LEARNING SYMMETRY-PRESERVING SAMPLING CIRCUITS FOR SAMPLE-BASED QUANTUM DIAGONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning quantum sampling circuits requires searching discrete architectures under physical constraints and a downstream scientific objective. We study this problem in sample-based quantum diagonalization (SQD), which estimates molecular ground-state energies by classically diagonalizing a Hamiltonian in a subspace sampled from a quantum circuit. A widely used sampler is the local unitary cluster Jastrow (LUCJ) ansatz parameterized from classical coupled-cluster amplitudes, which can become unreliable under strong correlation. We formulate sampler construction as a sequential decision process and introduce a neuro-symbolic reinforcement-learning approach that combines an actor–critic policy with a particle-number-preserving gate vocabulary and symbolic construction rules. A terminal reward based on the SQD energy and circuit length directly guides the search toward accurate, compact samplers. Across the tested molecular geometries of N_2, H_2O, and LiH, shallow searched circuits recover active-space ground-state energies in simulation. In the larger N_2 (14-electron, 10-orbital) active space, a searched circuit achieves chemical accuracy with 48 transpiled CZ gates, compared with 31,260 in the automatically constructed UCJ baseline. An ablation isolates the learned policy's contribution as improved search consistency rather than improved attainable accuracy. Executions on IBM Heron-class hardware retain 49–74% of measurements in the valid particle-number sector. The studied active spaces remain small enough that recovered energies do not discriminate subspace quality after configuration recovery. These results establish circuit efficiency and hardware feasibility for learned SQD samplers, while leaving their advantage in subspace quality at larger scales an open question.

1 INTRODUCTION

Determining the ground-state energy of a correlated electronic system is a central task in quantum chemistry, yet the cost of an exact solution grows combinatorially. Classical simulation of generic many-electron states is exponentially hard (Bauer et al., 2020), and exact diagonalization through full configuration interaction (FCI) is in practice confined to small active spaces. Even state-of-the-art distributed FCI, which reaches on the order of a trillion determinants (Gao et al., 2024), falls short of larger correlated active spaces. This gap motivates quantum-centric approaches, among them sample-based quantum diagonalization (SQD). SQD selects configurations by sampling a quantum state and then diagonalizes the Hamiltonian classically in the resulting subspace (Kanno et al., 2026; Robledo-Moreno et al., 2025), and it has recently reached active spaces beyond the scale of exact diagonalization on a quantum-centric supercomputer (Robledo-Moreno et al., 2025).

SQD departs from the variational quantum eigensolver (VQE), the standard near-term approach, in which a parameterized circuit is optimized by measuring the Hamiltonian expectation value on hardware (Kandala et al., 2017). Under noise the VQE energy is no longer guaranteed to be a variational upper bound, and the number of measurements needed to estimate it accurately scales prohibitively (Kanno et al., 2026; Robledo-Moreno et al., 2025). SQD avoids both problems by using the circuit only as a sampler and recovering a variational upper bound through classical diagonalization (Kanno et al., 2026). Its accuracy therefore depends critically on which configurations the circuit samples. In practice the sampling circuit is fixed to the local unitary cluster Jastrow (LUCJ)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ansatz (Motta et al., 2023), a local approximation of the unitary cluster Jastrow form (Matsuzawa & Kurashige, 2020), whose parameters are taken directly from a classical coupled-cluster (CCSD) calculation without further variational optimization (Robledo-Moreno et al., 2025).

The choice of sampling ansatz is a design decision rather than a fixed ingredient, and the standard LUCJ choice carries documented costs. SQD convergence can plateau, because sampling from a fixed ansatz repeatedly returns configurations that have already been seen, so that new determinants are discovered increasingly slowly as sampling continues (Reinholdt et al., 2025; Wray et al., 2025). Overcoming this plateau requires a higher expansion order of the LUCJ operator, which raises the two-qubit gate count linearly, and reaching chemical accuracy for N_2 has required expansion orders of four to eight (Wray et al., 2025). The sampled basis is also skewed: when the CCSD amplitudes are small, the LUCJ state becomes over-concentrated around the Hartree–Fock configuration (Robledo-Moreno et al., 2025), leaving high-excitation determinants under-sampled. Because the LUCJ parameters derive from CCSD amplitudes, the construction also inherits the breakdown of the single-reference description under strong correlation: CCSD is non-variational, and at stretched bond lengths the restricted Hartree–Fock reference becomes unstable toward a symmetry-broken solution, so CCSD over-correlates and its amplitudes become unreliable (Robledo-Moreno et al., 2025; Bulik et al., 2015). Even under full variational optimization, LUCJ parameter optimization for N_2 is pathological at both equilibrium and stretched geometries, where it exhibits energy cusps and oscillatory gradients (Motta et al., 2023). The CCSD-parameterized LUCJ sampler is thus least reliable in exactly the strongly correlated regime where a quantum method would be most valuable.

This has prompted work to improve the SQD sampler or its post-processing (Section 2.1). We pursue a complementary route: rather than fixing the sampling ansatz, we search for one whose samples yield the lowest SQD energy. Reinforcement learning (RL) has proven effective for quantum circuit architecture search (Ostaszewski et al., 2021; Patel et al., 2024; Foderà et al., 2024; Krumtünger et al., 2025), but existing methods score candidate circuits by a variational energy expectation and draw from gate pools that do not preserve electronic symmetries (Kandala et al., 2017). How a symmetry-preserving circuit populates the SQD subspace, and how efficiently, remains unexplored, as does a learned architecture search scored directly by SQD.

Viewed as a learning problem, ours is a sequential circuit-construction search with a quantum-evaluated reward. We contribute a neuro-symbolic reinforcement-learning (NSRL) framework that searches a symmetry-preserving gate vocabulary scored directly by the SQD energy. We evaluate across the dissociation curves of N_2 and H_2O and at equilibrium for LiH, and at representative geometries on IBM Heron-class hardware. An ablation separates what the search space contributes from what the learned policy contributes. To our knowledge this is the first learned architecture search coupled to an SQD objective.

Our primary study uses N_2 in a minimal STO-3G basis with a frozen-core $(10e, 8o)$ active space, together with H_2O $(8e, 6o)$ and LiH $(4e, 6o)$. We additionally characterize the sampler in the larger all-electron N_2 $(14e, 10o)$ space, where the sampling-budget dependence, the noise robustness, and the policy ablation are reported. We discuss the resulting limitations and the path to larger systems in Section 5.

