ROYAL SOCIETY
OF CHEMISTRY

PAPER

Check for updates

Efficient strategies for reducing sampling error in quantum Krylov subspace diagonalization

Gwonhak Lee, [ID] [a] Seonghoon Choi, [ID] [bc] Joonsuk Huh [ID] *[d]
and Artur F. Izmaylov [ID] *[bc]

Within the realm of early fault-tolerant quantum computing (EFTQC), quantum Krylov subspace diagonalization (QKSD) has emerged as a promising quantum algorithm for the approximate Hamiltonian diagonalization *via* projection onto the quantum Krylov subspace. However, the algorithm often requires solving an ill-conditioned generalized eigenvalue problem (GEVP) involving erroneous matrix pairs, which can significantly distort the solution. Since EFTQC assumes limited-scale error correction, finite sampling error becomes a dominant source of error in these matrices. This work focuses on quantifying sampling errors during the measurement of matrix element in the projected Hamiltonian examining two measurement approaches based on the Hamiltonian decompositions: the linear combination of unitaries and diagonalizable fragments. To reduce sampling error within a fixed budget of quantum circuit repetitions, we propose two measurement strategies: the shifting technique and coefficient splitting. The shifting technique eliminates redundant Hamiltonian components that annihilate either the bra or ket states, while coefficient splitting optimizes the measurement of common terms across different circuits. Numerical experiments with electronic structures of small molecules demonstrate the effectiveness of these strategies, reducing sampling costs by a factor of 20–500.

I. Introduction

Recent advancements in quantum computing devices, particularly in terms of the scale and coherence time,[1–7] have significantly heightened expectations for their ability to perform efficient quantum simulations. These advancements promise to deepen our understanding of many-body quantum systems, such as electronic structure in chemical systems.[8–10] This anticipation is driven by the expected stability, controllability, and scalability of universal quantum computers being developed across various platforms, including ion traps,[2,3] photons,[4,5] and superconductors.[6,7]

Currently, the field is progressing through the era of noisy and intermediate-scale quantum computers (NISQ).[11,12] This phase marks a regime of quantum computation which is hard to be simulated using classical computers, while quantum error correction is absent due to the limited scalability, inherent noises, and decoherence of current devices. Within this context, the variational quantum eigensolver (VQE) has been primarily

discussed as an algorithm for quantum simulation.[13,14] Based on the variational principle, VQE employs quantum-classical hybrid optimization of a parameterized ansatz, implementable within a shallow quantum circuit to approximate specific eigenstates of the target system. However, the expected quantum advantage—based on the classical hardness of simulating the ansatz—is negated by errors in estimating cost function for each optimization step, particularly associated with barren plateau problem.[15–19] Furthermore, the absence of the error correction results in the errors accumulating significantly, thus limiting the scalable quantum advantage in VQE.

This naturally shifts our attention towards early fault-tolerant quantum computing (EFTQC) as a viable next step beyond NISQ era. The feasibility of EFTQC is further supported by decreasing hardware error rates that are approaching the threshold for the error correction[20–22] and an emergence of a small-scale demonstration of logical qubits.[23] EFTQC is introduced within a framework of scale-limited quantum error correction, where the error rate for logical qubits increases with the size of the quantum circuit.[24] Consequently, unlike fully fault-tolerant quantum computing, EFTQC cannot arbitrarily scale the number of logical qubits or the use of non-Clifford operations, thereby limiting the practical implementation of quantum phase estimation.[25] To address these limitations, EFTQC algorithms typically aim to compromise between the circuit size and the number of repetitions. Quantum phase estimation requires $M = O(|\gamma_0|^{-2})$ repetitions of a circuit with real-time propagators ($e^{-i\hat{H}t_k}$) with the total

[a]SKKU Advanced Institute of Nanotechnology (SAINT), Sungkyunkwan University, Suwon 16419, Korea

[b]Chemical Physics Theory Group, Department of Chemistry, University of Toronto, Toronto, Ontario M5S 3H6, Canada. E-mail: artur.izmaylov@utoronto.ca

[c]Department of Physical and Environmental Sciences, University of Toronto Scarborough, Toronto, Ontario M1C 1A4, Canada

[d]Department of Chemistry, Yonsei University, Seoul 03722, South Korea. E-mail: joonsukhuh@gmail.com

propagation time, $\sum_k t_k = O(\varepsilon_{\text{alg}}^{-1})$, where $|\gamma_0|$ and ε_{alg} denote the initial overlap and algorithm accuracy, respectively.[25] Here, the real-time propagator is usually approximately implemented by Suzuki–Trotterization.[26,27] In contrast, EFTQC algorithms use shorter propagators, $t < O(\varepsilon_{\text{alg}}^{-1})$, but increase the number of repetitions, $M > O(|\gamma_0|^{-2})$.[28–33] An alternative approach within EFTQC utilizes the block encoding scheme,[34] which, while exact, demands significantly more resources than Trotterization with a minimal Trotter steps. Although reducing Trotterization errors necessitates more Trotter steps, thus approaching the resource demands of block encoding, the limitations of near-future hardware render small-scale Trotterization a more feasible option.

Within the domain of EFTQC algorithms, quantum Krylov subspace diagonalization (QKSD) is being explored as a promising candidate for quantum simulation algorithm.[32,33] QKSD employs quantum circuits to project the Hamiltonian onto the Krylov subspace, a reduced-dimensional space that is classically solvable. This approach is potentially feasible because the extremal eigenvalues of the projected matrix converge exponentially fast to those of the original Hamiltonian, provided the projection remains unperturbed and the overlap between the eigenstate and the initial state is large.[35] However, this advantage is counterbalanced by the challenge associated with ill-conditioning of the eigenvalue problem, where perturbations in the projected matrix can significantly distort the accuracy of the approximated eigenvalues.[35,36] These perturbations mainly arise from imperfect error correction, Trotterization error, and finite sampling error. As EFTQC stabilizes and expands, the first two factors can be suppressed. However, despite the discussion of the measurement problem in QKSD,[36] strategies to tackle this problem have not been suggested.

QKSD involves measuring the matrix elements, $\boldsymbol{H}_{kl} = \langle \phi_k | \hat{H} | \phi_l \rangle$, across a finite basis $\{ |\phi_k\rangle = \mathrm{e}^{-i\hat{H}k\Delta_t} |\phi_0\rangle \}$ that spans the Krylov subspace with a reference state $|\phi_0\rangle$. While other bases have been proposed, such as $|\phi_k\rangle = \hat{H}^k |\phi_0\rangle$[37] and $|\phi_k\rangle = \mathrm{e}^{-\hat{H}k\beta} |\phi_0\rangle$,[38] we focus on QKSD with the real-time evolution operator due to its simplicity and practical viability for EFTQC. Consequently, a primary objective in this scenario is to minimize the sampling error when measuring the matrix elements. Although this specific measurement problem has not been tackled, several measurement strategies have been proposed for the standard expectation values.[39–46] In general, the direct measurement of $\langle \psi | \hat{H} | \psi \rangle$ with a single circuit is impractical, as measurement is constrained to the Pauli-Z basis, which requires the implementation of a unitary operator that fully diagonalizes $\hat{H}$. As an alternative, $\hat{H}$ is decomposed into a linear combination of fragmented Hamiltonians, which can be efficiently diagonalized with implementable unitaries, and the results for each fragment are aggregated.[39,42,46] The goal then becomes to minimize the sampling error by optimally fragmenting $\hat{H}$ and allocating the number of circuit repetitions among these fragments. This optimization can be formulated as a combinatorial problem, akin to NP-hard clique covering problems,[42,46] for which a heuristic solution has been developed.[39] Additionally, this issue has been expanded into a continuous optimization problem that considers

approximated covariances between fragments.[40,44] On the other hand, randomized measurement strategies, known as classical shadow techniques, have been developed.[47–49] While these methods are superior when measuring multiple expectation values simultaneously, deterministic methods generally outperform them when measuring a single expectation value $\langle \hat{H} \rangle$.[40,41,50] Since our case involves measuring $\langle \phi_k | \hat{H} | \phi_l \rangle$, we focus on deterministic measurement strategies in this work.

These developments resolve the measurement problems associated with standard expectation values. In this paper, we aim to adapt these methods to the QKSD framework. To accomplish this, we analyze the measurement problem of the matrix elements with two decomposition scenarios for $\hat{H}$: linear combination of unitaries (LCU) and fragmented Hamiltonians (FH). Specifically, we focus on quantifying and mitigating the sampling errors for these scenarios, applying strategies originally designed for standard expectation values to enhance measurement accuracy in QKSD. Notably, the strategies that we propose can be applied to the general measurement of transitional amplitudes, $\langle \phi_k | \hat{H} | \phi_l \rangle$, which can be utilized for the design of algorithms beyond QKSD.

The organization of the paper is as follows. First, a brief preliminary of the QKSD is provided in Section II, followed by the analyses of the two measurement methods for the matrix elements and the associated errors in Section III. Section IV demonstrates how conventional methods[39,40,51,52] are converted to reduce the sampling errors for QKSD, highlighting that a method initially devised for reducing the simulation cost of LCU[51,52] is transferable to the measurement problem. Finally, we numerically validate these reduction methods by solving the electronic structure problems of small molecules as case studies in Section V.

II. QKSD

Before considering the measurement problem in QKSD, this section reviews the QKSD method for estimating the spectrum of a Hamiltonian $\hat{H}$, as originally introduced in ref. 32 and 33.

QKSD estimates approximated eigenstates of a Hamiltonian $\hat{H}$ with the following ansatz:

$$|\psi(\boldsymbol{w})\rangle := \frac{1}{\mathcal{N}} \sum_{k=0}^{n-1} w_k \hat{\boldsymbol{B}}^k |\phi_0\rangle, \qquad (1)$$

where $\mathcal{N}$ is the normalization factor, and $\boldsymbol{w} = (w_0, \ldots, w_{n-1}) \in \mathbb{C}^n$, and n is the Krylov order. This ansatz fully covers vectors in the Krylov subspace, $\mathcal{K}_n = \mathrm{span}(\{|\phi_0\rangle, \ldots |\phi_{n-1}\rangle\})$, where $|\phi_k\rangle = \hat{B}^k |\phi_0\rangle$ is defined by a reference state $|\phi_0\rangle$ and the base operator,

$$\hat{B} = \mathrm{e}^{-i\Delta_t \hat{H}}. \qquad (2)$$

This exponential function is approximated by the Trotterization. There are other choices for $\hat{B}$, such as $\hat{H}$, which is analogous to the classical Krylov method,[37] and imaginary time evolution $(\mathrm{e}^{-\beta \hat{H}})$.[38] Although we only focus on the real-time evolution operator, which is widely discussed, the methods that will be described in the Section IV can be expanded to the

other choices of $\hat{B}$ because they are approximated by linear combinations or products of real-time propagators.

Using the ansatz (eqn (1)) in the variational principle with $\boldsymbol{w}$ as optimized parameters leads to the following generalized eigenvalue problem (GEVP)

$$\boldsymbol{Hw} = \boldsymbol{SwE}^{(n)}, \tag{3}$$

where $E^{(n)}$ is an approximate eigenvalue, and the $n \times n$ Hamiltonian matrix $\boldsymbol{H}$ and overlap matrix $\boldsymbol{S}$ are defined as

$$\boldsymbol{H}_{kl} = \langle \phi_k | \hat{H} | \phi_l \rangle = \langle \phi_0 | \hat{B}^{\dagger k} \hat{H} \mathsf{B}^l | \phi_0 \rangle = \boldsymbol{H}_{0,l-k} \tag{4}$$

$$\boldsymbol{S}_{kl} = \langle \phi_k | \phi_l \rangle = \langle \phi_0 | \hat{B}^{\dagger k} \hat{B}^l | \phi_0 \rangle = \boldsymbol{S}_{0,l-k}. \tag{5}$$

These matrices are obtained by quantum algorithms using EFTQC circuits and measurements. Note that $\boldsymbol{S}$ has the structure of a Toeplitz matrix, $\boldsymbol{S}_{k,l} = \boldsymbol{S}_{0,l-k}$, because $\hat{B}$ is unitary. Furthermore, since $[\hat{H}, \hat{B}] = 0$, $\boldsymbol{H}$ becomes a Toeplitz matrix as well. Therefore, rather than n^2, only n elements are required to construct each matrices.

For systems with large energy gaps, the lowest solution of eqn (3) converges to the ground state energy of $\hat{H}$ exponentially fast with n.[35] However, the GEVP in eqn (3) can become ill-conditioned for larger n's, which makes QKSD sensitive to noise in matrix elements of $\boldsymbol{H}$ and $\boldsymbol{S}$.

III. Measurement of QKSD matrix elements

In the quantum subroutine that estimates the elements of $\boldsymbol{H}$ and $\boldsymbol{S}$, the quantum uncertainty predominantly induces the matrix perturbation. This perturbation, coupled with the ill-conditioned GEVP (eqn (3)), may introduce a significant error in the solution. In this section, we develop methods for estimating Hamiltonian matrix elements,

$$\boldsymbol{H}_{0k} = \langle \phi_0 | \hat{H} | \phi_k \rangle, \tag{6}$$

and the analysis of the associated sampling error.

