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.

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.

*These authors contributed equally to this work.
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 $E[R]$ 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 $R$ 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

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

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

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

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 $\text{H}_2$, which requires two qubits, to hydrogen selenide $\text{H}_2\text{Se}$, 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 $u_{cz}(k, n)$ by a factor of 3.5.

Lastly, we present data showing SORTED INSERTION outperforming the four conventional greedy
Given $\epsilon$, let $M_u$ and $M_g$ be the minimum number of measurements required to attain an accuracy $\epsilon$ 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 of collection $i$ is

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

Therefore, we have

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

where

$$\text{Var}[O_i] = \text{Cov}[O_i, O_i] \tag{6}$$

$$= \langle O_i^2 \rangle - \langle O_i \rangle^2. \tag{7}$$

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

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

where

$$\text{Var}[P_{ij}] = 1 - \langle P_{ij} \rangle^2. \tag{9}$$

The ratio $R$, defined as

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

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

$$R(\{\mathcal{A}, \mathcal{B}\}) \leq R(\{\mathcal{C}\}) \quad \text{for all } |\psi\rangle. \tag{11}$$
Proof. We write the operators associated with collections \( A \) and \( B \) as
\[
O_A = \sum_{j=1}^{k} a_j P_j, \quad (12)
\]
\[
O_B = \sum_{j=k+1}^{k+l} a_j P_j, \quad (13)
\]
for some \( k, l \) respectively, where the indexing of the sum in \( O_B \) is for later convenience. The operator associated with \( C \) can therefore be written as
\[
O_C = \sum_{j=1}^{k+l} a_j P_j. \quad (14)
\]
We define the \((k+l) \times (k+l)\) covariance matrix \( C \), which is symmetric and positive semi-definite, associated with collection \( C \) by its entries
\[
C_{ij} := \text{Cov}[P_i, P_j], \quad (15)
\]
with
\[
\text{Cov}[P_i, P_j] := \langle P_i P_j \rangle_{\psi} - \langle P_i \rangle_{\psi} \langle P_j \rangle_{\psi}. \quad (16)
\]
We write \( a \) 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 \( b \) 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 \( \{A, B\} \) than in the case of \( \{C\} \), i.e., that
\[
\sqrt{\text{Var}[O_A]} + \sqrt{\text{Var}[O_B]} \geq \sqrt{\text{Var}[O_C]}. \quad (17)
\]
With the above notation, this is equivalent to
\[
\sqrt{a^T Ca} + \sqrt{b^T Cb} \geq \sqrt{(a+b)^T C(a+b)}. \quad (18)
\]
We find
\[
\sqrt{a^T Ca} + \sqrt{b^T Cb} = \sqrt{(a^T Ca + b^T Cb)^2} = \sqrt{a^T Ca + b^T Cb + 2 \sqrt{(a^T Ca)(b^T Cb)}} \geq \sqrt{a^T Ca + b^T Cb + 2(a^T Cb)} = \sqrt{(a+b)^T C(a+b)}, \quad (19)
\]
where the fourth line follows due to the Cauchy-Schwarz inequality on a semi-inner product defined by \( \langle a, b \rangle := a^T Cb \) [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 \( \langle \alpha a + \beta b, \alpha a + \beta b \rangle = 0 \) [49, Example 1.4].

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
\[
\hat{R} := \left( \frac{\sum_{i=1}^{N} \sum_{j=1}^{m_i} |a_{ij}|}{\sum_{i=1}^{N} \sqrt{\sum_{j=1}^{m_i} |a_{ij}|^2}} \right)^2. \quad (20)
\]
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
\[
O = 4X_1 + 4X_2 + Z_2 + Z_1X_2. \quad (21)
\]
The operators cannot be assembled into a single commuting collection but there is a unique arrangement, \( G_1 \), consisting of two collections, given by
\[
G_1 := \{\{4X_1, Z_2\}, \{4X_2, Z_1X_2\}\}, \quad (22)
\]
which has \( \hat{R} = 1.47 \) to two decimal places. \( G_1 \) performs worse than an arrangement \( G_2 \), with three collections, given by
\[
G_2 := \{\{4X_1, 4X_2\}, \{Z_2\}, \{Z_1X_2\}\}, \quad (23)
\]
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
\[ O = \sum_{i=1}^{t} a_j P_i, \]  
where \( t \) is the total number of Pauli operators, the entire set \( \{ (a_j, P_i) \} \) 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.

