SQD-Enabled Circuit Compression for Resource-Efficient Quantum Chemistry

Kangyu Zheng
kyzheng@cse.cuhk.edu.hk
The Chinese University of Hong Kong
Hong Kong, China

Yidong Zhou
yidong.zhou@rutgers.edu
Rutgers University
New Brunswick, New Jersey, USA

Jinglei Cheng
jinglei@eigenai.com
Eigen AI
Palo Alto, California, USA

Zhemin Zhang
hisenzhang01@gmail.com
The Chinese University of Hong Kong
Hong Kong, China

Shaohua Li
shaohuali@cuhk.edu.hk
The Chinese University of Hong Kong
Hong Kong, China

Zhiding Liang
zliang@cse.cuhk.edu.hk
The Chinese University of Hong Kong
Hong Kong, China

Abstract

Sample-based Quantum Diagonalization (SQD) recovers ground-state energies by classically diagonalizing a Hamiltonian in the subspace spanned by quantum samples, requiring only bitstrings with sufficient ground-state overlap rather than an accurate variational energy. We reveal and exploit this underexplored robustness property: how much non-Clifford and variational expressivity can be removed from the sampling circuit before SQD accuracy degrades? We answer through two complementary compression techniques: gradient-based operator pruning, which discards low-impact excitation operators, and Clifford rounding, which snaps remaining parameters to the nearest Clifford angle. Both of these techniques can be applied to a VQE ansatz on a qubit-reduced Hamiltonian. A systematic ablation study across 21 molecules shows that median SQD error stays within chemical accuracy even at 50% compression on both axes, while simulation speedup reaches 33×. Hardware validation on 6 molecules on IBM quantum hardware confirms up to 2.8× transpiled-depth reduction with zero loss in SQD accuracy. Our implementation can be found at: https://github.com/zkysfls/cs-vqe-sqd

CCS Concepts

• **Computing methodologies** → **Modeling and simulation**; • **Applied computing** → **Chemistry**.

Keywords

Quantum simulation, Quantum chemistry

1 Introduction

Quantum computing (QC) presents a fundamentally different computational approach capable of addressing classically intractable problems, with applications in many areas [3, 12, 21, 31]. The current Noisy Intermediate-Scale Quantum (NISQ) era is characterized by systems with several hundred qubits that exhibit significant noise limitations [24]. Despite advances in error mitigation techniques to improve operational fidelity [2, 4, 7, 8, 17], performance remains inadequate for most practical applications [25].



This work is licensed under a Creative Commons Attribution 4.0 International License.

Variational Quantum Algorithms (VQAs) are well-suited for NISQ hardware implementation. These hybrid quantum-classical methods have shown utility across multiple domains, from molecular simulations [10, 23, 37, 38] to optimization problems [21]. VQAs utilize parameterized quantum circuits whose variables are optimized using classical computational methods to minimize problem-specific objective functions. While their adaptability to hardware characteristics is advantageous, their performance on current quantum processors often fails to meet precision requirements in fields such as quantum chemistry, especially at larger scales [12, 35].

VQAs face several persistent challenges, including optimization landscapes with barren plateaus, vulnerability to quantum hardware noise, limitations in implementable circuit depths, and difficulties in the optimal allocation of tasks between classical and quantum computational resources [26, 29]. These challenges intensify with increasing problem size and when addressing applications with high precision requirements.

To address these challenges, we propose a resource-efficient hybrid pipeline that reduces qubit count and circuit depth while preserving chemical accuracy. Our approach begins with a qubit-reduction step (Contextual-Subspace [36] in this work, though other methods may serve the same role): the Hamiltonian is split into a noncontextual component that is solved classically and a smaller contextual remainder targeted by VQE on hardware. This partitioning shrinks the qubit register and measurement load without discarding strongly correlated physics.

After pre-processing reduction, our pipeline constructs a problem-specific ansatz by performing a classical gradient analysis like ADAPT-VQE [10]. This allows us to select only the most significant excitation operators relevant to the reduced Hamiltonian, resulting in a low-depth, compact ansatz with few parameters. We further compress the circuit by rounding the parameters of the least significant operators to the nearest Clifford angle, converting costly non-Clifford gates into classically trackable Clifford operations.

A key observation motivating these aggressive simplifications is that SQD [27] does not require an accurate VQE energy, it only requires the circuit to produce bitstrings with sufficient overlap on the ground-state subspace. SQD then forms low-dimensional subspaces from these samples and classically diagonalizes the Hamiltonian to refine energies and eigenstates. This decoupling of circuit quality from final accuracy creates room for circuit compression that would be unacceptable in a standalone VQE.

In summary, our main contributions are:

- The identification that SQD's tolerance to circuit imperfection enables aggressive circuit compression—gradient pruning and Clifford rounding—without degrading final energy accuracy.
- We present a systematic ablation study across 21 molecules characterizing the two-dimensional accuracy: depth tradeoff of gradient pruning and Clifford rounding, demonstrating that the median error remains within chemical accuracy even at 50% compression along both axes.
- Hardware validation on 6 molecules on IBM quantum hardware, demonstrating up to 2.8× reduction in transpiled circuit depth with zero loss in SQD accuracy.

2 Background

2.1 Variational Quantum Algorithms and Classical Simulation

VQAs represent a leading approach for near-term quantum processors, employing a hybrid quantum-classical strategy [6]. A parameterized quantum circuit (the **Ansatz**) is executed on quantum hardware while a classical optimizer trains the parameters $\boldsymbol{\theta}$ to minimize a **Cost Function** $C(\boldsymbol{\theta})$. Variational Quantum Eigensolver (VQE) [23], one of the most prominent VQA variants, finds ground-state energies through minimization of the expectation value $C(\boldsymbol{\theta}) = \langle\psi(\theta)|H|\psi(\theta)\rangle$.