Since, rather than the matrix element, only the measurements of standard expectation values of $\langle \Phi | \hat{O}_j | \Phi \rangle$ with easily diagonalizable $\hat{O}_j$'s are possible at the circuit level, it is necessary to express eqn (6) in terms of $\langle \Phi | \hat{O}_j | \Phi \rangle$ using certain states $|\Phi\rangle$ and simple operators $\hat{O}_j$. To translate eqn (6) into standard expectations, we consider two approaches: the Hadamard and the extended swap tests. In both approaches, the problem is addressed by partitioning the Hamiltonian into diagonalizable Hermitian or implementable unitary operators. In the Hadamard test, the Hamiltonian is presented as an LCU: $\hat{H} = \sum_j \beta_j \hat{U}_j$, where each unitary $\hat{U}_j$ is implemented to estimate the overlap between $|\phi_0\rangle$ and $\hat{U}_j | \phi_k \rangle$ (Fig. 1a). In the extended swap test, $\hat{H}$ is decomposed as a sum of fragments: $\hat{H} = \sum_j \hat{V}_j^{\dagger} \hat{D}_j \hat{V}_j$, where $\hat{D}_j$ is diagonal and the corresponding diagonalizing unitary $\hat{V}_j$ can be implemented efficiently. Then, the extended swap test is conducted with the
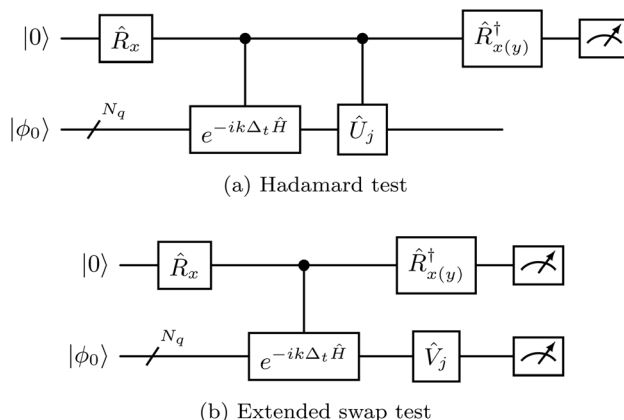

(a) Hadamard test

(b) Extended swap test

Fig. 1 Circuit diagrams for (a) the Hadamard test and (b) the extended swap test to estimate the j-th fragment of the Hamiltonian matrix element, $\langle \phi_0 | \hat{H} | \phi_k \rangle$ (see eqn (11) and (18)); here, $\hat{R}_{x(y)}$ operator rotates $\hat{\sigma}_z$ basis into $\hat{\sigma}_x$ ($\hat{\sigma}_y$) basis and is adopted to estimate the real (imaginary) part of the amplitude for the second operator applied to the ancilla qubit.

circuits of Hadamard tests which estimate $\langle \phi_0 | \phi_k \rangle$, while including the measurement of system qubits with the basis represented by $\hat{V}_j$ (Fig. 1b). Note that in the extended swap test, the overlap matrix elements can be measured simultaneously by measuring the ancilla qubit, which corresponds to a Hadamard test circuit that measures $S_{0,k} = \langle \phi_0 | \mathrm{e}^{-ik\Delta_t \hat{H}} | \phi_0 \rangle$. This simultaneous measurement capability is not available in the Hadamard test when measuring the Hamiltonian in its LCU form.

This work mainly focuses on analyzing and improving the sampling error determined by the decomposition of the observable, which is not required for the case of the overlap matrix element, $\boldsymbol{S}_{0k} = \langle \phi_0 | \hat{\mathbb{1}} | \phi_k \rangle$. In a previous work,[36] it was shown that the sampling variance of $\boldsymbol{S}_{0k}$ are determined identically across the test algorithms as:

$$\mathrm{Var}[\boldsymbol{S}_{0k}; m_R, m_I] = \frac{1}{M} \left(\frac{1 - \mathrm{Re}[\boldsymbol{S}_{0k}]^2}{m_R} + \frac{1 - \mathrm{Im}[\boldsymbol{S}_{0k}]^2}{m_I} \right), \tag{7}$$

where m_R and m_I are the fraction of total shots, M, allocated to measure real and imaginary parts of $\boldsymbol{S}_{0k}$, respectively ($m_R + m_I = 1$). The ideal allocation that minimizing the variance is $m_R^{(\mathrm{opt})} \propto (1 - \mathrm{Re}[\boldsymbol{S}_{0k}]^2)^{1/2}$ and $m_I^{(\mathrm{opt})} \propto (1 - \mathrm{Im}[\boldsymbol{S}_{0k}]^2)^{1/2}$, which is, however, infeasible to be estimated before measuring $\boldsymbol{S}_{0k}$. If we take Haar averaging of the states, $|\phi_0\rangle$ and $|\phi_k\rangle$, $m_R = m_I = 1/2$ is achieved from Lemma 2 in Appendix A. The corresponding variance is related to the amplitude of the matrix element as

$$\mathrm{Var}[\boldsymbol{S}_{0k}; 1/2, 1/2] = \frac{2}{M} \left(2 - |\boldsymbol{S}_{0k}|^2 \right). \tag{8}$$

In the rest of this section, we quantify the sampling error associated with the Hamiltonian matrix elements and examine how the decomposition of $\hat{H}$ affects this error.

A Hadamard test

An LCU decomposition of $\hat{H}$ is

$$\hat{H} = \sum_{j=1}^{N_\beta} \beta_j \hat{U}_j, \tag{9}$$

where $\hat{U}_j$ is unitary, and β_j is a real and positive coefficient. Such decomposition was originally motivated in the context of Hamiltonian simulation problem[53] and then expanded to the measurement problem.[54,55] The simulation cost in LCU based approach scales linearly with the L1 norm of the coefficients, $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|$. In the Hadamard test, we are going to show that the sampling cost to estimate eqn (6) within a certain level of accuracy is proportional to $\|\boldsymbol{\beta}\|_1^2$.

The matrix element in eqn (6) can be viewed as the weighted sum of overlaps between $|\phi_0\rangle$ and $\hat{U}_j|\phi_k\rangle$:

$$\langle\phi_0|\hat{H}|\phi_k\rangle = \sum_{j=1}^{N_\beta} \beta_j \langle\phi_0|\hat{U}_j|\phi_k\rangle. \tag{10}$$

Fig. 1a depicts the circuit that estimates $\langle\phi_0|\hat{U}_j|\phi_k\rangle$. To derive the sampling cost, eqn (6) is translated to standard expectations as

$$\langle\phi_0|\hat{H}|\phi_k\rangle = \sum_{j=1}^{N_\beta} \beta_j \langle\Phi_{0k;j}|\left(\hat{\sigma}_x + i\hat{\sigma}_y\right)\otimes\hat{\mathbb{I}}|\Phi_{0k;j}\rangle, \tag{11}$$

where

$$|\Phi_{0k;j}\rangle := \frac{1}{\sqrt{2}}\left(|0\rangle|\phi_0\rangle + |1\rangle\hat{U}_j|\phi_k\rangle\right) \tag{12}$$

is prepared with an additional qubit and the conditional evolution unitary. The additional qubit is measured in $\hat{\sigma}_x$ and $\hat{\sigma}_y$ bases, corresponding to the real and imaginary parts of $\langle\phi_0|\hat{U}_j|\phi_k\rangle$, respectively. Thus, $2N_\beta$ independent measurements of $\hat{O}_R = \hat{\sigma}_x\otimes\hat{\mathbb{I}}$ and $\hat{O}_I = \hat{\sigma}_y\otimes\hat{\mathbb{I}}$ with a set of states $\{|\Phi_{0k;j}\rangle\}_{j=1}^{N_\beta}$ complete the total estimation of a matrix element in eqn (11).

For measuring the expectation value of eqn (11), the variance is:

$$\text{Var}^{(\text{LCU})}[\boldsymbol{H}_{0k}; \boldsymbol{m}] = \frac{1}{M}\left(\sum_{j=1}^{N_\beta}\sum_{X\in\{R,I\}}\frac{\beta_j^2}{m_{jX}}\text{Var}\left[\hat{O}_X\right]_{\Phi_{0k;j}}\right), \tag{13}$$

where

$$\text{Var}\left[\hat{O}_X\right]_{\Phi_{0k;j}} = \left\langle\hat{O}_X^2\right\rangle_{\Phi_{0k;j}} - \left\langle\hat{O}_X\right\rangle_{\Phi_{0k;j}}^2$$

$$= \begin{cases} 1 - \text{Re}\left[\langle\phi_0|\hat{U}_j|\phi_k\rangle\right]^2, & X = R \\ 1 - \text{Im}\left[\langle\phi_0|\hat{U}_j|\phi_k\rangle\right]^2, & X = I \end{cases}. \tag{14}$$

Also, M denotes the total number of shots to measure $\boldsymbol{H}_{0k}$, and m_{jX} is the fraction of shots for $\langle\Phi_{0k;j}|\hat{O}_X|\Phi_{0k;j}\rangle$ $\left(\sum_{jX}m_{jX} = 1\right)$.

Given LCU decomposition, finding the optimal shot allocation $\boldsymbol{m}$ to minimize eqn (13) is a convex problem, which is analytically solved by

$$m_{jX}^{(\text{opt})} \propto \beta_j\text{Var}[\hat{O}_X]_{\Phi_{0k;j}}^{1/2}. \tag{15}$$

However, since $\text{Var}[\hat{O}_X]_{\Phi_{0k;j}}$ are not known in advance, they are estimated by taking Haar-averaging over the states $|\phi_0\rangle$ and $|\phi_k\rangle$. This results in a sub-optimal allocation instead, $m_{jX}^{(\text{subopt})} \propto \beta_j$, which is independent of the unknown variances. Furthermore, such allocation leads to the averaged variance as shown below:

$$V^{(\text{LCU})} := \mathbb{E}_{\phi_0,\phi_k}\left[\text{Var}^{(\text{LCU})}\left[\boldsymbol{H}_{0k}; \boldsymbol{m}^{(\text{subopt})}\right]\right]$$

$$= \frac{2\|\boldsymbol{\beta}\|_1^2}{M}\left(2 - \frac{1}{d}\right), \tag{16}$$

where d is the dimension of the Hilbert space, which is generally exponentially large. The proof of eqn (16) is provided in Appendix A. The number of total shots more than

$$M \approx \frac{4\|\boldsymbol{\beta}\|_1^2}{\varepsilon^2} \tag{17}$$

is required to make the total uncertainty less than ε with a high probability, which shows that the sampling cost is proportional to $\|\boldsymbol{\beta}\|_1^2$. This result indicates that an LCU decomposition of $\hat{H}$ with lower $\|\boldsymbol{\beta}\|_1$ can potentially reduce the number of samplings, while maintaining accuracy.

B Extended swap test

The extended swap test[32] estimates the matrix element through the measurement of

$$\langle\phi_0|\hat{H}|\phi_k\rangle = \langle\Phi_{0k}|(\hat{\sigma}_x + i\hat{\sigma}_y)\otimes\hat{H}|\Phi_{0k}\rangle, \tag{18}$$

where

$$|\Phi_{0k}\rangle := \frac{1}{\sqrt{2}}(|0\rangle|\phi_0\rangle + |1\rangle|\phi_k\rangle).$$

However, the direct measurement of eqn (18) is possible only when $\hat{H}$ is an Ising form. In general, $\hat{H}$ can be expressed as a sum of diagonalizable fragment Hamiltonians:

$$\hat{H} = \sum_{j=1}^{N_\gamma} \hat{H}_j = \sum_{j=1}^{N_\gamma} \hat{V}_j^\dagger \hat{D}_j \hat{V}_j, \tag{19}$$

where $\hat{V}_j$ can be efficiently implemented to diagonalize $\hat{H}_j$ onto the computational basis, yielding an Ising Hamiltonian $\hat{D}_j$. Then, the measurement can be performed for each $\hat{H}_j$. For example, if $\hat{H}_j$ is composed of mutually commuting Pauli operators, the unitary $\hat{V}_j$ can be efficiently determined and implemented as a Clifford circuit with the gate count of $O(N_q^2)$, where N_q is the number of qubits.[39,56,57]

After substituting $\hat{H}$ in eqn (18) by (19), the total variance of the estimator is

$$\text{Var}^{(\text{FH})}[\boldsymbol{H}_{0k}; \boldsymbol{m}] = \frac{1}{M}\left(\sum_{j=1}^{N_\gamma}\sum_{X\in\{R,I\}}\frac{\text{Var}\left[\hat{O}_{jX}\right]_{\Phi_{0k}}}{m_{jX}}\right), \tag{20}$$

where $\hat{O}_{j,R(I)} = \hat{\sigma}_{x(y)}\otimes\hat{H}_j$, and the quantum variance of $\hat{O}_j$ is

$$\text{Var}\left[\hat{O}_{jX}\right]_{\Phi_{0k}} = \begin{cases} \left\langle \hat{O}_{jR}^{\;2} \right\rangle_{\Phi_{0k}} - \text{Re}\left[\langle \phi_0|\hat{H}_j|\phi_k\rangle\right]^2 & X = R, \\ \left\langle \hat{O}_{jI}^{\;2} \right\rangle_{\Phi_{0k}} - \text{Im}\left[\langle \phi_0|\hat{H}_j|\phi_k\rangle\right]^2 & X = I, \end{cases} \quad (21)$$

with

$$\left\langle \hat{O}_{jR}^{\;2} \right\rangle_{\Phi_{0k}} = \left\langle \hat{O}_{jI}^{\;2} \right\rangle_{\Phi_{0k}} = \frac{1}{2}\left(\langle \phi_0|\hat{H}_j^{\;2}|\phi_0\rangle + \langle \phi_k|\hat{H}_j^{\;2}|\phi_k\rangle\right).$$

As in LCU decomposition, the optimal shot allocation, $m_{jX}^{(\text{opt})} \propto \text{Var}[\hat{O}_{jX}]_{\Phi_{0kj}}^{1/2}$, is hard to estimate in advance. In order to find a sub-optimal allocation, $\boldsymbol{m}^{(\text{subopt})}$, we perform Haar averaging on the variance over the states, $|\phi_0\rangle$ and $|\phi_k\rangle$. According to the result in Appendix A, this averaging produces $m_{jR(I)}^{(\text{subopt})} \propto (\text{Tr}[\hat{H}_j^{\;2}]/d)^{1/2}$. Furthermore, for the case of Pauli decomposition, the value of $(\text{Tr}[\hat{H}_j^{\;2}]/d)^{1/2}$ is efficiently determined as the L2 norm of Pauli coefficients in $\hat{H}_j$, as elaborated in Appendix B.

