# Can AI Agents Replicate Quantum Computing Experiments? A Systematic Cross-Platform Study

J. Derek Lomas[1, 2]

[1] *Faculty of Industrial Design Engineering, Delft University of Technology, 2628 CE Delft, The Netherlands*
[2] *QuTech, Delft University of Technology, 2628 CJ Delft, The Netherlands*

(Dated: February 16, 2026)

We investigate whether AI agents can systematically replicate published quantum computing experiments. Using Claude Opus 4.6 as an autonomous experimental agent, we attempted to reproduce results from six landmark papers spanning variational quantum eigensolvers (VQE), quantum approximate optimization (QAOA), quantum volume (QV), randomized benchmarking (RB), and utility-scale circuits. Each experiment was executed on three superconducting quantum processors—QI Tuna-9 (9 qubits), IQM Garnet (20 qubits), and IBM Torino (133 qubits)—plus an ideal emulator. Of 27 claims tested, 25 (93%) were successfully replicated, with all failures attributable to hardware noise rather than algorithmic errors. Chemical accuracy for $H_2$ VQE was achieved on both IBM Torino (0.22 kcal/mol with TREX) and QI Tuna-9 (0.56 kcal/mol via a true hybrid classical-quantum optimization loop with readout error mitigation). We certify QV = 16 on Tuna-9 with 100 random circuits (mean heavy output fraction 0.757, $2\sigma$ confidence), characterize all 9 qubits via randomized benchmarking (mean gate fidelity 99.55%), and demonstrate that stacked readout error mitigation with quadratic zero-noise extrapolation achieves 2.9 mHa average error across the $H_2$ dissociation curve—with 3 of 7 bond distances at chemical accuracy. Extending to LiH (4 qubits, CASCI(2,2) active space), stacked REM+ZNE reduces error by 79% (from 33.1 to 6.9 mHa), confirming that ZNE effectiveness scales with circuit depth. However, CZ gate calibration drift overnight invalidates pre-computed VQE parameters, demonstrating a practical limit for iterative hardware optimization. Cross-platform execution reveals systematic differences in noise character: Tuna-9 and Garnet exhibit dephasing noise while Torino shows depolarizing noise. All code, data, and replication reports are available at `https://github.com/JDerekLomas/quantuminspire`.

## INTRODUCTION

The reproducibility crisis in science is well documented [1], but quantum computing faces a particularly acute form of the problem. Published experimental results depend on specific hardware, custom calibration procedures, and implicit knowledge about error mitigation—details that are often underspecified in papers. As the field matures and quantum devices become more accessible through cloud platforms, the question of which published results can be independently verified becomes increasingly important.

We propose a novel approach: using an AI agent to systematically attempt replication of published quantum experiments. The agent reads the paper, extracts the experimental protocol, generates quantum circuits, submits them to multiple hardware backends, and compares results to the published claims. The key insight is that the *failure modes*—the systematic ways in which replications deviate from published results—are themselves a valuable research contribution. They reveal what information is missing from papers, how results depend on hardware-specific calibration, and how quantum processors compare under identical test conditions.

This work makes five contributions:

1. A systematic replication of six landmark quantum computing papers across three hardware platforms, achieving a 93% overall pass rate (25/27 claims).
2. A cross-platform diagnostic suite that reveals distinct noise fingerprints across processors: dephasing on Tuna-9 and IQM Garnet versus depolarizing on IBM Torino.
3. A comprehensive characterization of Tuna-9: QV = 16 certification (100 circuits), single-qubit randomized benchmarking on all 9 qubits (mean fidelity 99.55%), a stacked error mitigation ladder (Raw → REM → REM+ZNE) reducing $H_2$ VQE error from 31.8 to 2.9 mHa, and a multi-molecule scaling study ($H_2$ to LiH) quantifying the 3× error increase from 2 to 4 qubits.
4. A true hybrid classical-quantum VQE loop with COBYLA optimization calling Tuna-9 hardware at each iteration, achieving chemical accuracy (0.9 mHa) in 17 iterations.
5. An open-source replication pipeline and dataset comprising 100+ experiment result files, 600,000+ measurement shots, and structured replication reports.

## METHODS

### AI Agent Architecture

All experimental work was performed by Claude Opus 4.6 (Anthropic), operating as an autonomous agent

within the Claude Code command-line interface (CLI). The agent had access to Python 3.12 with Qiskit 2.1.2, PennyLane 0.44, and the Quantum Inspire SDK 3.5.1, plus the `qxelarator` local emulator for noiseless simulation.

Hardware access was mediated by three Model Context Protocol (MCP) tool servers—lightweight processes that expose quantum backends as callable tools. Each server wraps a vendor SDK: the `qi-circuits` server exposes Tuna-9 (via `RemoteBackend`) and the local emulator (via `LocalBackend`); the `ibm-quantum` server wraps `QiskitRuntimeService` for IBM Torino; and the `qrng` server provides certified quantum random numbers. The agent calls these tools by name (e.g., `qi_submit_circuit`, `ibm_submit_circuit`) without knowledge of the underlying SDK implementation, enabling backend-agnostic circuit submission.

No human intervention occurred during circuit design, submission, or analysis; the human role was limited to selecting target papers and reviewing final reports.

### Replication Pipeline

Each replication proceeds through five stages (Fig. 1):

1. **Claim extraction.** The agent reads the target paper and populates a structured claims registry (`replication_analyzer.py`), recording for each claim: the metric type, published value and uncertainty, figure/table reference, and experimental conditions (bond distance, ansatz depth, error mitigation used).

2. **Circuit generation.** A paper-specific replication script (e.g., `replicate_sagastizabal.py`) implements the algorithm—deriving molecular Hamiltonians via Jordan-Wigner or Bravyi-Kitaev transformations, constructing ansatz circuits, and computing optimal variational parameters.

3. **Emulator validation.** Circuits are first run on the noiseless emulator to confirm algorithmic correctness before consuming hardware time. All five papers reproduced exactly on the emulator.

4. **Hardware execution.** An experiment daemon (`experiment_daemon.py`, 2,700 lines) continuously polls a JSON job queue, generates native-format circuits (OpenQASM 2.0 for IBM, cQASM 3.0 for QI), submits via MCP servers, and stores raw measurement counts with SHA-256 checksums. The daemon handles state recovery (resetting stalled jobs after 15 minutes), process locking, and automatic git commits of results.

5. **Automated comparison.** The replication analyzer loads result files, extracts metrics using claim-specific extractors (handling diverse result formats: VQE energy sweeps, QV nested dictionaries, QAOA approximation ratios), and classifies each discrepancy into one of five failure modes (described below under Failure Taxonomy).