<table>
<thead>
<tr>
<th>Colouring Algorithm</th>
<th>Time Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Largest First</td>
<td>( O(t^2) )</td>
</tr>
<tr>
<td>Connected Sequential d.f.s.</td>
<td>( O(t^2) )</td>
</tr>
<tr>
<td>DSATUR</td>
<td>( O(t^2 \log t) )</td>
</tr>
<tr>
<td>Independent Set</td>
<td>( O(t^2) )</td>
</tr>
</tbody>
</table>

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 the 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 considering the operator \( O \), on two qubits, defined as
\[ O = X_1 + Z_1 + 2Z_2 \]  
and a grouping of the form
\[ G_3 := \{ \{X_1 + \alpha Z_2\}, \{Z_1 + (2 - \alpha) Z_2\} \} \]  
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, \( S'_{\text{start}} \), of \( m \) Paulis which can be represented as a binary \( 2n \times m \) matrix \( S'_{\text{start}} \). By Gaussian elimination on \( S'_{\text{start}} \), we can form a \( 2n \times k \) matrix \( S_{\text{start}} \) representing a set \( S_{\text{start}} \) of \( k \) independent Paulis drawn from \( S'_{\text{start}} \) where \( k \leq \min(n, m) \). Our goal is to transform \( S_{\text{start}} \), using certain allowed transformations, into a \( 2n \times n \) matrix \( S_{\text{end}} \) where
\[ S_{\text{end}} = \begin{pmatrix} 1_n \\ 0_n \end{pmatrix}. \]  

Let \( U \) denote the circuit consisting of one-qubit and two-qubit transformations in the order they were applied from \( S_{\text{start}} \rightarrow S_{\text{end}} \). 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 \( S'_{\text{start}} \) on \( |\psi\rangle \) simultaneously.

The allowed set, \( T \), 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 \( j \):
     \[ S_{ip} \leftarrow S_{ip} + S_{i+n,p}, \]
     \[ S_{jp} \leftarrow S_{jp} + S_{i+n,p}. \]
   
   - CNOT on target-qubit \( j \):
     \[ S_{ip} \leftarrow S_{ip} + S_{jp}, \]
     \[ S_{j+n,p} \leftarrow S_{j+n,p} + S_{i+n,p}. \]

   - X on row \( i \):
     \[ S_{ip} \leftarrow 2S_{ip} \]
- **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 \(j\):
  \[ S_{ip} \leftrightarrow S_{jp}, \quad S_{n+i,p} \leftrightarrow S_{n+j,p}. \]

3. Basis extension:

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

In the near term, operations in \(T\) 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 \(T\), given by

\[
S^\top J_{2n} S = 0_n, \tag{28}
\]

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

\[
J_{2n} = \begin{pmatrix} 0_n & I_n \\ I_n & 0_n \end{pmatrix}. \tag{29}
\]

We ignore any changes in sign of stabilisers under \(T\) 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

\[
g_i = X_i \prod_{j \in \text{nbd}(i)} Z_j, \quad i = 1, \ldots, n, \tag{30}
\]

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

\[
S_{\text{graph}} = \begin{pmatrix} A \\ I_n \end{pmatrix}, \tag{31}
\]

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 \(|\Phi_G\rangle = VHN^{2n}(0^n)\) 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 \(S_{\text{start}}\) to \(S_{\text{end}}\) via \(S_{\text{graph}}\) 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 \(k \leq n\) of \(S_{\text{start}}\).