2 Related Work

2.1 Sample-Based Quantum Diagonalization

SQD and its relatives recover ground states by sampling electronic configurations from a quantum circuit and diagonalizing the Hamiltonian classically in the resulting subspace (Robledo-Moreno et al., 2025; Kanno et al., 2026). Subsequent work has broadened the framework on its classical side, trading the depth of time-evolution circuits for convergence guarantees (Yu et al., 2025; Piccinelli et al., 2025), extending the subspace to low-lying excited states (Barison et al., 2025), and strengthening the classical post-processing through auxiliary-field quantum Monte Carlo applied to truncated SQD wavefunctions (Danilov et al., 2025). A parallel effort improves the input state itself, either by Hamiltonian time evolution (Mikkelsen & Nakagawa, 2025) or by adaptively growing the preparation circuit (Nakagawa et al., 2024). Even so, the sampler in the original and most widely used SQD experiments is the LUCJ ansatz (Motta et al., 2023), parameterized directly from a classical CCSD calculation and left optimization-free (Robledo-Moreno et al., 2025). We act on the circuit

instead of the classical pipeline: where input-state methods grow it greedily (Nakagawa et al., 2024) or determine it analytically, we learn the architecture, targeting shallow, sample-efficient samplers rather than convergence guarantees.

2.2 Learning-Based Quantum Architecture Search

Reinforcement learning has been used to construct quantum circuits: foundational architecture search (Ostaszewski et al., 2021), curriculum search under hardware noise (Patel et al., 2024), variational circuit design (Foderà et al., 2024), structured and modular synthesis (Turati et al., 2025), and gadget-based composition on real hardware (Kundu & Sarra, 2026); related adaptive methods grow an ansatz greedily, without a learned policy (Grimsley et al., 2019; Tang et al., 2021). Across this body of work the scoring objective is uniform: a candidate circuit is evaluated by a variational energy expectation, whether the VQE energy $\langle\psi(\theta)|H|\psi(\theta)\rangle$ (Ostaszewski et al., 2021; Foderà et al., 2024; Turati et al., 2025; Patel et al., 2024; Kundu & Sarra, 2026) or the QAOA cost (Khairy et al., 2020), with the circuit treated as a state to be variationally optimized. None scores by the quality of the sampled subspace, the objective that defines SQD. The one method that does, ADAPT-QSCI (Nakagawa et al., 2024), selects each gate greedily by immediate energy decrease and is not a learned policy. Our search differs on both axes at once: the selection rule is a learned policy rather than a fixed greedy criterion, and the search space is a symbolic, symmetry-preserving gate grammar rather than a flat pool of rotations and entanglers. Table 3 in Appendix A summarizes this positioning.

2.3 Symmetry-Preserving Circuits

Particle-number-preserving and Hamming-weight-preserving circuits are an established design principle for physically valid ansätze whose trainability can be characterized analytically (Gard et al., 2020; Anselmetti et al., 2021; Ayeni, 2025; Monbroussou et al., 2025). There, Krumtünger et al. (2025) show the failure mode without such constraints: a symmetry-breaking ansatz lets an RL agent lower the energy by leaving the physical subspace. Because every measurement from a symmetry-preserving sampler lands in the correct sector, no shots are spent on configurations recovery must later discard. This link between symmetry and sampling efficiency, rather than trainability, is what we exploit.

3 Methodology

We frame the design of an SQD sampling circuit as a sequential decision problem and solve it with reinforcement learning. Within the SQD pipeline, every stage except the sampling circuit is standard and fixed; the circuit is the single component we learn. Figure 1 places our contribution in that pipeline. A symbolic grammar restricts the agent at each step to a small vocabulary of particle-number-preserving operations, an actor–critic network proposes the next operation, and the completed circuit is scored by the energy that SQD recovers from its samples. Every circuit the agent can construct is therefore physically admissible, and the reward is the SQD energy rather than a variational expectation value.

3.1 Circuit construction as a Markov decision process

A candidate circuit is built one operation at a time: the state s_t summarizes the partial circuit, an action a_t appends one operation from a fixed vocabulary, and an episode ends on a stop action or at a depth limit. Only the completed circuit carries a reward, the negative SQD energy gap together with a small length penalty, so the task is a finite-horizon Markov decision process with a terminal reward. We adopt this reinforcement-learning formulation of circuit search from prior architecture-search work (Ostaszewski et al., 2021; Patel et al., 2024); our departure is the reward, defined below, and the symmetry-preserving search space.

3.2 SQD evaluation and the reward backend

The reward is supplied by the SQD pipeline, held fixed throughout. Its lowest eigenvalue E_R is a variational upper bound, $E_0 \leq E_R$, that survives statistical and physical noise because it follows from diagonalization in a fixed subspace rather than a hardware expectation value (Kanno

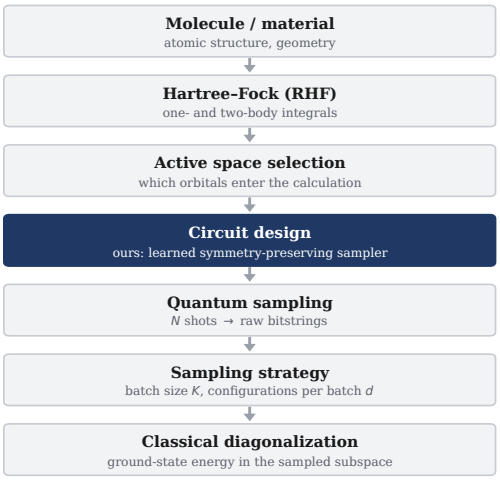

Figure 1: The SQD pipeline. A molecular Hamiltonian is reduced to an active space, a quantum circuit samples electronic configurations, and the Hamiltonian is diagonalized classically in the sampled subspace. Every stage is standard SQD except the circuit, which we replace with a learned, symmetry-preserving sampler (highlighted).

et al., 2026); a circuit sampling a better subspace yields a tighter bound. Given a circuit, we sample measurement outcomes in the computational basis and, following Robledo-Moreno et al. (2025), first apply configuration recovery, which repairs bitstrings whose particle number is incorrect by reference to the current occupation estimate. We then retain the configurations lying in the correct particle-number sector and diagonalize the molecular Hamiltonian in the subspace they span (Robledo-Moreno et al., 2025; Kanno et al., 2026). We report the gap of the resulting eigenvalue to the FCI energy of the active space, $\mathrm{gap} = E_{\mathrm{SQD}} - E_{\mathrm{FCI}}$, in milli-Hartree.