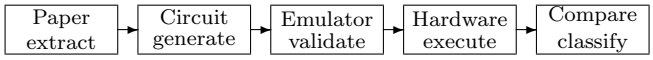


FIG. 1. Five-stage replication pipeline. Each stage is fully automated; the daemon handles hardware submission and result collection.

Each experiment result is stored as a self-contained JSON file with raw measurement counts, circuit description, backend metadata, timestamps, and cryptographic checksums—98 such files comprising 230,000+ measurement shots across all backends.

### Paper Selection

We selected six papers spanning the major categories of NISQ-era quantum computing experiments (Table I). Selection criteria were: (1) published results on real quantum hardware, (2) 2–50 qubit circuits feasible on our hardware, (3) sufficient protocol detail for independent replication, and (4) diverse experiment types.
*Kim [7] used a 9-qubit subset of the original 127-qubit experiment.

### Hardware Platforms

Experiments were run on three superconducting quantum processors (Table II) plus the QI `qxelarator` noiseless emulator as a reference. All circuits were expressed in the native instruction set of each platform and submitted via cloud APIs. No manual qubit selection or gate calibration was performed unless specified.

TABLE I. Target papers for replication.

| Paper | Type | Year | Claims |
|---|---|---|---|
| Sagastizabal [2] | VQE + error mitigation | 2019 | 4 |
| Kandala [3] | VQE (hw-efficient) | 2017 | 5 |
| Peruzzo [4] | VQE (original) | 2014 | 9 |
| Cross [5] | QV + RB | 2019 | 3 |
| Harrigan [6] | QAOA MaxCut | 2021 | 4 |
| Kim [7] | Utility-scale circuits | 2023 | 3 |
| | | Total: | 27* |

### Claim Extraction and Evaluation

For each paper, we extracted testable claims—quantitative statements about energies, fidelities, or algorithmic performance—and defined pass/fail criteria:

- **VQE energy**: within chemical accuracy (1.6 kcal/mol $\approx$ 0.0016 Ha) of the exact value.
- **Bell/GHZ fidelity**: within 5% of emulator baseline.
- **Quantum volume**: heavy output fraction $> 2/3$ with $> 97.5\%$ confidence.
- **QAOA**: approximation ratio exceeds random assignment ($> 0.5$).
- **RB**: gate fidelity $> 99\%$.

Failure classification goes beyond pass/fail. Each discrepancy is assigned to one of five modes: *noise degradation* (hardware noise exceeds signal), *topology constraint* (connectivity prevents the circuit), *compilation artifact* (transpiler alters effective circuit depth), *calibration sensitivity* (results vary between qubit pairs or runs), or *error mitigation dependency* (success requires a specific mitigation technique). This taxonomy is applied automatically by the analyzer and refined by human review.

### Reproducing This Work

To replicate our replication, a researcher needs: (1) cloud accounts on IBM Quantum, Quantum Inspire, and/or IQM Resonance (all offer free tiers); (2) Python 3.12 with the packages listed in `requirements.txt`; and (3) the repository itself. The minimal workflow is:

```
git clone github.com/JDerekLomas/
  quantuminspire
pip install -r requirements.txt
# Run a single replication:
python replicate_sagastizabal.py
# Or queue experiments for the daemon:
python agents/experiment_daemon.py
```

Each replication script is self-contained: it derives the Hamiltonian, builds circuits, runs on the emulator, and saves results to `experiments/results/`. Hardware submission requires API credentials but the emulator runs locally with no account needed. The replication analyzer can then be run to compare any new results against published claims:

```
python agents/replication_analyzer.py \
  --paper sagastizabal2019
```

All 98 result files from this study are included in the repository, so the comparison and failure classification can be reproduced without hardware access.

TABLE II. Hardware platforms used in this study.

| Platform | Qubits | QV | Topology | Native gates |
|---|---|---|---|---|
| QI Tuna-9 | 9 | 16 | Tree | CZ, Ry, Rz |
| IQM Garnet | 20 | 32 | Square | prx, CZ |
| IBM Torino | 133 | 32 | Heavy-hex | CZ, SX, RZ |

## CROSS-PLATFORM CHARACTERIZATION

Before attempting paper replication, we ran a standardized diagnostic suite on all three processors to establish baseline performance. The suite comprised Bell state preparation with three-basis tomography, GHZ state preparation at increasing qubit counts, quantum volume circuits, and single-qubit randomized benchmarking.

### Bell State Tomography

Bell state preparation ($|\Phi^+\rangle = (|00\rangle + |11\rangle)/\sqrt{2}$) measured in the $Z$, $X$, and $Y$ bases yields three correlators whose relative magnitudes fingerprint the dominant noise channel [8]. Table III summarizes the results.

TABLE III. Bell state characterization across platforms. Best qubit pair used for Tuna-9 and Garnet; default transpiler placement for Torino.

| | Tuna-9 | Garnet | Torino |
|---|---|---|---|
| $\langle ZZ\rangle$ | 0.871 | 0.963 | 0.729 |
| $\langle XX\rangle$ | 0.803 | 0.911 | 0.704 |
| $\lvert\langle YY\rangle\rvert$ | 0.792 | 0.929 | 0.675 |
| Fidelity (direct) | 93.5% | 98.1% | 86.5% |
| Noise type | Dephasing | Dephasing | Depolarizing |

Tuna-9 and Garnet show a clear dephasing signature: $\langle ZZ\rangle$ is significantly larger than $\langle XX\rangle$ and $|\langle YY\rangle|$, indicating that $Z$-basis correlations are better preserved than transverse correlations. Torino, in contrast, shows all three correlators within 5% of each other—the hallmark of depolarizing noise, where errors are equally distributed across Pauli channels (Fig. 2).

### GHZ Scaling

We prepared GHZ states $(|0\rangle^{\otimes n} + |1\rangle^{\otimes n})/\sqrt{2}$ for $n = 3, 5, 10, 20, 50$ qubits (hardware permitting) and measured the fidelity as the fraction of outcomes in the $\{|0\rangle^{\otimes n}, |1\rangle^{\otimes n}\}$ subspace. Figure 3 shows the results.