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

\[
\begin{align*}
S_{\text{start}} & \xrightarrow{\delta_1} \begin{pmatrix} A \\ B \end{pmatrix} \\
S_3 & \xrightarrow{\delta_3} \begin{pmatrix} C & D^\top \\ E & D^\top \\
I_k & 0 \\
0 & I_{n-k} \end{pmatrix} \\
S_{\text{end}} & \xrightarrow{\delta_5} 
\end{align*}
\]

Figure 1: Reductions used in our CZ-construction.

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

1. \(S_{\text{start}} \rightarrow S_1\). 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. \(S_1 \rightarrow S_2\). 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. \(S_2 \rightarrow S_3\). 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. \(S_3 \rightarrow S_4 = S_{\text{graph}}\). Column operations can eliminate \(F\), then Phase gates can ensure \(E\) has ze-
5. $S_4 \rightarrow S_{\text{end}}$. **Phase** and CZ gates can implement this final reduction as discussed above. The maximum number of CZ gates required to map $S_4$ to $S_{\text{end}}$ 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 $w_{\text{cz}}(k,n) = kn - k(k+1)/2$ due to sparsity of the upper half of $S_4$ which traces back to the structure of the additional operators appended in step $S_3 \rightarrow S_4$.

Note that in step $S_4 \rightarrow S_{\text{end}}$, 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

$$
\begin{align*}
S_4 & \xrightarrow{\text{rotation}} \left( \begin{array}{cc} 0_k & D^T \\ D & 0_{n-k} \\ M & 0 \\ 0 & I_{n-k} \end{array} \right) \rightarrow \left( \begin{array}{cc} I_k & D^T \\ D & I_{n-k} \\ I_k & 0 \\ 0 & I_{n-k} \end{array} \right) \\
& \xrightarrow{\text{CNOT}} \left( \begin{array}{cc} D_1 M_1 \end{array} \right) \rightarrow \left( \begin{array}{cc} I_k - D^T D \\ 0 \\ 0 \\ 0 \\ I_k \end{array} \right) \rightarrow \left( \begin{array}{cc} 0 \\ 0 \\ 0 \end{array} \right) \rightarrow \left( \begin{array}{cc} I_k \\ 0 \\ 0 \end{array} \right)
\end{align*}
$$

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 $S_{\text{end}}$ as shown in Fig. 2, using three blocks of CNOT gates.

5. $S_4 \rightarrow S_5$. 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 $E = \Lambda + M^T M$, with $\Lambda$ diagonal and $M$ invertible.

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

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

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

where

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

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

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 $w_{\text{cnot}}(k,n) = O(kn/\log k)$ via arguments of Patel, Markov, and Hayes [44]. The proof is given in Appendix E.

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

9. $S_8 \rightarrow S_{\text{end}}$. $I_k - D^T D$ 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 $S_4 \rightarrow S_5$, $S_6 \rightarrow S_7$, and $S_8 \rightarrow S_{\text{end}}$, 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 $\{|+, -\rangle\}$, $Y$ corresponds to control on $\{|+, -\rangle\}$, 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 $Y$. 

Accepted in `{Quantum} 2020-12-27, click title to verify. Published under CC-BY 4.0.`
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 quantum 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).

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.
operators, $N$, 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 mutually 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 $u_{\text{cnot}}(k,n) = kn - (k+1)/2$ two-qubit gates while the CNOT-construction uses a maximum of $u_{\text{cnot}}(k,n) = O(kn/\log k)$.

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 $3.5$.

---

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.