The total averaged variance is

$$\begin{aligned} V^{(\text{FH})} : &= \mathbb{E}_{\phi_0,\phi_k}\left[\text{Var}^{(\text{FH})}\left[\boldsymbol{H}_{0k}; \boldsymbol{m}^{(\text{subopt})}\right]\right] \\ &= \frac{2\|\boldsymbol{\gamma}\|_1^2}{M}\left(2 - \frac{1}{d}\right), \end{aligned} \quad (22)$$

where $\|\boldsymbol{\gamma}\|_1 := \sum_j (\text{Tr}[\hat{H}_j^{\;2}]/d)^{1/2}$. The corresponding measurement cost,

$$M \approx \frac{4\|\boldsymbol{\gamma}\|_1^2}{\varepsilon^2}, \quad (23)$$

scales quadratically with the $\|\boldsymbol{\gamma}\|_1$, which plays the same role as $\|\boldsymbol{\beta}\|_1$ in the LCU case.

We can find the similarity between the variances of LCU and FH, (see eqn (16) and (22)). Let us write the decompositions for both cases (eqn (9) and (19)) as

$$\hat{H} = \sum_{j=1}^{N_\zeta} \hat{N}_j, \quad (24)$$

where $\hat{N}_j \in \{\beta_j \hat{U}_j, \hat{H}_j\}$. Then, the partial variances, eqn (14) (multiplied by β_j) and (21), are generalized into a single form:

$$\text{Var}\left[X_{0k;j}\right] = \frac{1}{2}\left(\langle \phi_0|\hat{N}_j^\dagger \hat{N}|\phi_0\rangle + \langle \phi_k|\hat{N}_j^\dagger \hat{N}|\phi_k\rangle\right) - \mathbb{E}\left[X_{0k;j}\right]^2, \quad (25)$$

where $X_{0k;j} \in \{R_{0k;j}, I_{0k;j}\}$ is an estimator for the real or imaginary part of $\langle \phi_0|\hat{N}_j|\phi_k\rangle$.

Moreover, the decomposition norms, $\|\boldsymbol{\beta}\|_1$ and $\|\boldsymbol{\gamma}\|_1$ in eqn (16) and (22), are regarded as:

$$\|\boldsymbol{\zeta}\|_1 : = \frac{1}{\sqrt{d}}\sum_j \text{Tr}\left[\hat{N}_j^\dagger \hat{N}_j\right]^{1/2}. \quad (26)$$

Correspondingly, the sampling cost is given by:

$$M \approx \frac{4\|\boldsymbol{\zeta}\|_1^2}{\varepsilon^2}. \quad (27)$$

Thus, regardless of the test algorithms, $\|\boldsymbol{\zeta}\|_1$ serves as a metric of decomposition assessing the finite sampling error, akin to the measurements of standard expectation values.[39,55]

IV. Sampling cost reduction

In this section, we propose techniques to reduce the sampling cost discussed in the previous section. This is done by adapting the cost reduction techniques for the measurement of standard expectation, $\langle \phi|\hat{H}|\phi\rangle$,[39,40,44] to the measurement of matrix elements. Such adaptation is simply done by replacing the standard variance, $\langle \hat{O}_j^{\;2}\rangle - \langle \hat{O}_j\rangle^2$, by the variance for the matrix elements represented in eqn (14) and (21). We propose a method to optimize the decomposition, $\{\hat{U}_j\}$ or $\{\hat{H}_j\}$, to achieve smaller variance analogous to the approach in previous work.[39] Furthermore, the dependence of the sampling cost on $\|\boldsymbol{\zeta}\|_1$, as shown in eqn (27), allows us to use methods that reduce $\|\boldsymbol{\zeta}\|_1$.

A Shifting technique

Here, we introduce a technique to reduce the norm $\|\boldsymbol{\zeta}\|_1$, and consequently, lower expected sampling costs, by shifting the Hamiltonian. Before developing the technique, let's clarify the notation for the norm, $\|\boldsymbol{\zeta}_A(\hat{H})\|_1$, which indicates the norm of the decomposition achieved on $\hat{H}$ using a specific deterministic algorithm, A. This clarification is crucial because the decomposition applied to $\hat{H}$ is not unique without specifying the algorithm A. As an example of A, a greedy algorithm-based SORTED INSERTION heuristically finds a decomposition by Pauli operators that yields a relatively small norm.[39]

The shifting technique involves finding an operator $\hat{T}$ that shifts the Hamiltonian and minimizes the norm of the shifted decomposition:

$$\min_\tau \left\|\boldsymbol{\zeta}_A\left(\hat{H} - \hat{T}(\tau)\right)\right\|_1, \quad (28)$$

where A is a fixed polynomial time algorithm performing a decomposition, and the Hermitian operator $\hat{T}$ is parameterized by τ, enabling the use of classical optimization algorithms. Additionally, we impose a constraint on $\hat{T}(\tau)$, that is

$$\hat{T}(\tau)|\phi_0\rangle = t(\tau)|\phi_0\rangle, \quad (29)$$

for a known factor $t(\tau) \in \mathbb{R}$. Note that $\hat{T}$ is not required to commute with $\hat{H}$, unlike symmetry operators. The necessity of this constraint will be presented with the rest of procedure.

After the optimization of eqn (28), we then employ test algorithms explained in Section III to estimate the shifted Hamiltonian matrix $\boldsymbol{H} - \boldsymbol{T}$, consuming reduced cost ($\|\boldsymbol{\zeta}_A(\hat{H} - \hat{T})\|_1^2 \leq \|\boldsymbol{\zeta}_A(\hat{H})\|_1^2$). Here, $\boldsymbol{H} - \boldsymbol{T}$ is defined as a Toeplitz matrix satisfying

$$[\boldsymbol{H} - \boldsymbol{T}]_{kl} = \langle \phi_0|(\hat{H} - \hat{T})|\phi_{l-k}\rangle. \quad (30)$$

Then, the GEVP with the shifted Hamiltonian matrix is

$$(\boldsymbol{H} - \boldsymbol{T})\boldsymbol{w} = \boldsymbol{S}\boldsymbol{w}(E^{(n)} - t), \quad (31)$$

because the original matrix element is written in terms of the shifted matrix element as shown below:

$$
\begin{aligned}
H_{0k} &= \langle \phi_0 | \hat{H} | \phi_k \rangle \\
&= \langle \phi_0 | \hat{H} - \left(\hat{T} - t \right) | \phi_k \rangle \\
&= \langle \phi_0 | \left(\hat{H} - \hat{T} \right) | \phi_k \rangle + t \langle \phi_0 | \phi_k \rangle \\
&= [\boldsymbol{H} - \boldsymbol{T}]_{0k} + t\boldsymbol{S}_{0k},
\end{aligned}
$$

and $\boldsymbol{H} - \boldsymbol{T}$ is Toeplitz as defined in eqn (30). Thus, the constraints of eqn (29) and (30) enables the recovery of the solutions of the original GEVP, $E^{(n)}$ from that of shifted one by simply adding t.

Here, we give an example of designing and parameterizing $\hat{T}$ dedicated to the electronic structure problem. In many cases, the reference state, $|\phi_0\rangle$ is chosen as a simple state, such as Hartree Fock (HF) ground state or a configuration state function (CSF) which is a symmetry-adapted state composed of a small number of Slater determinants. Some or all orbitals in such reference states are separately occupied or unoccupied, which is represented as:

$$
|\phi_0\rangle = \left(\bigotimes_{q \in \text{occ}} |1\rangle_q \bigotimes_{q \in \text{virt}} |0\rangle_q \right) \otimes |\phi_0^{(r)}\rangle, \tag{32}
$$

where 'occ' and 'virt' denote the sets of occupied and virtual orbitals, respectively, and $|\phi_0^{(r)}\rangle$ is possibly entangled and in the remainder system, satisfying $|\phi_0^{(r)}\rangle\langle\phi_0^{(r)}| = \text{Tr}_{\text{occ,virt}}[|\phi_0\rangle\langle\phi_0|]$. In a HF ground state, all orbitals are either occupied or unoccupied while a CSF state may involve some entangled state $|\phi_0^{(r)}\rangle$. Furthermore, an electronic structure Hamiltonian is represented as

$$
\hat{H} = \sum_{p \leq q}^{N_{\text{orb}}} h_{pq} \hat{E}_{pq} + \sum_{pqrs} g_{pqrs} \left(\hat{E}_{pq} \hat{E}_{rs} + \text{h.c.} \right), \tag{33}
$$

where N_{orb} denotes the number of total orbitals and the notations of the symmetric excitation operators $\hat{E}_{rs} = \hat{a}_r^\dagger \hat{a}_s + \hat{a}_s^\dagger \hat{a}_r$ and the number operators $\hat{n}_q = \hat{a}_q^\dagger \hat{a}_q$ are employed.

Then, the shift operator can be designed to cover one- and two-body terms in the Hamiltonian and to satisfy eqn (29), which has the following form:

$$
\hat{T}\left(\boldsymbol{\tau}^{(1)}, \boldsymbol{\tau}^{(2)}\right) = \sum_{q \in \mathcal{F}} \left(\tau_q^{(1)} \hat{n}_q + \sum_{rs \in \mathcal{E}_q} \tau_{qrs}^{(2)} \hat{E}_{rs} \left(\hat{n}_q - \delta_{q \in \text{occ}} \right) \right). \tag{34}
$$

Here, the sets of orbital indices, $\mathcal{F} := \text{occ} \cup \text{virt}$ and $\mathcal{E}_q := \{(r,s): r, s \in [N_{\text{orb}}] \backslash \{q\}, r \leq s\}$ are defined, and $\delta_{q \in \text{occ}}$ is an identity if $q \in \text{occ}$, zero otherwise. The indices r and s range over the entire orbital set except q to make $\hat{T}$ Hermitian and to avoid duplication with the one-body number operator, since $\hat{n}_q^2 = \hat{n}_q$. Note that the two-body terms with $\hat{E}_{rs}\hat{n}_q$ for $q \in \text{virt}$ annihilate $|\phi_0\rangle$, as do $\hat{E}_{rs}(\hat{n}_q - 1)$ for $q \in \text{occ}$. Therefore, the corresponding shift factor is determined as

$$
t = \sum_{q \in \text{occ}} \tau_q^{(1)}. \tag{35}
$$

The optimal $\boldsymbol{\tau}$ can be found using iterative optimization algorithms like `Powell` or `BFGS` with the number of parameters of $|\boldsymbol{\tau}| = O(N_{\text{orb}}^3)$. However, the optimization overhead is reduced if we adopt a decomposition algorithm where each term in eqn (33)

is regarded as a fragment, as detailed in Appendix D. By using this reduced optimization, the optimal parameters are found as

$$
\begin{aligned}
\tau_q^{(1)} &= 2h_{qq} \\
\tau_{qrs}^{(2)} &= 4g_{rsqq} - 2g_{rqsq}.
\end{aligned} \tag{36}
$$

After this optimization, the entire number operators, along with a significant portion of one-body Hamiltonian, are discarded. Consequently, only a part of the two-body Hamiltonian needs to be measured.

The shifting technique is also applicable to other algorithms that require the measurement of $a_{\hat{O}}(\tau) := \langle \phi_0 | \hat{O} e^{-i\hat{H}\tau} | \phi_0 \rangle$ for some Hermitian operator $\hat{O}$. With the extended swap test, one can efficiently estimate $a_{\hat{O}}(\tau)$ from the measurement of $a_{\hat{O}-\hat{T}}(\tau)$ and $a_{\mathbb{1}}(\tau)$, which are always measurable simultaneously.

B Iterative coefficient splitting

In contrast to the shifting technique, which minimizes the state-averaged costs (eqn (16) and (22)), this section focuses on optimizing the state-dependent costs (eqn (13) and (20)). As one approach, we apply iterative coefficient splitting (ICS), initially designed for standard measurements,[40] to the problem of measuring the matrix elements.

Given a qubit Hamiltonian, $\hat{H} = \sum_p \alpha_p \hat{P}_p$, where $\alpha_p \in \mathbb{R}$ and $\hat{P}_p \in \{\hat{\mathbb{1}}, \hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z\}^{\otimes N_q}$, ICS seeks a decomposition into measurable operators that minimizes the total variance. Note that measurable operators here involve not only Hermitian operators but also scaled unitaries, which are used for the Hadamard test. As reviewed in Appendix B, such a decomposition in Pauli basis is written as:

$$
\hat{H} = \sum_{j=1}^{N_\zeta} \hat{N}_j = \sum_{j=1}^{N_\zeta} \sum_{p \in \mathcal{G}_j} \alpha_p^{(\mathcal{G}_j)} \hat{P}_p, \tag{37}
$$

where $\hat{N}_j$ are measurable operators and the corresponding sets of Pauli indices, $\mathcal{G}_j$, are predetermined by a decomposition algorithm like `SORTED INSERTION`. Importantly, these $\mathcal{G}_j$ sets may not be disjoint, meaning the same operator $\hat{P}_p$ can belong to multiple sets. Correspondingly, the coefficient α_p is split across these sets, satisfying the condition:

$$
\sum_{j: p \in \mathcal{G}_j} \alpha_p^{(\mathcal{G}_j)} = \alpha_p \quad \forall p. \tag{38}
$$

ICS leverages the flexibility of coefficient splitting to minimize the total variance, which is thus treated as a function of the split coefficients $\boldsymbol{\alpha}$ and the shot allocation $\boldsymbol{m}$:

$$
\text{Var}_{(\text{ICS})}(\boldsymbol{\alpha}, \boldsymbol{m}) = \frac{1}{M} \sum_{\substack{j=1,\ldots,N_\zeta \\ X \in \{R,I\}}} \frac{\text{Var}\left[X_{0k;j}; \boldsymbol{\alpha}^{(\mathcal{G}_j)}\right]}{m_{jX}}, \tag{39}
$$

where the vectors of split coefficients are defined as $\boldsymbol{\alpha} := (\boldsymbol{\alpha}^{(\mathcal{G}_1)}, \ldots, \boldsymbol{\alpha}^{(\mathcal{G}_{N_\zeta})})$ and $\boldsymbol{\alpha}^{(\mathcal{G}_j)} := \{\alpha_p^{\mathcal{G}_j} : p \in \mathcal{G}_j\}$. The partial variance for $\hat{N}_j$ (eqn (25)) is expressed as a quadratic form in terms of the split coefficients $\boldsymbol{\alpha}^{(\mathcal{G}_j)}$:

$$\text{Var}\left[X_{0k;j}; \alpha^{(\mathcal{G}_j)}\right] = \sum_{p,q \in \mathcal{G}_j} \alpha_p^{(\mathcal{G}_j)} \alpha_q^{(\mathcal{G}_j)} \text{Cov}^{(X)}\left(\hat{P}_p, \hat{P}_q\right)_{\Phi_{0k}}. \quad (40)$$

Here, the Pauli covariance for the real part is determined as:

$$\text{Cov}^{(R)}\left(\hat{P}_p, \hat{P}_q\right)_{\Phi_{0k}} = \langle \Phi_{0k}|\widehat{\mathbb{I}} \otimes \frac{1}{2}\left\{\hat{P}_p, \hat{P}_q\right\}|\Phi_{0k}\rangle$$
$$- \text{Re}\left[\langle\phi_0|\hat{P}_p|\phi_k\rangle\right]\text{Re}\left[\langle\phi_0|\hat{P}_q|\phi_k\rangle\right]. \quad (41)$$

The detailed derivation is provided in Appendix C. The covariance for the imaginary part is obtained by replacing Re with Im in eqn (41).