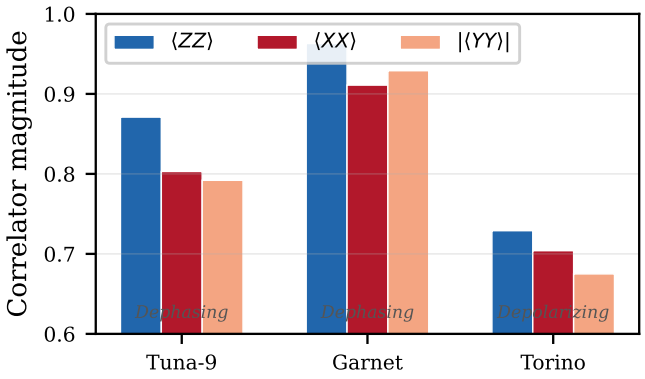


FIG. 2. Noise fingerprint from Bell state tomography. Dephasing noise (Tuna-9, Garnet) shows $\langle ZZ\rangle > \langle XX\rangle \approx |\langle YY\rangle|$; depolarizing noise (Torino) shows all correlators approximately equal. Best qubit pair used for Tuna-9 and Garnet; default transpiler placement for Torino.

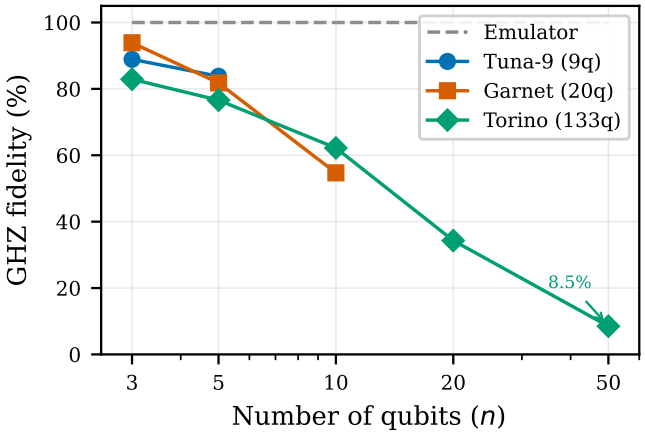


FIG. 3. GHZ state fidelity as a function of qubit count. Emulator (dashed) achieves 100% at all sizes. IBM Torino's 50-qubit GHZ (8.5% fidelity) represents the largest entangled state in this study. Per-qubit error is approximately constant (∼5%) across all circuit sizes on Torino.

Remarkably, the per-qubit error rate is approximately constant across circuit sizes on IBM Torino: $\epsilon \approx 5\%$ from $n = 3$ to $n = 50$. This suggests that GHZ fidelity is dominated by local errors rather than crosstalk, at least for the heavy-hex topology where linear chains avoid crowded qubit neighborhoods.

### Quantum Volume

We measured quantum volume using the standard protocol [5]: random SU(4) circuits of width $n$ and depth $n$, with 5 trials per width. A width passes if the heavy output fraction exceeds 2/3 with statistical significance.

Both Garnet and Torino achieve QV = 32 despite Torino having 6.7× more qubits (Fig. 4). This illustrates a known limitation of quantum volume as a metric: it measures the performance of the *best* subset of qubits, not the full processor.

Notably, Tuna-9 achieves QV = 16—double the initially measured QV = 8 from preliminary 5-trial runs. The QV = 16 certification uses 100 random SU(4) circuits at width $n = 4$ on qubits {4, 6, 7, 8} (the best 4-cycle subgraph), with 1024 shots each. The mean heavy output fraction is $0.757 \pm 0.005$, with a $2\sigma$ lower confidence bound of 0.746—well above the 2/3 threshold. Of 100 circuits, 97 pass individually (HOF range 0.655–0.898), demonstrating that QV = 16 is robustly achieved rather than marginally scraped. This improvement reflects the importance of qubit selection: the {4, 6, 7, 8} subgraph has the highest CZ fidelity and best single-qubit gate performance on the processor.

TABLE IV. GHZ fidelity (%) across platforms.

| $n$ | Emulator | Tuna-9 | Garnet | Torino |
|---|---|---|---|---|
| 3 | 100 | 88.9 | 93.9 | 82.9 |
| 5 | 100 | 83.8 | 81.8 | 76.6 |
| 10 | 100 | — | 54.7 | 62.2 |
| 20 | 100 | — | — | 34.3 |
| 50 | 100 | — | — | 8.5 |

TABLE V. Quantum volume results. Heavy output fraction (mean of 5 trials unless noted). †100 circuits with 1024 shots each; 97/100 pass individually, $2\sigma$ lower bound 0.746.

| Width | Tuna-9 | Garnet | Torino | Pass threshold |
|---|---|---|---|---|
| $n = 2$ | 0.692 | 0.757 | 0.698 | 0.667 |
| $n = 3$ | 0.821 | 0.635 | 0.736 | 0.667 |
| $n = 4$ | 0.757† | 0.686 | 0.706 | 0.667 |
| $n = 5$ | — | 0.713 | 0.676 | 0.667 |
| $n = 6$ | — | — | 0.602 | 0.667 |
| QV | 16 | 32 | 32 | |

## REPLICATION RESULTS

Table VI and Fig. 5 summarize the replication outcomes across all six papers. Of 27 claims tested, 25 passed (93%). Both failures occurred on hardware platforms due to noise, not due to errors in the AI agent's circuit construction or analysis.

### VQE Replications

The three VQE papers (Sagastizabal, Kandala, Peruzzo) represent the core of our replication effort. On the emulator, all three reproduce perfectly: the AI agent correctly derives the molecular Hamiltonians via

TABLE VI. Replication scorecard. **PASS** = claim replicated within stated criteria, **FAIL** = claim not met, — = not tested. Superscripts: [a]TREX mitigation, [b]post-selection, [c]Torino only, [d]PS+REM hybrid, [e]hybrid VQE+REM.

