires 38, finding that it leads to a 10 to 60 fold improvement in the number of measurements required. Note that we are defining a single measurement to consist of a measurement of all qubits, and so the number of measurements equals the number of ansatz state preparations.

We further present results of using SORTED INSERTION alongside our CZ-construction on molecules requiring up to 38 qubits to calculate the actual number of two-qubit gates required for real molecular systems. Our numerical results show that the typical number of two-qubit gates is fewer than the worst-case $u_{\mathrm{cz}}(k, n)$ by a factor of 3.5.

Lastly, we present data showing SORTED INSERTION outperforming the four conventional greedy

---

[*]We mention that both the collecting and rotation circuit synthesis problems have been addressed on an ad-hoc basis by experimentalists since Kandala et al. [5]. We refer readers to Table 2 of Ref. [20] for a good summary. We also mention that recent Ref. [23] and the less recent Ref. [24] allow for collecting of non-commuting Paulis. We found it interesting that such approaches are feasible, but found it hard to compare them to our work on a like-for-like basis.

colouring algorithms we tested, as measured by the metric $\hat{R}$. Our data strongly *challenges* the assumption that minimising the number of collections is best, as arrangements with the smallest number of collections do not typically perform the best with respect to $\hat{R}$.

We mention that an entirely different measurement strategy has been proposed recently [45] that is based on so-called "classical shadows". Ref. [45, Illustrative Example Applications] explicitly mentions that their method is unsuitable for estimating expectation values of global observables. Even for a Hamiltonian consisting of a single global Pauli operator $P := P_1 \otimes P_2 \otimes \cdots \otimes P_n$, with $P_i \in \{X, Y, Z\}$, their number of classical shadows must scale as $2^{\Omega(n)}$. In contrast, our method, as with the original VQE proposal, scales independently of n and would simply measure each P_i and multiply the outcomes.

Follow-up Ref. [46] sought to ameliorate this problem by using so-called "locally-biased classical shadows". However, as they mention in [46, Remark 1], the *analytical* upper bound on their estimator can still scale as $2^{\Omega(n)}$ for the same P above. Separately, Ref. [46] did not compare their techniques to ours, rather they observe that we require deeper circuits. However, the additional depth our method requires is modest *relative* to that of ansatz state preparation.

There have also been interesting developments [47, 48] on measurement reduction in VQE focusing on designing classical optimisers that intelligently choose how many measurements to use at each iteration – for example, fewer measurements may be used near the start versus near the end. These methods complement ours and can be combined with ours to reduce further the overall number of measurements.

2 Collecting strategies

In this section, we develop a method for collecting Pauli operators designed to minimise the number of measurements required to obtain an estimate of the expectation value of an operator to a given level of accuracy.

We have an operator, O, of the form

$$O = \sum_{i=1}^{N} O_i = \sum_{i=1}^{N} \sum_{j=1}^{m_i} a_{ij} P_{ij}, \qquad (3)$$

where N is the number of collections of mutually commuting operators, m_i is the number of operators in collection i, P_{ij} is the jth Pauli operator in the ith collection and $a_{ij} \in \mathbb{R}$ is its coefficient.

We assume that the P_{ij}s are distinct for different (i, j), i.e., we do not collect a single Pauli operator appearing in O into more than one collection by splitting its coefficient. This is discussed further at the end of this section.

Given ϵ, let M_u and M_g be the minimum number of measurements required to attain an accuracy ϵ in the uncollected and collected (as per Eq. 3) cases respectively. Finding M_u is a special case of finding M_g. To find M_g, we can solve the constrained optimisation problem that asks how a given number of measurements should be distributed in order to maximise accuracy. The solution to such a problem gives the optimal measurement strategy for a given operator and state. Using Lagrange multipliers [8, 25, 26], we find that the optimal number n_i of measurements of collection i is

$$n_i := \frac{1}{\epsilon^2} \sqrt{\mathrm{Var}[O_i]} \sum_{j=1}^{N} \sqrt{\mathrm{Var}[O_j]}. \qquad (4)$$

Therefore, we have

$$M_g = \sum_{i=1}^{N} n_i = \left(\frac{1}{\epsilon} \sum_{i=1}^{N} \sqrt{\mathrm{Var}[O_i]} \right)^2, \qquad (5)$$

where

$$\mathrm{Var}[O_i] = \mathrm{Cov}[O_i, O_i] \qquad (6)$$
$$= \langle O_i^2 \rangle - \langle O_i \rangle^2. \qquad (7)$$

Since M_u is just M_g evaluated with every operator in its own collection, we have

$$M_u = \left(\frac{1}{\epsilon} \sum_{i=1}^{N} \sum_{j=1}^{m_i} |a_{ij}| \sqrt{\mathrm{Var}[P_{ij}]} \right)^2, \qquad (8)$$

where

$$\mathrm{Var}[P_{ij}] = 1 - \langle P_{ij} \rangle^2. \qquad (9)$$

The ratio R, defined as

$$R := \frac{M_u}{M_g} = \left(\frac{\sum_{i=1}^{N} \sum_{j=1}^{m_i} |a_{ij}| \sqrt{\mathrm{Var}[P_{ij}]}}{\sum_{i=1}^{N} \sqrt{\mathrm{Var}[O_i]}} \right)^2, \qquad (10)$$

therefore acts as a natural metric for the performance of a particular arrangement of Paulis under the assumption that measurements are distributed optimally. The larger the value of R, the greater the saving that assembling operators into collections provides. In Theorem 1 below, we show that combining two collections into one can only improve R.

Theorem 1. *Consider two collections $\mathcal{A}$ and $\mathcal{B}$ of mutually commuting Paulis, where each Pauli is in at most one collection. Suppose that it is possible to combine $\mathcal{A}$ and $\mathcal{B}$ into a single commuting collection $\mathcal{C} := \mathcal{A} \cup \mathcal{B}$. Let $R(\{\mathcal{A}, \mathcal{B}\})$ and $R(\{\mathcal{C}\})$ denote the R metric, as defined in Eq. (10), for the two collections separated and combined respectively. Then*

$$R(\{\mathcal{A}, \mathcal{B}\}) \le R(\{\mathcal{C}\}) \quad \text{for all} \ |\psi\rangle. \qquad (11)$$

Proof. We write the operators associated with collections $\mathcal{A}$ and $\mathcal{B}$ as

$$O_{\mathcal{A}} = \sum_{j=1}^{k} a_j P_j, \qquad (12)$$

$$O_{\mathcal{B}} = \sum_{j=k+1}^{k+l} a_j P_j, \qquad (13)$$

for some k, l respectively, where the indexing of the sum in $O_{\mathcal{B}}$ is for later convenience. The operator associated with $\mathcal{C}$ can therefore be written as

$$O_{\mathcal{C}} = \sum_{j=1}^{k+l} a_j P_j. \qquad (14)$$

We define the $(k+l) \times (k+l)$ covariance matrix C, which is symmetric and positive semi-definite, associated with collection $\mathcal{C}$ by its entries

$$C_{ij} \coloneqq \mathrm{Cov}[P_i, P_j], \qquad (15)$$

with

$$\mathrm{Cov}[P_i, P_j] \coloneqq \langle P_i P_j \rangle_{\psi} - \langle P_i \rangle_{\psi} \langle P_j \rangle_{\psi}. \qquad (16)$$

We write $\mathbf{a}$ for the vector of length $k+l$ whose first k entries are given by $\{a_j\}_{j=1}^{k}$ and the remaining l entries are zero. Similarly, we write $\mathbf{b}$ for the vector of length $k+l$ whose first k entries are zero and the remaining l entries are given by $\{a_j\}_{j=k+1}^{k+l}$.

Since the numerator of R is constant for different arrangements, to prove our claim it suffices for us to show that the denominator of R is not smaller in the case of $\{\mathcal{A}, \mathcal{B}\}$ than in the case of $\{\mathcal{C}\}$, i.e., that

$$\sqrt{\mathrm{Var}[O_{\mathcal{A}}]} + \sqrt{\mathrm{Var}[O_{\mathcal{B}}]} \geq \sqrt{\mathrm{Var}[O_{\mathcal{C}}]}. \qquad (17)$$

With the above notation, this is equivalent to

$$\sqrt{\mathbf{a}^{\mathsf{T}} C \mathbf{a}} + \sqrt{\mathbf{b}^{\mathsf{T}} C \mathbf{b}} \geq \sqrt{(\mathbf{a}+\mathbf{b})^{\mathsf{T}} C (\mathbf{a}+\mathbf{b})}. \qquad (18)$$

We find

$$\begin{aligned}
&\sqrt{\mathbf{a}^{\mathsf{T}} C \mathbf{a}} + \sqrt{\mathbf{b}^{\mathsf{T}} C \mathbf{b}} \\
&= \sqrt{\left(\sqrt{\mathbf{a}^{\mathsf{T}} C \mathbf{a}} + \sqrt{\mathbf{b}^{\mathsf{T}} C \mathbf{b}}\right)^2} \\
&= \sqrt{\mathbf{a}^{\mathsf{T}} C \mathbf{a} + \mathbf{b}^{\mathsf{T}} C \mathbf{b} + 2\sqrt{(\mathbf{a}^{\mathsf{T}} C \mathbf{a})(\mathbf{b}^{\mathsf{T}} C \mathbf{b})}} \\
&\geq \sqrt{\mathbf{a}^{\mathsf{T}} C \mathbf{a} + \mathbf{b}^{\mathsf{T}} C \mathbf{b} + 2(\mathbf{a}^{\mathsf{T}} C \mathbf{b})} \\
&= \sqrt{(\mathbf{a}+\mathbf{b})^{\mathsf{T}} C (\mathbf{a}+\mathbf{b})},
\end{aligned} \qquad (19)$$

where the fourth line follows due to the Cauchy-Schwarz inequality on a semi-inner product defined by $\langle \mathbf{a}, \mathbf{b} \rangle \coloneqq \mathbf{a}^{\mathsf{T}} C \mathbf{b}$ [49, Example 1.1] and the fifth line follows as C is symmetric. This establishes our claim.