To proceed with the optimization of eqn (39), the covariances need to be estimated beforehand. A direct and precise calculation of eqn (41) requires a state, $|\phi_k\rangle = e^{-i\hat{H}t_k}|\phi_0\rangle$, which is classically difficult to obtain. Therefore, the covariance is approximated using configuration interaction single and double (CISD) ground state, $|\text{CISD}\rangle$, and its energy, E_{CISD}:

$$|\phi_k\rangle \approx e^{-iE_{\text{CISD}}t_k}|\text{CISD}\rangle. \quad (42)$$

Then, the optimization problem is given as:

$$(\boldsymbol{\alpha}^\star, \boldsymbol{m}^\star) = \underset{\substack{\boldsymbol{\alpha}:\alpha_p = \sum\limits_{j:p\in\mathcal{G}_j} \alpha_p^{(\mathcal{G}_j)} \forall p \\ \boldsymbol{m}:\sum\limits_{jX} m_{jX} = 1}}{\text{argmin}} \text{Var}_{(\text{ICS})}(\boldsymbol{\alpha}, \boldsymbol{m}). \quad (43)$$

Here, $\boldsymbol{\alpha}$ and $\boldsymbol{m}$, denote the split coefficients and the shot allocation, respectively. Although optimizing both $\boldsymbol{\alpha}$ and $\boldsymbol{m}$ does not have a closed-form solution, the each step of alternating optimization—by fixing one variable ($\boldsymbol{\alpha}$ or $\boldsymbol{m}$) while optimizing the other—is a convex problem. When $\boldsymbol{m}$ is held constant, constrained quadratic programming can be employed for optimizing $\boldsymbol{\alpha}$, because the variance is expressed as a quadratic form of $\boldsymbol{\alpha}$ as shown in eqn (40). Conversely, when optimizing $\boldsymbol{m}$ with a fixed $\boldsymbol{\alpha}$, the Lagrangian multiplier method is utilized, which results in

$$m_{jX}^{(\text{opt})} \propto \text{Var}\left[X_{0k;j}; \alpha^{(\mathcal{G}_j)}\right]^{1/2}. \quad (44)$$

Overall, adapting the ICS method to the measurement problem for the matrix elements involves three additional key features compared to ref. 40: (1) including scaled unitaries as measurable objects, (2) defining covariances between anti-commuting Pauli operators, (3) employing the CISD proxy state for the time-evolved state.

V. Numerical results

Here, we present numerical illustrations of our theoretical developments by examining the electronic structures of small molecules: H_2, H_4, LiH, BeH_2 and H_2O, using the STO-3G basis set. The fermionic Hamiltonians are transformed to qubit operators by the Bravyi–Kitaev mapping with two-qubit reduction.[58,59]

For the QKSD setting, we use the Hartree–Fock ground as the reference state $|\phi_0\rangle$. The time step for the propagator is chosen as $\Delta_t = \pi/\Delta E_1$, following the choice in ref. 35, Theorem 3.1, where ΔE_1 represents the first spectral gap. In practical scenario where the spectral gap is difficult to estimate in advance, a sufficiently large time step is often chosen to mitigate ill-conditioning, although this comes at the cost of increased circuit depth. The dependency of the conditioning on the time step has been numerically studied in ref. 37, Appendix A-2. Also, in order to focus on the finite sampling error, we assume that the exact propagator $\hat{B} = e^{-i\hat{H}\Delta_t}$ is available, which does not involve Trotterization error. The QKSD order is set to $n = N_q + 1$, where N_q is the number of qubits. In this setting, the overhead for the classical GEVP is exponentially small compare to the direct diagonalization, while the error induced by the projection onto the quantum Krylov space is exponentially small as shown in ref. 35, Theorem 3.1. We observed that the error caused by the Krylov projection is bounded as $|E_0^{(n)} - E_0| < 10^{-4}$ Ha in the electronic structures of our interest, where E_0 is the true ground state energy of $\hat{H}$.

The reduction in the norm achieved through the shifting operator in eqn (34) is presented in Table 1. Overall, the shifting technique reduced the norm more than 74%. The relative reductions are larger in the LCU case because the shifting removes large $\hat{Z}$-type Pauli operators. These operators cannot be grouped together in the LCU decomposition, leading to a larger norm when they were not removed. The details are provided in Appendix B. However, the resulting costs of LCU remain higher than those for FH, which implies that FH allows more efficient measurement.

The exact and empirical measurement costs for each scenario, both with and without the techniques described in Section IV, are tabulated in Table 2. The measurement costs obtained by the experiments approximate the exact costs described in eqn (13) and (20) within the error caused by the finite number of experiments. Generally, the cost reduction tends to increase with the system size. Also, a significant portion of the reduction is attributed to the shifting technique, which correlates closely with the squared reduction ratio in Table 1. This correlation suggests that $M\varepsilon^2 \propto \|\boldsymbol{\zeta}\|_1^2$, aligning with the relationship previously established in eqn (27).

Furthermore, we validated the approximation of Pauli covariance (eqn (41)) using the CISD proxy (eqn (42)) by comparing ICS results obtained with both the proxy and the true Krylov basis state, as shown in Table 2.

Table 1 LCU(β) and FH(γ) decomposition norms with and without shifting. `SORTED INSERTION` is adopted as decomposition algorithm. Shifting operators $\hat{T}$ are chosen as eqn (34) and optimized by the `POWELL` algorithm after assigning the parameters shown in eqn (36)

Norm (Hartree)	H_2	H_4	LiH	BeH_2	H_2O
$\|\beta_{\text{SI}}(\hat{H})\|_1$	0.8405	6.0055	9.9902	16.4482	57.3794
$\|\beta_{\text{SI}}(\hat{H} - \hat{T})\|_1$	0.1812	1.1278	0.4739	1.3582	2.0035
Reduction (LCU, %)	78.4	81.2	95.3	91.7	96.5
$\|\gamma_{\text{SI}}(\hat{H})\|_1$	0.7397	2.0310	2.5254	4.7003	21.9723
$\|\gamma_{\text{SI}}(\hat{H} - \hat{T})\|_1$	0.1812	0.5288	0.3268	0.7857	1.1727
Reduction (FH, %)	75.50	74.0	87.1	83.3	94.7

Table 2 The measurement costs $M\varepsilon^2$ for the (shifted) matrix elements $\langle\phi_0|\hat{H}(-\hat{T})|\phi_k\rangle$, averaged over k, are separately displayed for the cases with and without applying ICS and/or shift techniques from the measurement setting obtained from SORTED INSERTION (denoted as 'SI'). The costs are estimated by directly computing eqn (20) and (13). The values in the parenthesis are averaged empirical variances obtained by 1000 independent runs of QKSD algorithm for each setting. The sub-optimal shot allocations are used if ICS is not employed, while shot allocations of the ICS output are adopted otherwise. The result of ICS based on the true state $|\phi_k\rangle$ and phased CISD proxy are shown. Note that ICS by true state is not practically achievable

$M\varepsilon^2$ (Hartree^2)		H$_2$		H$_4$		LiH		BeH$_2$		H$_2$O	
LCU	SI	1.51	(1.50)	81.40	(82.53)	213.60	(215.62)	598.29	(600.75)	7265.04	(7290.16)
	ICS (True)	0.88	(0.87)	69.84	(69.77)	185.04	(182.27)	534.46	(534.83)	6534.53	(6508.27)
	ICS (CISD)	0.92	(0.90)	70.11	(71.19)	185.34	(184.59)	536.47	(529.47)	6550.79	(6561.18)
	Shift	0.13	(0.13)	5.08	(5.01)	0.89	(0.91)	7.37	(7.38)	16.04	(15.77)
	Shift, ICS	0.13	(0.13)	4.82	(4.85)	0.80	(0.79)	6.97	(7.03)	14.53	(14.50)
FH	SI	2.18	(2.24)	34.66	(34.27)	50.19	(50.25)	151.65	(149.91)	2284.67	(2287.81)
	ICS (True)	1.29	(1.29)	18.32	(18.32)	26.43	(26.29)	88.00	(87.51)	1528.96	(1529.87)
	ICS (CISD)	1.42	(1.44)	18.74	(18.60)	26.48	(26.18)	89.52	(89.90)	1535.81	(1523.23)
	Shift	0.13	(0.13)	1.61	(1.60)	0.67	(0.67)	3.01	(3.01)	6.44	(6.44)
	Shift, ICS	0.13	(0.13)	0.80	(0.79)	0.37	(0.36)	1.46	(1.47)	3.25	(3.25)

However, we observed that for the LCU cases, the ICS method performed less effectively than the FH cases. As shown in Appendix B, a scaled unitary fragment, $\beta_j\hat{U}_j$, is constructed by grouping mutually anticommuting Pauli operators, while commuting ones form a Hamiltonian fragment, $\hat{H}_j$. In general, because the anticommutation between Pauli operators occurs less frequently, there are fewer opportunities for the Pauli operators to be grouped to form a unitary. This makes the coefficient splitting with LCU less effective.

As shown in Fig. 2, we compare the perturbed ground state energies of the electronic Hamiltonian of H$_2$O, obtained by

Fig. 2 Histogram of the error of the estimated ground state energies $(\tilde{E}_0^{(n\to n')} - \text{EFCI})$ of the electronic structure Hamiltonian of H$_2$O with different measurement settings obtained from QKSD algorithm with the thresholding by eqn (45) and the finite number of shots of $M = 10^8$. The horizontal axis represents the errors in the atomic unit (mHa), and the vertical one denotes the frequency of the each histogram bin. The histogram is plotted using the perturbed QKSD energies from 10 000 independent and identical experiments. Here, the optimal thresholding of eqn (45) is applied to mitigate the sampling error, further reducing the Krylov dimension from $n = 9$ to $n' = 3$. In the FH case, two different values of $n' = 2, 3$ were observed across the random experiments, resulting in the two peaks in the histogram. $\tilde{E}^{(n\to n')}$ and $E^{(n\to n')}$ denote the QKSD ground state energies with and without considering the effect of sampling error, respectively. $E^{(n)}$ represents the QKSD ground state energy without error or thresholding.

the QKSD algorithm, with and without applying the reduction techniques. We observed the unperturbed QKSD energy, $E^{(n)}$, is close to the full configuration interaction (FCI) energy ($|E^{(n)} - E_{\text{FCI}}| \approx 0.1$ mHa).

We also employed a classical postprocessing called basis thresholding to alleviate the numerical instability of GEVP.[35,36] Since small singular values in S significantly amplify the perturbation to the eigenvalue $E^{(n)}$, the thresholding technique further projects the GEVP onto the singular basis of S with corresponding singular values larger than a certain value $\varepsilon_{\text{th}} > 0$. However, the thresholding also discards the information about the eigenstates, not only the error, which biases thresholded QKSD energy. Thus, by adjusting ε_{th}, the thresholding establishes a trade-off, reducing the effect of the statistical error in H and S while introducing additional projection error. Within this trade-off, the optimal ε_{th} was heuristically found in ref. 36 to be:

$$\varepsilon_{\text{th}} = \tilde{O}\left(n\Big/\sqrt{M_S}\right), \tag{45}$$

where M_S is the number of shots used to construct S with the Hadamard or extended swap test. Note that for the case of errors other than the sampling error are present, an automated thresholding[60] can be adopted. We denote n' as the dimension of the thresholded problem and $E^{(n\to n')}$ as corresponding eigenvalue.

Despite the bias caused by the projection error, the effect of the matrix perturbation is minimized. In the FH case, the application of the shifting technique and ICS resulted in the perturbed QKSD solution being concentrated within chemical accuracy ($|\tilde{E}_0^{(n\to n')} - E_{\text{FCI}}| < 1.6$ mHa), whereas most results without these techniques deviated beyond chemical accuracy.