A fundamental challenge in VQE is balancing classical tractability with quantum expressivity [11, 32]. Clifford-assisted methods [26, 29] restrict ansätze to Clifford gates for guaranteed classical simulability, but this severely limits expressivity for capturing complex correlations. Recent optimization approaches [15, 16] focus on reducing energy evaluations through sequential optimization or pulse-level refinement, yet treat parameter optimization as independent of ansatz design without addressing the underlying expressivity-tractability trade-off [1].

Rather than fixing an ansatz structure a priori, adaptive ansatz selection methods [28, 34] improve expressivity efficiency by dynamically constructing circuits tailored to specific problems. ADAPT-VQE [10] exemplifies this approach by iteratively building ansätze through greedy selection of operators that maximize energy reduction at each step. This adaptive strategy improves upon standard VQE by focusing computational resources on the most impactful circuit elements, rather than using a generic fixed structure. However, existing ADAPT-VQE implementations face a critical limitation: they either rely on full-space VQE evaluations or sacrifice expressivity through restrictive operator pools, treating ansatz design independently from the classical simulation efficiency constraints [11, 30].

When VQEs are implemented on classical simulators, cost function evaluation becomes a computational bottleneck [19]. Full statevector simulation scales exponentially as $O(2^n)$ [32], rendering large systems intractable. This limitation has motivated development of specialized classical simulators that exploit circuit structure to achieve polynomial-time cost function evaluation.

2.2 Stabilizer Tensor Network Simulators

Stabilizer Tensor Network (STN) simulators [18] combine the stabilizer formalism [9] with tensor network contraction to enable polynomial-time classical evaluation of quantum expectation values for Clifford-dominated circuits. Clifford gates, which are unitaries that map Pauli operators to Pauli operators, can be tracked classically in $O(n^3)$ time via stabilizer tableaux. Non-Clifford gates are handled by a tensor network whose contraction cost scales with the number of such gates. Consequently, reducing the non-Clifford gate count directly lowers the simulation cost, which is the motivation for our Clifford rounding technique (Section 3.2).

2.3 Qubit Reduction and Contextual Subspace Methods

Ground states of molecular systems concentrate amplitude on a low-dimensional manifold, enabling qubit reduction by projecting the Hamiltonian onto a smaller subspace [5, 20]. Contextual subspace methods exploit this structure by identifying low-energy configurations and intelligently projecting the optimization problem onto carefully chosen subspaces $\mathcal{H}_{\text{sub}} \subseteq \mathcal{H}_{\text{full}}$ that are compatible with physical symmetries and electronic structure properties [14], reducing the Hilbert space from 2^n to $2^{n'}$ dimensions where $n' \ll n$. Our pipeline uses Contextual Subspace [36] for this step, though other reduction methods such as active space selection or entanglement forging could serve the same role.

2.4 Sample-based Quantum Diagonalization

SQD [27], which builds on the quantum-selected configuration interaction (QSCI) framework [13], uses a quantum circuit purely as a bitstring sampler: sampled configurations are used to construct a reduced configuration interaction subspace, which is then classically diagonalized to refine the ground-state energy. Because SQD performs its own classical eigen-solve, it does not require the VQE circuit to produce an accurate energy estimate—only bitstrings with sufficient overlap on the ground-state subspace. This property is central to our work: it means the preceding circuit can be aggressively simplified (fewer operators, Clifford-rounded parameters) without degrading the final SQD accuracy.

3 Method

3.1 Core Insight

Quantum chemistry simulation reveals a fundamental contradiction: the full electronic structure problem requires exploring exponentially large Hilbert spaces, yet near-term quantum processors cannot maintain coherence for the deep circuits necessary to explore such spaces. Classical simulation conversely can execute arbitrarily deep circuits but faces exponential overhead.

Our core insight is that SQD changes what the quantum circuit must achieve. In standalone VQE, the circuit must produce an accurate energy expectation value, demanding high expressivity and deep circuits. With SQD as a post-processing step, the circuit only needs to produce bitstrings that overlap sufficiently with the ground-state subspace which is a much weaker requirement. This gap between what standalone VQE demands and what SQD actually needs creates room for two circuit compression techniques:

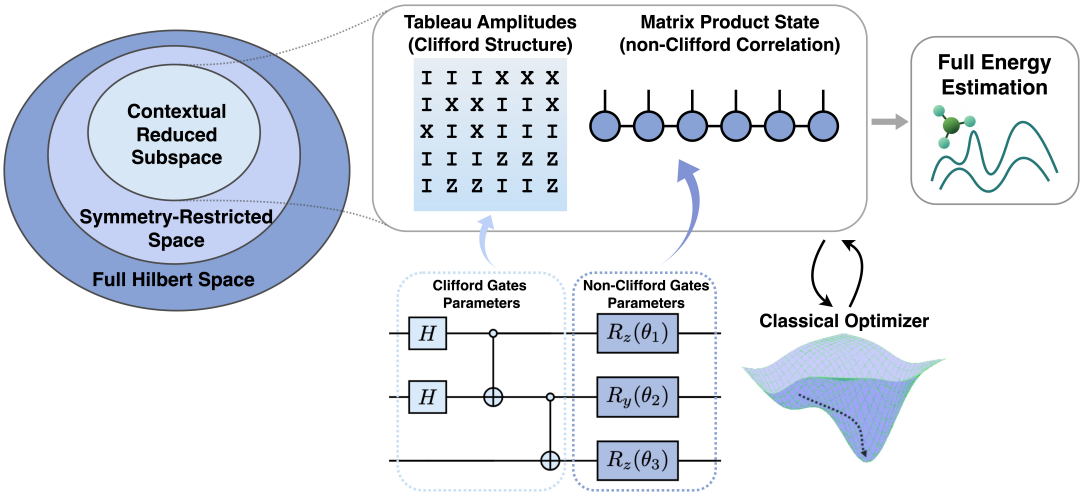


Figure 1: Overview of the proposed pipeline. The full Hilbert space is compressed into a contextual reduced subspace capturing the dominant low-energy structure of the molecule. Inside this subspace, the quantum state is expressed via a dual representation: a stabilizer-tableau component encoding Clifford-compatible structure and an MPS component capturing non-Clifford correlations. During variational optimization, Clifford gates act only on the tableau representation and non-Clifford gates act only on the MPS component, coordinated by a classical optimizer. The optimized reduced-space state is finally lifted back to the full orbital space using a precomputed inverse mapping, enabling accurate reconstruction of molecular observables and ground-state energies.