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Largest First</th>
<th>Connected Sequential d.f.s.</th>
<th>Independent Set</th>
<th>DSATUR</th>
<th>Sorted Insertion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N$</td>
<td>$\bar{N}$</td>
<td>$\bar{R}$</td>
<td>$N$</td>
<td>$\bar{R}$</td>
</tr>
<tr>
<td>H$_2$</td>
<td>2</td>
<td>1.76</td>
<td>1.76</td>
<td>2</td>
<td>1.76</td>
</tr>
<tr>
<td>H$_2^+$</td>
<td>10</td>
<td>10.25</td>
<td>10.30</td>
<td>10</td>
<td>10.30</td>
</tr>
<tr>
<td>LiH</td>
<td>9</td>
<td>23.87</td>
<td>23.33</td>
<td>9</td>
<td>23.33</td>
</tr>
<tr>
<td>OH$^-$</td>
<td>40</td>
<td>8.41</td>
<td>8.41</td>
<td>40</td>
<td>8.41</td>
</tr>
<tr>
<td>HF</td>
<td>38</td>
<td>8.05</td>
<td>8.07</td>
<td>38</td>
<td>8.07</td>
</tr>
<tr>
<td>H$_2$O</td>
<td>57</td>
<td>29.86</td>
<td>30.66</td>
<td>57</td>
<td>30.66</td>
</tr>
<tr>
<td>BH$_3$</td>
<td>66</td>
<td>18.80</td>
<td>18.70</td>
<td>66</td>
<td>18.70</td>
</tr>
<tr>
<td>NH$_3$</td>
<td>122</td>
<td>174</td>
<td>13.97</td>
<td>126</td>
<td>13.97</td>
</tr>
<tr>
<td>CH$_4$</td>
<td>122</td>
<td>5.84</td>
<td>18.93</td>
<td>114</td>
<td>18.93</td>
</tr>
<tr>
<td>O$_2$</td>
<td>62</td>
<td>13.62</td>
<td>19.95</td>
<td>42</td>
<td>19.95</td>
</tr>
<tr>
<td>N$_2$</td>
<td>62</td>
<td>15.00</td>
<td>21.15</td>
<td>39</td>
<td>21.15</td>
</tr>
<tr>
<td>CO</td>
<td>124</td>
<td>20.70</td>
<td>20.67</td>
<td>89</td>
<td>20.67</td>
</tr>
<tr>
<td>HCl</td>
<td>117</td>
<td>2.16</td>
<td>10.29</td>
<td>98</td>
<td>10.29</td>
</tr>
<tr>
<td>NaH</td>
<td>121</td>
<td>8.78</td>
<td>12.40</td>
<td>149</td>
<td>12.40</td>
</tr>
<tr>
<td>H$_2$S</td>
<td>122</td>
<td>8.81</td>
<td>12.45</td>
<td>147</td>
<td>12.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Arrangement</th>
<th>$\bar{R}_{\text{min}}$</th>
<th>$\bar{R}_{\text{mean}}$</th>
<th>$\bar{R}_{\text{max}}$</th>
<th>$\hat{R}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_2$</td>
<td>1.09</td>
<td>1.93</td>
<td>4.60</td>
<td>1.76</td>
</tr>
<tr>
<td>$H_2^+$</td>
<td>3.76</td>
<td>11.92</td>
<td>33.04</td>
<td>10.25</td>
</tr>
<tr>
<td>LiH</td>
<td>19.60</td>
<td>24.91</td>
<td>34.74</td>
<td>29.97</td>
</tr>
<tr>
<td>OH$^-$</td>
<td>6.32</td>
<td>8.90</td>
<td>12.86</td>
<td>8.51</td>
</tr>
<tr>
<td>HF</td>
<td>6.07</td>
<td>8.57</td>
<td>12.27</td>
<td>8.21</td>
</tr>
<tr>
<td>H$_2$O</td>
<td>7.68</td>
<td>11.27</td>
<td>16.96</td>
<td>10.67</td>
</tr>
<tr>
<td>BH$_3$</td>
<td>17.21</td>
<td>20.93</td>
<td>32.13</td>
<td>20.05</td>
</tr>
<tr>
<td>NH$_3$</td>
<td>12.65</td>
<td>15.96</td>
<td>26.93</td>
<td>15.31</td>
</tr>
<tr>
<td>CH$_4$</td>
<td>16.96</td>
<td>21.63</td>
<td>29.33</td>
<td>20.27</td>
</tr>
</tbody>
</table>

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.
6 Acknowledgement

We thank Oscar Higgott for informing us of the ancilla-based simultaneous measurement method and Ref. [60, Fig. 14]. We also thank the editor Ronald de Wolf and two anonymous reviewers for their comments and suggestions.