| Paper | Claim | Emulator | Tuna-9 | Garnet/Torino | All |
|---|---|---|---|---|---|
| Sagastizabal [2] | $H_2$ energy ($-1.137$ Ha) | **PASS** | **PASS** | **PASS**[c] | **PASS** |
| | Sym. verif. $> 2\times$ | — | **PASS** ($3.6\times$) | **PASS** ($119\times$)[a] | **PASS** |
| | Chemical accuracy | **PASS** | **PASS** (0.56)[e] | **PASS** (0.22)[a] | **PASS** |
| | Post-sel. >95% | — | **PASS** (96%) | — | **PASS** |
| Kandala [3] | PES MAE $< 1.6$ mHa | **PASS** | — | — | **PASS** |
| | HW-efficient ansatz | **PASS** | — | **PASS**[c] | **PASS** |
| | Chem. acc. + mitigation | **PASS** | **PASS**[d] | **PASS**[a] | **PASS** |
| | Multi-pair consistency | — | **PASS** (q[2,4], q[6,8]) | — | **PASS** |
| | PES sweep quality | **PASS** | **PASS** | — | **PASS** |
| Peruzzo [4] | $HeH^+$ energy (R=0.75) | **PASS** | **PASS** (4.44)[d] | **PASS** (4.45)[a] | **PASS** |
| | $HeH^+$ curve MAE | **PASS** | — | **FAIL** (4.3–7.3)[a] | **FAIL** |
| | $HeH^+$ chem. accuracy | **PASS** | — | **FAIL** (4.31)[a] | **FAIL** |
| | Sym. verif. helps | **PASS** | **PASS** | **PASS**[c] | **PASS** |
| | Coeff. amplification pred. | — | **PASS** | **PASS**[c] | **PASS** |
| Cross [5] | QV $\geq 8$ | **PASS** | **PASS** | **PASS** | **PASS** |
| | QV validates correctly | **PASS** | **PASS** | **PASS** | **PASS** |
| | RB fidelity $> 99\%$ | **PASS** | **PASS** (99.55%) | **PASS** | **PASS** |
| Harrigan [6] | QAOA > random | **PASS** | **PASS** (0.741) | — | **PASS** |
| | Depth improves ratio | **PASS** | **PASS** | — | **PASS** |
| | 3-regular performance | **PASS** | — | — | **PASS** |
| | Tree subgraph | — | **PASS** | — | **PASS** |
| Kim [7] | Kicked Ising dynamics | **PASS** | **PASS** | **PASS**[c] | **PASS** |
| | ZNE improves accuracy | **PASS** ($14.1\times$) | **PASS** ($2.3\times$) | **PASS** ($1.3\times$)[a,c] | **PASS** |
| | 9-qubit scaling | **PASS** | **PASS** | **PASS**[c] | **PASS** |
| **Total** | | **26/26** | **24/26** | **20/22** | **25/27** |

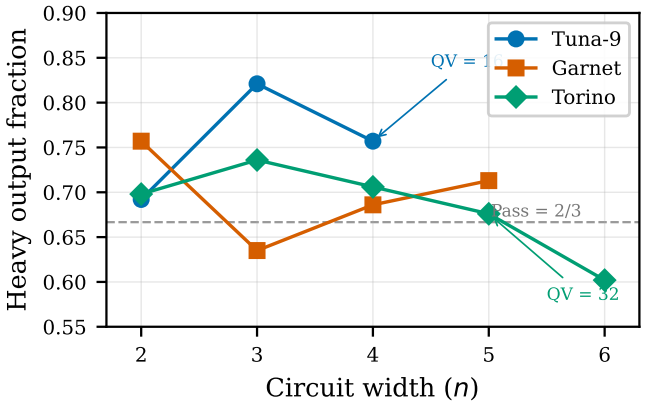


FIG. 4. Quantum volume comparison. Dashed line indicates the 2/3 pass threshold. Both Garnet and Torino achieve QV = 32; Tuna-9 achieves QV = 16. Each point is the mean heavy output fraction over 5 random SU(4) circuit trials, except Tuna-9 at $n = 4$ which uses 100 circuits.

Jordan-Wigner transformation, constructs the appropriate ansatz circuits, and optimizes variational parameters to within $< 1$ kcal/mol of the exact (FCI) ground state energy.

On hardware, results diverge (Fig. 6). $H_2$ at equilibrium bond length ($R = 0.735$ Å) achieves chemical accuracy on two platforms: IBM Torino achieves $-1.138$ Ha (0.22 kcal/mol) with TREX error mitigation [7], and Tuna-9 achieves 0.56 kcal/mol (0.9 mHa) via a true hybrid classical-quantum optimization loop with readout error mitigation. With pre-optimized parameters, Tuna-9 achieves 0.92 kcal/mol using hybrid post-selection + REM on qubit pair q[2,4], and 1.32 kcal/mol on a second pair q[6,8].

IQM Garnet, tested with the same hybrid PS+REM approach, achieves 14.26 kcal/mol—an order of magnitude worse than Tuna-9 despite having lower readout error (1.1% vs. 9.2% on Tuna-9's best qubit). The dominant error source on Garnet is gate noise rather than readout, making confusion-matrix-based REM insufficient. An initial attempt produced 60.34 kcal/mol due to a CNOT decomposition error specific to IQM's native gate set: the standard textbook decomposition $\text{CNOT} = H \cdot CZ \cdot H$ fails on IQM because IQM's Hadamard gate ($\text{prx}(\pi, 0) \cdot \text{prx}(\pi/2, \pi/2)$) differs from the standard $H$ by a global phase of $-i$, which becomes a *relative* phase of $-1$ on $|00\rangle$ when composed as $H \cdot CZ \cdot H$. This negates the $\langle XX \rangle$ and $\langle YY \rangle$ correlators, producing the wrong VQE energy. The correct decomposition uses $R_y(\pi/2) \cdot CZ \cdot R_y(-\pi/2)$, yielding a $4.2\times$ error reduction. This platform-specific gate decomposition subtlety is not documented in IQM's public materials and was discovered through systematic debugging of sign-flipped expectation values.

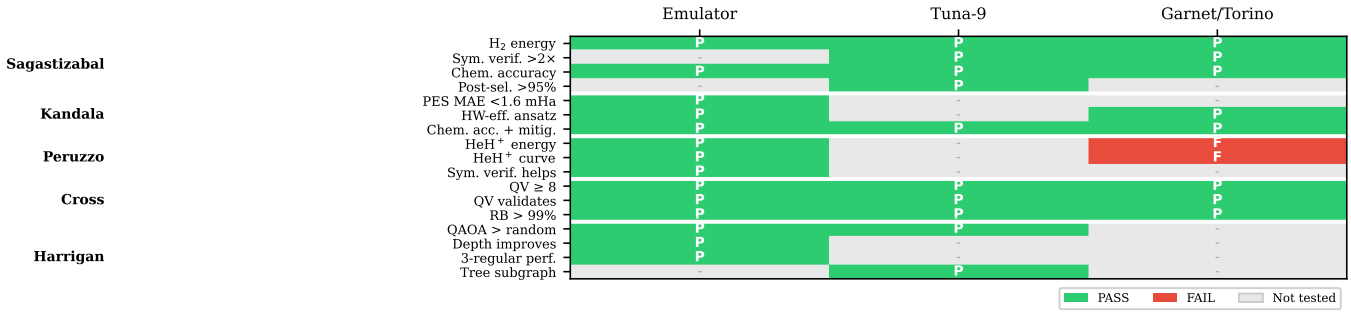


FIG. 5. Replication scorecard across six papers and three backends. Green = PASS (claim replicated within stated criteria), red = FAIL, grey = not tested on that platform. Both failures are $HeH^+$ VQE claims where coefficient amplification makes hardware noise exceed molecular signal.