Gradient pruning removes excitation operators with small gradient magnitude, reducing the number of ansatz parameters. **Clifford rounding** snaps the parameters of the least significant remaining operators to the nearest Clifford angle ($k\pi/2$), converting non-Clifford gates into classically trackable Clifford operations. Both techniques degrade standalone VQE accuracy but, as we demonstrate in Section 4.2, leave SQD accuracy essentially unchanged.

The overall pipeline (Figure 1) proceeds in three stages: (1) qubit reduction via contextual subspace projection, (2) compressed VQE with gradient pruning and Clifford rounding, and (3) SQD recovery.

3.2 Stage 1: Qubit Reduction

The reduction stage applies qubit tapering followed by contextual-subspace projection. Beginning with a molecular Hamiltonian in Jordan-Wigner representation, we first exploit conservation laws to reduce the qubit count, yielding the tapered Hamiltonian H_{taper} operating on $n_{\text{taper}} \ll n_{\text{full}}$ qubits with no approximation error.

Next, we construct stabilizer generators defining a low-energy subspace, projecting the tapered Hamiltonian onto this stabilizer eigenspace to yield the contextual-subspace Hamiltonian H_{cs} on n_{cs} qubits. We extract excitation operators by projecting UCCSD into the CS space, retaining only dominant Pauli terms to keep circuit depths manageable. We also precompute a mapping table relating CS bitstrings to the original Jordan-Wigner space, stored for use in Stage 3 recovery. All Stage 1 computations are performed once offline and cached.

3.3 Stage 2: Compressed Variational Optimization

The ansatz is parameterized as $|\psi(\theta)\rangle = \prod_{j=1}^{N} e^{-i\theta_j P_j}|\text{HF}\rangle$, where P_j are single-term Pauli excitations and $|\text{HF}\rangle$ is the Hartree-Fock reference prepared with Clifford gates.

Gradient Pruning: We first compute gradients $\partial E/\partial\theta_j$ for all operators using a shallow reference circuit, evaluating each excitation operator independently at $\theta_j = 0$ with all other parameters fixed at zero. This single-term gradient evaluation is efficient: it requires N STN evaluations rather than the $O(N^2)$ evaluations that a full ADAPT-VQE gradient step would need, since we do not account for inter-operator interactions. We then select the top-ranked operators by gradient magnitude up to a specified fraction f_g of the full pool, discarding operators whose gradients fall below the threshold. The discarded operators' parameters are fixed at zero, effectively removing those gates from the circuit entirely. This reduces parameter count without sacrificing expressivity for the dominant effects.

Clifford Rounding: After VQE convergence, we identify the $\lfloor f_c \cdot N_{\text{active}} \rfloor$ parameters closest to a Clifford angle $k\pi/2$ and snap them to that angle, where f_c is the Clifford fraction controlling what proportion of active parameters are rounded. These gates become Clifford operations that the STN simulator handles at negligible cost, directly reducing both simulation time and hardware circuit depth. The remaining $N_{\text{active}} - \lfloor f_c \cdot N_{\text{active}} \rfloor$ parameters retain their optimized non-Clifford values, preserving expressivity where it matters most.

Concretely, the rounding uses a min-diff strategy: for each parameter θ_j, we compute $\delta_j = \min_k |\theta_j - k\pi/2|$, the distance to the nearest Clifford angle. We sort parameters by δ_j in ascending order

and round the $\lfloor f_c \cdot N_{\text{active}} \rfloor$ parameters with the smallest δ_j. This ensures that we round the parameters that are already closest to a Clifford angle, minimizing the perturbation to the wavefunction. Parameters far from any Clifford angle, typically those encoding the strongest correlations, are left untouched. When $f_c = 0$ no parameters are rounded (fully non-Clifford); when $f_c = 1$ all parameters are rounded (fully Clifford). Note that Clifford rounding is applied after VQE convergence: the optimizer runs with full non-Clifford freedom, and rounding is a post-hoc compression step. This is important because it means the rounding does not affect the optimization landscape, only the final circuit used for SQD sampling.

The cost function is evaluated using the STN simulator, which exploits the Clifford-dominance of the circuit, computing the expectation value in polynomial time by tracking the stabilizer state. Fewer non-Clifford gates directly lower the STN contraction cost. Upon convergence, we sample bitstrings from the final state in the computational basis. These samples serve as input to Stage 3.

3.4 Stage 3: SQD Recovery

Bitstrings sampled from the VQE state are lifted to the original Jordan-Wigner space using the precomputed mapping table, accounting for weighted contributions from the eliminated degrees of freedom. The lifted bitstrings are then passed to SQD's fermion-solver, which builds a selected-CI subspace from the sampled determinants and solves it via classical diagonalization to refine energy estimates.

In each iteration, SQD generates fresh samples, constructs configuration space approximations to the full Hamiltonian using sampled determinants, and solves these approximations exactly via classical diagonalization. Over iterations, the algorithm converges toward the ground state. Specifically, each SQD iteration proceeds as follows: (i) bitstrings are postselected to the correct electron number in each spin sector; (ii) the postselected bitstrings are subsampled into multiple batches to reduce variance; (iii) for each batch, a configuration interaction (CI) matrix is constructed from the PySCF one- and two-electron integrals restricted to the sampled determinants; (iv) a Davidson eigensolver diagonalizes each CI matrix to yield a batch energy estimate; (v) the per-batch occupancy profiles are averaged and fed back to a configuration recovery step that biases the next iteration's sampling toward the most occupied orbitals. The best energy across all batches and iterations is reported as the final SQD estimate.