References


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


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

$$H = -XX - YY + ZZ + IZ + ZI,$$  \hspace{1cm} (35)

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

$$C = \begin{pmatrix}
1 & 1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0
\end{pmatrix}.$$  \hspace{1cm} (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

$$\{-XX, -YY, ZZ\}, \{IZ, ZI\}.$$  \hspace{1cm} (37)

For these collections of Paulis, the coefficient vectors are

$$a = \begin{pmatrix}
-1 \\
-1 \\
0 \\
0 \\
0
\end{pmatrix}, \quad c = \begin{pmatrix}
0 \\
0 \\
1 \\
1 \\
1
\end{pmatrix}.$$  \hspace{1cm} (38)

The number of measurements to achieve an accuracy of $\epsilon$ is

$$M_g = \frac{1}{\epsilon^2} \left( \sqrt{a^T Ca} + \sqrt{c^T Cc} \right)^2 = \frac{1}{\epsilon^2} \left( \sqrt{4} + \sqrt{0} \right)^2 = \frac{4}{\epsilon^2}.$$  \hspace{1cm} (39)

Now, let us consider breaking up the first collection into

$$\{-XX\}, \{-YY, ZZ\}.$$  \hspace{1cm} (40)

In this case, the coefficient vectors are

$$a = \begin{pmatrix}
-1 \\
0 \\
0 \\
0 \\
0
\end{pmatrix}, \quad b = \begin{pmatrix}
0 \\
-1 \\
1 \\
0 \\
0
\end{pmatrix}, \quad c = \begin{pmatrix}
0 \\
0 \\
1 \\
1 \\
1
\end{pmatrix}.$$  \hspace{1cm} (41)

The number of measurements required to attain an accuracy $\epsilon$ is therefore

$$M_g = \frac{1}{\epsilon^2} \left( \sqrt{a^T Ca} + \sqrt{b^T Cb} + \sqrt{c^T Cc} \right)^2 = \frac{1}{\epsilon^2} \left( \sqrt{T + \sqrt{T + \sqrt{0}}} \right)^2 = \frac{4}{\epsilon^2}.$$  \hspace{1cm} (42)

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 $\langle \alpha a + \beta b, \alpha a + \beta b \rangle = 0$.

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

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

Proof. The variance of a single Pauli operator is

$$\text{Var}[P_1] := 1 - \langle P_1 \rangle^2.$$  \hspace{1cm} (44)

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

$$\mathbb{E}[\text{Var}[P_1]] = 1 - \mathbb{E}[\langle P_1 \rangle^2] = 1 - \int [\langle P_1 | \psi \rangle^2] d\psi = 1 - \alpha_n,$$

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

$$\mathbb{E}[\text{Cov}[P_i, P_j]] = 0,$$  \hspace{1cm} (46)

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

C  Binary representation

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

$$\mathcal{P}_n = \{ i^k \sigma_1 \otimes \cdots \otimes \sigma_n | \sigma_i \in \{I, X, Y, Z\}, \ k \in \{0, 1, 2, 3\} \}.$$  \hspace{1cm} (47)

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
\[
\begin{align*}
\sigma_{00} &:= I \rightarrow (0,0), \\
\sigma_{01} &:= X \rightarrow (0,1), \\
\sigma_{10} &:= Z \rightarrow (1,0), \\
\sigma_{11} &:= Y \rightarrow (1,1),
\end{align*}
\] (48)