The inequality of Eq. 11 holds with equality if and only if there exist $0 \neq \alpha, \beta \in \mathbb{C}$, such that $\langle \alpha\mathbf{a} + \beta\mathbf{b}, \alpha\mathbf{a} + \beta\mathbf{b} \rangle = 0$ [49, Example 1.4]. $\qquad \square$

Theorem 1 shows that it is impossible to mitigate covariances by splitting collections under the optimal measurement strategy. This is in contrast to Refs [3, 20], which showed that it is possible using a sub-optimal measurement strategy. In Appendix A, we re-do precisely their example using the optimal measurement strategy. Note that Theorem 1 simply implies that it is never better to break a single collection into two, and not that the minimum number of collections is necessarily the best; indeed, we provide a counter-example below.

If all of the variances in R are replaced by their expectation values over the uniform spherical measure (see Ref. [50, Ch. 7]), we obtain another metric, $\hat{R}$, given by

$$\hat{R} \coloneqq \left(\frac{\sum_{i=1}^{N} \sum_{j=1}^{m_i} |a_{ij}|}{\sum_{i=1}^{N} \sqrt{\sum_{j=1}^{m_i} |a_{ij}|^2}} \right)^2. \qquad (20)$$

The derivation of $\hat{R}$ is given in Appendix B. The same proof as in Theorem 1 can be used to show that breaking a collection into two is never better when measured by $\hat{R}$; the only difference is the covariance matrix must be replaced by its expectation over the uniform spherical measure.

$\hat{R}$ is a particularly useful metric because it approximates R, but can be calculated analytically. When ignorant of the quantum state $|\psi\rangle$, a good collecting strategy should aim to maximise $\hat{R}$. As mentioned above, this is emphatically not the same as minimising the number of collections. For example, let us consider the operator O, on two qubits, defined as

$$O = 4 X_1 + 4 X_2 + Z_2 + Z_1 X_2. \qquad (21)$$

The operators cannot be assembled into a single commuting collection but there is a unique arrangement, $\mathcal{G}_1$, consisting of two collections, given by

$$\mathcal{G}_1 \coloneqq \left\{ \{4 X_1, \ Z_2\}, \{4 X_2, \ Z_1 X_2\} \right\}, \qquad (22)$$

which has $\hat{R} = 1.47$ to two decimal places. $\mathcal{G}_1$ performs worse than an arrangement $\mathcal{G}_2$, with three collections, given by

$$\mathcal{G}_2 \coloneqq \left\{ \{4 X_1, 4 X_2\}, \{Z_2\}, \{Z_1 X_2\} \right\}, \qquad (23)$$

which has $\hat{R} = 1.71$ to two decimal places.

Reflecting upon such examples and the nature of the square root function in the denominator of $\hat{R}$ leads us to propose a collecting algorithm that prioritises collecting Paulis with large coefficients.

This algorithm, which we name SORTED INSER-

TION, can be described as follows. Given an operator

$$O = \sum_{i=1}^{t} a_j P_j, \qquad (24)$$

where t is the total number of Pauli operators, the entire set $\{(a_j, P_j)\}_{j=1}^{t}$ is sorted by the absolute value of coefficients a_j, so that $|a_1| \geq \cdots \geq |a_t|$. Then, in the order $i = 1, \ldots t$, it is checked whether P_i commutes with all elements in an existing collection. If it does, it is added to that collection. If not, a new collection is created and P_i is inserted there. The collections are checked in order of their creation[*]. The collections formed are tracked and outputted at the end once the final P_t has been inserted.

SORTED INSERTION has worst-case complexity $O(nt^2)$, where we recall that n is the number of qubits. In contrast, greedy colouring algorithms, as implemented in Ref. [16], require pre-generating the *full* commutation graph which has complexity $\Theta(nt^2)$ even in the best case. The colouring algorithms then run on this graph adding further complexity – see Table 1. Therefore, SORTED INSERTION is at least as efficient as these greedy colouring algorithms.

Colouring Algorithm	Time Complexity
Largest First	$O(t^2)$
Connected Sequential d.f.s.	$O(t^2)$
DSATUR	$O(t^2 \log t)$
Independent Set	$O(t^3)$

Table 1: Time complexities of the greedy colouring algorithms we compare with SORTED INSERTION in Sec. 4 *after* pre-generating the commutation graph [51].

Note that SORTED INSERTION is a heuristic algorithm for maximising $\hat{R}$. It works well in practice, as demonstrated by Table 3 in Sec. 4, but may fail to output collections that actually maximise $\hat{R}$. In fact, it is unlikely we can go beyond heuristics as the problem of maximising $\hat{R}$ is NP-hard in general. We can show this by combining the reduction from MIN-CLIQUE-COVER in Ref. [20, Appendix A] with the NP-hardness of $|V|^{1-\epsilon}$-approximating the MIN-CLIQUE-COVER [52], where $|V|$ is the number of vertices and $\epsilon = 0.5$. We appeal to the hardness of *approximation* because, as we have stressed, maximising $\hat{R}$ is not the same as minimising the number of collections, which corresponds to minimising number of cliques in a cover. For a full proof, see Appendix F.

Lastly, we mention that our choice not to put a single Pauli into multiple collections by splitting its coefficient may be sub-optimal as can be seen by con-

sidering the operator O, on two qubits, defined as

$$O = X_1 + Z_1 + 2Z_2 \qquad (25)$$

and a grouping of the form

$$\mathcal{G}_3 := \{\{X_1 + \alpha Z_2\}, \{Z_1 + (2-\alpha)Z_2\}\} \qquad (26)$$

for $\alpha \in [0,2]$. It can be verified that $\hat{R}$, defined as before even with coefficient splitting, is maximised at $\alpha = 1$. We leave maximising $\hat{R}$ with consideration of coefficient splitting for future work.

3 Rotation constructions

In this section, we present two methods of calculating a rotation circuit which enables measurements of all operators in a mutually commuting collection to be obtained simultaneously. We assume familiarity of the reader with the stabiliser formalism, especially the $2n$-bit binary representation of n-qubit Paulis [43, 53–55]. We follow the convention that the upper and lower halves of the binary matrix encode Z and X operators respectively. This representation is reviewed in Appendix C. We also reserve symbols I_m and 0_m for the $m \times m$ identity and all-zero matrices respectively.

Our starting point is a commuting collection, $\mathcal{S}'_{\text{start}}$, of m Paulis which can be represented as a binary $2n \times m$ matrix S'_{start}. By Gaussian elimination on S'_{start}, we can form a $2n \times k$ matrix S_{start} representing a set $\mathcal{S}_{\text{start}}$ of k independent Paulis drawn from $\mathcal{S}'_{\text{start}}$ where $k \leq \min(n, m)$. Our goal is to transform S_{start}, using certain allowed transformations, into a $2n \times n$ matrix S_{end} where

$$S_{\text{end}} = \begin{pmatrix} I_n \\ 0_n \end{pmatrix}. \qquad (27)$$

Let U denote the circuit consisting of one-qubit and two-qubit transformations in the order they were applied from $S_{\text{start}} \to S_{\text{end}}$. Then applying U to any state $|\psi\rangle$, measuring in the computational basis, and classically post-processing allows us to obtain measurements of all operators in $\mathcal{S}'_{\text{start}}$ on $|\psi\rangle$ simultaneously.

The allowed set, $\mathcal{T}$, of transformations on a binary $2n \times m$ matrix S is, where p ranges over all columns, r ranges over all rows, and addition is mod 2:

1. One-qubit and two-qubit quantum gates, corresponding to row operations, specifically:

   ■ CZ on qubits i and j:
   $S_{ip} \leftarrow S_{ip} + S_{j+n,p}$,
   $S_{jp} \leftarrow S_{jp} + S_{i+n,p}$.

   ■ CNOT on control-qubit i and target-qubit j:
   $S_{ip} \leftarrow S_{ip} + S_{jp}$,
   $S_{j+n,p} \leftarrow S_{j+n,p} + S_{i+n,p}$.

---

[*]We note that other collection orderings could also be considered, for example, ordering the collections by their values of $\sum_{j=1}^{m_i} |a_{ij}|^2$, where the a_{ij} are the coefficients of the Paulis contained in collection i at the point of ordering.

- HADAMARD (H) on qubit i:
  $S_{ip} \leftrightarrow S_{i+n,p}$.
- PHASE (P) on qubit i:
  $S_{ip} \leftarrow S_{ip} + S_{i+n,p}$.

2. Classical post-processing:

- Products of eventual single-qubit computational-basis measurements: right-multiplication by invertible $m \times m$ matrix.
- Relabelling of qubits i and j:
  $S_{ip} \leftrightarrow S_{jp}$,
  $S_{n+i,p} \leftrightarrow S_{n+j,p}$.

3. Basis extension:

- Addition of further stabiliser: appending of new column $S_{r,m+1}$.

In the near term, operations in $\mathcal{T}$ have justifiably different costs. Two-qubit gates are much more costly than one-qubit gates, which are more costly than classical post-processing. Basis extension has no cost. In the context of one-qubit and two-qubit quantum gates, cost can refer to either fidelity or gate-time [27–41]. Therefore, we have aimed to minimise the number of two-qubit gates in the U resulting from our constructions. This means, for example, we choose to perform a row swap using classical post-processing rather than the two-qubit SWAP gate.

In presenting our constructions, we shall refer to the commutativity condition, preserved under $\mathcal{T}$, given by

$$S^{\top} J_{2n} S = 0_m, \qquad (28)$$

where S is the $2n \times m$ matrix encoding the Paulis and

$$J_{2n} = \begin{pmatrix} 0_n & I_n \\ I_n & 0_n \end{pmatrix}. \qquad (29)$$

We ignore any changes in sign of stabilisers under $\mathcal{T}$ as this can be easily accounted for by classical post-processing. Readers interested in this and other details are referred to Appendix D, where we work through our CZ-construction with a specific example.

3.1 CZ-construction

Important to our first approach is the special class of stabiliser states known as graph states. Consider any graph G on n vertices. The graph state $|\Phi_G\rangle$ is then defined by n independent stabiliser generators

$$g_i = X_i \prod_{j \in \mathrm{nbd}(i)} Z_j, \; i = 1, \ldots, n, \qquad (30)$$

where $\mathrm{nbd}(i)$ is the set of neighbours of vertex i in G. The binary representation of these stabilisers is

$$S_{\mathrm{graph}} = \begin{pmatrix} A \\ I_n \end{pmatrix}, \qquad (31)$$

where A, an $n \times n$ symmetric matrix with 0s on its diagonal, is exactly the adjacency matrix of G.