3.3 A PARTICLE-NUMBER-PRESERVING ACTION VOCABULARY

The agent selects from a discrete vocabulary of Givens rotations, a stop action, and two collective "brick" actions described at the end of this section. Under a fermion-to-qubit mapping, each electronic configuration is a computational-basis state whose Hamming weight equals the electron number. Each Givens rotation acts on one occupied–virtual orbital pair within a single spin sector at a fixed angle. For the primary N_2 $(10e, 8o)$ active space there are five occupied and three virtual orbitals per spin sector, hence fifteen occupied–virtual pairs in each of the α and β sectors, giving thirty single-rotation actions; with the two bricks and the stop action the full vocabulary contains thirty-three actions. The larger all-electron N_2 $(14e, 10o)$ space yields a correspondingly larger vocabulary under the same rule. Figure 3 in Appendix C shows the agent selecting from this vocabulary under the grammar.

Each Givens rotation is generated by an $XX+YY$ interaction, which is block-diagonal in Hamming weight, so every action maps a fixed-electron-number state to another state of the same electron number. Under ideal execution such a circuit cannot leave the correct particle-number sector, so postselection discards none of its weight on symmetry grounds; on hardware, gate and readout errors still move population out of the sector (Section 4.6). Symmetry constrains which configurations are reachable, not the amplitudes that set the sampling distribution. A vocabulary of raw single- and two-qubit gates would forfeit this guarantee, leaving the agent to rediscover particle-number conservation from reward alone.

We fix the rotation angle at $\pi/8$, which keeps the action space discrete. The angle is not incidental: it sets the configurational spread of the brick action. Each Givens rotation transfers amplitude to its excited configuration in an amount that grows with the angle, until at $\pi/2$ it becomes a basis-state permutation, collapsing the spread over determinants. The recovered energy for N_2 $(10e, 8o)$ is insensitive to this over the lower range, 0.000 mHa at $\pi/8$ and 0.000–0.628 mHa at $\pi/4$, and degrades

to between 127 and 372 mHa across geometries at $\pi/2$. The two brick actions each apply every Givens rotation of one spin sector, in a fixed occupied-major order, populating many configurations in a single step; because the rotations share qubits they are applied in sequence, not in parallel. The learned policies for the primary systems typically place a single brick early in the circuit; in the larger N_2 $(14e, 10o)$ space we additionally study a Givens-and-stop variant without bricks.

3.4 Symbolic grammar

A symbolic grammar filters the vocabulary at every step so the network chooses only among operations yielding a valid, hardware-reasonable circuit. It enforces a maximum depth, forbids two identical operations in succession, requires at least two operations before the stop action becomes available, and forces termination once the depth limit is reached. The depth limit keeps circuits within a range that is realistic to transpile and execute, and the remaining rules remove degenerate sequences that waste search effort.

3.5 Policy and value network

The agent is an actor–critic network with a shared representation. The state s_t is encoded as a fixed-length vector recording the normalized depth of the partial circuit, the reward returned at the previous step, a one-hot indicator of the most recent action, and a normalized histogram of the actions taken so far. Because intermediate rewards are zero (Section 3.6), the second entry is zero at every non-terminal step. A single hidden layer with a rectified-linear activation maps the state to a shared representation, from which a policy head produces logits over the action vocabulary and a value head produces a scalar estimate of the state value. The grammar of Section 3.4 masks the policy logits to the admissible set, and the next action is sampled from the resulting distribution, which keeps the procedure on-policy.

3.6 Reward

The reward is terminal. When an episode ends, the completed circuit is evaluated by the SQD pipeline of Section 3.2, and the agent receives

$$r \;=\; -\,\frac{\mathrm{gap}_{\mathrm{mHa}}}{\sigma} \;-\; \lambda\,|c|, \tag{1}$$

where $\mathrm{gap}_{\mathrm{mHa}}$ is the SQD energy gap to FCI in milli-Hartree, $|c|$ is the number of operations in the circuit, σ scales the energy term, and λ weights the length penalty. The energy term is the objective; the length penalty is a tie-breaker that favours the shorter of two circuits of comparable energy. In the large-space runs reported below every episode reaches the depth limit, so $\lambda|c|$ is constant across circuits and the ranking is set by the energy term alone. Intermediate steps return zero reward, so the full-episode return reduces to the discounted terminal reward. In the reported runs $\sigma = 50$, which maps the energy gap into a well-conditioned range, and $\lambda = 0.003$, small enough that $\lambda|c|$ never exceeds a few hundredths.

3.7 Training

We train the network with on-policy advantage actor–critic (A2C) using Monte-Carlo returns. Because the reward is terminal, we estimate the return of each step as the full discounted sum of rewards to the end of the episode, without bootstrapping from the value head; the value head serves as a baseline. For a batch of episodes collected under the current policy, the advantage of each step is the Monte-Carlo return minus the baseline, the policy is updated by the corresponding policy-gradient term, the value head is regressed onto the returns, and an entropy bonus on the policy encourages exploration. Exploration comes entirely from the stochastic policy and this entropy bonus; we do not use epsilon-greedy exploration. Updates are applied once per fixed number of collected episodes. Because the training reward is evaluated at a reduced sampling budget for efficiency, the search can favor circuits whose advantage is specific to that budget. To guard against this, we retain the best K circuits encountered during training and re-evaluate each at the full validation budget, selecting the circuit with the lowest validation gap as the reported policy. All hyperparameters are listed in Appendix B.

Table 1: Energy gap to active-space FCI across molecular geometries, searched sampler versus auto-UCJ, at gate angle $\pi/8$. Searched-circuit gaps are over three seeds for N_2 and H_2O, one for LiH. Auto-UCJ is deterministic given the CCSD amplitudes; its ranges are the spread of its SQD energy over the same seeds, which vary the sampling draws.

Molecule	R (Å)	NSRL gap (mHa)	auto-UCJ gap (mHa)	NSRL 2q gates
N_2 $(10e, 8o)$	1.1	0.000	1.33–1.57	16–17
N_2 $(10e, 8o)$	1.8	0.000	1.68–3.46	16–17
N_2 $(10e, 8o)$	2.5	0.000	0.57–2.46	16
H_2O $(8e, 6o)$	0.96	0.000	0.111	9–10
H_2O $(8e, 6o)$	1.5	0.000	0.047	9–10
H_2O $(8e, 6o)$	2.0	0.000	0.028	9–10
LiH $(4e, 6o)$	1.6	0.000	0.167	10

The circuits for the primary systems (Sections 4.2, 4.6) come from an earlier version of this search with the same environment, vocabulary, grammar and reward but an off-policy update: replay buffer, bootstrapped critic target, ϵ-greedy exploration, and no top-K validation (Appendix B). Every N_2 $(14e, 10o)$ result, including the ablation, uses the on-policy formulation above.