An \(n\)-qubit Pauli
\[
\sigma_{a_1v_1} \otimes \cdots \otimes \sigma_{a_nv_n}
\] (49)
is then represented by the \(2n\)-dimensional binary vector
\[
(u_1, \ldots, u_n, v_1, \ldots, v_n)\trans.
\] (50)
In this representation, two \(n\)-qubit Paulis with binary vectors \(a\) and \(b\) commute if and only if
\[
a\trans J_{2n} b = 0,
\] (51)
where \(J_{2n}\) denotes the \(2n \times 2n\) matrix
\[
J_{2n} := \begin{pmatrix}
0 & I_n \\
I_n & 0
\end{pmatrix}.
\] (52)

Given a set \(S\) 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 \(S\) mutually commute if and only if
\[
S\trans J_{2n} S = 0_m,
\] (53)
which recovers Eq. (28) in the main text. We say that the set \(S\) 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.,
\[
S = \begin{pmatrix}
S^{(Z)} \\
S^{(X)}
\end{pmatrix}.
\] (54)

The conjugation action of quantum gates on \(S\) 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 \(j\):**
  \[
  S_{ip} \leftarrow S_{ip} + S_{i+n,p},
  S_{jp} \leftarrow S_{jp} + S_{i+n,p}.
  \]

- **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 \(i\):**
  \[
  S_{ip} \leftarrow S_{i+n,p}.
  \]

- **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 \(S_{\text{start}}'\) of six-qubit Paulis given by
\[
\begin{align*}
P_1 &= Z_1 Z_2 Z_3 Z_4, \\
P_2 &= X_1 X_2 Y_3 Y_4, \\
P_3 &= Y_1 Y_2 X_3 X_4, \\
P_4 &= Y_2 X_3, \\
P_5 &= Y_1 X_4, \\
P_6 &= X_1 Z_2 Z_3 Y_4.
\end{align*}
\] (55)

We can represent these Paulis in a matrix \(S_{\text{start}}'\) with
\[
S_{\text{start}}' = \begin{pmatrix}
1 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 1
\end{pmatrix}.
\] (56)

By Gaussian elimination, we find the reduced row echelon form of \(S_{\text{start}}'\) to be
\[
\begin{pmatrix}
1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.
\] (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 \(S_{\text{start}}'\) can be constructed. Therefore, we can write \(S_{\text{start}}' = S_{\text{start}}' R_0^{-1}\).
where

$$S_{\text{start}} := \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad R_0^{-1} := \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$  (58)

Note that the inverse on $R_0^{-1}$ is purely notational.

Now, the lower half $S^{(X)}_{\text{start}}$ of $S_{\text{start}}$ has column echelon form

$$S^{(X)}_{\text{start}} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$  (59)

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

$$S_1 := Q_1 S_{\text{start}} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$  (60)

where

$$Q_1 := \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$  (61)

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

$$S_2 := Q_1 S_{\text{start}} R_1 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$  (62)

where

$$R_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$  (63)

We now extend $S_2$ to a rank $n = 4$ matrix by adding a column that corresponds to a fourth Pauli $P_{\text{ext}}$. In the main text, this is the crucial basis extension step from $S_2 \rightarrow S_3$ 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

$$P(S_2) := \begin{pmatrix} Y & I & I & Z \\ I & Y & Z & I \\ Z & Z & X & X \end{pmatrix},$$  (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 $P_{\text{ext}}$ (and nowhere else) to ensure $P_{\text{ext}}$ 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 $P_{\text{ext}}$ 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 $P_{\text{ext}} = Z_1 I_2 I_3 X_4$. Therefore, $S_2$ is extended to

$$S_3 := \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix},$$  (65)

and

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

where

$$R_2^{-1} := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$  (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

$$R_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$  (68)
and
\[ S_3 R_3 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \] (69)

Finally, we apply \textsc{Phase} to qubits 1 and 2 so that
\[ S_4 = Q_2 S_3 R_3 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \] (70)
where
\[ Q_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \] (71)