VI. Conclusion

In this work, we analyzed the finite sampling errors that arise when projecting the Hamiltonian onto the quantum Krylov subspace with quantum algorithms. The measurement cost analyses of two scenarios, LCU and FH decomposition, converge to a unified perspective, where the decomposition rules and circuit construction are different. We also showed that the expected cost is analogous to the LCU 1-norm, which has the
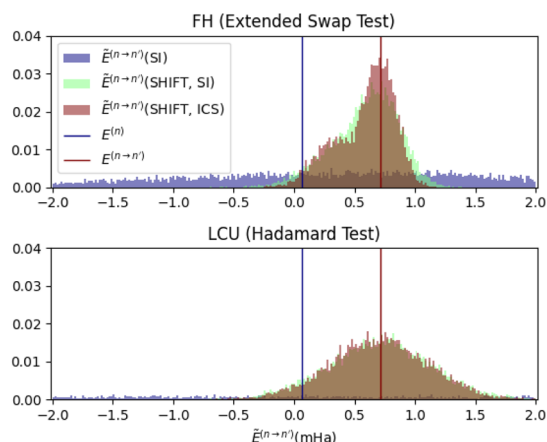
same definition in the context of block encoding.[53] Such analogies enable the translation of methods originally motivated to reduce the costs in LCU simulation[51,52] and the expectation measurement[39,40] to the problem of measuring matrix elements. Especially, adopting the symmetry shift[51,52] to the measurement problem eases the constraint on the shift operator, and thus provides larger cost reduction. Although the shifting technique is more effective in the LCU case, the measurement cost is observed to be lower in the FH case.

Despite of the effort to reduce the measurement cost, achieving the chemical accuracy by QKSD in practice remains still challenging, as the corresponding GEVP is often ill-conditioned. Note that in our application of QKSD to the H_2O system, the results were fitted within the chemical accuracy by using 10^8 shots, which can be considered expensive. In classical Krylov subspace diagonalization, the perturbation on the GEVP matrices mainly depends on the round off or floating point error, which decreases exponentially to the number of bits of the data type, although the calculation of the matrix elements takes exponentially long time. On the other hand, governed by the standard quantum limit, the matrix perturbation in QKSD decays with the square root of the number of measurements, which is much slower than the classical KSD, while each measurement takes polynomial time. Therefore, if exponential precision for the matrix element is required because of the ill-conditioning, it is not yet obvious that QKSD is superior to the classical counterpart in terms of running time to achieve a certain precision in the estimated eigenvalues.

As discussed in Section III, we observed that the extended swap test, enabled by FH decomposition, allows for the simultaneous measurement of both the overlap and Hamiltonian matrices:

$$\boldsymbol{S}_{0k} = \langle \Phi_{0k} | \hat{\mathbb{I}} \otimes (\hat{\sigma}_x + i\hat{\sigma}_y) | \Phi_{0k} \rangle,$$
$$\boldsymbol{H}_{0k} = \langle \Phi_{0k} | \hat{H} \otimes (\hat{\sigma}_x + i\hat{\sigma}_y) | \Phi_{0k} \rangle,$$

due to commutativity between $\hat{\mathbb{I}} \otimes \hat{\sigma}_{x(y)}$ and $\hat{H} \otimes \hat{\sigma}_{x(y)}$. This idea can be extended to the other EFTQC algorithms. Based on our knowledge, recently developed EFTQC algorithms other than QKSD focus on extracting the spectrum only from the autocorrelation function, $a(t) := \langle \phi_0 | e^{-i\hat{H}t} | \phi_0 \rangle$. However, those algorithms can be more refined by adopting the simultaneous measurement of $a_{\hat{O}}(t) := \langle \phi_0 | \hat{O} e^{-i\hat{H}t} | \phi_0 \rangle$ for some observable $\hat{O}$. For instance, the first derivative of $a(t)$ can be directly calculated from $a_{\hat{H}}(t)$. This measurement, which can be done precisely if the techniques introduced in this work are adopted, only requires overheads of $O(N_q^2)$ Clifford operations. These operations do not need additional logical qubits and present an endurable cost for EFTQCs. Given this rationale, our future research direction involves exploring EFTQC algorithms including $a_{\hat{O}}(t)$ in the spectrum extraction.

Code availability

The data and code that support the findings of this study are openly available in **https://github.com/snow0369/ofex**. The repository contains the full source code to reproduce the results presented in this work. All scripts and related data are licensed under MIT License, allowing for reuse and adaptation with proper attribution.

Data availability

The Python package for the techniques introduced in this work (shifting and iterative coefficient splitting) is available at ref. 61. Furthermore, the raw data for numerical results are available at ref. 62 along with the reproducing scripts.

Author contributions

GL: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft preparation, writing – review & editing, visualization, funding acquisition. SC: methodology, software. JH: conceptualization, investigation, resources, writing – review & editing, supervision, project administration, funding acquisition. AFI: conceptualization, methodology, formal analysis, investigation, resources, writing – review & editing, visualization, supervision, project administration, funding acquisition.

Conflicts of interest

There are no conflicts of interest to declare.

Appendices

Appendix A: averaged variance on H_{0k}

In this section, we derive the state-independent variance, represented by eqn (16) and (22), by averaging the states $|\phi_0\rangle$ and $|\phi_k\rangle$ over the independent uniform Haar distributions ($\phi_0 \sim \mathcal{H}(d), \phi_k \sim \mathcal{H}(d)$). Before proceeding with the derivation, we introduce three lemmas that will be instrumental in this process. Throughout this section, $\mathbb{E}_{\phi \sim \mathcal{H}(d)}[\cdot]$ is abbreviated as $\mathbb{E}_\phi[\cdot]$ unless otherwise mentioned.

Lemma 1: for any normal operator $\hat{A} \in \mathbb{C}^{d \times d}$, the following equality holds:

$$\mathbb{E}_{\phi_1, \phi_2}\left[\left| \langle \phi_1 | \hat{A} | \phi_2 \rangle \right|^2 \right] = \frac{1}{d^2} \mathrm{Tr}\left[\hat{A}^\dagger \hat{A} \right]. \tag{A1}$$

Proof: for any operator $\hat{X} \in \mathbb{C}^{d \times d}$, the averaged conjugation is known as

$$\mathbb{E}_{\hat{U} \sim \mathcal{U}(d)}\left[\hat{U}^\dagger \hat{X} \hat{U} \right] = \frac{\hat{I}}{d} \mathrm{Tr}\left[\hat{X} \right], \tag{A2}$$

as a consequence of Schur's lemma and the left and right invariance of the Haar measure, where $\mathcal{U}(d)$ is the uniform Haar distribution over the unitary group of dimension d. Therefore, we can state

$$\mathbb{E}_\phi\left[\langle \phi | \hat{X} | \phi \rangle \right] = \frac{\mathrm{Tr}\left[\hat{X} \right]}{d}. \tag{A3}$$

Finally, by applying eqn (A3) consecutively, we have:

$$\mathbb{E}_{\phi_1,\phi_2}\left[\left|\langle\phi_1|\hat{A}|\phi_2\rangle\right|^2\right] = \mathbb{E}_{\phi_1,\phi_2}\left[\langle\phi_1|\hat{A}|\phi_2\rangle\langle\phi_2|\hat{A}^\dagger|\phi_1\rangle\right]$$

$$= \frac{1}{d}\mathbb{E}_{\phi_2}\left[\mathrm{Tr}\left[\hat{A}|\phi_2\rangle\langle\phi_2|\hat{A}^\dagger\right]\right]$$

$$= \frac{1}{d}\mathbb{E}_{\phi_2}\left[\langle\phi_2|\hat{A}^\dagger\hat{A}|\phi_2\rangle\right]$$

$$= \frac{1}{d^2}\mathrm{Tr}\left[\hat{A}^\dagger\hat{A}\right].$$

Lemma 2: for any normal operator $\hat{A} \in \mathbb{C}^{d\times d}$, the following equality holds:

$$\mathbb{E}_{\phi_1,\phi_2}\left[\mathrm{Re}\left[\langle\phi_1|\hat{A}|\phi_2\rangle\right]^2\right] = \mathbb{E}_{\phi_1,\phi_2}\left[\mathrm{Im}\left[\langle\phi_1|\hat{A}|\phi_2\rangle\right]^2\right] \quad (A4)$$

Proof: because of the unitary-invariant property of Haar measure, $\mathbb{E}_\phi[f(|\phi\rangle)]$ is always identical to $\mathbb{E}_\phi[f(e^{i\pi/2}|\phi\rangle)] = \mathbb{E}_\phi[f(i|\phi\rangle)]$ for any function f. Therefore, the following proves eqn (A4):

$$\mathbb{E}_{\phi_1,\phi_2}\left[\mathrm{Re}\left[\langle\phi_1|\hat{A}|\phi_2\rangle\right]^2 - \frac{1}{2}\left|\langle\phi_1|\hat{A}|\phi_2\rangle\right|^2\right] \quad (A5)$$

$$= \frac{1}{4}\mathbb{E}_{\phi_1,\phi_2}\left[\langle\phi_1|\hat{A}|\phi_2\rangle^2 + \langle\phi_2|\hat{A}^\dagger|\phi_1\rangle^2\right] \quad (A6)$$

$$= \frac{1}{4}\mathbb{E}_{\phi_1,\phi_2}\left[\langle\phi_1|i\hat{A}|\phi_2\rangle^2 + \langle\phi_2|(-i)\hat{A}^\dagger|\phi_1\rangle^2\right] \quad (A7)$$

$$= \mathbb{E}_{\phi_1,\phi_2}\left[\mathrm{Im}\left[\langle\phi_1|\hat{A}|\phi_2\rangle\right]^2 - \frac{1}{2}\left|\langle\phi_1|\hat{A}|\phi_2\rangle\right|^2\right], \quad (A8)$$

where eqn (A6) and (A8) are obtained using $\mathrm{Re}[z]^2 = \frac{1}{4}[z^2 + z^{*2} + 2|z|^2]$ and $\mathrm{Im}[z]^2 = \frac{1}{4}[-z^2 - z^{*2} + 2|z|^2]$, respectively, for $z \in \mathbb{C}$. Additionally, eqn (A7) is derived by replacing $|\phi_2\rangle$ with $i|\phi_2\rangle$.

Lemma 3: for any normal operator $\hat{A} \in \mathbb{C}^{d\times d}$, the following equality holds:

$$\mathbb{E}_\phi\left[\left|\langle\phi|\hat{A}|\phi\rangle\right|^2\right] = \frac{\mathrm{Tr}\left[\hat{A}^\dagger\hat{A}\right] + \left|\mathrm{Tr}\left[\hat{A}\right]\right|^2}{d(d+1)}. \quad (A9)$$

Proof: because $\hat{A}$ is normal, we can perform the eigendecomposition to an arbitrary state as $|\phi\rangle = \sum_{i=1}^d \gamma_i|\psi_i\rangle$, where $\hat{A}|\psi_i\rangle = a_i|\psi_i\rangle$ and $\gamma_i = \langle\psi_i|\phi\rangle$. Then the Haar averaging over $|\phi\rangle$ is identical to the averaging $\{\gamma_i\}_{i=1}^d$ over d-dimensional complex unit sphere, $\mathbb{E}_\phi[\cdot] = \mathbb{E}_{\{\gamma_i\}\sim\mathbb{CP}^{d-1}}[\cdot]$, where $\mathbb{CP}$ denotes the complex projective space. Therefore, it can be shown that

$$\mathbb{E}_\phi\left[\left|\langle\phi|\hat{A}|\phi\rangle\right|^2\right] \quad (A10)$$

$$= \mathbb{E}_{\{\gamma_i\}\sim\mathbb{CP}^{d-1}}\left[\left|\sum_{i=1}^d a_i|\gamma_i|^2\right|^2\right] \quad (A11)$$

$$= \sum_i |a_i|^2\mathbb{E}_\gamma\left[|\gamma_i|^4\right] + \sum_{i\neq j} a_i^* a_j\mathbb{E}_\gamma\left[|\gamma_i|^2|\gamma_j|^2\right] \quad (A12)$$

$$= \frac{1}{d(d+1)}\left(2\sum_i |a_i|^2 + \sum_{i\neq j} a_i^* a_j\right) \quad (A13)$$

$$= \frac{1}{d(d+1)}\left(\sum_i |a_i|^2 + \left(\sum_i a_i^*\right)\left(\sum_i a_i\right)\right) \quad (A14)$$

$$= \frac{1}{d(d+1)}\left(\mathrm{Tr}\left[\hat{A}^\dagger\hat{A} + \mathrm{Tr}\left[\hat{A}^\dagger\right]\mathrm{Tr}\left[\hat{A}\right]\right]\right), \quad (A15)$$

where $\mathbb{E}_\gamma[\cdot]$ denotes $\mathbb{E}_{\{\gamma_i\}\sim\mathbb{CP}^{d-1}}[\cdot]$ and eqn (A13) holds because of ref. 63, Theorem 2.6.

For the case of LCU decomposition, the Haar-averaged partial variance in eqn (14) is identically determined over X and j as

$$\mathbb{E}_{\phi_0,\phi_k}\left[\mathrm{Var}\left[\hat{O}_X\right]_{\Phi_{0k;j}}\right] = 1 - \frac{1}{2d}, \quad (A16)$$

by applying Lemmas 1 and 2 with $\hat{A} = \hat{U}_j$. Therefore, the sub-optimal shot allocation,

$$m_{jX}^{(\mathrm{subopt})} = \frac{\beta_j}{2\|\boldsymbol{\beta}\|_1}, \quad (A17)$$

is obtained by replacing $\mathrm{Var}[\hat{O}_X]\phi_{0k;j}$ in eqn (15) with the expected variance, where the denominator is set to satisfy the normalization constraint, $\sum_{jX} m_{jX}^{(\mathrm{subopt})} = 1$.

The variance with the sub-optimal shot allocation is derived from eqn (13) by assigning eqn (A17), which is

$$\mathrm{Var}^{(\mathrm{LCU})}\left[\boldsymbol{H}_{0k};\boldsymbol{m}^{(\mathrm{subopt})}\right] = \frac{1}{M}\sum_{j=1}^{N_\beta} 2\|\boldsymbol{\beta}\|_1^2\left(2 - \left|\langle\phi_0|\hat{U}_j|\phi_k\rangle\right|^2\right). \quad (A18)$$

Finally, by Lemma 1, the Haar averaging of $|\langle\phi_0|\hat{U}_j|\phi_k\rangle|^2$ both over ϕ_0 and ϕ_k results in eqn (16).