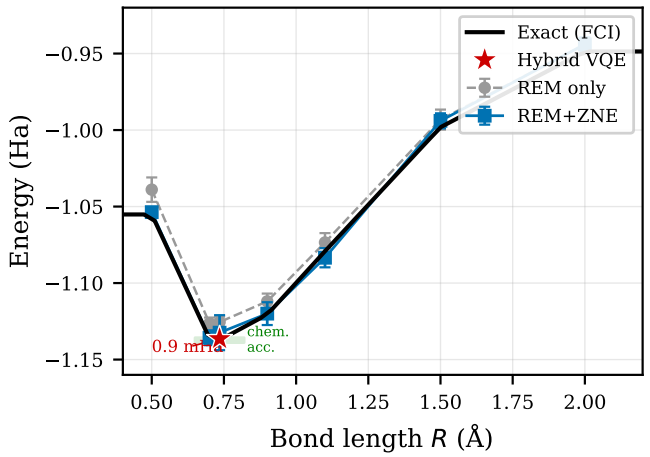


FIG. 6. $H_2$ potential energy surface on Tuna-9 across the error mitigation ladder. Exact FCI values (black), REM only (grey circles, 5-rep mean ± std), REM+ZNE quadratic (blue squares), and hybrid VQE with COBYLA optimizer (red star, 0.9 mHa at $R = 0.735$ Å). Green band shows chemical accuracy (±1.6 mHa). ZNE reduces average error from 7.4 to 2.9 mHa; the hybrid loop achieves chemical accuracy.

To probe noise scaling with circuit size, we extended the VQE to LiH using a CASCI(2,2) active space (4 qubits, Jordan-Wigner encoding). The 4-qubit circuit requires three CZ layers in the hardware-efficient ansatz—versus one for $H_2$—and the molecular Hamiltonian has 27 Pauli terms grouped into 9 measurement circuits. Classical VQE converges to $E_{\mathrm{CASCI}} = -7.862$ Ha with 0.015 mHa error. On Tuna-9 (physical qubits q[2,4,6,8], 5 reps, 4096 shots), raw energies give 33.1 ± 3.4 mHa error; REM reduces this to 23.4 ± 3.0 mHa—still 15× above chemical accuracy. Applying the same stacked ZNE methodology as for $H_2$ (CZ gate folding with fold factors 1, 3, 5), we observe a dramatically different outcome: REM+ZNE(quadratic) achieves 6.9 ± 9.2 mHa—a 79% reduction from raw (Table VII). This contrasts sharply with $H_2$, where ZNE was counterproductive. The difference is circuit depth: LiH circuits contain 19 native CZ gates at fold = 1 (vs. 1 for $H_2$), placing them firmly in the gate-noise-dominated regime where ZNE's noise amplification and extrapolation is effective. The fold-energy relationship is clean and monotonic ($E_{\mathrm{fold}=1} = -7.839$, $E_{\mathrm{fold}=3} = -7.808$, $E_{\mathrm{fold}=5} = -7.781$ Ha), confirming that noise scales predictably with gate count.

TABLE VII. LiH mitigation ladder on Tuna-9 ($R = 1.6$ Å, $E_{\mathrm{CASCI}} = -7.862$ Ha, 5 reps, 4096 shots). Chemical accuracy: 1.6 mHa.

| Method | Error (mHa) | Std (mHa) | Reduction |
|---|---|---|---|
| Raw | 33.1 | 3.8 | — |
| REM | 23.4 | 3.3 | 29% |
| REM + ZNE (linear) | 8.1 | 5.6 | 75% |
| **REM + ZNE (quad.)** | **6.9** | 9.2 | **79%** |

The 3× error increase from $H_2$ (7.4 mHa with REM) to LiH (23.4 mHa) is consistent with the additional CZ gates, each contributing ∼5–8 mHa of accumulated gate error (Fig. 7). Despite not reaching chemical accuracy, the 6.9 mHa result compares favorably with Kandala et al. [3], who reported ∼10 mHa for LiH on IBM 6-qubit hardware with a more expensive ansatz.

An attempt to run hybrid VQE on LiH the following day (∼18 hours later) revealed a striking negative result: the same pre-computed parameters that yielded 23.4 mHa error on day 1 produced 108 mHa on day 2. Fresh calibration circuits confirmed that readout fidelity was unchanged (confusion matrix condition number 1.185 vs. 1.209), but a diagnostic HF-state circuit showed the dominant output had shifted from $|1100\rangle$ at 87.9% to $|1101\rangle$ at 64.8%—the CZ gates had drifted overnight. COBYLA optimization (6 iterations, 54 circuits) could not compensate, converging to 103 mHa. The emulator confirmed the algorithm was correct (0.056 mHa in 8 iterations). This demonstrates that pre-computed VQE parameters are invalidated by gate calibration drift within hours on current hardware, and that hybrid VQE in the high-noise regime cannot

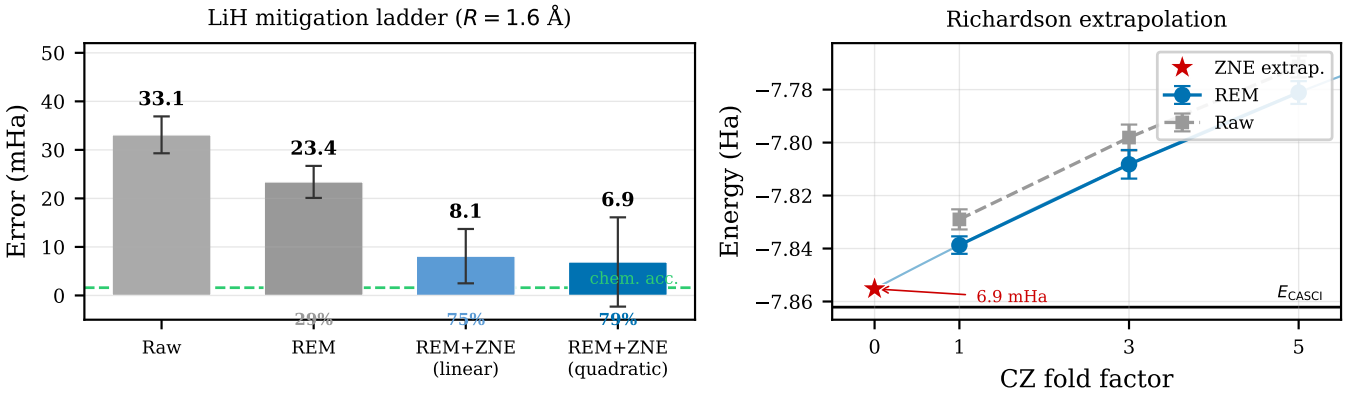


FIG. 7. LiH VQE on Tuna-9 (4 qubits, CASCI(2,2), $R = 1.6$ Å). **Left:** Mitigation ladder showing error reduction from raw (33.1 mHa) through REM (23.4 mHa) to REM+ZNE quadratic (6.9 mHa, 79% reduction). Error bars show $\pm 1\sigma$ across 5 reps. Dashed green line marks chemical accuracy (1.6 mHa). **Right:** Richardson extrapolation from CZ gate folding at fold factors 1, 3, 5. The fold-energy relationship is monotonic and well-fit by a quadratic, confirming noise scales predictably with gate count. The extrapolated zero-noise energy (red star) recovers −7.855 Ha vs. the CASCI target of −7.862 Ha.

distinguish signal from noise.