\( S_4 \) is of the form of a graph state and represents the following Paulis:
\[
\begin{align*}
\tilde{P}_1 &= X_1 I_2 I_3 Z_4, \\
\tilde{P}_2 &= I_1 X_2 Z_3 I_4, \\
\tilde{P}_3 &= I_1 Z_2 X_3 I_4, \\
\tilde{P}_4 &= Z_1 I_2 I_3 X_4.
\end{align*}
\] (72)

We now have
\[ S'_{\text{start}} = Q^{-1} S_4 R^{-1}, \] (73)
where
\[ Q^{-1} := Q_1^{-1} Q_2^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \] (74)

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 \((+1 +1 +1 -1 -1 -1)\). Therefore, we can construct measurements of the original Pauli strings as follows:
\begin{itemize}
  \item \( P_1 \) from product of measurements of qubits 3 and 4,
  \item \( P_2 \) from product of measurements of qubits 1 to 4,
  \item \( P_3 \) from product of measurements of qubits 1 and 2,
  \item \( P_4 \) from the negative of measurement of qubit 2,
  \item \( P_5 \) from the negative of measurement of qubit 1,
  \item \( P_6 \) from the negative product of measurements of qubits 1, 3 and 4.
\end{itemize}

\[ R^{-1} := R_3^{-1} R_2^{-1} R_1^{-1} R_0^{-1} = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}. \] (75)

\[ \text{Figure 4: The rotation circuit } U \text{ for the CZ-construction walk-through example.} \]

\[ q_1 \quad S \quad H \quad - \quad - \quad - \]
\[ q_2 \quad S \quad H \quad - \quad - \quad - \]
\[ q_3 \quad H \quad H \quad - \quad - \quad - \]
\[ q_4 \quad H \quad H \quad - \quad - \quad - \]

\begin{IEEEproof}[E] Proof of \( O(\frac{kn}{\log k}) \)

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 \textsc{CNOT} gates, as explained in detail in Ref. [44].

\textbf{Claim 2. Let } \( M \) \text{ be a } n \times n \text{ matrix with block form}
\[
\begin{pmatrix}
I_k & A \\
0 & I_{n-k}
\end{pmatrix},
\] (76)

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

\textbf{Proof.} Let \( m \) be a constant we later choose. Partition \( A \) into \( l := (n-k)/m \) consecutive column-blocks \( A_l \), 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

$$
\begin{pmatrix}
B & 0 \\
0 & I_{n-k}
\end{pmatrix},
$$

(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

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

(78)

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

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

(79)

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

## 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 $P_i = \bigotimes_{j=1}^n P_{i,j}$ is a tensor product of $n$ single-qubit Paulis, so that

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

(80)

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

For an arrangement $A$ 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

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

(81)

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

$$
N \leq D(A) \leq \sqrt{n} \sqrt{\sum_{i=1}^N m_i} \leq \sqrt{N} \sqrt{n},
$$

(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 $A_0$ and let $N_1$ denote the number of commuting collections in an arrangement $A_1$ that maximises $\hat{R}$. By definition, $N_0 \leq N_1$. Since maximising $\hat{R}$ is equivalent to minimising $D(A)$, we also have

$$
N_1 \leq D(A_1) \leq D(A_0) \leq \sqrt{n} N_0,
$$

(83)

using Eq. (82) for the first and third inequalities. Hence, $D(A_1)$ 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 $D(A_1)$ 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 $\epsilon > 0$. 