4 Results

4.1 Experimental setup

We study three molecules in STO-3G active spaces. The primary systems, used for the dissociation-curve and hardware experiments, are N_2 $(10e, 8o)$, 16 qubits; H_2O $(8e, 6o)$, 12 qubits; and LiH $(4e, 6o)$, 12 qubits. To characterize the sampler in greater depth, we additionally use the larger all-electron N_2 active space $(14e, 10o)$, 20 qubits, at bond length $R = 1.2$ Å, for the sampling-budget study, the noise sweep, and the policy ablation. Each candidate circuit is evaluated through an identical SQD solver.

The circuits reported for the primary systems are drawn from the full vocabulary including the two brick actions, and the learned policies typically place a single brick early in the circuit. The larger-space $(14e, 10o)$ circuits use the Givens-and-stop vocabulary. Circuit sizes are logical two-qubit gate counts, or transpiled CZ counts where a device compilation is referenced; the two are never conflated. Searched circuits use a fixed gate angle of $\pi/8$ unless stated otherwise, and gaps are in milli-Hartree to three decimals, a gap below the 10^{-3} mHa resolution written 0.000. Our baseline is the spin-balanced UCJ ansatz built from the CCSD amplitudes with no restriction on orbital interactions; "auto-UCJ" denotes the untruncated construction, as distinct from truncations at $L = 1$ and $L = 2$. It is the unrestricted parent of the local LUCJ ansatz used in SQD (Robledo-Moreno et al., 2025): more expressive but deeper, so our gate-count ratios exceed what a connectivity-restricted LUCJ would give. Gate counts in Sections 4.3 and 4.5 are transpiled to `cz`/`rz`/`sx`/`x` with all-to-all connectivity and no routing, so they exclude the SWAP overhead that Section 4.6 incurs on the device.

4.2 Gate count across molecular geometries

We compare the searched sampler with the auto-UCJ ansatz across geometries, including the stretched ones where the single-reference CCSD amplitudes that parameterize LUCJ become unreliable. At every geometry of Table 1, spanning all three molecules, the searched circuit recovers the active-space FCI energy to within 0.000 mHa using between 9 and 17 logical two-qubit gates, while the auto-UCJ ansatz retains a finite gap at every geometry. At the most stretched LiH geometry we examined, $R = 2.5$ Å, two of three seeds instead returned gaps of 0.069 and 0.647 mHa at the same episode budget, and we exclude that point from Table 1. Compact symmetry-preserving samplers therefore exist and are constructible across geometries; what the policy itself contributes is isolated in Section 4.4. Table 1 reports the comparison at $\pi/8$, over three seeds per point for N_2 and H_2O and one for LiH.

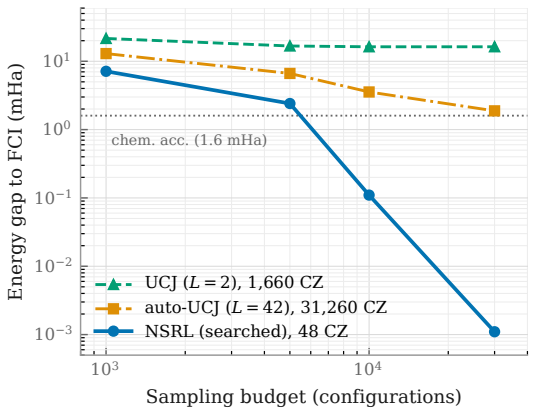

Figure 2: Energy gap to FCI versus sampling budget for N_2 $(14e, 10o)$ at $R = 1.2$ Å (log–log axes). The searched circuit (48 CZ gates) reaches 0.001 mHa by 3×10^4 samples, while the auto-UCJ ansatz improves only gradually despite 31,260 CZ gates, and the truncated UCJ ansatz (L=2, shown) converges to a fixed gap that further sampling does not reduce; the L=1 ansatz reaches the same plateau from a larger initial gap. The dotted line marks chemical accuracy (1.6 mHa).

The searched circuit recovers the ground state at every geometry, including the stretched N_2 geometries $R = 1.8$ and $R = 2.5$ Å in the strongly correlated regime, at a shallow, fixed gate count. Because each of these active spaces is small (the particle-number-conserving dimension is $D = 3{,}136$ for N_2, 225 for H_2O, and 225 for LiH, all below the sampling budget), the recovery of FCI by the searched circuit is a statement of gate efficiency and of constructibility across geometries, and not of subspace quality; the latter is addressed in Section 5.

4.3 Sampling-budget dependence

In the larger all-electron N_2 $(14e, 10o)$ space at $R = 1.2$ Å ($E_{\mathrm{FCI}} = -107.6773$ Ha), we resolve the energy gap as a function of sampling budget. The searched circuit, of 24 logical gates (48 transpiled CZ gates), reaches a gap of 7.13 mHa at 1,000 samples and decreases monotonically to 2.41, 0.11, and 0.001 mHa at 5,000, 10,000, and 30,000 samples. The auto-UCJ ansatz reaches 1.88 mHa at the largest budget using 31,260 CZ gates, improving only gradually with budget, while the truncated UCJ ansätze at $L = 1$ and $L = 2$ reach a common gap of 16.29 mHa that further sampling does not reduce, $L = 1$ approaching it more slowly (30.04 mHa at 1,000 samples against 21.56 for $L = 2$). Uniform sampling within the correct sector also recovers FCI at every budget tested here (Section 5); it is a diagnostic of subspace size, not a competing sampler, since it carries no circuit a device could execute. The comparison among preparable circuits is therefore one of circuit cost at fixed recovered energy, and the searched circuit reaches it with more than two orders of magnitude fewer two-qubit gates, shallow enough to run on hardware. Figure 2 shows the gap as a function of sampling budget for all samplers.

The searched circuit reaches chemical accuracy with 48 transpiled CZ gates. The auto-UCJ ansatz sits on the threshold: over 31 independent draws at this geometry and budget its gap spans 1.29–1.88 mHa, below 1.6 mHa in 13 of them. Ours is therefore a comparison of gate count at comparable recovered energy, not of accuracy the baseline cannot reach.

4.4 Ablation: learned policy versus undirected search

To separate the learned policy's contribution from that of the gate space, we compare the full policy (actor–critic enabled) against undirected search over the same space (actor–critic disabled), over ten random seeds each, holding vocabulary and grammar fixed. Because both arms sit in a regime where the recovered energy saturates (Section 5), this ablation measures the reliability of the search, not the accuracy it can ultimately reach. Both produce circuits of 24 logical gates (48 CZ gates); for each seed we record the gap to FCI of the validated best circuit at a 30,000-sample budget. The

Table 2: Searched circuits executed on ibm_marrakesh (Heron r2). Gaps are relative to the active-space FCI energy.