4 Evaluation

4.1 Experiment Setup

Problem Instances. We study a set of 21 small molecules and ions in the STO-3G minimal basis, including H_2O, N_2, HCl, and map them to qubits via the Jordan–Wigner transformation. For each system, we generate its Pauli-sum Hamiltonian, record qubit and electron counts, the Hartree–Fock reference state vector, and other molecular attributes from PySCF [33]. Across the set, the full JW register spans 6–20 qubits, depending on the system.

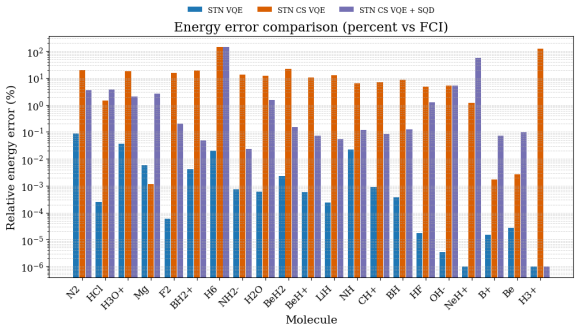


Figure 2: Relative energy error versus FCI comparison under full parameters mode

Contextual Subspace. To reduce problem size while preserving ground-state accuracy, we apply qubit tapering with the HF reference and then project the tapered Hamiltonian into a contextual subspace [36]. For each molecule we sweep the subspace size and select the smallest number of contextual qubits whose subspace ground energy is within approximately 1.6 mHa of full configuration interaction (FCI). We also compute a mapping from the contextual subspace back to the full Jordan–Wigner basis with associated weights, so that sampled contextual configurations can be consistently distributed over the full basis during the SQD step. Within the contextual subspace we initialize the HF determinant and build a single-term, UCCSD-inspired excitation pool by projecting the tapered UCC operator and retaining the dominant Pauli term from each composite operator.

VQE setting. For the VQE used in our experiments, we choose COBYLA optimizer with initial parameters set to zero and reports electronic energies including nuclear repulsion. For the simulation estimator, we adopt and modify the implementation from [18] to make our own STN estimator. For hardware experiments, circuits are transpiled to the `ibm_rensselaer` instruction set at optimization level 3 and sampled with 100k shots via IBM Qiskit SamplerV2.

SQD setting. SQD post-processing proceeds by sampling bitstrings from the optimized contextual ansatz. On simulation, we sample from the exact statevector distribution; on real hardware, we measure with IBM Qiskit SamplerV2. Sampled contextual bitstrings are lifted to JW via the precomputed CS to JW map, postselected to the correct electron numbers. We then perform a subspace eigen-solve per batch, which builds a selected-CI subspace using the PySCF one- and two-electron integrals and solves it with a Davidson routine. Unless noted otherwise, SQD runs for 5 outer iterations; we cap Davidson iterations at 2000 and use a fixed random seed 42.

Compression settings. We evaluate gradient pruning at fractions $f_g \in \{1.00, 0.75, 0.50, 0.25\}$ and Clifford rounding at fractions $f_c \in \{1.00, 0.75, 0.50, 0.25, 0.00\}$, forming a 4×5 ablation grid. On hardware, we test four representative configurations: baseline (f_g=1, f_c=1), gradient only (f_g=0.5), Clifford only (f_c=0.5), and the combination of gradient and clifford method (f_g=0.5, f_c=0.5).

4.2 Experiment Results

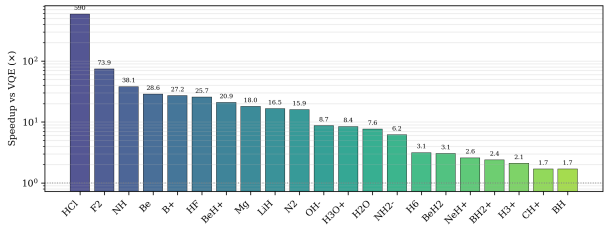


Figure 3: Runtime speedup of CS-VQE+SQD vs. standard VQE across 21 molecules (uncompressed baseline, f_g=1, f_c=1). Molecules are sorted by descending speedup. HCl achieves 590× due to aggressive qubit reduction (20→3 qubits); molecules with large contextual subspaces (BH, CH^+) show modest gains.

Table 1: Median SQD energy error (mHa) across 21 molecules. Bold entries are within chemical accuracy (1.6 mHa). Rows: gradient fraction f_g. Columns: Clifford fraction f_c.

$f_g \downarrow$ / $f_c \rightarrow$	1.00	0.75	0.50	0.25	0.00
1.00	**0.73**	**0.59**	**1.19**	**0.81**	**0.73**
0.75	**1.15**	**1.19**	**1.21**	**0.81**	37.16
0.50	3.70	3.62	3.79	3.82	37.16
0.25	8.75	8.75	7.54	9.06	37.16

Baseline Performance. We first establish the baseline by comparing three methods without any circuit compression: standard VQE on the full Hamiltonian, CS-VQE (qubit-reduced, no SQD), and our full pipeline CS-VQE+SQD. Figure 2 shows the relative energy error versus FCI. The baseline VQE consistently achieves very low error, around 10^{-3}%. CS-VQE alone shows much higher error in many cases, around $1-10$% or worse, because truncating to a small contextual subspace sacrifices expressive power. When SQD is added, the error in most molecules falls back into the $\leq 10^{-2}$% region. This confirms that SQD effectively recovers the accuracy lost through subspace truncation—the property we exploit for circuit compression.

Figure 3 shows the runtime speedup of our pipeline (CS-VQE+SQD) relative to the standard VQE baseline across all 21 molecules. The speedups vary substantially: HCl achieves 590× with 20 JW qubits reduced to 3 CS qubits, F_2 achieves 74×, NH 38×, and Be/B^+ around 28×. At the lower end, BH and CH^+ achieve only 1.7× because their contextual subspaces remain large: 8 CS qubits out of 12 JW qubits. Of the 21 molecules, 15 achieve SQD error below the 1.6 mHa chemical accuracy threshold. The outliers: N_2 (33.8 mHa, 14 CS qubits), Mg (17.5 mHa), H_3O^+ (16.7 mHa) require large contextual subspaces, limiting the compression benefit. These molecules are included in all subsequent ablation results to ensure our compression techniques are evaluated on both easy and hard cases.