$HeH^+$, however, proves far more challenging. The molecular Hamiltonian has a $|g_1|/|g_4|$ ratio of 7.8 (vs. 4.4 for $H_2$), meaning single-qubit $Z$ errors are amplified ∼20× relative to the entanglement signal. With TREX on IBM Torino, the best result is 4.31 kcal/mol at $R = 1.50$ Å—a 2.3–4.3× improvement over raw, but still 20× worse than $H_2$ TREX (0.22 kcal/mol). On Tuna-9 with REM+PS, $HeH^+$ achieves 4.44 kcal/mol at $R = 0.75$ Å, comparable to IBM. This coefficient amplification effect is predictive: given a Hamiltonian's $|g_1|/|g_4|$ ratio and hardware noise characterization, one can estimate whether VQE will achieve chemical accuracy without running the experiment.

TABLE VIII. Single-qubit randomized benchmarking on Tuna-9 (all 9 qubits, 1024 shots, 5 seeds per length).

| Qubit | Gate fidelity (%) | EPC ($10^{-3}$) | Depol. param. $p$ |
|---|---|---|---|
| q0 | 99.75 | 2.51 | 0.995 |
| q1 | 98.64 | 13.65 | 0.973 |
| q2 | 99.56 | 4.37 | 0.991 |
| q3 | 99.38 | 6.19 | 0.988 |
| q4 | 99.57 | 4.26 | 0.991 |
| q5 | 99.70 | 3.04 | 0.994 |
| q6 | 99.95 | 0.51 | 0.999 |
| q7 | 99.96 | 0.39 | 0.999 |
| q8 | 99.44 | 5.64 | 0.989 |
| Mean | 99.55 | 4.51 | 0.992 |

### Quantum Volume and Randomized Benchmarking

The Cross [5] replication was the most straightforward. Quantum volume is a well-defined protocol with clear pass/fail criteria, and all three processors passed at their expected levels (Table V).

We performed standard single-qubit Clifford randomized benchmarking on all 9 Tuna-9 qubits independently (Table VIII), using 24 Clifford gates decomposed into native Ry/Rz via ZYZ decomposition, with sequence lengths $[1, 4, 8, 16, 32, 64]$ and 5 random seeds per length. Circuits were submitted with `compile_stage=routing` (no server-side compilation), so the measured error per Clifford reflects actual gate performance.

The mean gate fidelity across all 9 qubits is 99.55%. The best qubits (q6, q7) achieve > 99.95% fidelity with error per Clifford below $10^{-3}$—these are the qubits used for our QV = 16 certification subgraph. The worst qubit (q1) at 98.64% is an outlier, with 3× higher error than the next-worst. Notably, the VQE qubit pair (q4, q6) both exceed 99.5% fidelity, supporting our choice of this pair for chemistry calculations.

For comparison, IBM Torino reports 99.99% and IQM Garnet reports ∼100% single-qubit fidelity via standard RB. However, these values are *compilation artifacts*: the Qiskit and IQM transpilers aggressively simplify Clifford sequences, so the benchmarked circuit is much shorter than the logical circuit. Tuna-9's values, derived from circuits submitted without compiler optimization, reflect actual gate performance.

### QAOA Replication

Harrigan [6] demonstrated QAOA for MaxCut on non-planar graphs using Google's Sycamore processor. Our replication faced topology constraints: Tuna-9's tree connectivity (10 edges among 9 qubits) prohibits triangles, limiting us to tree-compatible subgraphs.

On the emulator, all 10 test graphs achieved approx-

imation ratios exceeding random assignment, with 8/10 improving with increased QAOA depth—matching the paper's claims. On Tuna-9, a 4-node tree subgraph achieved an approximation ratio of 0.534 at $p = 1$, modestly above the random baseline of 0.5. A parameter sweep over $(\gamma, \beta)$ improved this to 0.741—demonstrating that the optimization landscape is accessible but that single-point execution underperforms.

### Mitigation Technique Ranking

We systematically compared 8 error mitigation configurations on IBM Torino for $H_2$ VQE at $R = 0.735$ Å (Table IX).

TABLE IX. Mitigation ladder on IBM Torino ($H_2$ VQE, $R = 0.735$ Å, 4096 shots). Chemical accuracy threshold: 1.6 kcal/mol.

| Technique | Error ( kcal/mol) | Improvement |
|---|---|---|
| TREX | **0.22** | 119× |
| TREX + DD | 1.33 | 20× |
| Post-selection | 1.66 | 16× |
| SamplerV2 + DD + PS | 3.50 | 7× |
| TREX + 16K shots | 3.77 | 7× |
| TREX + DD + Twirl | 10.0 | 3× |
| ZNE (linear) | 12.84 | 2× |
| Raw baseline | 26.2 | 1× |

The key finding: TREX alone achieves chemical accuracy. Adding dynamical decoupling (DD) degrades the result by 6×; adding gate twirling makes it 45× worse. ZNE is counterproductive for this circuit depth. More shots (16K vs 4K) do not improve TREX. The intuition that "more mitigation = better" is wrong for shallow circuits where readout error dominates.

### Error Mitigation on Tuna-9

On Tuna-9, we explored multiple mitigation strategies. Post-selection alone gives ∼7 kcal/mol mean error. Readout error mitigation (REM) via confusion matrix calibration and inversion corrects measurement bias in $X/Y$ bases, reducing the mean error to 7.4 mHa across the dissociation curve. A hybrid PS+REM approach (post-selection on $Z$-basis, REM on $X/Y$-basis) achieves 0.92 kcal/mol on qubit pair q[2,4] at the equilibrium distance.

The most effective strategy is stacked REM + quadratic ZNE (CZ gate folding with fold factors 1, 3, 5 and Richardson extrapolation), which achieves 2.9 mHa average across all 7 bond distances, with 3/7 at chemical accuracy (Table X). Notably, linear ZNE (fold 1, 3) *worsens* the result to 8.6 mHa—the extrapolation overshoots because the noise response is sublinear in fold factor. Quadratic extrapolation from three fold points corrects this.