Furthermore, because $|\phi_k\rangle = \hat{B}^k|\phi_0\rangle$ is an evolved state from $|\phi_0\rangle$, we can consider averaging the variance over single state $|\phi_0\rangle$, fixing the evolution operator, $\hat{B}^k$. Then from eqn (A18), $|\langle\phi_0|\hat{U}_j|\phi_k\rangle|^2 = |\langle\phi_0|\hat{U}_j\hat{B}^k|\phi_0\rangle|^2$ needs to be averaged over $|\phi_0\rangle$. Using Lemma 3 with $\hat{A} = \hat{U}_j\hat{B}^k$, we can show that

$$\mathbb{E}_{\phi_0}\left[\left|\langle\phi_0|\hat{U}_j\hat{B}^k|\phi_0\rangle\right|\right] = \frac{d + \left|\mathrm{Tr}\left[\hat{U}_j\hat{B}^k\right]\right|^2}{d(d+1)},$$

and thus

$$\frac{1}{d+1} \leq \mathbb{E}_{\phi_0}\left[\left|\langle\phi_0|\hat{U}_j\hat{B}^k|\phi_0\rangle\right|\right] \leq 1,$$

because $0 \leq |\mathrm{Tr}[\hat{U}_j\hat{B}^k]|^2 \leq d^2$. Finally, the total averaged variance is bounded by

$$\frac{2\|\boldsymbol{\beta}\|_1{}^2}{M} \le \mathbb{E}_{\phi_0}\left[\text{Var}^{(\text{LCU})}\right] \le \frac{2\|\boldsymbol{\beta}\|_1{}^2}{M}\left(2 - \frac{1}{d+1}\right), \quad (A19)$$

which is approximately less than the expected variance considering the two states independently (eqn (16).)

In FH case, the shots for real and imaginary parts are identically allocated ($m_{jR} = m_{jI} =: m_j/2$) because of Lemma 2, similar to the LCU case. Therefore, the variances for the real and imaginary parts in eqn (21) can be directly added, yielding

$$\text{Var}^{(\text{FH})} = \sum_{j=1}^{N_\gamma} \frac{2}{m_j}\left(\text{Var}\left[\hat{O}_{jR}\right]_{\Phi_{0,k}} + \text{Var}\left[\hat{O}_{jI}\right]_{\Phi_{0,k}}\right)$$

$$= \sum_{j=1}^{N_\gamma} \frac{2}{m_j}\left(2\left\langle \hat{O}_{jX}{}^2\right\rangle_{\Phi_{0k}} - \left|\langle\phi_0|\hat{H}_j|\phi_k\rangle\right|^2\right). \quad (A20)$$

In eqn (21), the expected second moment is determined as $\mathbb{E}_{\phi_0,\phi_k}[\langle\hat{O}_{jX}{}^2\rangle_{\Phi_{0k}}] = \text{Tr}[\hat{H}_j{}^2]/d$, while the last term, $\mathbb{E}_{\phi_0,\phi_k}[|\langle\phi_0|\hat{H}|\phi_k\rangle|^2] = \text{Tr}[\hat{H}_j{}^2]/d^2$ is obtained using Lemma 1. Finally, the expected variance of jth fragment is derived as below:

$$\mathbb{E}_{\phi_0,\phi_k}\left[\text{Var}\left[\hat{O}_{jR}\right]_{\Phi_{0,k}} + \text{Var}\left[\hat{O}_{jI}\right]_{\Phi_{0,k}}\right]$$

$$= \frac{1}{d}\text{Tr}\left[\hat{H}_j{}^2\right]\left(2 - \frac{1}{d}\right), \quad (A21)$$

which results in a total averaged variance of eqn (22) with the sub-optimal allocation $m_j \propto \text{Tr}[\hat{H}_j{}^2]^{1/2}$.

Similar to the LCU case, if we take the expectation only on $|\phi_0\rangle$, we have

$$\mathbb{E}_{\phi_0}\left[\text{Var}^{(\text{FH})}\right] \le \frac{2\|\boldsymbol{\gamma}\|^2}{M}\left(2 - \frac{1}{d+1}\right). \quad (A22)$$

Appendix B: grouping Pauli operators

In this section, we review the partitioning of qubit Hamiltonian in a form of LCU and FH. Also, the expressions of $\|\boldsymbol{\beta}\|_1$ and $\|\boldsymbol{\gamma}\|_1$ are presented with respect to the Pauli coefficients.

A qubit Hamiltonian $\hat{H}$ is expressed as

$$\hat{H} = \sum_{p=1}^{N_P} \alpha_p \hat{P}_p, \quad (B1)$$

where $\hat{P}_p \in \{\hat{I}, \hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z\}^{\otimes N_q}$ is an N_q-qubit Pauli operator and its coefficient is $\alpha_p \in \mathbb{R}$.

First, we review the derivation of LCU in Pauli basis,[54,55] which is described as

$$\hat{H} = \sum_{j=1}^{N_\beta} \beta_j \hat{U}_j = \sum_{j=1}^{N_\beta} \sum_{p\in\mathcal{A}_j} \alpha_p{}^{(\mathcal{A}_j)} \hat{P}_p. \quad (B2)$$

Here, a Pauli term $\alpha_p \hat{P}_p$ can be separated to the multiple groups $\{\mathcal{A}_j\}$. Thus, $\alpha_p = \sum_{j:p\in\mathcal{A}_j} \alpha_p{}^{(\mathcal{A}_j)}$ for all $1 \le p \le N_p$, should be imposed for the partitioned coefficients. Also, if the anti-commutation conditions of different Pauli operators within a group holds:

$$\left\{\hat{P}_p, \hat{P}_q\right\} = 2\delta_{pq}\hat{I} \quad \forall p, q \in \mathcal{A}_j, \forall j, \quad (B3)$$

each partition, $\sum_{p\in\mathcal{A}_j} \alpha_p{}^{(\mathcal{A}_j)} \hat{P}_p$, becomes a scaled unitary, $\beta_j \hat{U}_j$. Then, $\beta_j{}^2 = \sum_{p\in\mathcal{A}_j} \alpha_p{}^{(\mathcal{A}_j)2}$ is found consequently, and the norm becomes

$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{N_\beta}\left(\sum_{p\in\mathcal{A}_j} \alpha_p{}^{(\mathcal{A}_j)2}\right)^{1/2}. \quad (B4)$$

Furthermore, a circuit realization of controlled $\hat{U}_j$ is known, which requires $O(N_q|\mathcal{A}_j|)$ two-qubit gates.[55]

Now, for the case of FH, Pauli operators are grouped together if all pairs of the operators in a group commute:

$$\hat{H} = \sum_{j=1}^{N_\gamma} \hat{H}_j = \sum_{j=1}^{N_\gamma} \sum_{p\in\mathcal{C}_j} \alpha_p{}^{(\mathcal{C}_j)} \hat{P}_p, \quad (B5)$$

where

$$\left[\hat{P}_p, \hat{P}_q\right] = 0 \quad \forall p, q \in \mathcal{C}_j, \forall j, \quad (B6)$$

and $\alpha_p = \sum_{j:p\in\mathcal{C}_j} \alpha_p{}^{(\mathcal{C}_j)}$. For each of $\hat{H}_j$, one can construct a diagonalizing Clifford circuit, $\hat{V}_j$, taking $O(N_q \min(N_q, |\mathcal{C}_j|))$ two-qubit gates. Furthermore, one can find the FH decomposition norm, $\|\boldsymbol{\gamma}\|_1$ in terms of Pauli coefficients:

$$\|\boldsymbol{\gamma}\|_1 = \frac{1}{\sqrt{d}}\sum_{j=1}^{N_\gamma} \text{Tr}\left[\hat{H}_j{}^2\right]^{1/2} = \sum_{j=1}^{N_\gamma}\left(\sum_{p\in\mathcal{C}_j} \alpha_p{}^{(\mathcal{C}_j)2}\right)^{1/2}. \quad (B7)$$

Here, note that we used $\frac{1}{d}\text{Tr}[\hat{P}_p\hat{P}_q] = \delta_{p,q}$ because the trace of Pauli operators except the identity is zero.

The two norms described in eqn (B4) and (B7) shares a common expression as the sum of the L2 norms of Pauli coefficients. We can consider decomposing eqn (B1) without grouping, $\hat{H}_j = \beta_j \hat{U}_j = \alpha_j \hat{P}_j$, which is also an LCU and an FH simultaneously. However, in such case, the norm $\|\boldsymbol{\beta}\|_1 = \|\boldsymbol{\gamma}\|_1 = \|\boldsymbol{\alpha}\|_1$ becomes larger than that of grouped Pauli case because the sum of L2 norms of grouped Pauli coefficients is smaller than the L1 norm.

Pauli groupings are not unique and can be translated to a clique covering problem on the (anti-)commutation graph. The commutation graph, denoted as $G_\mathcal{C} = (V, E_\mathcal{C})$, encodes the commutativity between Pauli operators. Specifically, the nodes in V correspond to individual Pauli operators: $V = \{v_p : 1 \le p \le N_P\}$ with a node weight function $w(v_p) = \alpha_p$. Also, undirected and unweighted edges connect nodes whose corresponding operators commute ($E_\mathcal{C} = \{(v_p, v_q) : [\hat{P}_p, \hat{P}_q] = 0\}$). The anti-commutation graph, $G_\mathcal{A} = (V, E_\mathcal{A})$, shares the same node set with $G_\mathcal{C}$ but connects nodes with anticommuting operators ($E_\mathcal{A} = \{(v_p, v_q) : \{\hat{P}_p, \hat{P}_q\} = 0\}$). Because any two Pauli operators either commute or anticommute, these graphs are complements, meaning that $E_\mathcal{A}^c = E_\mathcal{C}$.

In such setting, minimizing eqn (B4) and (B7) translates to finding a clique covering that minimizes the sum of the clique weights. Each weight is defined as a L2 norm of node weights covered by each clique. Like other clique covering problems, this is an NP-hard problem. However, a heuristic and greedy algorithm called SORTED INSERTION often outperform other

heuristic algorithms as demonstrated in ref. 39. The original work on SORTED INSERTION aimed to minimize the cost of measuring standard expectation values, so it only considered the commutation graph. However, we extend the algorithm to the anticommutation case to reduce the cost of the LCU measurements by simply modifying its grouping condition.

In the numerical results of Section V, the $\hat{Z}$-type operators, which dominate the example Hamiltonians, correspond to the number operators in the fermionic representation. In the LCU decomposition, these operators are not grouped due to their mutual commutativity, resulting in a larger LCU norm than FH norm ($\|\boldsymbol{\beta}_{\mathrm{SI}}(\hat{H})\| \geq \|\boldsymbol{\gamma}_{\mathrm{SI}}(\hat{H})\|$) because large coefficients are treated separately. However, the shifting technique effectively eliminates these operators, significantly reducing the norm $\|\boldsymbol{\beta}_{\mathrm{SI}}(\hat{H} - \hat{T})\|$. In the FH decomposition, $\hat{Z}$-type operators are naturally grouped together, which contribute to the norm less significantly than the LCU case. Consequently, the impact of the shifting technique is less pronounced in the FH case compared to the LCU case. Nevertheless, the shifted FH decomposition still achieves a smaller norm than the shifted LCU decomposition.

Appendix C: Pauli covariance

In this section derives the Pauli covariance for measuring the matrix element (eqn (41)). Starting from the partial variance in eqn (25), and substituting $\hat{N}_j = \sum_{p \in \mathcal{G}_j} \alpha_p^{(\mathcal{G}_j)} \hat{P}_p$, which is analogous to eqn (B2) or (B5), we obtain:

$$\mathrm{Var}\big[R_{0k;j}\big] = \sum_{p,q \in \mathcal{G}_j} \alpha_p^{(\mathcal{G}_j)} \alpha_q^{(\mathcal{G}_j)} \left[\frac{1}{2}\left(\langle\phi_0|\hat{P}_p\hat{P}_q|\phi_0\rangle + \langle\phi_k|\hat{P}_p\hat{P}_q|\phi_k\rangle\right) - \mathrm{Re}\left[\langle\phi_0|\hat{P}_p|\phi_k\rangle\right]\mathrm{Re}\left[\langle\phi_0|\hat{P}_q|\phi_k\rangle\right]\right]. \quad \text{(C1)}$$

Comparing eqn (40) and (C1) suggests defining the covariance as the expression within the bracket. However, this definition violates the symmetric property of the covariance because $\hat{P}_p\hat{P}_q$ is not necessarily identical to $\hat{P}_q\hat{P}_p$. Therefore, to ensure the symmetry, we superpose the product of operators in both orders, (p, q) and (q, p). This results in the definition of the covariance as eqn (41).

Appendix D: efficient shifting technique for electronic structure problem

Here, we describe an efficient procedure of optimizing eqn (28) for an electronic structure Hamiltonian. If $\hat{T}$ is given as eqn (34), the number of real parameters determining $\hat{T}$ is:

$$N_{\mathrm{param}} = (N_{\mathrm{occ}} + N_{\mathrm{virt}})\left(1 + \binom{N_{\mathrm{orb}} - 1}{2}\right), \quad \text{(D1)}$$

where N_{occ} and N_{virt} denote the number of occupied and virtual orbitals, respectively. This number scales as $O(N_{\mathrm{orb}}^3)$ and makes the optimization computationally expensive. In the rest of the

section, we show an alternative and efficient method for the minimization of the norm by the shift technique.