Ablation Study: Gradient Pruning × Clifford Rounding. The central experiment is a systematic ablation over both compression axes. Figure 4 and Table 1 show the median energy error and speedup across all 21 molecules for each (f_g, f_c) combination.

Several findings emerge. First, Clifford rounding at $f_c \geq 0.25$ has almost no effect on accuracy for a given f_g: error stays flat across columns until $f_c = 0.00$ where it jumps sharply. This confirms that most VQE parameters are already near Clifford angles and rounding them is nearly free. Second, gradient pruning is the dominant source of error increase: reducing f_g from 1.00 to 0.25 raises median error from 0.73 to 8.75 mHa, regardless of f_c. Third, the sweet spot for chemical accuracy is $f_g \geq 0.75$ with any $f_c \geq 0.25$, achieving a median simulation speedup of 1.2–2.5× from Clifford rounding alone and up to 9.4× when combined with gradient pruning.

Table 2: Median simulation speedup (×) vs baseline across 21 molecules. Rows: gradient fraction f_g. Columns: Clifford fraction f_c.

$f_g \downarrow$ / $f_c \rightarrow$	1.00	0.75	0.50	0.25	0.00
1.00	1.00×	1.17×	1.52×	2.06×	1.01×
0.75	1.19×	1.37×	1.63×	2.47×	9.37×
0.50	1.49×	1.79×	2.04×	2.51×	15.22×
0.25	2.81×	2.57×	3.78×	3.80×	33.09×

Table 2 shows the corresponding median simulation speedups. The two compression axes contribute multiplicatively: gradient pruning at $f_g = 0.25$ alone gives 2.8×, Clifford rounding at $f_c = 0.25$ alone gives 2.1×, and combining them yields 3.8×. The $f_c = 0.00$ column achieves the largest speedups (9–33×) because the STN simulator handles fully Clifford circuits at negligible cost, but as Table 1 shows, this comes at the price of accuracy. The practical operating point is therefore $f_c \in [0.25, 0.50]$, which captures most of the speedup benefit while staying well within chemical accuracy.

We note that the ablation grid reveals molecule-dependent behavior. Molecules with few CS qubits and few excitation operators (e.g., Be, B^+, HCl with 3 CS qubits and 6–16 operators) are robust to aggressive compression, maintaining chemical accuracy even at $f_g = 0.25$, because SQD has sufficient bitstring diversity from their small Hilbert spaces. Conversely, molecules with large contextual subspaces (e.g., N_2 with 14 CS qubits, H_3O^+ with 10 CS qubits) are more sensitive to pruning because the ground state requires many excitation operators to achieve adequate overlap. For these systems, $f_g \geq 0.75$ is recommended to maintain accuracy, while Clifford rounding remains safe at any $f_c \geq 0.25$.

Individual Sweep Analysis. Sweeping each axis in isolation confirms these trends. Reducing f_g from 1.00 to 0.50 keeps the median error below chemical accuracy while the median speedup grows to 1.5×; at $f_g = 0.25$ the speedup reaches 2.8× but error rises above the threshold for several molecules, with the steepest degradation seen in molecules that rely on many excitation operators. Clifford rounding shows error essentially flat from $f_c = 1.00$ to 0.25 (variance across molecules stays low in this range) while speedup climbs to 2.1×; only at $f_c = 0.00$, where the circuit loses all non-Clifford expressivity, does accuracy degrade sharply. Both axes thus offer a favorable accuracy–speedup tradeoff, with Clifford rounding the cheaper of the two in accuracy cost.

An interesting anomaly appears at $f_c = 0.00$ in Table 1: the median error at (f_g=1.00, f_c=0.00) is 0.73 mHa, identical to the uncompressed baseline (f_g=1.00, f_c=1.00). Yet it jumps to 37.16 mHa for $f_g \leq 0.75$. This reveals an interaction between the two compression axes. When all UCCSD operators are present (f_g=1.00), the fully

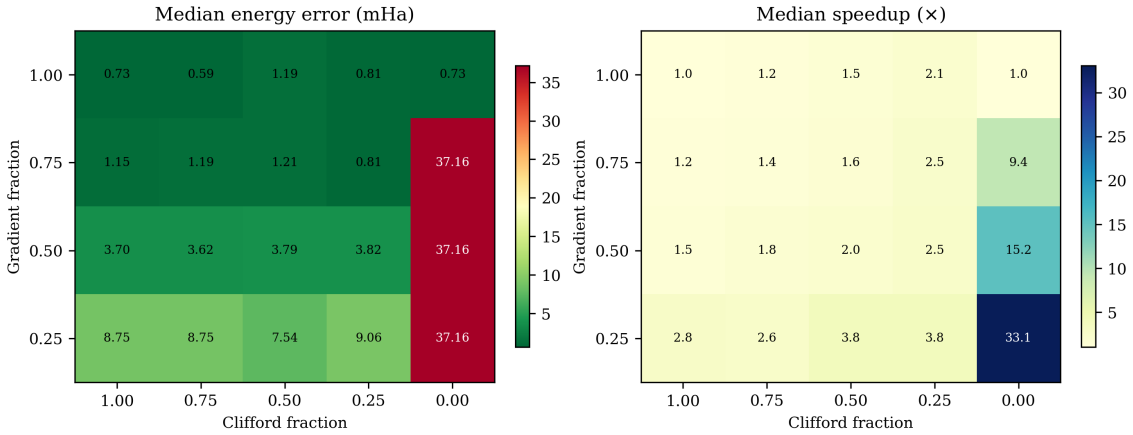


Figure 4: Ablation heatmap across 21 molecules. Left: median SQD energy error (mHa). Right: median simulation speedup (×). Rows vary gradient fraction f_g; columns vary Clifford fraction f_c. Error remains within chemical accuracy (1.6 mHa, bold in Table 1) for the top-left region; speedup increases toward the bottom-right. The $f_c = 0.00$ column shows a phase transition: accuracy collapses when no non-Clifford gates remain, but speedup peaks.