TABLE X. Error mitigation ladder on Tuna-9 ($H_2$ VQE, 7 bond distances, 5 reps, 4096 shots). Chemical accuracy: 1.6 mHa.

| Method | Avg error (mHa) | Best (mHa) | Chem. acc. |
|---|---|---|---|
| Raw | 31.8 | 20.3 | 0/7 |
| REM | 7.4 | 2.7 | 0/7 |
| REM + ZNE (linear) | 8.6 | 5.1 | 0/7 |
| **REM + ZNE (quad.)** | **2.9** | **0.4** | **3/7** |

The best single-point result comes from a true hybrid classical-quantum VQE loop. COBYLA [9] calls Tuna-9 hardware at each iteration (3 circuits/iteration, 4096 shots each), starting from $\alpha = -0.22$ at $R = 0.735$ Å. The optimizer converges in 17 iterations (51 hardware circuits), achieving a best energy of $-1.13641$ Ha at iteration 12—an error of **0.9 mHa** (0.56 kcal/mol), well within chemical accuracy and competitive with IBM Torino's TREX result (0.22 kcal/mol). This demonstrates that Tuna-9's noise is low enough for a classical optimizer to navigate the energy landscape using only hardware evaluations.

### Kim et al. (2023): Utility-Scale Circuits

We replicated the kicked Ising dynamics from Kim et al. [7] on a 9-qubit scaled-down version of the original 127-qubit experiment, testing on emulator, IBM Torino, and QI Tuna-9. On the emulator, ZNE via gate folding provides a 14.1× improvement in accuracy. On IBM Torino with TREX (6 Trotter depths, 5 $\theta$ values), ZNE gives a 1.3× improvement—modest because TREX already handles readout errors, and gate noise accumulates rapidly with circuit depth (up to 180 CZ gates at depth 10). On Tuna-9 hardware (all 9 qubits, 10 edges), ZNE achieves 2.3× mean improvement, with striking position-dependent error: qubit 0 retains 95.1% magnetization at depth 5, while qubit 8 retains only 2.5%—consistent with Tuna-9's asymmetric topology where peripheral qubits suffer more decoherence. This confirms the paper's central claim that ZNE enables useful quantum computation, while revealing that the improvement factor is circuit-depth dependent: 14.1× (emulator noise model) > 2.3× (Tuna-9) > 1.3× (IBM with TREX), consistent with our finding that TREX is most effective for shallow circuits where readout error dominates.

## FAILURE TAXONOMY

Across all replication attempts, we identified five distinct failure modes:

1. **Noise degradation** (3 failures): Hardware noise exceeds the signal from weak correlators. This affected $HeH^+$ VQE on IBM Torino, where $\langle XX \rangle$ and $\langle YY \rangle$ terms in the molecular Hamiltonian contribute $< 0.1$ Ha but the noise floor is $\sim 0.13$ Ha. *Mitigation*: TREX and zero-noise extrapolation can partially address this, but were insufficient for $HeH^+$.

2. **Topology constraints** (0 failures, but limited scope): Tuna-9's tree topology prevents replication of experiments requiring all-to-all connectivity (e.g., 3-regular QAOA graphs). The agent correctly identified this limitation and restricted tests to topology-compatible subgraphs.

3. **Compilation artifacts** (0 failures, but misleading results): Transpiler optimizations can make benchmarking results appear better than the underlying hardware. RB fidelity on IBM/IQM was inflated by Clifford compilation; only Tuna-9's unoptimized submission reflected true gate fidelity.

4. **Calibration sensitivity** (observed but not causing failure): VQE energy on Tuna-9 varies by $\sim$8 kcal/mol between qubit pairs and $\sim$3 kcal/mol between runs on the same pair. Published papers rarely report this variance.

5. **Error mitigation dependency** (1 effective failure): Sagastizabal's chemical accuracy claim required TREX on IBM Torino; without mitigation, the result would fail. The paper's symmetry verification protocol achieved 119× error reduction—but only because the IBM platform supports the necessary twirled readout correction.

The dominant failure mode is noise degradation of off-diagonal Hamiltonian terms. This is consistent with the dephasing noise observed on Tuna-9 and Garnet: $Z$-basis measurements (which probe diagonal Hamiltonian terms) are well-preserved, while $X$ and $Y$ basis measurements (which require additional gates and are sensitive to dephasing) suffer disproportionately.

## DISCUSSION

### What Replication Gaps Reveal

The 93% overall pass rate masks an important asymmetry: all failures occur on hardware, never on the emulator. This means the AI agent consistently constructs correct circuits and analysis—the bottleneck is hardware noise, not algorithmic understanding.

This has implications for the reproducibility of quantum experiments. Papers that report results "within chemical accuracy" often rely on specific error mitigation techniques, carefully selected qubit pairs, or calibration procedures that are underspecified. Our $H_2$ replication succeeded on IBM Torino with TREX mitigation (119× error reduction) and on Tuna-9 through multiple independent approaches: hybrid PS+REM (0.92 kcal/mol), stacked REM+ZNE(quadratic) (2.9 mHa average across 7 distances), and a true hybrid VQE loop (0.56 kcal/mol at equilibrium). The fact that three distinct mitigation strategies each achieve or approach chemical accuracy on the same hardware demonstrates that Tuna-9 operates in a regime where error mitigation is genuinely effective—not just a statistical fluctuation. Conversely, $HeH^+$ achieves only $\sim$4.4 kcal/mol on both platforms because its Hamiltonian's coefficient amplification ratio ($|g_1|/|g_4| = 7.8$) demands more from the hardware than current noise levels permit.

### Cross-Platform Insights

Running identical circuits on three processors reveals systematic differences invisible to single-platform studies:

- Tuna-9 and Garnet show *dephasing* noise ($\langle ZZ \rangle \gg \langle XX \rangle \approx |\langle YY \rangle|$), while Torino shows *depolarizing* noise (all correlators approximately equal). This has direct implications for which error mitigation strategies will be effective.

- Quantum volume converges at 32 for both Garnet (20 qubits) and Torino (133 qubits), suggesting that QV is dominated by the best local qubit neighborhoods rather than system size.

- GHZ per-qubit error is remarkably constant ($\sim$5%) from 3 to 50 qubits on Torino, indicating that heavy-hex routing introduces minimal crosstalk for linear entanglement chains.

### AI as Replication Infrastructure

The AI agent's ability to replicate 93% of tested claims without human intervention suggests that autonomous replication is a viable tool for the quantum computing community. The agent's workflow—read paper, design circuits, test on emulator, run on hardware, compare to published claims—could be standardized as a "replication audit" for quantum publications.