It is well-known that $|\Phi_G\rangle = V \mathrm{H}^{\otimes n} |0^n\rangle$ where V is a product of CZ gates and H is the HADAMARD gate. More specifically, V applies CZ between qubits i and j if and only if vertex i neighbours j in G. Van den Nest, Dehaene, and De Moor [42] tell us that *any* stabiliser state can be transformed to a graph state by a product of single-qubit Clifford gates and classical post-processing. It is therefore clear that we can transform any S_{start} to S_{end} via S_{graph} using at most $n(n-1)/2$ two-qubit (CZ) gates, as this is the maximum number of edges on an n-vertex graph. The interesting question is whether we can do better by exploiting the potential low rank $k \le n$ of S_{start}.

Our answer is in the affirmative and we now present an explicit and efficient algorithm that constructs U with at most $u_{\mathrm{cz}}(k, n) = kn - k(k+1)/2$ two-qubit gates.

Figure 1: Reductions used in our CZ-construction.

In Fig. 1, we illustrate the sequence of reductions that allow us to reach S_{graph}, and so S_{end}, from S_{start}. We now describe the salient aspects of each step.

1. $S_{\mathrm{start}} \to S_1$. Following Ref. [43, Lemma 6], we can apply HADAMARD gates so that B has rank k. By classical row-swaps (relabelling of qubits), we can ensure that the first k rows of B have full-rank.

2. $S_1 \to S_2$. Since the upper $k \times k$ submatrix of B has full-rank, column operations corresponding to classical post-processing can reduce it to I_k.

3. $S_2 \to S_3$. Additional columns, corresponding to further operators, are appended. We can directly verify that the extension to S_3 is valid by Eq. (28). Clearly S_3 has full column-rank n. The sparsity of our chosen extension shall play a crucial role in the reduced two-qubit gate size of U when $k < n$ in both the CZ- and CNOT-constructions.

4. $S_3 \to S_4 = S_{\mathrm{graph}}$. Column operations can eliminate F, then PHASE gates can ensure E has ze-

ros on its diagonal. Importantly, S_4 represents a graph state S_{graph}.

5. $S_4 \to S_{\text{end}}$. PHASE and CZ gates can implement this final reduction as discussed above. The maximum number of CZ gates required to map S_4 to S_{end} equals the maximum number of off-diagonal 1s in the upper half of S_4. When $n = k$, this is $n(n-1)/2 = O(n^2)$. When $k \neq n$, this is $w_{cz}(k,n) = kn - k(k+1)/2$ due to sparsity of the upper half of S_4 which traces back to the structure of the additional operators appended in step $S_3 \to S_4$.

Note that in step $S_4 \to S_{\text{end}}$, we can first try to reduce the upper half of S_4 by single-qubit gates before applying CZ. One way to do this is to reduce the number of edges in the graph whose adjacency matrix is specified by the upper half of S_4 by the so-called "local complementation" operation [42, 56, 57]. This corresponds precisely to reducing the number of CZ gates in our CZ-construction.

3.2 CNOT-construction

Figure 2: Reductions used in our CNOT-construction starting at S_4 of our CZ-construction.

We start from S_4 above, which we reached without using two-qubit gates. Now, instead of using one block of CZ gates, we reduce to S_{end} as shown in Fig. 2, using three blocks of CNOT gates.

5. $S_4 \to S_5$. Note that E must be symmetric by the commutativity condition given in Eq. (28). Then, following Ref. [43, Lemma 7], we can eliminate E using single-qubit and $O(k^2/\log k)$ CNOT gates. This is accomplished by noting that any symmetric binary E can be Cholesky decomposed as $E = \Lambda + M^{\mathsf{T}}M$, with Λ diagonal and M invertible.

6. $S_5 \to S_6$. Reduce M to I_k by column operations, then add 1s on the top diagonal by phase gates.

7. $S_6 \to S_7$. Now, the upper $n \times n$ matrix can be block-Cholesky decomposed as

$$\begin{pmatrix} I_k & D^{\mathsf{T}} \\ D & I_{n-k} \end{pmatrix} = M_1^{\mathsf{T}} D_1 M_1, \qquad (32)$$

where

$$M_1 := \begin{pmatrix} I_k & 0 \\ D & I_{n-k} \end{pmatrix}, \qquad (33)$$

$$D_1 := \begin{pmatrix} I_k - D^{\mathsf{T}}D & 0 \\ 0 & I_{n-k} \end{pmatrix}. \qquad (34)$$

Next, we apply CNOT gates corresponding to M_1. The number of CNOT gates required here equals the number of row operations required to reduce M_1 to I_n. We find this is at most $u_{\text{cnot}}(k,n) = O(kn/\log k)$ via arguments of Patel, Markov, and Hayes [44]. The proof is given in Appendix E.

8. $S_7 \to S_8$. Multiply by M_1^{-1} on the right.

9. $S_8 \to S_{\text{end}}$. $I_k - D^{\mathsf{T}}D$ is a $k \times k$ symmetric matrix and so can be again eliminated via the Cholesky decomposition using $O(k^2/\log k)$ CNOT gates.

Note that in the three steps $S_4 \to S_5$, $S_6 \to S_7$, and $S_8 \to S_{\text{end}}$, we have used blocks of CNOT gates. The method we used to synthesise these blocks is size-optimal [44, Lemma 1], but we could have alternatively used methods in Ref. [58], that built on Ref. [59], to achieve optimal space-depth tradeoff, where space refers to extra ancilla qubits.

To end our discussion of constructing rotation circuits, we briefly mention a third, ancilla-based construction with two-qubit gate size at most kn. This construction is well-known in the context of syndrome measurement [55, 60] in quantum error correction but does not seem to have been mentioned in the context of measuring a Pauli decomposition of an operator, as in VQE. To measure k commuting Paulis $\{P_i\}_{i=1}^k$, this "ancilla-construction" uses k ancilla and involves k consecutive blocks of generalised-CNOT gates, each targeted at a different ancilla. The controls in block $b \leq k$ are activated or deactivated by the $+1$ or -1 eigenstates of the single-qubit Paulis forming P_b [*]. k single-qubit measurements are performed on the ancilla at the end of each block to exactly give measurements of P_i. Unfortunately, this construction requires k extra ancilla qubits (or else a single extra ancilla qubit that needs to be measured and reset k times) and has worse worst-case two-qubit gate size than both of our constructions.

---

[*] So Z corresponds to standard control on $\{|0\rangle, |1\rangle\}$, X corresponds to control on $\{|+\rangle, |-\rangle\}$, Y corresponds to control on $\{|+i\rangle, |-i\rangle\}$, and I corresponds to not having a generalised-CNOT. These generalised-CNOT gates can be implemented by the standard CNOT conjugated by H in case of X, or SH in case of Y.

4 Application to VQE

In this section, we present numerical results of the collecting method discussed in Sec. 2, alongside the CZ-construction of Sec. 3.1 to construct the rotation circuits for given commuting collections. In particular, we have applied our methods to the Hamiltonians of simple molecules so as to demonstrate their use in the context of VQE. The full results are given in Table 4 of Appendix G, with a subset shown in Table 2. In all cases, we used OpenFermion [61] to obtain Hamiltonians in the STO-3G basis, at approximately the equilibrium geometry of the molecules, with the symmetry conserving Bravyi-Kitaev transformation [62, 63]. In order to reduce the number of two-qubit gates required, we considered qubits on which all operators in a collection locally commute separately – a one-qubit rotation per locally commuting qubit is all that is required to do so.

In Fig. 3(a), we plot the average collection size against the number of qubits, n, for the molecular Hamiltonians. We can see that the average collection size increases with increasing n, and the increase does not appear to be slowing down. We therefore conclude that our sorting method works well on systems up to at least size $n = 38$, and looks likely to work for larger systems. However, the key advantage of assembling a Hamiltonian into collections of mutually commuting operators is a reduction in the number of measurements required to obtain an energy expectation to a certain level of accuracy, and collection size alone does not directly quantify this reduction. For a given Hamiltonian and quantum state, the reduction is instead given by R, as in Eq. (10).

We therefore calculated the value of R for 100 different quantum states, generated using 100 random sets of ansatz parameters with a hardware efficient ansatz of depth 1, for the nine smallest molecular systems. We show the mean, minimum and maximum values for each molecule. In practice, the value of R can at best be obtained approximately by making measurements on the quantum computer and so cannot be used to determine the expected advantage of a particular arrangement a priori. The metric $\hat{R}$, given by Eq. (20), on the other hand, depends only on the coefficients of the terms in the Hamiltonian. From Table 2, we can see that $\hat{R}$ closely approximates the average of R over many ansatz parameters, for the ansatz we have considered, but can be calculated analytically without the need for simulations. Further investigation of the relationship between $\hat{R}$ and R will be useful if considering a different ansatz. In Fig. 3(b), we show $\hat{R}$ as a function of the number of qubits for our full selection of molecules. We can see that it is highly molecule dependent, with systems of similar size having very different values.