Clifford circuit still distributes bitstring probability across configurations that SQD can exploit. But when operators are also pruned, the combined loss of both operator diversity and non-Clifford expressivity pushes the circuit's output distribution too far from the ground-state subspace, and SQD can no longer recover. This finding underscores that the two techniques are not independently safe at their extremes: practitioners should avoid combining aggressive pruning with full Clifford rounding.

Hardware Validation. We validate on `ibm_rensselaer` with 6 molecules spanning 3–6 CS qubits. Table 3 presents the results. The most striking finding is that **SQD energy error is identical across all four configurations for every molecule**. Gradient pruning, Clifford rounding, and their combination produce exactly the same SQD accuracy as the uncompressed baseline. This is the direct experimental confirmation of our core insight: SQD's tolerance to circuit imperfection makes these compressions free in terms of accuracy.

Meanwhile, the transpiled circuit depth drops substantially (Figure 5). For molecules with more than a few parameters, the combined configuration reduces depth by 2.4–2.8×: LiH drops from 711 to 252, BeH^+ from 306 to 116, HCl from 161 to 58, and H_2O from 4437 to 1690. Gradient pruning contributes the larger share (median 1.8×), while Clifford rounding adds an additional 1.2× on top. Small molecules (Be, B^+ with 6 parameters and depth ∼30) show no benefit, as their circuits are already shallow. The depth reduction is most significant for H_2O (168 parameters), where the combined approach removes 2747 layers of circuit depth, directly translating to reduced decoherence and gate errors on hardware.

The invariance of SQD accuracy across all four hardware configurations arises because all four circuits, despite varying compression levels, still produce bitstrings with high overlap on the ground-state subspace: the CS qubit counts here are small enough that even a significantly perturbed circuit distributes probability over the correct region, letting SQD's eigensolve extract the same optimal

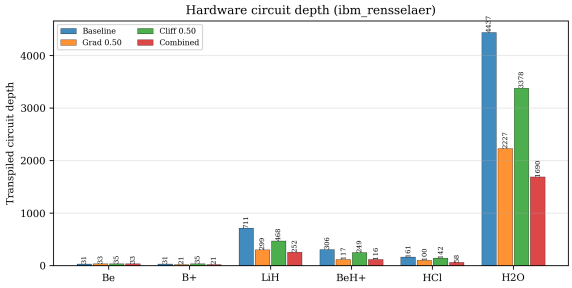


Figure 5: Absolute transpiled circuit depth on `ibm_rensselaer` across four configurations. The depth reduction scales with molecule size: H_2O (168 parameters) drops from 4437 to 1690; small molecules (Be, B^+) are already shallow and show no change.

energy from every sample set. This suggests SQD-based pipelines are particularly well-suited to molecules where qubit reduction yields a small contextual space, since the needed bitstring diversity is easily achievable even with compressed circuits.

From a hardware-cost perspective, the depth reduction directly impacts circuit fidelity: on `ibm_rensselaer` (Eagle r3, 127 qubits) with typical two-qubit gate error rates of $\sim 10^{-2}$, a circuit of depth 4437 (H_2O baseline) accumulates substantially more noise than one of depth 1690 (combined). Both produce identical SQD energy estimates in our experiments, but the compressed circuits provide additional headroom for noisier backends or larger molecules where accumulated errors could eventually degrade bitstring quality below SQD's tolerance threshold.

5 Related Work

Clifford-assisted VQA. CAFQA [26] restricts ansätze to Clifford gates for classical simulability via Bayesian optimization, and [29]

Table 3: Hardware results on `ibm_rensselaer`. Transpiled circuit depth and SQD energy error across four configurations. All runs use 100k shots.

	Transpiled Depth				SQD Error (mHa)			
Mol.	Baseline	Grad 0.50	Cliff 0.50	Combined	Baseline	Grad 0.50	Cliff 0.50	Combined
Be	31	33	35	33	0.33	0.33	0.33	0.33
B^+	31	21	35	21	0.38	0.38	0.38	0.38
LiH	711	299	468	252	1.30	1.30	1.30	1.30
BeH^+	306	117	249	116	0.32	0.32	0.32	0.32
HCl	161	100	142	58	0.54	0.54	0.54	0.54
H_2O	4437	2227	3378	1690	1.14	1.14	1.14	1.14
Median red.	1.0×	1.8×	1.2×	2.6×				

extends this to mixed Clifford/non-Clifford circuits. These methods use Clifford structure to make optimization cheaper; our Clifford rounding instead exploits it to compress an already-optimized circuit for SQD sampling. Unlike CAFQA, our approach retains non-Clifford parameters where they matter most and relies on SQD to compensate where they are rounded away.

Adaptive ansatz construction. ADAPT-VQE [10] iteratively grows the ansatz by selecting operators that maximize energy reduction. Our gradient pruning is related in spirit, also using gradient information to select operators, but ADAPT-VQE builds the ansatz from scratch while we prune a pre-defined UCCSD pool, and it targets VQE energy directly while our pruning targets a weaker objective that permits more aggressive reduction.

VQE parameter optimization. SOAP [15] reduces energy evaluations via sequential parabolic approximation, and NAPA [16] optimizes at the pulse level; both improve efficiency in reaching a VQE solution but do not address what happens after. Our work is complementary: any VQE optimizer can be used in Stage 2, with compression applied post-convergence.