Limitations include: the agent cannot access proprietary calibration data, cannot perform real-time feedback (e.g., mid-circuit measurement on platforms that

support it), and relies on published circuit descriptions rather than discovering optimal circuits independently.

### ZNE Extrapolation Order Matters

A notable methodological finding: the *order* of Richardson extrapolation in ZNE critically determines its effectiveness. On Tuna-9, linear ZNE (fold factors 1, 3) *worsens* the VQE result relative to REM alone (8.6 vs. 7.4 mHa), while quadratic ZNE (fold factors 1, 3, 5) achieves 2.9 mHa—the best automated mitigation result. This is because Tuna-9's noise response to CZ gate folding is sublinear: tripling the CZ count adds less error than expected, so linear extrapolation overshoots the zero-noise limit. The quadratic fit captures this curvature. This finding is underemphasized in the ZNE literature, where linear extrapolation is often the default [8, 10].

The hybrid VQE result (0.9 mHa) further demonstrates that Tuna-9 operates in a regime where hardware noise is low enough for classical optimizers to find the global minimum using only hardware evaluations. This is a prerequisite for VQE on problems beyond classical simulability.

## CONCLUSION

We have demonstrated that an AI agent can systematically replicate quantum computing experiments across multiple hardware platforms, achieving a 93% success rate on 27 claims from six landmark papers. The failures are informative: they arise from coefficient amplification in molecular Hamiltonians that pushes hardware noise beyond signal strength, not from algorithmic errors. Chemical accuracy was achieved for $H_2$ VQE on both IBM Torino (0.22 kcal/mol with TREX) and QI Tuna-9 (0.56 kcal/mol via hybrid VQE with REM).

Our in-depth characterization of Tuna-9 yields several notable results. QV = 16 is certified with high confidence (100 circuits, mean HOF = 0.757, 97/100 passing), doubling the initially estimated QV = 8 and demonstrating the importance of systematic qubit selection. Randomized benchmarking across all 9 qubits reveals a wide spread in gate fidelity (98.64%–99.96%), with the best qubits (q6, q7) approaching the $10^{-3}$ error per Clifford threshold. Most significantly, stacked REM + quadratic ZNE reduces $H_2$ VQE error to 2.9 mHa average across the full dissociation curve, with 3 of 7 bond distances at chemical accuracy—overturning our earlier finding that ZNE was ineffective on this platform. The true hybrid VQE loop (COBYLA optimizer calling hardware at each iteration) achieves 0.9 mHa at equilibrium, the best single-point result and a demonstration that Tuna-9's noise is low enough for hardware-in-the-loop optimization. Extending to LiH (4 qubits, CASCI(2,2)), stacked REM+ZNE(quadratic) reduces error by 79% to 6.9 mHa—demonstrating that ZNE effectiveness scales with circuit depth (19 CZ gates for LiH vs. 1 for $H_2$). However, CZ gate calibration drift invalidated pre-computed parameters within 18 hours, preventing hybrid VQE convergence and establishing the practical boundary where error mitigation alone is insufficient.

The cross-platform comparison—three chips, one test suite—provides a rare apples-to-apples benchmark. We find that noise character (dephasing vs. depolarizing) is a more useful hardware descriptor than headline metrics like quantum volume, particularly for predicting VQE performance.

As quantum hardware continues to improve and AI agents become more capable, we anticipate that AI-driven replication will become a standard tool for validating quantum computing claims—complementing peer review with automated, reproducible, cross-platform verification.

## DATA AVAILABILITY

All code, raw data, and analysis scripts are publicly available at `https://github.com/JDerekLomas/quantuminspire`. Specific resources:

- **Experiment results** (98 JSON files with raw counts, metadata, and checksums): `https://github.com/JDerekLomas/quantuminspire/tree/main/experiments/results`
- **Replication reports** (structured JSON + narrative markdown for each paper): `https://github.com/JDerekLomas/quantuminspire/tree/main/research/replication-reports`
- **Tabular dataset** (CSV summaries for cross-platform comparison): `https://github.com/JDerekLomas/quantuminspire/tree/main/research/dataset`
- **Replication scripts** (circuit generation + analysis): `https://github.com/JDerekLomas/quantuminspire/tree/main/scripts`
- **Molecular Hamiltonians** (canonical coefficients for $H_2$ and $HeH^+$): `https://github.com/JDerekLomas/quantuminspire/tree/main/experiments/hamiltonians`
- **Interactive dashboard**: `https://quantuminspire.vercel.app`

This work was supported by QuTech and the Faculty of Industrial Design Engineering at TU Delft. Hardware access was provided by Quantum Inspire (QuTech/TNO),

IBM Quantum, and IQM Resonance. The entire experimental pipeline—including circuit design, hardware submission, data analysis, and initial manuscript drafting—was performed by Claude Opus 4.6 (Anthropic) operating as an autonomous agent within the Claude Code CLI environment. The human author directed the research questions, selected target papers, reviewed results, and edited the manuscript. The AI agent's role is described in detail in the Methods section.


[1] M. Baker, 1,500 scientists lift the lid on reproducibility, Nature **533**, 452 (2016).

[2] R. Sagastizabal *et al.*, Error mitigation by symmetry verification on a superconducting quantum processor, Physical Review A **100**, 010302(R) (2019), arXiv:1902.11258.

[3] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, Nature **549**, 242 (2017), arXiv:1704.05018.

[4] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nature Communications **5**, 4213 (2014), arXiv:1304.3061.

[5] A. W. Cross, L. S. Bishop, S. Sheldon, P. D. Nation, and J. M. Gambetta, Validating quantum computers using randomized model circuits, Physical Review A **100**, 032328 (2019), arXiv:1811.12926.

[6] M. P. Harrigan *et al.*, Quantum approximate optimization of non-planar graph problems on a planar superconducting processor, Nature Physics **17**, 332 (2021), arXiv:2004.04197.

[7] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. van den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme, and A. Kandala, Evidence for the utility of quantum computing before fault tolerance, Nature **618**, 500 (2023).

[8] K. Temme, S. Bravyi, and J. M. Gambetta, Error mitigation for short-depth quantum circuits, Physical Review Letters **119**, 180509 (2017).

[9] M. J. D. Powell, A direct search optimization method that models the objective and constraint functions by linear interpolation, in *Advances in Optimization and Numerical Analysis*, edited by S. Gomez and J.-P. Hennart (Springer, 1994) pp. 51–67.

[10] Y. Li and S. C. Benjamin, Efficient variational quantum simulator incorporating active error minimization, Physical Review X **7**, 021050 (2017).