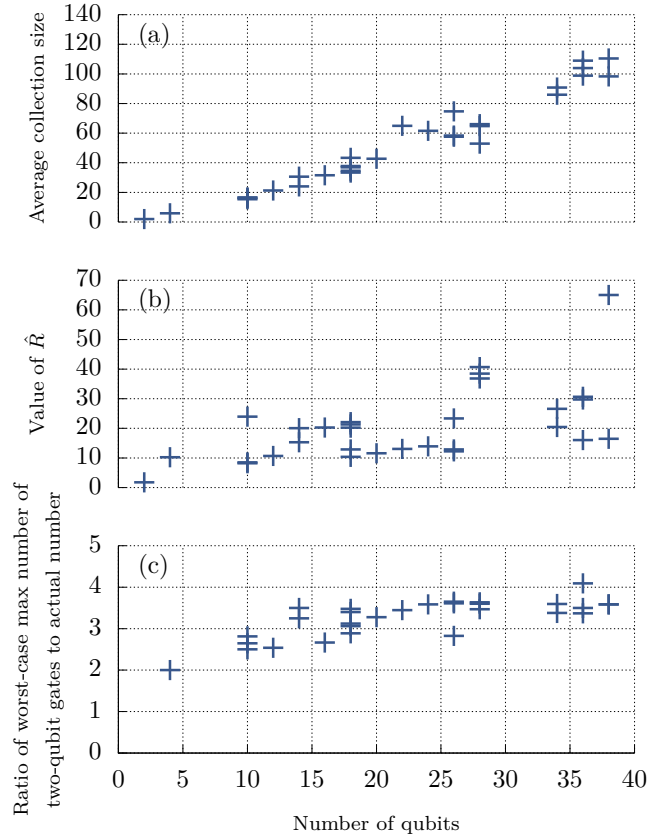The reduction in the number of measurements required comes at the cost of applying additional quan-

Figure 3: The results of numerical simulations discussed in the text. We show (a) the average collection size, (b) the value of $\hat{R}$, given by Eq. (20), and (c) the ratio of the worst-case maximum number of two-qubit gates in a single rotation circuit to the actual number as a function of the number of qubits for a range of simple molecules. A portion of the data is shown in Table 2, and the full data is shown in Table 4.

tum gates before the qubits are measured, the most costly of which are two-qubit gates. For the CZ-construction, we demonstrated in Sec. 3.1 that the maximum number of additional two-qubit gates required for a collection with k independent terms is $nk - k(k + 1)/2$. We would like to know, in practice, how many additional two-qubit gates are required at a maximum, as this is the quantum resource that is most limiting. Assuming for a given Hamiltonian that at least one collection has rank n, obtaining a measurement of all terms in a Hamiltonian on n qubits may therefore require applying an additional $n(n - 1)/2$ gates in a single circuit. However, for the molecules we have considered, we find that the largest number of two-qubit gates required is in fact far lower than this, typically by a factor of about 3.5, as can be seen in Fig. 3(c).

Given the close relationship between the average value of R and the value of $\hat{R}$, we propose using $\hat{R}$ as a metric for the quality of a collecting method and compare different methods of collecting the operators with this metric in mind. The results are shown in Table 3, along with the number of collections of op-

Molecule	qubits n	Paulis t	Arrangement			Ratios $R, \hat{R}$				Rotation Circuit $2q$-size		
			N	$\overline{m_i}$	$\overline{k_i}$	R min	R mean	R max	$\hat{R}$	theory max	true max	mean
H_2	2	4	2	2.00	1.50	1.09	1.93	4.60	1.76	0	0	0
H_3^+	4	59	10	5.90	3.50	3.76	11.92	33.04	10.25	6	3	0.80
LiH	10	630	41	15.37	6.85	19.60	24.91	34.74	23.97	45	18	5.29
OH^-	10	630	38	16.58	7.29	6.32	8.90	12.86	8.51	45	17	5.63
HF	10	630	39	16.15	6.97	6.07	8.57	12.27	8.21	45	16	5.74
H_2O	12	1085	51	21.27	9.04	7.68	11.27	16.96	10.67	66	26	7.37
BH_3	14	1584	66	24.00	10.36	17.21	20.93	32.13	20.05	91	26	9.56
NH_3	14	3608	118	30.58	11.34	12.65	15.96	26.93	15.31	91	28	10.26
CH_4	16	3887	123	31.60	13.39	16.96	21.63	29.33	20.27	120	45	16.75

Table 2: A reduced set of results of the numerical simulations discussed in the main text, and shown in full in Table 4. For each molecule, we show a number of results related to the collecting of Hamiltonian terms, how the arrangement reduces the number of measurements required using the metrics R and $\hat{R}$, and the number of two-qubit gates in the resulting rotation circuits. Note that the mean value of R and $\hat{R}$ are very similar.

Molecule	Largest First		Connected Sequential d.f.s.		Independent Set		DSATUR		Sorted Insertion	
	N	$\hat{R}$	N	$\hat{R}$	N	$\hat{R}$	N	$\hat{R}$	N	$\hat{R}$
H_2	2	**1.76**	2	**1.76**	2	**1.76**	2	**1.76**	2	**1.76**
H_3^+	10	4.86	10	10.25	10	**10.30**	9	4.10	10	10.25
LiH	39	23.87	45	23.33	30	5.72	29	10.47	41	**23.97**
OH^-	40	8.27	41	8.41	21	3.00	28	3.23	37	**8.51**
HF	38	8.05	41	8.07	21	2.80	28	3.23	38	**8.21**
H_2O	57	2.98	55	**10.66**	42	3.87	51	3.18	51	**10.66**
BH_3	66	4.80	85	18.70	60	7.85	72	4.11	68	**20.05**
NH_3	124	6.50	174	13.97	126	4.03	137	2.92	117	**15.31**
CH_4	122	5.84	176	18.93	114	9.88	110	4.90	125	**20.27**
O_2	62	13.62	85	19.95	42	6.79	52	7.91	67	**20.23**
N_2	62	15.00	86	21.15	39	8.37	49	5.80	78	**22.10**
CO	124	20.70	155	20.67	89	6.03	106	4.55	128	**21.31**
HCl	117	2.16	141	10.29	98	3.52	104	2.04	123	**10.36**
NaH	121	8.78	181	12.40	149	3.44	145	3.65	135	**12.90**
H_2S	122	8.81	180	**12.45**	147	3.80	145	3.66	147	11.60

Table 3: Comparison of the arrangements produced by the greedy colouring algorithms "Largest First", "Connected Sequential d.f.s." (depth first search), "Independent Set" and "DSATUR" as implemented by the Python package NetworkX [64] with our method Sorted Insertion. For each method, the number of collections produced, N, and the metric $\hat{R}$ given by Eq. (20), are presented. The best or joint best methods are highlighted in bold for each molecule.

erators, N, that each method produces. Out of the methods, "Independent Set" was best at approximating the minimum clique cover – it found the cover with the fewest cliques in all but one case. However, the minimum clique cover does not necessarily result in the fewest measurements, with "Independent Set" only performing best once with respect to $\hat{R}$. Overall, our Sorted Insertion was best at maximising $\hat{R}$, performing best or joint best in all but two cases.

5 Conclusion

We have addressed two problems related to the efficient measurement of Pauli operators on a quantum computer in the presence of finite sampling error. The first is how to assemble a set of Paulis into collections in which they mutually commute, and the second is how to construct rotation circuits that enable mutu-

ally commuting Paulis to be measured simultaneously.

For the first problem, we contribute two natural metrics, R and $\hat{R}$, that justifiably measure the effectiveness of an arrangement, followed by a collecting strategy motivated by $\hat{R}$ that we call Sorted Insertion. For the second problem, we contribute two rotation circuit constructions, CZ and CNOT. The CZ-construction uses a maximum of $u_{cz}(k,n) = kn - k(k+1)/2$ two-qubit gates while the CNOT-construction uses a maximum of $u_{cnot}(k,n) = O(kn/\log k)$.

We have applied our theoretical work to the task of estimating energies of molecules in the context of VQE. Comparison to other collecting methods shows that while Sorted Insertion does not normally result in the smallest number of collections, it nearly always results in the best value of $\hat{R}$. We also find that, for the CZ-construction, the largest number of two-qubit gates required is typically less than the theoretical worst-case by a factor of about 3.5.

6 Acknowledgement

References

[1] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5(1):4213, 2014. DOI: 10.1038/ncomms5213.

[2] John Preskill. Quantum computing in the NISQ era and beyond. *Quantum*, 2:79, August 2018. ISSN 2521-327X. DOI: 10.22331/q-2018-08-06-79.

[3] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, February 2016. DOI: 10.1088/1367-2630/18/2/023023.

[4] P. J. J. O'Malley, R. Babbush, I. D. Kivlichan, J. Romero, J. R. McClean, R. Barends, J. Kelly, P. Roushan, A. Tranter, N. Ding, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. G. Fowler, E. Jeffrey, E. Lucero, A. Megrant, J. Y. Mutus, M. Neeley, C. Neill, C. Quintana, D. Sank, A. Vainsencher, J. Wenner, T. C. White, P. V. Coveney, P. J. Love, H. Neven, A. Aspuru-Guzik, and J. M. Martinis. Scalable quantum simulation of molecular energies. *Phys. Rev. X*, 6:031007, July 2016. DOI: 10.1103/PhysRevX.6.031007.

[5] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M. Chow, and Jay M. Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, 2017. DOI: 10.1038/nature23879.

[6] Sam McArdle, Suguru Endo, Alán Aspuru-Guzik, Simon C. Benjamin, and Xiao Yuan. Quantum computational chemistry. *Rev. Mod. Phys.*, 92:015003, Mar 2020. DOI: 10.1103/RevModPhys.92.015003.

[7] Ilya G. Ryabinkin, Scott N. Genin, and Artur F. Izmaylov. Constrained variational quantum eigensolver: Quantum computer search engine in the Fock space. *Journal of Chemical Theory and Computation*, 15(1):249–255, 01 2019. DOI: 10.1021/acs.jctc.8b00943.

[8] Jonathan Romero, Ryan Babbush, Jarrod R McClean, Cornelius Hempel, Peter J Love, and Alán Aspuru-Guzik. Strategies for quantum computing molecular energies using the unitary coupled cluster ansatz. *Quantum Science and Technology*, 4(1):014008, Oct 2018. DOI: 10.1088/2058-9565/aad3e4.

[9] Daochen Wang, Oscar Higgott, and Stephen Brierley. Accelerated variational quantum eigensolver. *Phys. Rev. Lett.*, 122:140504, Apr 2019. DOI: 10.1103/PhysRevLett.122.140504.

[10] Jarrod R. McClean, Mollie E. Kimchi-Schwartz, Jonathan Carter, and Wibe A. de Jong. Hybrid quantum-classical hierarchy for mitigation of decoherence and determination of excited states. *Phys. Rev. A*, 95:042308, Apr 2017. DOI: 10.1103/PhysRevA.95.042308.

[11] Raffaele Santagati, Jianwei Wang, Antonio A. Gentile, Stefano Paesani, Nathan Wiebe, Jarrod R. McClean, Sam Morley-Short, Peter J. Shadbolt, Damien Bonneau, Joshua W. Silverstone, David P. Tew, Xiaoqi Zhou, Jeremy L. O'Brien, and Mark G. Thompson. Witnessing eigenstates for quantum simulation of Hamiltonian spectra. *Science Advances*, 4(1), 2018. DOI: 10.1126/sciadv.aap9646.