Subspace methods and SQD. The sample-and-diagonalize framework was first proposed as QSCI [13]; SQD [27] extends it with configuration recovery and large-scale demonstrations on nitrogen and iron–sulfur clusters using fixed, carefully optimized circuits. ADAPT-QSCI [22] adaptively grows the input circuit through iterative operator selection, sharing our goal of a compact sampling circuit; we instead prune a fixed UCCSD pool post hoc in a qubit-reduced space. Subsequent work has used SQD with standard VQE circuits as an accuracy booster for a given circuit quality, but no prior work has asked the inverse question: *given that SQD will post-process the output, how much can the preceding circuit be degraded?* Our contribution answers this directly, showing that circuit quality can be substantially degraded via pruning and rounding without affecting SQD's output, reframing SQD not merely as an accuracy refinement tool but as an enabler of circuit compression.

Contextual subspace methods. Contextual subspace VQE [36] reduces qubit count by projecting onto stabilizer-defined subspaces; in our pipeline it provides the qubit reduction of Stage 1 but is not the main contribution, and any qubit-reduction method could serve the same role (Section II-C). Our novelty lies in the Stage 2 circuit compression and the demonstration that SQD tolerates it.

Circuit compilation and depth reduction. Transpiler-level optimizations [8] reduce circuit depth through gate cancellation, routing, and scheduling, independently of the algorithm. Our compression is complementary: gradient pruning and Clifford rounding simplify the logical circuit before transpilation, so the 2.6× median depth reduction we observe (Table 3) stacks on top of whatever the transpiler achieves at optimization level 3. Combining algorithm- and transpiler-level compression is a promising direction for maximizing circuit fidelity on near-term hardware.

6 Conclusion

We showed that SQD's tolerance to circuit imperfection creates a previously unexploited opportunity for circuit compression. By introducing gradient pruning and Clifford rounding, we reduce transpiled circuit depth by up to 2.8× on IBM hardware while SQD preserves identical energy accuracy. A systematic ablation across 21 molecules characterizes the two-dimensional tradeoff: Clifford rounding is nearly free in accuracy cost, gradient pruning dominates the error budget, and their combination yields the largest depth savings, a finding confirmed by hardware experiments on 6 molecules where all configurations produce the same energy error as the uncompressed baseline. These results suggest that future SQD-based pipelines should deliberately design for shallow, imprecise circuits rather than optimizing VQE accuracy, since SQD will recover the accuracy regardless.

Limitations and future work. Our evaluation uses the STO-3G minimal basis with 3–14 CS qubits that remain classically simulable; extending to larger basis sets beyond 20 CS qubits, generalizing the compression principle beyond UCCSD to other ansätze, co-optimizing parameters with a rounding-aware objective, and combining with error mitigation are promising directions for future work.

References

[1] Andrew Arrasmith, Zoë Holmes, Marco Cerezo, and Patrick J Coles. 2022. Equivalence of quantum barren plateaus to cost concentration and narrow gorges. *Quantum Science and Technology* 7, 4 (2022), 045015.

[2] George S Barron and Christopher J Wood. 2020. Measurement error mitigation for variational quantum algorithms. *arXiv preprint arXiv:2010.08520* (2020).

[3] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. 2017. Quantum machine learning. *Nature* 549, 7671 (2017).

[4] Ludmila Botelho, Adam Glos, Akash Kundu, Jarosław Adam Miszczak, Özlem Salehi, and Zoltán Zimborás. 2022. Error mitigation for variational quantum algorithms through mid-circuit measurements. *Physical Review A* 105, 2 (2022), 022441.

[5] Yudong Cao, Jonathan Romero, Jonathan P Olson, Matthias Degroote, Peter D Johnson, Mária Kieferová, Ian D Kivlichan, Tim Menke, Borja Peropadre, Nicolas PD Sawaya, et al. 2019. Quantum chemistry in the age of quantum computing. *Chemical reviews* 119, 19 (2019), 10856–10915.

[6] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. 2021. Variational quantum algorithms. *Nature Reviews Physics* 3, 9 (Aug. 2021), 625–644. doi:10.1038/s42254-021-00348-9

[7] Piotr Czarnik, Andrew Arrasmith, Patrick J Coles, and Lukasz Cincio. 2021. Error mitigation with Clifford quantum-circuit data. *Quantum* 5 (2021), 592.

[8] Yongshan Ding, Pranav Gokhale, Sophia Fuhui Lin, Richard Rines, Thomas Propson, and Frederic T Chong. 2020. Systematic crosstalk mitigation for superconducting qubits via frequency-aware compilation. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 201–214.

[9] Daniel Gottesman. 1997. *Stabilizer codes and quantum error correction*. California Institute of Technology.

[10] Harper R Grimsley, Sophia E Economou, Edwin Barnes, and Nicholas J Mayhall. 2019. An adaptive variational algorithm for exact molecular simulations on a quantum computer. *Nature communications* 10, 1 (2019), 3007.

[11] Zoë Holmes, Kunal Sharma, Marco Cerezo, and Patrick J Coles. 2022. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX quantum* 3, 1 (2022), 010313.

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

[13] Keita Kanno, Masaya Kohda, Ryosuke Imai, Sho Koh, Kosuke Mitarai, Wataru Mizukami, and Yuya O. Nakagawa. 2026. Quantum-selected configuration interaction: Classical diagonalization of Hamiltonians in subspaces selected by quantum computers. *Physical Review Research* 8 (2026), 023268.

[14] Sami Khairy, Ruslan Shaydulin, Lukasz Cincio, Yuri Alexeev, and Prasanna Balaprakash. 2020. Learning to optimize variational quantum circuits to solve combinatorial problems. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 2367–2375.

[15] Weitang Li, Yufei Ge, Shi-Xin Zhang, Yu-Qin Chen, and Shengyu Zhang. 2024. Efficient and robust parameter optimization of the unitary coupled-cluster ansatz. *Journal of Chemical Theory and Computation* 20, 9 (2024), 3683–3696.

[16] Zhiding Liang, Jinglei Cheng, Hang Ren, Hanrui Wang, Fei Hua, Zhixin Song, Yongshan Ding, Frederic T Chong, Song Han, Xuehai Qian, et al. 2024. Napa: intermediate-level variational native-pulse ansatz for variational quantum algorithms. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 43, 6 (2024), 1834–1847.