Molecule	Geometry	CZ	depth	gap (mHa)	valid (%)
N_2 $(10e, 8o)$	R=1.1	59	128	0.000	65.4
N_2 $(10e, 8o)$	R=1.1 (repeat)	59	129	0.099	50.3
N_2 $(10e, 8o)$	R=1.8	59	129	6.022	49.4
H_2O $(8e, 6o)$	R=0.96	26	76	0.000	66.4
H_2O $(8e, 6o)$	R=0.96 (no mit.)	26	76	0.000	72.0
LiH $(4e, 6o)$	R=1.0	30	80	0.000	74.1

learned policy attains a mean gap of 0.006 mHa (standard deviation 0.010, range 0.000–0.030 over ten seeds), against 0.038 mHa (standard deviation 0.036, range 0.000–0.104) for undirected search. The 95% confidence intervals on the two means are $[0.000, 0.013]$ and $[0.012, 0.064]$ mHa; per-seed values are listed in Appendix B. The learned policy thus yields a lower mean gap and a smaller variance, with a worst-case gap about three times smaller; a Welch's t-test over the ten seeds per arm gives $p \approx 0.02$. Both arms terminate at the depth limit rather than by the stop action, which is why circuit length is identical across configurations. We attribute the improvement in consistency to the learned policy. The advantage of the symmetry-preserving gate space itself is not isolated by this ablation, which varies the policy within that space; its effect on sampling efficiency is the subject of Section 4.5.

4.5 Noise resilience

The SQD solver retains only those measured samples whose total electron number is correct; samples whose Hamming weight has been altered by noise are discarded prior to diagonalization. We use the valid sample fraction p_{hw}, the proportion of samples with the correct Hamming weight, to quantify the usable signal a circuit produces under noise. We apply a depolarizing model with rates representative of ibm_marrakesh (single-qubit gate error $g_1 = 2 \times 10^{-4}$, readout error $\approx 10^{-2}$) and vary the two-qubit gate error g_2 over $\{0.001, 0.003, 0.005, 0.010\}$, a range bracketing the device median $g_2 \approx 0.003$.

Across $g_2 = 0.001$ to 0.010 the searched circuit's valid sample fraction falls from 0.771 to 0.603, while the truncated UCJ circuits fall from 0.355 to 0.037 (L=1) and 0.200 to 0.018 (L=2). At the device median $g_2 = 0.003$ the searched circuit retains 0.727 against 0.148 and 0.069; its advantage over L=1 grows from roughly fivefold there to more than an order of magnitude at $g_2 = 0.010$, following the dependence of p_{hw} on two-qubit gate count (Figure 4, Appendix D).

Computed before diagonalization, this quantity characterizes the circuit alone. Without configuration recovery the searched circuit is not favoured: its recovery-free gap stays near 8 mHa up to $g_2 = 0.005$, while the truncated UCJ circuits reach 0.40–1.21 (L=1) and 0.52–34.55 mHa (L=2), their surviving samples concentrated on the CCSD-dominant configurations. We claim no recovery-free energy advantage; our claims are circuit depth and valid-sector retention.

4.6 Hardware experiments

We executed searched circuits on ibm_marrakesh (Heron r2, 156 qubits) for the three molecules, in the frozen-core active spaces of Section 4.1. Each circuit is a searched policy of 26 to 59 transpiled CZ gates, executed at $10{,}000$ shots with readout twirling and dynamical decoupling except where indicated, and evaluated with five SQD iterations, of which we report the last. The transpiled count exceeds twice the logical two-qubit count because routing to the device's heavy-hex connectivity inserts SWAP layers, at a ratio of up to 3.7 here against 2.0 in Section 4.3. Table 2 reports the results; the corresponding IBM job identifiers are listed in Appendix D.

The three circuits executed on hardware each act within a single spin sector (β for N_2, α for H_2O and LiH), and the solver runs in closed-shell mode ($S^2 = 0$), which symmetrizes the α and β determinant lists. The complementary sector is therefore populated by the solver, not the circuit, and separating the two contributions requires a sampler acting on both. This is a property of these

circuits, not of the search space: the N_2 $(14e, 10o)$ policy of Sections 4.3 and 4.4 uses eight α and sixteen β rotations of its twenty-four gates.

In four of the six executions of Table 2, spanning all three molecules, the searched circuits recover the active-space FCI energy to within 0.000 mHa on hardware while retaining 65 to 74% of samples in the correct particle-number sector; the two remaining N_2 executions are taken up below. The LiH circuit was executed at $R = 1.0$ Å, a compressed geometry, and so does not correspond to the LiH row of Table 1. Figure 5 in Appendix D shows why this is possible for the H_2O circuit: the measured bitstrings concentrate on the eight-electron sector, and the stray weight falls into the adjacent sectors rather than across the spectrum, the signature of single bit-flip errors. Postselection therefore discards little.

The same H_2O circuit also recovers the FCI energy without error mitigation, retaining 72% of valid shots, consistent with a depth low enough to run unmitigated. A repeated execution of the N_2 circuit at $R = 1.1$ Å yielded a gap of 0.099 mHa, against 0.000 mHa for the first execution, which we attribute to calibration drift between jobs; both values are reported. At the stretched geometry $R = 1.8$ Å, the gap increased to 6.022 mHa with 49% of shots valid, which we take to indicate the current limit of the hardware experiments.

5 Limitations

The active spaces studied here are small: the particle-number-conserving dimension is at most $D = \binom{10}{7}^2 = 14{,}400$ for the $(14e, 10o)$ N_2 space, comparable to the largest sampling budget and dominated by far fewer determinants than that. Any sampler restricted to the correct sector, uniform sampling included, therefore reaches the FCI energy once configuration recovery is applied; the finite UCJ gaps above are those of the UCJ subspace at the stated budget, not a floor recovery cannot remove. In this regime the recovered energy does not discriminate subspace quality between samplers, and we make no claim that the searched circuit yields a better subspace than an optimized UCJ circuit on the energy–dimension axis of Robledo-Moreno et al. (2025). Our claims are confined to circuit cost, constructibility across geometries, and usable-sample fraction under noise.

Training evaluates the reward through a fixed SQD pipeline at a reduced sampling budget, and the gate angle is held fixed at a value justified empirically in Section 3.3 rather than searched jointly with the architecture. Because the gap to FCI differs from the SQD energy by a constant, the policy gradient is unchanged if E_{SQD} is used directly; the FCI reference enters only for reporting and is unavailable at the scales targeted. The hardware demonstration covers equilibrium geometries cleanly and reaches its limit at strongly stretched bonds, where both the gap and the valid-sample fraction degrade.