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

<table>
<thead>
<tr>
<th>Molecule</th>
<th>$n$ qubits</th>
<th>$t$ Paulis</th>
<th>Arrangement Ratios $R$, $\hat{R}$</th>
<th>Rotation Circuit $2q$-size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N$</td>
<td>$m$</td>
<td>$\overline{k}$</td>
<td>$R_{\text{min}}$</td>
</tr>
<tr>
<td>H$_2$</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2.00</td>
</tr>
<tr>
<td>H$_2^+$</td>
<td>4</td>
<td>59</td>
<td>10</td>
<td>5.90</td>
</tr>
<tr>
<td>LiH</td>
<td>10</td>
<td>630</td>
<td>41</td>
<td>15.37</td>
</tr>
<tr>
<td>OH$^-$</td>
<td>10</td>
<td>630</td>
<td>38</td>
<td>16.58</td>
</tr>
<tr>
<td>HF</td>
<td>10</td>
<td>630</td>
<td>39</td>
<td>16.15</td>
</tr>
<tr>
<td>H$_2$O</td>
<td>10</td>
<td>1085</td>
<td>51</td>
<td>21.27</td>
</tr>
<tr>
<td>BH$_3$</td>
<td>14</td>
<td>1584</td>
<td>66</td>
<td>24.00</td>
</tr>
<tr>
<td>NH$_3$</td>
<td>14</td>
<td>3608</td>
<td>118</td>
<td>30.58</td>
</tr>
<tr>
<td>CH$_4$</td>
<td>16</td>
<td>3887</td>
<td>123</td>
<td>31.60</td>
</tr>
<tr>
<td>O$_2$</td>
<td>18</td>
<td>2238</td>
<td>67</td>
<td>33.40</td>
</tr>
<tr>
<td>N$_2$</td>
<td>18</td>
<td>2950</td>
<td>78</td>
<td>37.82</td>
</tr>
<tr>
<td>CO</td>
<td>18</td>
<td>4426</td>
<td>128</td>
<td>34.58</td>
</tr>
<tr>
<td>HCl</td>
<td>18</td>
<td>4538</td>
<td>123</td>
<td>36.89</td>
</tr>
<tr>
<td>NaH</td>
<td>18</td>
<td>5850</td>
<td>135</td>
<td>43.33</td>
</tr>
<tr>
<td>O$_3$</td>
<td>20</td>
<td>6277</td>
<td>147</td>
<td>42.70</td>
</tr>
<tr>
<td>PH$_3$</td>
<td>22</td>
<td>19746</td>
<td>304</td>
<td>64.95</td>
</tr>
<tr>
<td>SiH$_4$</td>
<td>24</td>
<td>18713</td>
<td>304</td>
<td>61.56</td>
</tr>
<tr>
<td>NaF</td>
<td>26</td>
<td>10538</td>
<td>287</td>
<td>57.62</td>
</tr>
<tr>
<td>LiCl</td>
<td>26</td>
<td>17044</td>
<td>292</td>
<td>58.37</td>
</tr>
<tr>
<td>KF</td>
<td>26</td>
<td>24290</td>
<td>325</td>
<td>74.74</td>
</tr>
<tr>
<td>F$_2$</td>
<td>28</td>
<td>11429</td>
<td>216</td>
<td>52.91</td>
</tr>
<tr>
<td>NO$_2$</td>
<td>28</td>
<td>20541</td>
<td>317</td>
<td>64.80</td>
</tr>
<tr>
<td>NO$_3$</td>
<td>28</td>
<td>20549</td>
<td>311</td>
<td>66.07</td>
</tr>
<tr>
<td>Cl$_2$</td>
<td>34</td>
<td>34334</td>
<td>378</td>
<td>90.83</td>
</tr>
<tr>
<td>NaCl</td>
<td>34</td>
<td>42826</td>
<td>498</td>
<td>86.00</td>
</tr>
<tr>
<td>SF$_2$</td>
<td>36</td>
<td>56025</td>
<td>567</td>
<td>98.81</td>
</tr>
<tr>
<td>HBr</td>
<td>36</td>
<td>62589</td>
<td>602</td>
<td>103.97</td>
</tr>
<tr>
<td>SO$_2$</td>
<td>36</td>
<td>75315</td>
<td>691</td>
<td>108.99</td>
</tr>
<tr>
<td>NO$_3^-$</td>
<td>38</td>
<td>61132</td>
<td>622</td>
<td>92.28</td>
</tr>
<tr>
<td>H$_2$Se</td>
<td>38</td>
<td>69684</td>
<td>631</td>
<td>110.43</td>
</tr>
</tbody>
</table>

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, $m$, and the average rank of the collections, $k$, 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.