[17] Haoran Liao, Derek S Wang, Iskandar Sitdikov, Ciro Salcedo, Alireza Seif, and Zlatko K Minev. 2024. Machine learning for practical quantum error mitigation. *Nature Machine Intelligence* (2024), 1–9.

[18] Sergi Masot-Llima and Artur Garcia-Saez. 2024. Stabilizer tensor networks: universal quantum simulator on a basis of stabilizer states. *Physical Review Letters* 133, 23 (2024), 230601.

[19] Alexander McCaskey, Eugene Dumitrescu, Mengsu Chen, Dmitry Lyakh, and Travis Humble. 2018. Validating quantum-classical programming models with tensor network simulations. *PloS one* 13, 12 (2018), e0206704.

[20] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. 2016. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics* 18, 2 (2016), 023023.

[21] Nikolaj Moll, Panagiotis Barkoutsos, Lev S Bishop, Jerry M Chow, Andrew Cross, Daniel J Egger, Stefan Filipp, Andreas Fuhrer, Jay M Gambetta, Marc Ganzhorn, et al. 2018. Quantum optimization using variational algorithms on near-term quantum devices. *Quantum Science and Technology* 3, 3 (2018), 030503.

[22] Y. O. Nakagawa, M. Kamoshita, W. Mizukami, S. Sudo, and Y.-y. Ohnishi. 2024. ADAPT-QSCI: Adaptive construction of an input state for quantum-selected configuration interaction. *Journal of Chemical Theory and Computation* 20, 24 (2024), 10817–10825.

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

[24] John Preskill. 2018. Quantum computing in the NISQ era and beyond. *Quantum* 2 (2018), 79.

[25] Yihui Quek, Daniel Stilck França, Sumeet Khatri, Johannes Jakob Meyer, and Jens Eisert. 2024. Exponentially tighter bounds on limitations of quantum error mitigation. *Nature Physics* 20, 10 (2024), 1648–1658.

[26] Gokul Subramanian Ravi, Pranav Gokhale, Yi Ding, William Kirby, Kaitlin Smith, Jonathan M Baker, Peter J Love, Henry Hoffmann, Kenneth R Brown, and Frederic T Chong. 2022. CAFQA: A classical simulation bootstrap for variational quantum algorithms. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*. 15–29.

[27] Javier Robledo-Moreno, Mario Motta, Holger Haas, Ali Javadi-Abhari, Petar Jurcevic, William Kirby, Simon Martiel, Kunal Sharma, Sandeep Sharma, Tomonori Shirakawa, Iskandar Sitdikov, Rong-Yang Sun, Kevin J. Sung, Maika Takita, Minh C. Tran, Seiji Yunoki, and Antonio Mezzacapo. 2025. Chemistry beyond the scale of exact diagonalization on a quantum-centric supercomputer. *Science Advances* 11, 25 (June 2025). doi:10.1126/sciadv.adu9991

[28] Manuel S Rudolph, Jacob Miller, Danial Motlagh, Jing Chen, Atithi Acharya, and Alejandro Perdomo-Ortiz. 2023. Synergistic pretraining of parametrized quantum circuits via tensor networks. *Nature Communications* 14, 1 (2023), 8367.

[29] Lennart Maximilian Seifert, Siddharth Dangwal, Frederic T. Chong, and Gokul Subramanian Ravi. 2025. Clapton: Clifford Assisted Problem Transformation for Error Mitigation in Variational Quantum Algorithms. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 4* (Hilton La Jolla Torrey Pines, La Jolla, CA, USA) *(ASPLOS '24)*. Association for Computing Machinery, New York, NY, USA, 47–62. doi:10.1145/3622781.3674178

[30] Vlad O Shkolnikov, Nicholas J Mayhall, Sophia E Economou, and Edwin Barnes. 2023. Avoiding symmetry roadblocks and minimizing the measurement overhead of adaptive variational quantum eigensolvers. *Quantum* 7 (2023), 1040.

[31] Peter W Shor. 1999. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review* 41, 2 (1999), 303–332.

[32] Sukin Sim, Peter D Johnson, and Alán Aspuru-Guzik. 2019. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies* 2, 12 (2019), 1900070.

[33] Qiming Sun et al. 2020. Recent developments in the PySCF program package. *The Journal of Chemical Physics* 153, 2 (2020), 024109. doi:10.1063/5.0006074

[34] Ho Lun Tang, VO Shkolnikov, George S Barron, Harper R Grimsley, Nicholas J Mayhall, Edwin Barnes, and Sophia E Economou. 2021. qubit-adapt-vqe: An adaptive algorithm for constructing hardware-efficient ansätze on a quantum processor. *PRX Quantum* 2, 2 (2021), 020310.

[35] Samson Wang, Piotr Czarnik, Andrew Arrasmith, Marco Cerezo, Lukasz Cincio, and Patrick J Coles. 2024. Can error mitigation improve trainability of noisy variational quantum algorithms? *Quantum* 8 (2024), 1287.

[36] Tim Weaving, Alexis Ralli, Peter J. Love, Sauro Succi, and Peter V. Coveney. 2025. Contextual subspace variational quantum eigensolver calculation of the dissociation curve of molecular nitrogen on a superconducting quantum computer. *npj Quantum Information* 11, 1 (feb 2025), 25. doi:10.1038/s41534-024-00952-4

[37] Kangyu Zheng, Tianfan Fu, and Zhiding Liang. 2025. QCS-ADME: Quantum Circuit Search for Drug Property Prediction with Imbalanced Data and Regression Adaptation. arXiv:2503.01927

[38] Yidong Zhou, Jintai Chen, Jinglei Cheng, Gopal Karemore, Marinka Zitnik, Frederic T Chong, Junyu Liu, Tianfan Fu, and Zhiding Liang. 2024. Quantum-machine-assisted drug discovery: Survey and perspective. *arXiv preprint arXiv:2408.13479* (2024).

accepted 11 July 2026