Future work. The next step is an active space whose configuration count exceeds the sampling budget by orders of magnitude, where subspace quality rather than recovered energy discriminates: N_2 in 6-31G with a $(10e, 16o)$ space, 32 qubits and $D = \binom{16}{5}^2 \approx 1.9 \times 10^7$. A sampler acting on both spin sectors and a joint search over the gate angle are the other open directions; Appendix E sketches richer grammars.

6 Conclusion

We recast SQD sampling-circuit design as a reinforcement-learning problem: a neuro-symbolic agent searching a symmetry-preserving vocabulary scored by the SQD energy finds shallow circuits that recover the active-space ground state at every geometry reported, reach chemical accuracy in the larger N_2 space with more than two orders of magnitude fewer two-qubit gates than auto-UCJ, and run on IBM Heron-class hardware with half to three-quarters of measurements in the valid sector. These spaces remain small enough that the recovered energy does not by itself separate samplers on subspace quality; obtaining that separation is the object of the larger-space experiments above.

AI USE STATEMENT

We used a large language model for writing assistance: editing prose for clarity and concision and checking the manuscript for internal consistency. It was not used to generate research ideas, to design the method or the experiments, or to produce the experimental results, all of which are the authors' own work. The authors reviewed all AI-assisted text and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The search method is specified in full in Section 3: the action vocabulary and its construction rule in Section 3.3, the symbolic grammar in Section 3.4, the network architecture and state encoding in Section 3.5, the reward in Section 3.6, and the training procedure in Section 3.7. The SQD evaluation pipeline, held fixed throughout, is described in Section 3.2. All hyperparameters are listed in Appendix B. Active spaces, qubit counts, geometries, sampling budgets, and the gate-count reporting convention are given in Section 4.1; hardware execution settings in Section 4.6.

REFERENCES

Gian-Luca R Anselmetti, David Wierichs, Christian Gogolin, and Robert M Parrish. Local, expressive, quantum-number-preserving VQE ansätze for fermionic systems. *New Journal of Physics*, 23(11):113010, 2021.

Babatunde M Ayeni. Efficient particle-conserving symmetric quantum circuits. *Physical Review A*, 111(2):022629, 2025.

Stefano Barison, Javier Robledo Moreno, and Mario Motta. Quantum-centric computation of molecular excited states with extended sample-based quantum diagonalization. *Quantum Science and Technology*, 10(2):025034, 2025.

Bela Bauer, Sergey Bravyi, Mario Motta, and Garnet Kin-Lic Chan. Quantum algorithms for quantum chemistry and quantum materials science. *Chemical Reviews*, 120(22):12685–12717, 2020.

Ireneusz W Bulik, Thomas M Henderson, and Gustavo E Scuseria. Can single-reference coupled cluster theory describe static correlation? *Journal of Chemical Theory and Computation*, 11(7): 3171–3179, 2015.

Don Danilov, Javier Robledo-Moreno, Kevin J Sung, Mario Motta, and James Shee. Enhancing the accuracy and efficiency of sample-based quantum diagonalization with phaseless auxiliary-field quantum Monte Carlo. *Journal of Chemical Theory and Computation*, 21(22):11585–11594, 2025.

Simone Foderà, Gloria Turati, Riccardo Nembrini, Maurizio Ferrari Dacrema, and Paolo Cremonesi. Reinforcement learning for variational quantum circuits design. *arXiv preprint arXiv:2409.05475*, 2024.

Hong Gao, Satoshi Imamura, Akihiko Kasagi, and Eiji Yoshida. Distributed implementation of full configuration interaction for one trillion determinants. *Journal of Chemical Theory and Computation*, 20(3):1185–1192, 2024.

Bryan T Gard, Linghua Zhu, George S Barron, Nicholas J Mayhall, Sophia E Economou, and Edwin Barnes. Efficient symmetry-preserving state preparation circuits for the variational quantum eigensolver algorithm. *npj Quantum Information*, 6(1):10, 2020.

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

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

Keita Kanno, Masaya Kohda, Ryosuke Imai, Sho Koh, Kosuke Mitarai, Wataru Mizukami, and Yuya O Nakagawa. Quantum-selected configuration interaction: Classical diagonalization of Hamiltonians in subspaces selected by quantum computers. *Physical Review Research*, 8(2): 023268, 2026.

Sami Khairy, Ruslan Shaydulin, Lukasz Cincio, Yuri Alexeev, and Prasanna Balaprakash. Learning to optimize variational quantum circuits to solve combinatorial problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2367–2375, 2020.

Maureen Krumtünger, Alissa Wilms, Paul K. Faehrmann, Jens Eisert, Jakob Kottmann, Paolo Andrea Erdman, and Sumeet Khatri. Reinforcement learning of quantum circuit architectures for molecular potential energy curves. *arXiv preprint arXiv:2511.16559*, 2025.

Akash Kundu and Leopoldo Sarra. Reinforcement learning with learned gadgets to tackle hard quantum problems on real hardware. *Communications Physics*, 2026. In press.

Yuta Matsuzawa and Yuki Kurashige. Jastrow-type decomposition in quantum chemistry for low-depth quantum circuits. *Journal of Chemical Theory and Computation*, 16(2):944–952, 2020.

Mathias Mikkelsen and Yuya O Nakagawa. Quantum-selected configuration interaction with time-evolved state. *Physical Review Research*, 7(4):043043, 2025.

Léo Monbroussou, Eliott Z Mamon, Jonas Landman, Alex B Grilo, Romain Kukla, and Elham Kashefi. Trainability and expressivity of Hamming-weight preserving quantum circuits for machine learning. *Quantum*, 9:1745, 2025.

Mario Motta, Kevin J Sung, K Birgitta Whaley, Martin Head-Gordon, and James Shee. Bridging physical intuition and hardware efficiency for correlated electronic states: the local unitary cluster Jastrow ansatz for electronic structure. *Chemical Science*, 14(40):11213–11227, 2023.

Yuya O Nakagawa, Masahiko Kamoshita, Wataru Mizukami, Shotaro Sudo, and Yu-ya Ohnishi. ADAPT-QSCI: Adaptive construction of an input state for quantum-selected configuration interaction. *Journal of Chemical Theory and Computation*, 20(24):10817–10825, 2024.

Mateusz Ostaszewski, Lea M Trenkwalder, Wojciech Masarczyk, Eleanor Scerri, and Vedran Dunjko. Reinforcement learning for optimization of variational quantum circuit architectures. *Advances in Neural Information Processing Systems*, 34:18182–18194, 2021.

Yash J. Patel, Akash Kundu, Mateusz Ostaszewski, Xavier Bonet-Monroig, Vedran Dunjko, and Onur Danaci. Curriculum reinforcement learning for quantum architecture search under hardware errors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rINBD8jPoP.