We consider that the Hamiltonian $\hat{H}$ is given as an electronic structure Hamiltonian in eqn (33), whose the indices are determined for the unique terms:

$$\hat{H} = \sum_{r \leq s}^{N_{\mathrm{orb}}} h_{rs}\hat{E}_{rs} + \sum_{(p,q,r,s)\in\mathcal{P}} g_{pqrs}\left(\hat{E}_{pq}\hat{E}_{rs} + \hat{E}_{rs}\hat{E}_{pq}\right), \quad \text{(D2)}$$

where

$$\mathcal{P} := \left\{\begin{array}{l} (p,q,r,s) : (1 \leq p \leq q \leq N_{\mathrm{orb}}) \\ \wedge(1 \leq r \leq s \leq N_{\mathrm{orb}}) \\ \wedge((p < r) \vee ((p = r) \wedge (q \leq s))) \\ \wedge \neg(p = q = r = s) \end{array}\right\}.$$

The set $\mathcal{P}$ with the size of $|\mathcal{P}| \approx N_{\mathrm{orb}}^4/8$ represents the indices of Hermitian two-body operators avoiding the duplication from the following eight-fold symmetries:

$$g_{pqrs} = g_{qprs} = g_{pqsr} = g_{qpsr} = g_{rspq} = g_{srpq} = g_{rsqp} = g_{srqp},$$

and squared number operators ($p = q = r = s$), which are absorbed into one-body number operators due to the idempotent property ($\hat{n}_p^2 = \hat{n}_p$).

In order to realize the efficient optimization, we find a part affected by the shift among the terms in eqn (D2). The terms relevant to eqn (34) are selected to construct the partial Hamiltonian, $\hat{H}^{(s)}$:

$$\hat{H}^{(s)} = \sum_{r\in\mathcal{F}} 2\tilde{h}_{rr}\hat{n}_r + \sum_{r<s}^{N_{\mathrm{orb}}} \tilde{h}_{rs}\hat{E}_{rs} + \sum_{q\in\mathcal{F}}\sum_{rs\in\mathcal{E}_q} \tilde{g}_{rsqq}\hat{E}_{rs}\hat{n}_q, \quad \text{(D3)}$$

where the modified coefficients are

$$\tilde{h}_{rs} = h_{rs} + \delta_{r\neq s}\sum_{q\in[N_{\mathrm{orb}}]\backslash\{r,s\}} g_{rqsq}$$
$$\tilde{g}_{rsqq} = 4g_{rsqq} - 2g_{rqsq}.$$

Such coefficients are determined by the following transformation to make the form of the Hamiltonian consistent to the shift operator:

$$g_{rqsq}\hat{E}_{rq}\hat{E}_{sq} + g_{rsqq}\hat{E}_{rs}\hat{E}_{qq} + \text{h.c.}$$
$$= g_{rqsq}\hat{E}_{rs} + \left(4g_{rsqq} - 2g_{rqsq}\right)\hat{E}_{rs}\hat{n}_q, \quad \text{(D4)}$$

for all $r \neq q$ and $s \neq q$. Because the rest of the Hamiltonian ($\hat{H} - \hat{H}^{(s)}$) is invariant to the shifting with respect to the term-wise grouping algorithm, denoted as $\boldsymbol{T}$, we focus on the following norm minimization:

$$\min_{\tau}\left\|\gamma_T\left(\hat{H}^{(s)} - \hat{T}(\tau)\right)\right\|. \tag{D5}$$

Here, $\hat{H}^{(s)} - \hat{T}(\tau)$ is given as

$$\hat{H}^{(s)} - \hat{T}(\tau) = \sum_{r \in \mathcal{F}}\left(2\tilde{h}_{rr} - \tau_r^{(1)}\right)\hat{n}_r$$

$$+ \sum_{r<s}\left(\tilde{h}_{rs} + \sum_{q \in \text{occ}\backslash\{r,s\}}\tau_{qrs}^{(2)}\right)\hat{E}_{rs}$$

$$+ \sum_{q \in \mathcal{F}}\sum_{rs \in \mathcal{E}_q}\left(\tilde{g}_{rsqq} - \tau_{qrs}^{(2)}\right)\hat{E}_{rs}\hat{n}_q. \tag{D6}$$

Note that the each operators in eqn (D6) contributes to the norm as shown below:

$$\frac{1}{\sqrt{d}}\text{Tr}\left[\hat{n}_r^2\right]^{1/2} = \frac{1}{\sqrt{2}},$$

$$\frac{1}{\sqrt{d}}\text{Tr}\left[\hat{E}_{rs}^2\right]^{1/2} = \sqrt{\frac{2}{d}}\text{Tr}\left[\left(\hat{E}_{rs}\hat{n}_q\right)^2\right]^{1/2} = \sqrt{\frac{(3\delta_{rs}+1)}{2}}, \tag{D7}$$

for $r \neq q$ and $s \neq q$. Therefore, the norm with term-wise grouping is determined as

$$\left\|\gamma_T\left(\hat{H}^{(s)} - \hat{T}(\tau)\right)\right\| = \sum_{r \in \mathcal{F}}\frac{1}{\sqrt{2}}\left|2\tilde{h}_{rr} - \tau_r^{(1)}\right|$$

$$+ \sum_{r<s}\frac{1}{\sqrt{2}}\left|\tilde{h}_{rs} + \sum_{q \in \text{occ}\backslash\{r,s\}}\tau_{qrs}^{(2)}\right|$$

$$+ \sum_{q \in \mathcal{F}}\sum_{rs \in \mathcal{E}_q}\sqrt{\frac{3\delta_{rs}+1}{2}}\left|\tilde{g}_{rsqq} - \tau_{qrs}^{(2)}\right|. \tag{D8}$$

The first summation becomes zero by assigning

$$\tau_r^{(1)} = 2\tilde{h}_{rr} \quad \forall r \in \mathcal{F}. \tag{D9}$$

Furthermore, because the variables, $\tau_{qrs}^{(2)}$ for $(q,r,s) \in \mathcal{T}$ such that

$$\mathcal{T} = \{(q,r,s) : 1 \le r = s \le N_{\text{orb}}, q \in \mathcal{F}\backslash\{r\}\}$$
$$\cup\{(q,r,s): 1 \le r \le s \le N_{\text{orb}}, q \in \text{virt}\backslash\{r,s\}\}, \tag{D10}$$

are only included in the third summation, they are also determined as

$$\tau_{qrs}^{(2)} = \tilde{g}_{rsqq} \quad \forall (q,r,s) \in \mathcal{T}. \tag{D11}$$

However, the rest of the variables, $\tau_{qrs}^{(2)}$ for $(q,r,s) \in \mathcal{T}^c$ occur both in the second and the third summations, where

$$\mathcal{T}^c = \{(q,r,s) : 1 \le r < s \le N_{\text{orb}}, q \in \text{occ}\backslash\{r,s\}\}. \tag{D12}$$

After the assignment of eqn (D9) and (D11), the minimization problem of eqn (D8) is then reduced to

$$\left\|\gamma_T\left(\hat{H}^{(s)} - \hat{T}'\left(\tau_{\mathcal{T}^c}^{(2)}\right)\right)\right\| = \frac{1}{\sqrt{2}}\sum_{r<s}\left[\left|\tilde{h}_{rs} + \sum_{q \in \text{occ}\backslash\{r,s\}}\tau_{qrs}^{(2)}\right|\right.$$

$$\left. + \sum_{q \in \text{occ}\backslash\{r,s\}}\left|\tilde{g}_{rsqq} - \tau_{qrs}^{(2)}\right|\right] \tag{D13}$$

where $\hat{T}'(\tau_{\mathcal{T}^c}^{(2)})$ denotes the shift operator $\hat{T}$ with the partial assignment and $\tau_{\mathcal{T}^c}^{(2)}$ denotes the set of variables, $\tau_{qrs}^{(2)}$ with $(q,r,s) \in \mathcal{T}^c$. Furthermore, minimizing eqn (D13) is identical to the separated optimizations:

$$\min_{\boldsymbol{\mu}^{(rs)} \in \mathbb{R}^{|\mathcal{U}_{rs}|}}\left(P_{rs}\left(\boldsymbol{\mu}^{(rs)}\right) + \|\boldsymbol{\mu}^{rs}\|_1\right) \tag{D14}$$

for each (r,s) with respect to the variables, $\boldsymbol{\mu}^{(rs)} = \{\mu_q^{(rs)} = \tilde{g}_{rsqq} - \tau_{qrs}^{(2)}: q \in \mathcal{U}_{rs}\}$, where $\mathcal{U}_{rs} = \text{occ}\backslash\{r,s\}$ and

$$P_{rs}\left(\boldsymbol{\mu}^{(rs)}\right) = \left|A_{rs} - \sum_{q \in \mathcal{U}_{rs}}\mu_q^{(rs)}\right|, \tag{D15}$$

(a) Pictorial analysis for a fixed $\boldsymbol{\mu}^\star$ (b) Optima with a fixed 1-norm   (c) Final optimal solutions
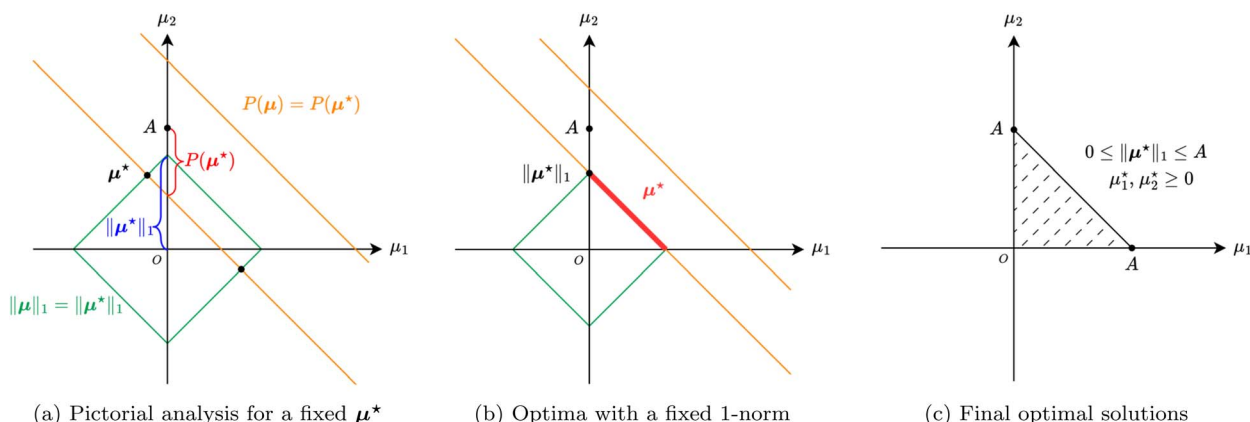
Fig. 3 A pictorial approach for the optimization of eqn (D14). For the simplicity, the super- and subscripts (rs) are omitted, and it is assumed that $\boldsymbol{\mu}$ is two-dimensional and $A > 0$. In figure (a), for a fixed $\boldsymbol{\mu}^\star$, the green diamond-shaped plot and the orange parallel lines denote all the points of $\boldsymbol{\mu}$ such that $\|\boldsymbol{\mu}\|_1 = \|\boldsymbol{\mu}^\star\|_1$ and $P(\boldsymbol{\mu}) = P(\boldsymbol{\mu}^\star)$, respectively. It is shown that $P(\boldsymbol{\mu}^\star) + \|\boldsymbol{\mu}^\star\|_1 \ge A$ must be satisfied for the intersections between two plots to be established, and thus $\boldsymbol{\mu}^\star$ exists. In figure (b), for a fixed $\|\boldsymbol{\mu}\|_1$, the optimal points are found, as indicated by the red line, with the corresponding optimal value of $P(\boldsymbol{\mu}^\star) + \|\boldsymbol{\mu}^\star\|_1 = A$. By varying $\|\boldsymbol{\mu}\|_1$, these points are expanded to the shaded area in figure (c). For the case of $A \le 0$, the plots are symmetrically transposed about the origin, resulting in the optimal points of $0 \le \|\boldsymbol{\mu}^\star\|_1 \le -A$ and $\mu_1^\star, \mu_2^\star \le 0$.

and $A_{rs} = \tilde{h}_{rs} + \sum_{q \in \mathcal{U}_{rs}} \tilde{g}_{rsqq}$. We provide a pictorial procedure to solve eqn (D14) in Fig. 3. Although the solution for 2-dimensional problem is shown, this can be extended to $|\mathcal{U}_{rs}|$-dimensional problem by considering hyperplane instead of lines in the figure. Every $\tau_{\mathcal{T}^c}^{(2)\star}$ satisfying

$$\sum_{q \in \mathcal{U}_{rs}} \left| \tau_{qrs}^{(2)\star} - \tilde{g}_{rsqq} \right| \leq |A_{rs}|,$$
$$\text{sign}\left(\tau_{qrs}^{(2)\star} - \tilde{g}_{rsqq} \right) = \text{sign}(A_{rs}) \tag{D16}$$

for all r and s, identically results in the optimal value of eqn (D14), which is

$$\left\| \gamma_T \left(\hat{H}^{(s)} - \hat{T}'\left(\tau_{\mathcal{T}^c}^{(2)\star} \right) \right) \right\| = \frac{1}{\sqrt{2}} \sum_{r<s} \left| \tilde{h}_{rs} + \sum_{q \in \text{occ}\backslash\{r,s\}} \tilde{g}_{rsqq} \right|. \tag{D17}$$

Note that assigning $\tau_{qrs}^{(2)\star} = \tilde{g}_{rsqq}$ for all $r, s, q \in \mathcal{T}^c$ satisfies eqn (D16), which is analogous to eqn (D11).

Therefore, we conclude that the parameters

$$\tau_r^{(1)} = 2h_{rr} \tag{D18}$$

$$\tau_{qrs}^{(2)} = 4g_{rsqq} - 2g_{rqsq} \tag{D19}$$

lead to the optimal shift with respect to the term-wise grouping algorithm. However, strictly to say, the above simplification only holds for the fragment Hamiltonian, as $\hat{n}_r$ and $\hat{E}_{rs}\hat{n}_q$ are Hermitian, not unitary. Although $\hat{n}_q$ can be written as a unitary operator ($\hat{r}_q = 2\hat{n}_q - 1$), our current understanding of representing a unitary operator as a linear combination of one-body excitation operators remains insufficient to establish an analogous reduction for the LCU case. Therefore, this work adopts the same shift operator for both LCU decomposition as an interim solution, leaving the parameter reduction for the LCU case as a future work.