[12] J. I. Colless, V. V. Ramasesh, D. Dahlen, M. S. Blok, M. E. Kimchi-Schwartz, J. R. McClean, J. Carter, W. A. de Jong, and I. Siddiqi. Computation of molecular spectra on a quantum processor with an error-resilient algorithm. *Phys. Rev. X*, 8:011021, Feb 2018. DOI: 10.1103/PhysRevX.8.011021.

[13] Kentaro Heya, Ken M Nakanishi, Kosuke Mitarai, and Keisuke Fujii. Subspace variational quantum simulator. *arXiv e-prints*, Apr 2019.

[14] Tyson Jones, Suguru Endo, Sam McArdle, Xiao Yuan, and Simon C. Benjamin. Variational quantum algorithms for discovering Hamiltonian spectra. *Phys. Rev. A*, 99:062304, Jun 2019. DOI: 10.1103/PhysRevA.99.062304.

[15] Oscar Higgott, Daochen Wang, and Stephen Brierley. Variational quantum computation of excited states. *Quantum*, 3:156, July 2019. ISSN 2521-327X. DOI: 10.22331/q-2019-07-01-156.

[16] Vladyslav Verteletskyi, Tzu-Ching Yen, and Artur F. Izmaylov. Measurement optimization in the variational quantum eigensolver using a minimum clique cover. *The Journal of Chemical Physics*, 152(12):124114, 2020. DOI: 10.1063/1.5141458.

[17] Andrew Jena, Scott Genin, and Michele Mosca. Pauli partitioning with respect to gate sets. *arXiv e-prints*, July 2019.

[18] Tzu-Ching Yen, Vladyslav Verteletskyi, and Artur F. Izmaylov. Measuring all compatible operators in one series of single-qubit measurements using unitary transformations. *Journal of Chemical Theory and Computation*, 16(4):2400–2409, 04 2020. DOI: 10.1021/acs.jctc.0c00008.

[19] William J. Huggins, Jarrod McClean, Nicholas Rubin, Zhang Jiang, Nathan Wiebe, K. Birgitta Whaley, and Ryan Babbush. Efficient and noise resilient measurements for quantum chemistry on near-term quantum computers. *arXiv e-prints*, July 2019.

[20] Pranav Gokhale, Olivia Angiuli, Yongshan Ding, Kaiwen Gui, Teague Tomesh, Martin Suchara, Margaret Martonosi, and Frederic T. Chong. Minimizing state preparations in variational quantum eigensolver by partitioning into commuting families. *arXiv e-prints*, July 2019.

[21] Andrew Zhao, Andrew Tranter, William M. Kirby, Shu Fay Ung, Akimasa Miyake, and Peter J. Love. Measurement reduction in variational quantum algorithms. *Phys. Rev. A*, 101:062322, Jun 2020. DOI: 10.1103/PhysRevA.101.062322.

[22] P. Gokhale, O. Angiuli, Y. Ding, K. Gui, T. Tomesh, M. Suchara, M. Martonosi, and F. T. Chong. $O(N^3)$ measurement cost for variational quantum eigensolver on molecular Hamiltonians. *IEEE Transactions on Quantum Engineering*, 1: 1–24, 2020. DOI: 10.1109/TQE.2020.3035814.

[23] Artur F. Izmaylov, Tzu-Ching Yen, Robert A. Lang, and Vladyslav Verteletskyi. Unitary partitioning approach to the measurement problem in the variational quantum eigensolver method. *Journal of Chemical Theory and Computation*, 16(1):190–195, 01 2020. DOI: 10.1021/acs.jctc.9b00791.

[24] Artur F. Izmaylov, Tzu-Ching Yen, and Ilya G. Ryabinkin. Revising the measurement process in the variational quantum eigensolver: is it possible to reduce the number of separately measured operators? *Chem. Sci.*, 10:3746–3755, 2019. DOI: 10.1039/C8SC05592K.

[25] Dave Wecker, Matthew B. Hastings, and Matthias Troyer. Progress towards practical quantum variational algorithms. *Phys. Rev. A*, 92:042303, Oct 2015. DOI: 10.1103/PhysRevA.92.042303.

[26] Nicholas C Rubin, Ryan Babbush, and Jarrod McClean. Application of fermionic marginal constraints to hybrid quantum algorithms. *New Journal of Physics*, 20(5):053020, May 2018. DOI: 10.1088/1367-2630/aab919.

[27] Alexandre Blais, Jay Gambetta, A. Wallraff, D. I. Schuster, S. M. Girvin, M. H. Devoret, and R. J. Schoelkopf. Quantum-information processing with circuit quantum electrodynamics. *Phys. Rev. A*, 75:032329, Mar 2007. DOI: 10.1103/PhysRevA.75.032329.

[28] Anthony Laing, Alberto Peruzzo, Alberto Politi, Maria Rodas Verde, Matthaeus Halder, Timothy C. Ralph, Mark G. Thompson, and Jeremy L. O'Brien. High-fidelity operation of quantum photonic circuits. *Applied Physics Letters*, 97(21): 211109, 2010. DOI: 10.1063/1.3497087.

[29] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, C. Neill, P. O'Malley, P. Roushan, A. Vainsencher, J. Wenner, A. N. Korotkov, A. N. Cleland, and John M. Martinis. Superconducting quantum circuits at the surface code threshold for fault tolerance. *Nature*, 508(7497):500–503, 2014. DOI: 10.1038/nature13171.

[30] C. J. Ballance, T. P. Harty, N. M. Linke, M. A. Sepiol, and D. M. Lucas. High-fidelity quantum logic gates using trapped-ion hyperfine qubits. *Phys. Rev. Lett.*, 117:060504, Aug 2016. DOI: 10.1103/PhysRevLett.117.060504.

[31] Joseph L. Allen, Robert Kosut, Jaewoo Joo, Peter Leek, and Eran Ginossar. Optimal control of two qubits via a single cavity drive in circuit quantum electrodynamics. *Phys. Rev. A*, 95:042325, Apr 2017. DOI: 10.1103/PhysRevA.95.042325.

[32] Norbert M. Linke, Dmitri Maslov, Martin Roetteler, Shantanu Debnath, Caroline Figgatt, Kevin A. Landsman, Kenneth Wright, and Christopher Monroe. Experimental comparison of two quantum computing architectures. *Proceedings of the National Academy of Sciences*, 114(13):3305–3310, 2017. ISSN 0027-8424. DOI: 10.1073/pnas.1618020114.

[33] G Wendin. Quantum information processing with superconducting circuits: a review. *Reports on Progress in Physics*, 80(10):106001, sep 2017. DOI: 10.1088/1361-6633/aa7e1a.

[34] Matthew Reagor, Christopher B. Osborn, Nikolas Tezak, Alexa Staley, Guenevere Prawiroatmodjo, Michael Scheer, Nasser Alidoust, Eyob A. Sete, Nicolas Didier, Marcus P. da Silva, Ezer Acala, Joel Angeles, Andrew Bestwick, Maxwell Block, Benjamin Bloom, Adam Bradley, Catvu Bui, Shane Caldwell, Lauren Capelluto, Rick Chilcott, Jeff Cordova, Genya Crossman, Michael Curtis, Saniya Deshpande, Tristan El Bouayadi, Daniel Girshovich, Sabrina Hong, Alex Hudson, Peter Karalekas, Kat Kuang, Michael Lenihan, Riccardo Manenti, Thomas Manning, Jayss Marshall, Yuvraj Mohan, William O'Brien, Johannes Otterbach, Alexander Papageorge, Jean-Philip Paquette, Michael Pelstring, Anthony Polloreno, Vijay Rawat, Colm A. Ryan, Russ Renzas, Nick Rubin, Damon Russel, Michael Rust, Diego Scarabelli, Michael Selvanayagam, Rodney Sinclair, Robert Smith, Mark Suska, Ting-Wai To, Mehrnoosh Vahidpour, Nagesh Vodrahalli, Tyler Whyland, Kamal Yadav, William Zeng, and Chad T. Rigetti. Demonstration of universal parametric

Accepted in ⟨ ⟩uantum 2020-12-27, click title to verify. Published under CC-BY 4.0.

11

entangling gates on a multi-qubit lattice. *Science Advances*, 4(2), 2018. DOI: 10.1126/sciadv.aao3603.

[35] V. M. Schäfer, C. J. Ballance, K. Thirumalai, L. J. Stephenson, T. G. Ballance, A. M. Steane, and D. M. Lucas. Fast quantum logic gates with trapped-ion qubits. *Nature*, 555(7694):75–78, 2018. DOI: 10.1038/nature25737.

[36] A. E. Webb, S. C. Webster, S. Collingbourne, D. Bretaud, A. M. Lawrence, S. Weidt, F. Mintert, and W. K. Hensinger. Resilient entangling gates for trapped ions. *Phys. Rev. Lett.*, 121:180501, Nov 2018. DOI: 10.1103/PhysRevLett.121.180501.

[37] Harry Levine, Alexander Keesling, Ahmed Omran, Hannes Bernien, Sylvain Schwartz, Alexander S. Zibrov, Manuel Endres, Markus Greiner, Vladan Vuletić, and Mikhail D. Lukin. High-fidelity control and entanglement of Rydberg-atom qubits. *Phys. Rev. Lett.*, 121:123603, Sep 2018. DOI: 10.1103/PhysRevLett.121.123603.

[38] Y. He, S. K. Gorman, D. Keith, L. Kranz, J. G. Keizer, and M. Y. Simmons. A two-qubit gate between phosphorus donor electrons in silicon. *Nature*, 571(7765):371–375, 2019. DOI: 10.1038/s41586-019-1381-2.

[39] W. Huang, C. H. Yang, K. W. Chan, T. Tanttu, B. Hensen, R. C. C. Leon, M. A. Fogarty, J. C. C. Hwang, F. E. Hudson, K. M. Itoh, A. Morello, A. Laucht, and A. S. Dzurak. Fidelity benchmarks for two-qubit gates in silicon. *Nature*, 569 (7757):532–536, 2019. DOI: 10.1038/s41586-019-1197-0.