Samuele Piccinelli, Alberto Baiardi, Max Rossmannek, Almudena Carrera Vazquez, Francesco Tacchino, Stefano Mensa, Edoardo Altamura, Ali Alavi, Mario Motta, Javier Robledo-Moreno, et al. Quantum chemistry with provable convergence via randomized sample-based quantum diagonalization. *arXiv preprint arXiv:2508.02578*, 2025.

Peter Reinholdt, Karl Michael Ziems, Erik Rosendahl Kjellgren, Sonia Coriani, Stephan PA Sauer, and Jacob Kongsted. Critical limitations in quantum-selected configuration interaction methods. *Journal of Chemical Theory and Computation*, 21(14):6811–6822, 2025.

Javier Robledo-Moreno, Mario Motta, Holger Haas, Ali Javadi-Abhari, Petar Jurcevic, William Kirby, Simon Martiel, Kunal Sharma, Sandeep Sharma, Tomonori Shirakawa, et al. Chemistry beyond the scale of exact diagonalization on a quantum-centric supercomputer. *Science Advances*, 11(25):eadu9991, 2025.

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

Gloria Turati, Simone Foderà, Riccardo Nembrini, Maurizio Ferrari Dacrema, and Paolo Cremonesi. Automated design of structured variational quantum circuits with reinforcement learning. *arXiv preprint arXiv:2507.16001*, 2025.

L Andrew Wray, Cheng-Ju Lin, Vincent Su, and Hrant Gharibyan. Convergence of sample-based quantum diagonalization on a variable-length cuprate chain. *arXiv preprint arXiv:2512.04962*, 2025.

Jeffery Yu, Javier Robledo Moreno, Joseph T Iosue, Luke Bertels, Daniel Claudino, Bryce Fuller, Peter Groszkowski, Travis S Humble, Petar Jurcevic, William Kirby, et al. Quantum-centric algorithm for sample-based Krylov diagonalization. *arXiv preprint arXiv:2501.09702*, 2025.

A Positioning against prior work

Table 3: Positioning against prior circuit-construction methods. "Score" is the objective by which a candidate circuit is evaluated; "Architecture" is whether the circuit structure is fixed, grown greedily, or searched. Ours is the only method that pairs a searched architecture with an SQD objective over a symmetry-preserving gate set.

Method	Score	Architecture	Symmetry	Validation
LUCJ-SQD (Robledo-Moreno et al., 2025)	SQD	fixed	preserving	hardware
ADAPT-QSCI (Nakagawa et al., 2024)	SQD	greedy	preserving	simulation
Ostaszewski et al. (2021); Patel et al. (2024)	VQE	searched	no	simulation
Foderà et al. (2024); Turati et al. (2025); Krumtünger et al. (2025)	VQE	searched	no	simulation
Khairy et al. (2020)	QAOA	searched	no	simulation
This work	SQD	searched	preserving	hardware

B Training hyperparameters

Table 4 lists the hyperparameters used in all reported runs. The same values are used for the learned policy and for the undirected baseline of Section 4.4; the two differ only in whether the actor–critic update is applied.

Configuration recovery is initialized from a uniform occupancy estimate and re-seeded from the lowest-energy batch at each SQD iteration; the fermionic solver is called in closed-shell mode with $S^2 = 0$. Circuits in Sections 4.3 and 4.5 are transpiled to `cz`/`rz`/`sx`/`x` at optimization level 0 with all-to-all connectivity; the hardware circuits of Section 4.6 are transpiled to the device coupling map at optimization level 3 with a pinned transpiler seed, and executed with readout twirling and dynamical decoupling except where stated. Software versions: `ffsim` 0.0.53, `qiskit` 2.4.1, `qiskit-aer` 0.17.2, `qiskit-addon-sqd` 0.9.0, `pyscf` 2.7.0, `pytorch` 2.11.0.

Per-seed validation gaps for the ablation of Section 4.4 (mHa, 30,000 samples). Learned policy: 0.000, 0.003, 0.002, 0.002, 0.001, 0.000, 0.000, 0.030, 0.018, 0.001. Undirected search: 0.059, 0.060, 0.002, 0.030, 0.003, 0.000, 0.060, 0.104, 0.063, 0.002.

C Agent architecture

Figure 3 shows one construction step: the state representation of the partial circuit feeds both heads, the symbolic grammar of Section 3.4 masks the actor's choice to the admissible set, and the completed circuit is returned to the SQD backend for scoring.

D Hardware execution details

Figure 4 gives the noise sweep of Section 4.5 in full, and Figure 5 the Hamming-weight distribution behind the valid-sample fractions of Table 2. Table 5 lists the IBM job identifier for each hardware execution.

Table 4: Hyperparameters. The large-space column applies to the sampling-budget study, the noise sweep, and the policy ablation on N_2 $(14e, 10o)$; the primary-system column applies to the geometry scan and the hardware circuits.

Hyperparameter	N_2 $(14e, 10o)$	Primary systems
Search		
Update rule	on-policy A2C, MC returns	off-policy, TD(0) target
Replay buffer / minibatch	none	10,000 / 32
Exploration	entropy bonus 0.02	ϵ: $1.0 \to 0.05$ over 70%
Final policy selection	best of validated top-K	best training reward
Episodes per gradient update	5	every step once buffer full
Network and optimization		
Hidden layer width	128	128 (64 for LiH)
Optimizer / learning rate	Adam / 10^{-3}	Adam / 10^{-3}
Value loss coefficient	0.5	0.5
Advantage normalization	per batch	none
Gradient-norm clip	1.0	1.0
Discount factor γ	0.99	0.99
Total training episodes	300	500
Depth limit (max. operations)	24	16
Validation pool size K	10	—
Reward and evaluation		
Gate angle	$\pi/8$	$\pi/8$
Energy scale σ / length weight λ	50 / 0.003	50 / 0.003
Training shots / SQD iterations	10,000 / 1	10,000 / 1
Training samples per batch	800	800
Validation shots / SQD iterations	30,000 / 1	30,000 / 1
Validation samples per batch	1600	1600
SQD batches per iteration	10	10