Acknowledgements

References

1 K. Wintersperger, F. Dommert, T. Ehmer, A. Hoursanov, J. Klepsch, W. Mauerer, et al., Neutral atom quantum computing hardware: performance and end-user perspective, EPJ Quantum Technol., 2023, 10(1), 32.

2 S. Debnath, N. M. Linke, C. Figgatt, K. A. Landsman, K. Wright and C. Monroe, Demonstration of a small programmable quantum computer with atomic qubits, Nature, 2016, 536(7614), 63–66.

3 M. F. Brandl, M. W. van Mourik, L. Postler, A. Nolf, K. Lakhmanskiy, R. R. Paiva, et al., Cryogenic setup for trapped ion quantum computing, Rev. Sci. Instrum., 2016, 87(11), 113103.

4 X. L. Wang, L. K. Chen, W. Li, H. L. Huang, C. Liu, C. Chen, et al., Experimental Ten-Photon Entanglement, Phys. Rev. Lett., 2016, 117, 210502.

5 X. Qiang, X. Zhou, J. Wang, C. M. Wilkes, T. Loke, S. O'Gara, et al., Large-scale silicon quantum photonics implementing arbitrary two-qubit processing, Nat. Photonics, 2018, 12(9), 534–539.

6 J. M. Chow, J. M. Gambetta, A. D. Córcoles, S. T. Merkel, J. A. Smolin, C. Rigetti, et al., Universal Quantum Gate Set Approaching Fault-Tolerant Thresholds with Superconducting Qubits, Phys. Rev. Lett., 2012, 109, 060501.

7 G. Wendin, Quantum information processing with superconducting circuits: a review, Rep. Prog. Phys., 2017, 80(10), 106001.

8 Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferová, et al., Quantum Chemistry in the Age of Quantum Computing, Chem. Rev., 2019, 119(19), 10856–10915, PMID: 31469277.

9 S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin and X. Yuan, Quantum computational chemistry, Rev. Mod. Phys., 2020, 92, 015003.

10 B. Bauer, S. Bravyi, M. Motta and G. K. L. Chan, Quantum Algorithms for Quantum Chemistry and Quantum Materials Science, Chem. Rev., 2020, 120(22), 12685–12717, PMID: 33090772.

11 J. Preskill, Quantum Computing in the NISQ era and beyond, Quantum, 2018, 2, 79.

12 K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, et al., Noisy intermediate-scale quantum algorithms, Rev. Mod. Phys., 2022, 94, 015004.

13 M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, et al., Variational quantum algorithms, Nat. Rev. Phys., 2021, 3(9), 625–644.

14 J. Tilly, H. Chen, S. Cao, D. Picozzi, K. Setia, Y. Li, et al., The Variational Quantum Eigensolver: A review of methods and best practices, Phys. Rep., 2022, 986, 1–128.

15 J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush and H. Neven, Barren plateaus in quantum neural network training landscapes, *Nat. Commun.*, 2018, **9**(1), 4812.

16 M. Cerezo, A. Sone, T. Volkoff, L. Cincio and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, *Nat. Commun.*, 2021, **12**(1), 1791.

17 S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, *et al.*, Noise-induced barren plateaus in variational quantum algorithms, *Nat. Commun.*, 2021, **12**(1), 6961.

18 M. Cerezo and P. J. Coles, Higher order derivatives of quantum neural networks with barren plateaus, *Quantum Sci. Technol.*, 2021, **6**(3), 035006.

19 A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio and P. J. Coles, Effect of barren plateaus on gradient-free optimization, *Quantum*, 2021, **5**, 558.

20 X. Xue, M. Russ, N. Samkharadze, B. Undseth, A. Sammak, G. Scappucci, *et al.*, Quantum logic with spin qubits crossing the surface code threshold, *Nature*, 2022, **601**(7893), 343–347.

21 R. Blume-Kohout, J. K. Gamble, E. Nielsen, K. Rudinger, J. Mizrahi, K. Fortier, *et al.*, Demonstration of qubit operations below a rigorous fault tolerance threshold with gate set tomography, *Nat. Commun.*, 2017, **8**(1), 14485.

22 L. Postler, S. Heußen, I. Pogorelov, M. Rispler, T. Feldker, M. Meth, *et al.*, Demonstration of fault-tolerant universal quantum gate operations, *Nature*, 2022, **605**(7911), 675–680.

23 D. Bluvstein, S. J. Evered, A. A. Geim, S. H. Li, H. Zhou, T. Manovitz, *et al.*, Logical quantum processor based on reconfigurable atom arrays, *Nature*, 2024, **626**(7997), 58–65.

24 A. Katabarwa, K. Gratsea, A. Caesura and P. D. Johnson, Early Fault-Tolerant Quantum Computing, *PRX Quantum*, 2024, **5**(2), 020101.

25 D. S. Abrams and S. Lloyd, Quantum Algorithm Providing Exponential Speed Increase for Finding Eigenvalues and Eigenvectors, *Phys. Rev. Lett.*, 1999, **83**, 5162–5165.

26 M. Suzuki, Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations, *Phys. Lett. A*, 1990, **146**(6), 319–323.

27 *Finding Exponential Product Formulas of Higher Orders*, ed. Hatano N., Suzuki M., Das A. and K Chakrabarti B., Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 37–68.

28 R. D. Somma, Quantum eigenvalue estimation *via* time series analysis, *New J. Phys.*, 2019, **21**(12), 123025.

29 R. Zhang, G. Wang and P. Johnson, Computing Ground State Properties with Early Fault-Tolerant Quantum Computers, *Quantum*, 2022, **6**, 761.

30 Z. Ding and L. Lin, Even Shorter Quantum Circuit for Phase Estimation on Early Fault-Tolerant Quantum Computers with Applications to Ground-State Energy Estimation, *Phys. Rev. X*, 2023, **4**, 020331.

31 Z. Ding and L. Lin, Simultaneous estimation of multiple eigenvalues with short-depth quantum circuit on early fault-tolerant quantum computers, *Quantum*, 2023, **7**, 1136.

32 R. M. Parrish and P. L. McMahon, Quantum Filter Diagonalization: Quantum Eigendecomposition without Full Quantum Phase Estimation, *arXiv*, 2019, arXiv:1909.08925, DOI: 10.48550/arXiv.1909.08925.

33 N. H. Stair, R. Huang and F. A. Evangelista, A Multireference Quantum Krylov Algorithm for Strongly Correlated Electrons, *J. Chem. Theory Comput.*, 2020, **16**(4), 2236–2245.

34 W. Kirby, M. Motta and A. Mezzacapo, Exact and efficient Lanczos method on a quantum computer, *Quantum*, 2023, **7**, 1018.

35 E. N. Epperly, L. Lin and Y. Nakatsukasa, A Theory of Quantum Subspace Diagonalization, *SIAM J. Matrix Anal. Appl.*, 2022, **43**(3), 1263–1290.

36 G. Lee, D. Lee and J. Huh, Sampling Error Analysis in Quantum Krylov Subspace Diagonalization, *Quantum*, 2024, **8**, 1477.

37 K. Seki and S. Yunoki, Quantum Power Method by a Superposition of Time-Evolved States, *Phys. Rev. X*, 2021, **2**, 010333.

38 M. Motta, C. Sun, A. T. K. Tan, M. J. O'Rourke, E. Ye, A. J. Minnich, *et al.*, Determining eigenstates and thermal states on a quantum computer using quantum imaginary time evolution, *Nat. Phys.*, 2020, **16**(2), 205–210.

39 O. Crawford, B. Straaten, D. Wang, T. Parks, E. Campbell and S. Brierley, Efficient quantum measurement of Pauli operators in the presence of finite sampling error, *Quantum*, 2021, **5**, 385.

40 T. C. Yen, A. Ganeshram and A. F. Izmaylov, Deterministic improvements of quantum measurements with grouping of compatible operators, non-local transformations, and covariance estimates, *NPJ Quantum Inf.*, 2023, **9**(1), 14.

41 S. Choi and A. F. Izmaylov, Measurement Optimization Techniques for Excited Electronic States in Near-Term Quantum Computing Algorithms, *J. Chem. Theory Comput.*, 2023, **19**(11), 3184–3193, PMID: 37224265.

42 T. C. Yen, V. Verteletskyi and A. F. Izmaylov, Measuring All Compatible Operators in One Series of Single-Qubit Measurements Using Unitary Transformations, *J. Chem. Theory Comput.*, 2020, **16**(4), 2400–2409, PMID: 32150412.

43 S. Choi, T. C. Yen and A. F. Izmaylov, Improving Quantum Measurements by Introducing "Ghost" Pauli Products, *J. Chem. Theory Comput.*, 2022, **18**(12), 7394–7402, PMID: 36332111.

44 S. Choi, I. Loaiza and A. F. Izmaylov, Fluid fermionic fragments for optimizing quantum measurements of electronic Hamiltonians in the variational quantum eigensolver, *Quantum*, 2023, **7**, 889.

45 T. C. Yen and A. F. Izmaylov, Cartan Subalgebra Approach to Efficient Measurements of Quantum Observables, *Phys. Rev. X*, 2021, **2**, 040320.

46 V. Verteletskyi, T. C. Yen and A. F. Izmaylov, Measurement optimization in the variational quantum eigensolver using a minimum clique cover, *J. Chem. Phys.*, 2020, **152**(12), 124114.

47 H. Y. Huang, R. Kueng and J. Preskill, Predicting many properties of a quantum system from very few

measurements, *Nat. Phys.*, 2020, **16**(10), 1050–1057, DOI: **10.1038/s41567-020-0932-7**.

48 C. Hadfield, S. Bravyi, R. Raymond and A. Mezzacapo, Measurements of Quantum Hamiltonians with Locally-Biased Classical Shadows, *arXiv*, 2020, arXiv:2006.15788, DOI: **10.48550/arXiv.2006.15788**.

49 H. Y. Huang, R. Kueng and J. Preskill, Efficient Estimation of Pauli Observables by Derandomization, *Phys. Rev. Lett.*, 2021, **127**, 030503, DOI: **10.1103/PhysRevLett.127.030503**.

50 Y. Nakamura, Y. Yano and N. Yoshioka, Adaptive measurement strategy for quantum subspace methods, *New J. Phys.*, 2024, **26**(3), 033028, DOI: **10.1088/1367-2630/ad2c3b**.

51 I. Loaiza and A. F. Izmaylov, Block-Invariant Symmetry Shift: Preprocessing Technique for Second-Quantized Hamiltonians to Improve Their Decompositions to Linear Combination of Unitaries, *J. Chem. Theory Comput.*, 2023, **19**(22), 8201–8209, PMID: 37939198.

52 I. Loaiza, A. M. Khah, N. Wiebe and A. F. Izmaylov, Reducing molecular electronic Hamiltonian simulation cost for linear combination of unitaries approaches, *Quantum Sci. Technol.*, 2023, **8**(3), 035019.

53 A. M. Childs and N. Wiebe, Hamiltonian simulation using linear combinations of unitary operations, *Quantum Inf. Comput.*, 2012, **12**(11–12), 901–924.

54 A. F. Izmaylov, T. C. Yen, R. A. Lang and V. Verteletskyi, Unitary Partitioning Approach to the Measurement Problem in the Variational Quantum Eigensolver Method, *J. Chem. Theory Comput.*, 2020, **16**(1), 190–195, PMID: 31747266.

55 A. Zhao, A. Tranter, W. M. Kirby, S. F. Ung, A. Miyake and P. J. Love, Measurement reduction in variational quantum algorithms, *Phys. Rev. A*, 2020, **101**, 062322.

56 Y.-C. Zheng, C.-Y. Lai, T. A. Brun and L.-C. Kwek, Depth reduction for quantum Clifford circuits through Pauli measurements, *arXiv*, 2018, arXiv:1805.12082, DOI: **10.48550/arXiv.1805.12082**.

57 P. Gokhale, O. Angiuli, Y. Ding, K. Gui, T. Tomesh, M. Suchara, M. Martonosi and F. T. Chong, Minimizing State Preparations in Variational Quantum Eigensolver by Partitioning into Commuting Families, *arXiv*, 2019, arXiv:1907.13623, DOI: **10.48550/arXiv.1907.13623**.

58 J. T. Seeley, M. J. Richard and P. J. Love, The Bravyi-Kitaev transformation for quantum computation of electronic structure, *J. Chem. Phys.*, 2012, **137**(22), 224109.

59 S. Bravyi, J. M. Gambetta, A. Mezzacapo and K. Temme, Tapering off qubits to simulate fermionic Hamiltonians, *arXiv*, 2017, arXiv:1701.08213, DOI: **10.48550/arXiv.1701.08213**.

60 N. Yoshioka, M. Amico, W. Kirby, P. Jurcevic, A. Dutt, B. Fuller, S. Garion, H. Haas, I. Hamamura, A. Ivrii, R. Majumdar, Z. Minev, M. Motta, B. Pokharel, P. Rivero, K. Sharma, C. J. Wood, A. Javadi-Abhari and A. Mezzacapo, Diagonalization of large many-body Hamiltonians on a quantum processor, *arXiv*, 2024, arXiv:2407.14431, DOI: **10.48550/arXiv.2407.14431**.

61 G. Lee, Openfermion Expansion Toolkit, *GitHub*, 2024, **https://github.com/snow0369/ofex**.

62 G. Lee, Measurement Strategies for QKSD, *GitHub*, 2024, **https://github.com/snow0369/eff_qksd_script**.

63 A. Roy and S. Suda, Complex Spherical Designs and Codes, *J. Combin. Designs*, 2014, **22**(3), 105–148.