[40] Reinhold Blumel, Nikodem Grzesiak, and Yunseong Nam. Power-optimal, stabilized entangling gate between trapped-ion qubits. *arXiv e-prints*, May 2019.

[41] Y. He, S. K. Gorman, D. Keith, L. Kranz, J. G. Keizer, and M. Y. Simmons. A two-qubit gate between phosphorus donor electrons in silicon. *Nature*, 571(7765):371–375, 2019. DOI: 10.1038/s41586-019-1381-2.

[42] Maarten Van den Nest, Jeroen Dehaene, and Bart De Moor. Graphical description of the action of local Clifford transformations on graph states. *Phys. Rev. A*, 69:022316, Feb 2004. DOI: 10.1103/PhysRevA.69.022316.

[43] Scott Aaronson and Daniel Gottesman. Improved simulation of stabilizer circuits. *Phys. Rev. A*, 70:052328, Nov 2004. DOI: 10.1103/PhysRevA.70.052328.

[44] Ketan N. Patel, Igor L. Markov, and John P. Hayes. Optimal synthesis of linear reversible circuits. *Quantum Info. Comput.*, 8(3):282–294, March 2008. ISSN 1533-7146. DOI: 10.26421/QIC8.3-4.

[45] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quan-

tum system from very few measurements. *Nature Physics*, 16(10):1050–1057, Jun 2020. ISSN 1745-2481. DOI: 10.1038/s41567-020-0932-7.

[46] Charles Hadfield, Sergey Bravyi, Rudy Raymond, and Antonio Mezzacapo. Measurements of quantum Hamiltonians with locally-biased classical shadows. *arXiv e-prints*, June 2020.

[47] Jonas M. Kübler, Andrew Arrasmith, Lukasz Cincio, and Patrick J. Coles. An adaptive optimizer for measurement-frugal variational algorithms. *Quantum*, 4:263, May 2020. ISSN 2521-327X. DOI: 10.22331/q-2020-05-11-263.

[48] Andrew Arrasmith, Lukasz Cincio, Rolando D. Somma, and Patrick J. Coles. Operator sampling for shot-frugal optimization in variational algorithms. *arXiv e-prints*, April 2020.

[49] J.B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York, 1994. ISBN 9780387972459.

[50] John Watrous. *The Theory of Quantum Information*. Cambridge University Press, 2018.

[51] Adrian Kosowski and Krzysztof Manuszewski. Classical coloring of graphs. In Marek Kubale, editor, *Graph Colorings*, chapter 1. American Mathematical Society, 2004. ISBN 978-0-8218-7942-9. DOI: 10.1090/conm/352.

[52] David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. In *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing (STOC)*, pages 681–690, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595931341. DOI: 10.1145/1132516.1132612.

[53] A. R. Calderbank, E. M. Rains, P. W. Shor, and N. J. A. Sloane. Quantum error correction and orthogonal geometry. *Phys. Rev. Lett.*, 78:405–408, Jan 1997. DOI: 10.1103/PhysRevLett.78.405.

[54] Daniel Gottesman. *Stabilizer codes and quantum error correction*. PhD thesis, California Institute of Technology, Jan 1997.

[55] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, New York, NY, USA, 2011. ISBN 978-1107002173.

[56] M. Hein, J. Eisert, and H. J. Briegel. Multiparty entanglement in graph states. *Phys. Rev. A*, 69:062311, Jun 2004. DOI: 10.1103/PhysRevA.69.062311.

[57] André Bouchet. Recognizing locally equivalent graphs. *Discrete Mathematics*, 114(1):75 – 86, 1993. ISSN 0012-365X. DOI: 10.1016/0012-365X(93)90357-Y.

[58] Jiaqing Jiang, Xiaoming Sun, Shang-Hua Teng, Bujiao Wu, Kewen Wu, and Jialin Zhang. Optimal space-depth trade-off of CNOT circuits in

quantum logic synthesis. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 213–229, 2020. DOI: 10.1137/1.9781611975994.13.

[59] Cristopher Moore and Martin Nilsson. Parallel quantum computation and quantum codes. *SIAM J. Comput.*, 31(3):799–815, March 2002. ISSN 0097-5397. DOI: 10.1137/S0097539799355053.

[60] Eric Dennis, Alexei Kitaev, Andrew Landahl, and John Preskill. Topological quantum memory. *Journal of Mathematical Physics*, 43(9): 4452–4505, 2002. DOI: 10.1063/1.1499754.

[61] Jarrod R McClean, Nicholas C Rubin, Kevin J Sung, Ian D Kivlichan, Xavier Bonet-Monroig, Yudong Cao, Chengyu Dai, E Schuyler Fried, Craig Gidney, Brendan Gimby, Pranav Gokhale, Thomas Häner, Tarini Hardikar, Vojtěch Havlíček, Oscar Higgott, Cupjin Huang, Josh Izaac, Zhang Jiang, Xinle Liu, Sam McArdle, Matthew Neeley, Thomas O'Brien, Bryan O'Gorman, Isil Ozfidan, Maxwell D Radin, Jhonathan Romero, Nicolas P D Sawaya, Bruno Senjean, Kanav Setia, Sukin Sim, Damian S Steiger, Mark Steudtner, Qiming Sun, Wei Sun, Daochen Wang, Fang Zhang, and Ryan Babbush. OpenFermion: the electronic structure package for quantum computers. *Quantum Science and Technology*, 5(3):034014, June 2020. DOI: 10.1088/2058-9565/ab8ebc.

[62] Sergey B. Bravyi and Alexei Yu. Kitaev. Fermionic quantum computation. *Annals of Physics*, 298(1):210 – 226, 2002. ISSN 0003-4916. DOI: 10.1006/aphy.2002.6254.

[63] Sergey Bravyi, Jay M. Gambetta, Antonio Mezzacapo, and Kristan Temme. Tapering off qubits to simulate fermionic Hamiltonians. *arXiv e-prints*, Jan 2017.

[64] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, 2008.

[65] V.L. Arlazarov, E.A. Dinic, M.A. Kronod, and I.A. Faradez. On economical construction of the transitive closure of an oriented graph. *Soviet Mathematics Doklady*, pages 1209–10, 1970.

[66] Robert M. Parrish, Lori A. Burns, Daniel G. A. Smith, Andrew C. Simmonett, A. Eugene DePrince, Edward G. Hohenstein, Uğur Bozkaya, Alexander Yu. Sokolov, Roberto Di Remigio, Ryan M. Richard, Jérôme F. Gonthier, Andrew M. James, Harley R. McAlexander, Ashutosh Kumar, Masaaki Saitow, Xiao Wang, Benjamin P. Pritchard, Prakash Verma, Henry F. Schaefer, Konrad Patkowski, Rollin A. King, Edward F. Valeev, Francesco A. Evangelista, Justin M. Turney, T. Daniel Crawford, and C. David Sherrill. Psi4 1.1: An open-source electronic structure program emphasizing automation, advanced libraries, and interoperability. *Journal of Chemical Theory and Computation*, 13(7):3185–3197, 07 2017. DOI: 10.1021/acs.jctc.7b00174.

A Example to demonstrate that combining two collections into one never reduces R

Consider the example in Refs [3, 20] where we consider measuring the energy, given by the Hamiltonian

$$H = -XX - YY + ZZ + IZ + ZI, \qquad (35)$$

of the state $|\psi\rangle = |01\rangle$. The covariance matrix is

$$C = \begin{pmatrix} 1 & 1 & 0 & \mathbf{0} & \mathbf{0} \\ 1 & 1 & 0 & \mathbf{0} & \mathbf{0} \\ 0 & 0 & 0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 & 0 \end{pmatrix}. \qquad (36)$$

The covariance between the non-commuting operators in the upper right and lower left blocks is not defined. We have set them to equal zero for convenience (highlighted in bold).

First, we consider collecting the Paulis into

$$\{-XX, -YY, ZZ\}, \; \{IZ, ZI\}. \qquad (37)$$

For these collections of Paulis, the coefficient vectors are

$$\mathbf{a} = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \; \mathbf{c} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}. \qquad (38)$$

The number of measurements to achieve an accuracy of ϵ is

$$\begin{aligned} M_g &= \frac{1}{\epsilon^2} \left(\sqrt{\mathbf{a}^\mathsf{T} C \mathbf{a}} + \sqrt{\mathbf{c}^\mathsf{T} C \mathbf{c}} \right)^2 \\ &= \frac{1}{\epsilon^2} \left(\sqrt{4} + \sqrt{0} \right)^2 \qquad (39) \\ &= \frac{4}{\epsilon^2}. \end{aligned}$$

Now, let us consider breaking up the first collection into

$$\{-XX\}, \; \{-YY, ZZ\}. \qquad (40)$$

In this case, the coefficient vectors are

$$\mathbf{a} = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \; \mathbf{b} = \begin{pmatrix} 0 \\ -1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \; \mathbf{c} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}. \qquad (41)$$

The number of measurements required to attain an accuracy ϵ is therefore

$$\begin{aligned} M_g &= \frac{1}{\epsilon^2} \left(\sqrt{\mathbf{a}^\mathsf{T} C \mathbf{a}} + \sqrt{\mathbf{b}^\mathsf{T} C \mathbf{b}} + \sqrt{\mathbf{c}^\mathsf{T} C \mathbf{c}} \right)^2 \\ &= \frac{1}{\epsilon^2} \left(\sqrt{1} + \sqrt{1} + \sqrt{0} \right)^2 \qquad (42) \\ &= \frac{4}{\epsilon^2}. \end{aligned}$$

Therefore, under the optimal measurement strategy it is not preferable to break the $\{-XX, -YY, ZZ\}$ collection into $\{-XX\}$ and $\{-YY, ZZ\}$. In this specific example we have equality because for $\alpha = -\beta$ we have $\langle \alpha\mathbf{a} + \beta\mathbf{b}, \alpha\mathbf{a} + \beta\mathbf{b} \rangle = 0$.