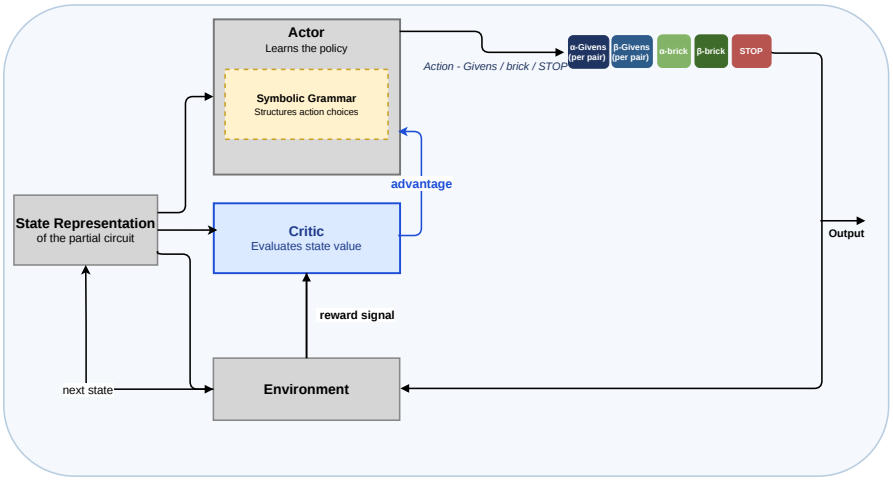

Figure 3: The NSRL agent. At each step the actor–critic network proposes an operation from the particle-number-preserving vocabulary, the symbolic grammar masks the choice to the admissible set, and the operation is appended to the circuit. The completed circuit is scored by the SQD energy, which is returned as the reward.

E Extending the symbolic grammar

The present grammar encodes one chemistry-aware constraint, particle-number preservation. The same mechanism admits richer chemical structure: the action vocabulary and grammar can be ex-

Table 5: IBM Quantum job identifiers for the executions reported in Table 2.

Molecule	Geometry	IBM job identifier
N_2 $(10e, 8o)$	R=1.1	`d8g7sik2upec739jd5qg`
N_2 $(10e, 8o)$	R=1.1 (repeat)	`d8g9elhe8nrc73bfl3f0`
N_2 $(10e, 8o)$	R=1.8	`d8g9of9e8nrc73bfldt0`
H_2O $(8e, 6o)$	R=0.96	`d8g7ndc2upec739jcvt0`
H_2O $(8e, 6o)$	R=0.96 (no mit.)	`d8gaelc2upec739jful0`
LiH $(4e, 6o)$	R=1.0	`d8g7q4s2upec739jd2vg`

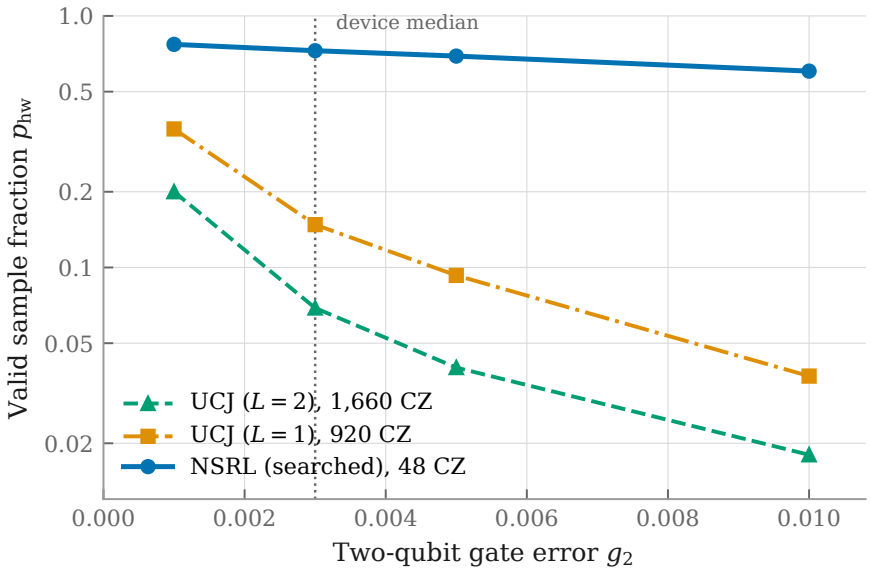

Figure 4: Valid sample fraction p_{hw} under depolarizing noise versus two-qubit gate error g_2 for N_2 $(14e, 10o)$ (3,000 samples, log-y axis). The symmetry-preserving searched circuit (48 CZ gates) retains most of its samples across the range, while the truncated UCJ ansätze lose theirs rapidly as g_2 grows. The advantage over $L = 1$ widens from roughly fivefold at the device median to more than an order of magnitude at $g_2 = 0.010$.

tended to prioritize excitations by orbital locality or by the magnitude of the corresponding Hamiltonian integrals, to encode spin or point-group symmetries beyond particle number, or to seed the search from the dominant coupled-cluster amplitudes instead of a uniform policy. Each of these incorporates additional chemical prior knowledge through the same symbolic interface, and studying whether such rules sharpen the search is a natural continuation of this work.

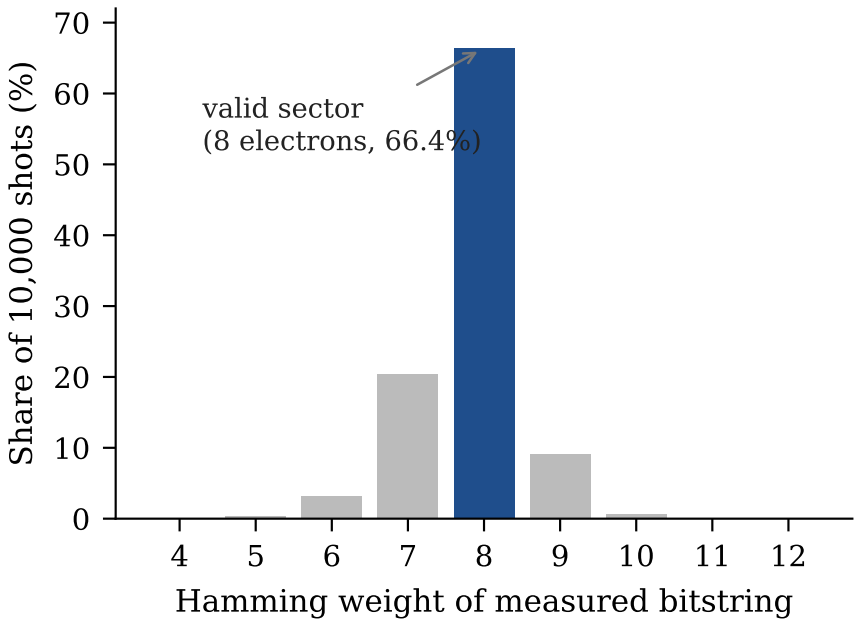

Figure 5: Distribution of measured bitstrings by Hamming weight for the H_2O $(8e, 6o)$ circuit executed on ibm_marrakesh (10,000 shots). The valid eight-electron sector concentrates two-thirds of all shots, with the remaining weight in the adjacent single-error sectors rather than spread across the spectrum, which is what makes postselection effective on hardware.