B Derivation of $\hat{R}$ formula

Claim 1. *For R as defined in Eq. (10), if all variances and covariances are replaced with their expectation value over uniform spherical distribution, we obtain a new metric, $\hat{R}$, given by*

$$\hat{R} = \left(\frac{\sum_{i=1}^{N} \sum_{j=1}^{N_i} |a_{ij}|}{\sum_{i=1}^{N} \sqrt{\sum_{j=1}^{N_i} a_{ij}^2}} \right)^2. \qquad (43)$$

Proof. The variance of a single Pauli operator is

$$\mathrm{Var}[P_i] := 1 - \langle P_i \rangle^2. \qquad (44)$$

The expectation of this variance for all $P_i \neq I$ is

$$\begin{aligned} \mathbb{E}[\mathrm{Var} P_i] &= 1 - \mathbb{E}[\langle P_i \rangle^2] \\ &= 1 - \int [\langle \psi | P_i | \psi \rangle]^2 \, \mathrm{d}\psi \qquad (45) \\ &= 1 - \alpha_n, \end{aligned}$$

where $\alpha_n := 1/(2^n + 1)$, with n the number of qubits, is independent of P_i [50, Exercise 7.3]. Trivially, $\mathrm{Var}[I] = 0$. In addition, it was shown in Ref. [20] that

$$\mathbb{E}[\mathrm{Cov}[P_i, P_j]] = 0, \qquad (46)$$

for all $P_i \neq P_j$. Substitution of these results yields Eq. (43). $\qquad \square$

C Binary representation

The Pauli subset $\mathcal{P}_n$ on n-qubits is a collection of 4^{n+1} elements defined by

$$\begin{aligned} \mathcal{P}_n = \{ i^k \, \sigma_1 \otimes \cdots \otimes \sigma_n \mid \sigma_i \in \{I, X, Y, Z\}, \\ k \in \{0, 1, 2, 3\} \}. \qquad (47) \end{aligned}$$

The binary representation, first introduced by Calderbank, Rains, Shor, and Sloane [53], is a representation of $\mathcal{P}_n$ as binary vectors. In this representation, Paulis differing only in phase i^k are represented in the same way.

Single-qubit Paulis are represented by 2-dimensional binary vectors, so that

$$
\begin{aligned}
\sigma_{00} &:= I \mapsto (0,0), \\
\sigma_{01} &:= X \mapsto (0,1), \\
\sigma_{10} &:= Z \mapsto (1,0), \\
\sigma_{11} &:= Y \mapsto (1,1).
\end{aligned}
\tag{48}
$$

An n-qubit Pauli

$$
\sigma_{u_1 v_1} \otimes \cdots \otimes \sigma_{u_n v_n}
\tag{49}
$$

is then represented by the $2n$-dimensional binary vector

$$
(u_1, \ldots, u_n, v_1, \ldots, v_n)^\mathsf{T}.
\tag{50}
$$

In this representation, two n-qubit Paulis with binary vectors a and b commute if and only if

$$
a^\mathsf{T} J_{2n} b = 0,
\tag{51}
$$

where J_{2n} denotes the $2n \times 2n$ matrix

$$
J_{2n} := \begin{pmatrix} 0 & I_n \\ I_n & 0 \end{pmatrix}.
\tag{52}
$$

Given a set $\mathcal{S}$ of m n-qubit Paulis, we can write down a corresponding $2n \times m$ binary matrix S where each column represents a Pauli. Then, from Eq. (51), we deduce that all Paulis in $\mathcal{S}$ mutually commute if and only if

$$
S^\mathsf{T} J_{2n} S = 0_m,
\tag{53}
$$

which recovers Eq. (28) in the main text. We say that the set $\mathcal{S}$ of Paulis is independent if the matrix S has rank m.

We shall often find it helpful to write S in terms of its upper half $S^{(Z)}$ and lower half $S^{(X)}$, separated by a horizontal line for visual-aid, i.e.,

$$
S = \left(\frac{S^{(Z)}}{S^{(X)}} \right).
\tag{54}
$$

The conjugation action of quantum gates on $\mathcal{S}$ can be represented as transformations to the matrix S. For example, we document the transformations on S that represent four common quantum gates. In the following, addition is mod 2 and p ranges over all columns $\{1, \ldots, m\}$.

- CZ on qubits i and j:
  $S_{ip} \leftarrow S_{ip} + S_{j+n,p}$,
  $S_{jp} \leftarrow S_{jp} + S_{i+n,p}$.

- CNOT on control-qubit i and target-qubit j:
  $S_{ip} \leftarrow S_{ip} + S_{jp}$,
  $S_{j+n,p} \leftarrow S_{j+n,p} + S_{i+n,p}$.

- HADAMARD (H) on qubit i:
  $S_{ip} \leftrightarrow S_{i+n,p}$.

- PHASE (P) on qubit i:
  $S_{ip} \leftarrow S_{ip} + S_{i+n,p}$.

These rules can be directly verified by conjugating X_i, X_j, Z_i, Z_j by the listed gates. They are also reproduced in Sec. 3 of the main text.

D CZ-construction example

We walk through our CZ-construction for a specific example. In this example, we would like to obtain measurements simultaneously of a collection $\mathcal{S}'_{\text{start}}$ of six four-qubit Paulis given by

$$
\begin{aligned}
P_1 &= Z_1 Z_2 Z_3 Z_4, \\
P_2 &= X_1 X_2 Y_3 Y_4, \\
P_3 &= Y_1 Y_2 X_3 X_4, \\
P_4 &= Y_2 X_3, \\
P_5 &= Y_1 X_4, \\
P_6 &= X_1 Z_2 Z_3 Y_4.
\end{aligned}
\tag{55}
$$

We can represent these Paulis in a matrix S'_{start} with

$$
S'_{\text{start}} = \left(\begin{array}{cccccc}
1 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 1 \\
\hline
0 & 1 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 1
\end{array} \right).
\tag{56}
$$

By Gaussian elimination, we find the reduced row echelon form of S'_{start} to be

$$
\left(\begin{array}{cccccc}
1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{array} \right).
\tag{57}
$$

The pivot columns are numbers 1, 2 and 4 which tells us that P_1, P_2 and P_4 are the three independent Paulis from which the remaining Paulis in $\mathcal{S}'_{\text{start}}$ can be constructed. Therefore, we can write $S'_{\text{start}} = S_{\text{start}} R_0^{-1}$,

where

$$S_{\text{start}} := \left(\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \hline 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{array}\right), \quad R_0^{-1} := \left(\begin{array}{cccccc} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{array}\right). \tag{58}$$

Note that the inverse on R_0^{-1} is purely notational. Now, the lower half $S_{\text{start}}^{(X)}$ of S_{start} has column echelon form

$$S_{\text{start}}^{(X)} = \left(\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{array}\right), \tag{59}$$

and so the first two rows are pivot rows. In order to give the lower half of S_{start} a rank of $k = 3$, we therefore apply a HADAMARD to the rows corresponding to qubits 3 and 4 so that

$$S_1 := Q_1 S_{\text{start}} = \left(\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ \hline 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{array}\right), \tag{60}$$

where

$$Q_1 := \left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right). \tag{61}$$

The lower half of S_1 now has rank 3, and performing Gaussian elimination on it, we find

$$S_2 := Q_1 S_{\text{start}} R_1 = \left(\begin{array}{ccc} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{array}\right), \tag{62}$$

where

$$R_1 = \left(\begin{array}{ccc} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{array}\right)^{-1} = \left(\begin{array}{ccc} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{array}\right). \tag{63}$$

We now extend S_2 to a rank $n = 4$ matrix by adding a column that corresponds to a fourth Pauli P_{ext}. In the main text, this is the crucial basis extension step from $S_2 \to S_3$ which might have seemed fortuitous. In fact, the extension was systematically obtained as follows.

To make our reasoning clearer, let us represent S_2 alternatively by the matrix

$$P(S_2) := \left(\begin{array}{cccc} Y & I & I & Z \\ I & Y & Z & I \\ Z & Z & X & X \end{array}\right), \tag{64}$$

where each row corresponds to a Pauli operator given by a column of S_2. Looking at the form of $P(S_2)$, we see that we can place X in the 4th qubit position of P_{ext} (and nowhere else) to ensure P_{ext} is independent of the other Paulis. Then we observe that the left 3-by-3 sub-matrix of $P(S_2)$ has X/Y on the diagonal and I/Z everywhere else. This means we can place I/Z in the other qubit positions of P_{ext} depending on whether the X in its 4th qubit position commutes with the 4th position terms of the other Paulis.

By this prescription, we find $P_{ext} = Z_1 I_2 I_3 X_4$. Therefore, S_2 is extended to

$$S_3 := \left(\begin{array}{cccc} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right), \tag{65}$$

and

$$Q_1 S_{\text{start}} R_1 = S_3 R_2^{-1}, \tag{66}$$

where

$$R_2^{-1} := \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{array}\right). \tag{67}$$

Note that the inverse on R_2^{-1} is also purely notational. The lower half of S_3 is full-rank and so we can take its inverse to find

$$R_3 = \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right)^{-1} = \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right) \tag{68}$$

and

$$S_3 R_3 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{69}$$

Finally, we apply PHASE to qubits 1 and 2 so that

$$S_4 = Q_2 S_3 R_3 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{70}$$

where

$$Q_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{71}$$

S_4 is of the form of a graph state and represents the following Paulis:

$$\begin{aligned} \tilde{P}_1 &= X_1\, I_2\, I_3\, Z_4, \\ \tilde{P}_2 &= I_1\, X_2\, Z_3\, I_4, \\ \tilde{P}_3 &= I_1\, Z_2\, X_3\, I_4, \\ \tilde{P}_4 &= Z_1\, I_2\, I_3\, X_4. \end{aligned} \tag{72}$$

We now have

$$S'_{\text{start}} = Q^{-1} S_4 R^{-1}, \tag{73}$$

where

$$Q^{-1} := Q_1^{-1} Q_2^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{74}$$
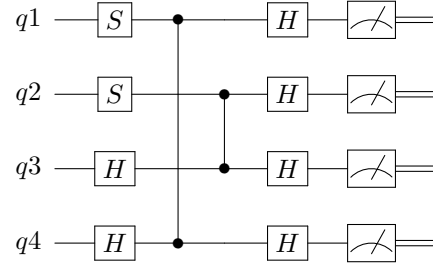

Figure 4: The rotation circuit U for the CZ-construction walk-through example.

and

$$R^{-1} := R_3^{-1} R_2^{-1} R_1^{-1} R_0^{-1} = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{75}$$

The rotation circuit is shown in Fig. 4. Using Q^{-1}, S_4 and R^{-1}, we can work out that the phases for the six original operators are $\begin{pmatrix} +1 & +1 & +1 & -1 & -1 & -1 \end{pmatrix}$. Therefore, we can construct measurements of the original Pauli strings as follows:

- P_1 from product of measurements of qubits 3 and 4,

- P_2 from product of measurements of qubits 1 to 4,

- P_3 from product of measurements of qubits 1 and 2,

- P_4 from the negative of measurement of qubit 2,

- P_5 from the negative of measurement of qubit 1,

- P_6 from the negative product of measurements of qubits 1, 3 and 4.

E Proof of $O(kn/\log k)$

We prove the following Claim 2 via arguments of Patel, Markov, and Hayes [44]. As acknowledged in Ref. [44], these arguments originate from the "Method of Four Russians" [65]. Note that row operations correspond to CNOT gates, as explained in detail in Ref. [44].

Claim 2. *Let M be a $n \times n$ matrix with block form*

$$\begin{pmatrix} I_k & A \\ 0 & I_{n-k} \end{pmatrix}, \tag{76}$$

where A is any $k \times (n-k)$ matrix. Then $O(kn/\log k)$ row operations suffice to reduce M to identity I_n.

Proof. Let m be a constant we later choose. Partition A into $l := (n-k)/m$ consecutive column-blocks A_i, each containing m columns.

Start at A_1 and eliminate any duplicate rows using at most k row operations. There then remain at most 2^m unique rows in A_1 which can be eliminated by at most $m2^{m-1}$ row operations that add rows from I_{n-k}. A_1 is now zero.

For each of $A_2, \ldots, A_l$ perform the same operation as was done to A_1. M then becomes

$$\begin{pmatrix} B & 0 \\ 0 & I_{n-k} \end{pmatrix}, \qquad (77)$$

where B is some $k \times k$ matrix that must be invertible. B can then be row-reduced to I_k using $O(k^2/\log k)$ by the result of Ref. [44].

The total number N of row operations is therefore

$$N = \left(k + m2^{m-1}\right) \frac{n-k}{m} + O\left(\frac{k^2}{\log k}\right). \qquad (78)$$

Choosing $m = \alpha \log k$, we find

$$N = \frac{k(n-k)}{\alpha \log k} + \frac{k^\alpha(n-k)}{2} + O\left(\frac{k^2}{\log k}\right), \qquad (79)$$

which is $O(kn/\log k)$ provided $\alpha < 1$. $\qquad \square$

F NP-hardness of maximising $\hat{R}$

In this Appendix, we show that maximising $\hat{R}$ is NP-hard by a reduction argument.

Let $G = (V, E)$ be a simple graph with vertices $V := \{1, 2, \ldots, n\}$ and edges E. Following Ref. [20, Appendix A], we then define the operator $O = \sum_{i=1}^n P_i$ on n-qubits, where $P_i = \otimes_{j=1}^n P_i^{(j)}$ is a tensor product of n single-qubit Paulis, so that

$$P_i^{(j)} := \begin{cases} Z & \text{if } i = j, \\ X & \text{if } j > i \text{ and } (i,j) \notin E, \\ I & \text{otherwise.} \end{cases} \qquad (80)$$

It can be easily seen [20, Appendix A] that a commuting collection of the P_is corresponds exactly to a clique in G.

For an arrangement $\mathcal{A}$ of O into N commuting collections with sizes $m_i \geq 1$, we consider the term appearing in the denominator of $\hat{R}$, and see that

$$D(\mathcal{A}) := \sum_{i=1}^N \sqrt{\sum_{j=1}^{m_i} |a_{ij}|^2} = \sum_{i=1}^N \sqrt{m_i}, \qquad (81)$$

where the second equality is because the coefficients of P_is in O are all 1s and we do not consider splitting these coefficients. We have

$$N \leq D(\mathcal{A}) \leq \sqrt{n} \sqrt{N} \leq \sqrt{n} \, N, \qquad (82)$$

where the second inequality follows from Cauchy-Schwarz and $\sum_{i=1}^N m_i = n$.

Now, let N_0 denote the minimum number of commuting collections corresponding to some arrangement $\mathcal{A}_0$ and let N_1 denote the number of commuting collections in an arrangement $\mathcal{A}_1$ that maximises $\hat{R}$. By definition, $N_0 \leq N_1$. Since maximising $\hat{R}$ is equivalent to minimising $D(\mathcal{A})$, we also have

$$N_1 \leq D(\mathcal{A}_1) \leq D(\mathcal{A}_0) \leq \sqrt{n} \, N_0, \qquad (83)$$

using Eq. (82) for the first and third inequalities. Hence, $D(\mathcal{A}_1)$ is a $\sqrt{n}$-approximation to N_0. But N_0 is also the solution to the MIN-CLIQUE-COVER problem on G by construction. So maximising $\hat{R}$ involves computing a $\sqrt{n}$-approximation to MIN-CLIQUE-COVER which is NP-hard by Ref. [52, Theorem 1.2].

In fact, Ref. [52, Theorem 1.2] gives the stronger result that computing a $n^{1-\epsilon}$ approximation to MIN-CLIQUE-COVER is NP-hard for any $\epsilon > 0$. This implies that it is NP-hard to compute a $n^{1/2-\epsilon}$ approximation to $D(\mathcal{A}_1)$ for any $\epsilon > 0$. In other words, it is NP-hard to compute a $1/n^{1/2-\epsilon}$ approximation to the maximum $\hat{R}$ for any $\epsilon > 0$.

G Full numerical results

In this Appendix, we present the full version of Table 2 in the main text.

Molecule	n qubits	t Paulis	Arrangement			Ratios R, $\hat{R}$				Rotation Circuit $2q$-size		
			N	$\overline{m_i}$	$\overline{k_i}$	R min	R mean	R max	$\hat{R}$	max theory	max true	mean
H_2	2	4	2	2.00	1.50	1.09	1.93	4.60	1.76	0	0	0
H_3^+	4	59	10	5.90	3.50	3.76	11.92	33.04	10.25	6	3	0.80
LiH	10	630	41	15.37	6.85	19.60	24.91	34.74	23.97	45	18	5.29
OH^-	10	630	38	16.58	7.29	6.32	8.90	12.86	8.51	45	17	5.63
HF	10	630	39	16.15	6.97	6.07	8.57	12.27	8.21	45	16	5.74
H_2O	12	1085	51	21.27	9.04	7.68	11.27	16.96	10.67	66	26	7.37
BH_3	14	1584	66	24.00	10.36	17.21	20.93	32.13	20.05	91	26	9.56
NH_3	14	3608	118	30.58	11.34	12.65	15.96	26.93	15.31	91	28	10.26
CH_4	16	3887	123	31.60	13.39	16.96	21.63	29.33	20.27	120	45	16.75
O_2	18	2238	67	33.40	13.48	-	-	-	20.23	153	44	21.57
N_2	18	2950	78	37.82	13.91	-	-	-	22.10	153	53	20.42
CO	18	4426	128	34.58	13.48	-	-	-	21.31	153	50	20.23
HCl	18	4538	123	36.89	13.87	-	-	-	10.36	153	49	20.35
NaH	18	5850	135	43.33	14.73	-	-	-	12.90	153	45	21.44
H_2S	20	6277	147	42.70	16.06	-	-	-	11.60	190	58	25.98
PH_3	22	19746	304	64.95	18.77	-	-	-	13.05	231	67	28.02
SiH_4	24	18713	304	61.56	20.98	-	-	-	13.94	276	77	36.03
NaF	26	16538	287	57.62	20.44	-	-	-	23.36	325	90	42.28
LiCl	26	17044	292	58.37	20.46	-	-	-	12.22	325	89	39.42
KH	26	24290	325	74.74	22.30	-	-	-	12.87	325	115	45.18
CO_2	28	11429	216	52.91	21.02	-	-	-	38.47	378	104	44.51
F_2O	28	20541	317	64.80	22.83	-	-	-	36.82	378	105	46.12
NO_2	28	20549	311	66.07	22.93	-	-	-	40.69	378	109	46.25
Cl_2	34	34334	378	90.83	28.09	-	-	-	26.58	561	156	73.24
NaCl	34	42826	498	86.00	28.54	-	-	-	20.46	561	166	74.34
SF_2	36	56025	567	98.81	31.96	-	-	-	30.65	630	180	78.67
HBr	36	62589	602	103.97	31.88	-	-	-	16.03	630	154	78.98
SO_2	36	75315	691	108.99	33.21	-	-	-	29.75	630	187	66.90
NO_3^-	38	61132	622	92.28	31.23	-	-	-	65.01	703	182	86.40
H_2Se	38	69684	631	110.43	33.91	-	-	-	16.49	703	196	86.88

Table 4: The full set of results of the numerical simulations discussed in the main text. The molecular geometry is approximately that of the equilibrium configuration. A number of results related to the collecting of Hamiltonian terms, how the arrangement reduces the number of measurements required, and the number of two-qubit gates in the resulting rotation circuits are shown. We used OpenFermion [61] and Psi4 [66] to obtain Hamiltonians in the STO-3G basis and under the symmetry conserving Bravyi-Kitaev transformation [62, 63]. Using our collecting method, SORTED INSERTION, the number of collections, N, the average number of terms per collection, $\overline{m_i}$, and the average rank of the collections, $\overline{k_i}$ are shown, for molecules with n qubits and t Pauli operators, excluding the identity, in the Hamiltonian sum. For the smallest nine molecules, we calculated the resulting ratio R, as given in Eq. (10), for 100 randomly selected trial states, prepared by choosing random sets of parameters for a hardware efficient ansatz preparation circuit of depth 1. For each molecule, we show the mean, minimum and maximum values of R obtained from the 100 runs. We also show the value of $\hat{R}$, given by Eq. (20), obtained, which is close to the mean value of R where this has been calculated. A key result of interest is the maximum number of two-qubit gates required to obtain a measurement of all the operators in a given Hamiltonian. We show the theoretical maximum, given by the largest value of $kn - \frac{1}{2}k(k+1)$ for any collection, and the true largest value for any collection. The ratio of these two numbers is shown in Fig. 3(c). We also show the mean number of two-qubit gates required in a rotation circuit, averaged across all collections for a given molecule.