Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Approximate Sampling for Randomization Tests: A Total-Variation Certification Framework

Mohamed Faisal Sindhi

sindhi@usf.edu

University of South Florida https://orcid.org/0009-0000-1348-0566

Muhammad Hammaz

University of South Florida

Research Article

Keywords: approximate sampling, randomization tests, total variation distance, covariate-constrained randomization, cluster randomization, quantum optimization

Posted Date: October 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11322841/v2

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

Additional Declarations: The authors declare no competing interests.

Certifying approximate samplers for constrained cluster randomization

Mohamed Faisal Sindhi, Muhammad Hammaz
Muma College of Business, School of Information Systems, University of South Florida, Tampa, Florida, USA

Short title: Certifying approximate samplers for randomization tests

Correspondence: Muhammad Hammaz, Muma College of Business, University of South Florida, Tampa, FL, USA.

Abstract

Background: Covariate-constrained randomization draws an assignment uniformly from an accepted set of balanced allocations, and randomization-based inference uses that same set as its reference distribution. When an approximate sampler replaces the exact law, the test changes with it. **Methods:** On a 16-county benchmark from the Dickinson cluster-randomization example, we enumerated the exact target (1,287 of 12,870 allocations) and certified sampler outputs by total variation (TV) distance, which bounds the p-value error of any fixed test event. We compared classical and quantum approximate optimization (QAOA) samplers and ran a randomization-inference simulation (50,000 replicates per cell) with an unadjusted difference in cluster means (T1) and a covariate-adjusted estimator (T2). **Results:** The exact target rejected a true null at 4.98% (T1) and 5.05% (T2). A biased classical sampler inflated T1 type-I error to 15.87%. An XY-QAOA output at TV 0.834 gave 0.20% under T1 but 4.53% under T2, and disagreed with the exact p-value on 36.7% (T1) versus 2.9% (T2) of replicates, so distance alone did not predict inferential damage. Minimizing TV directly could not take the XY ansatz below 0.828, and a Grover-mixer variant reached 0.686. A penalty QAOA output placed 52% of its mass on invalid allocations. **Conclusions:** Auditing a sampler's output law by exact TV, together with the behavior of the planned test statistics, is a practical check for small cluster trials. The

study shows no quantum advantage or hardware benefit; on eight qubits, device noise erased the circuit structure within three layers.

Keywords: covariate-constrained randomization; cluster randomized trials; randomization tests; total variation; sampler certification; QAOA

1 Introduction

Small cluster randomized trials often restrict randomization to allocations that meet a covariate-balance criterion, then draw one allocation uniformly from the accepted set[1,2]. Randomization-based inference uses that same set as its reference distribution. The sampler that searches the set is therefore part of the inferential procedure: if it draws from a different law, the test changes with it.

The statistics literature supplies the design and its theory. Morgan and Rubin[3] introduced rerandomization, Li, Ding and Rubin[4] developed its asymptotics, and Wang and Li[5] treat covariate adjustment under constrained randomization. Software such as fastrerandomize[6] speeds up large accepted sets. Lu, Zhang and Zhu[7] propose a Metropolis–Hastings sampler with validity results, including a total-variation (TV) bound for their generalized sampler. These works analyze particular samplers.

A second literature restores exact validity by construction when a null can only be sampled approximately[8–11]. Those constructions rely on exchangeability or on a known sampling kernel. A one-shot sampler, such as a QAOA readout or a descent heuristic, has neither property, so its output law has to be checked directly.

This paper takes the complementary view of an external, sampler-agnostic audit. Given an exactly known target P* and any sampler whose output law Q can be computed or estimated, how far is Q from P*, and what does that distance imply for p-values, type-I error and power?

We answer this on one fixed benchmark, the 16-county Dickinson example, where the accepted set can be enumerated exactly. The audited samplers are classical controls plus ideal-statevector quantum circuits (a Grover-style amplification control and two QAOA arms), chosen because their output laws can be computed exactly.

The benchmark is small enough to enumerate, and pragmatic-trial guidance already recommends enumerating all allocations when fewer than about 20 clusters are randomized[12]. We therefore present it as a controlled stress test, not as a case where sampling is needed. Its value is that a sampler can be validated against known truth before anyone trusts it where enumeration is impossible.

The paper makes three contributions.

1. **A certification workflow.** An exact accepted-set target (1,287 of 12,870 balanced allocations, exact rational scoring, canonical tie-break) and TV certificates with thresholds fixed before execution: close at $TV \leq 0.02$, far at $TV \geq 0.10$. Statevector samplers are certified exactly; sampled outputs use a distribution-free finite-shot bound.
2. **An event-level bridge.** For any fixed outcome vector and fully defined test event, the p-value under a sampler differs from the exact p-value by at most TV. This is a standard inequality, used here as the link from distance to inference, not as a new theorem. For samplers that emit invalid assignments, we report an envelope between two fully defined extension rules.
3. **Downstream measurements.** Type-I error and power for two statistics on simulated outcomes anchored on the observed covariates, plus a frozen synthetic extension as a robustness check. The two statistics respond differently to the same distortion, so the worst-case TV bound does not predict the error of any particular test.

The paper does not claim quantum advantage, hardware performance, noise robustness, or clinical benefit.

2 Related work

This paper sits between three bodies of work: constrained randomization and rerandomization, valid inference under approximate sampling, and quantum and fair sampling for balance problems. Our claim is positional: none found, with the searches we ran, combines an exactly enumerated accepted-set target, exact TV certificates for classical and quantum samplers, and a type-I error and power study. Forward-citation searches were not completed, so this is a statement about our searches, not a proof of absence.

2.1 Rerandomization and constrained randomization

Morgan and Rubin[3] define rerandomization by Mahalanobis-distance acceptance, and Li, Ding and Rubin[4] give its asymptotic theory. For cluster trials, cvcrand[2] enumerates small allocation spaces and otherwise simulates a large number of schemes, and Wang and Li[5] study covariate adjustment under covariate-constrained randomization. Fastrerandomize[6] uses accelerated computing for large accepted sets, and pair-switching rerandomization[13] and BRAIN[14] are further fast or approximate samplers. Lu, Zhang and Zhu[7] give a Metropolis–Hastings framework; they supply an exactly uniform envelope-rejection sampler and, in an appendix, a TV bound for a generalized sampler, while uniformity of their practical algorithm is supported empirically. We do not claim that existing samplers lack certification. Our contribution is an external audit that applies to any sampler and is tied to randomization-test consequences.

2.2 Valid inference when the null is sampled approximately

Monte Carlo and MCMC significance tests restore validity by construction. Besag and Clifford[8] give exact tests from Markov-chain samples, and Phipson and Smyth[9] show that random permutation p-values need the +1 correction. Howes[10] gives valid p-values when a null can only be sampled approximately, with no mixing-time assumption, through exchangeability. Ramdas,

Barber, Candès and Tibshirani[11] show that a randomized p-value is valid for an arbitrary non-uniform distribution over permutations, using the sampling kernel in the construction. Toulis[15] bounds the size drift of approximate randomization tests whose statistics are estimated. Mehta, Patel and Wei[16] is the classical precedent that analysis must match a restricted randomization rule.

These constructions require exchangeability or a known kernel. A one-shot sampler such as a QAOA readout or a steepest-descent heuristic does not provide either, so there is nothing to repair; its law has to be measured. That is the case our audit addresses. We did not locate a source that states the p-value bound in Section 4.4 as a named result, but the inequality follows directly from the definition of TV, and we make no novelty claim for it.

2.3 Quantum and fair sampling for balance problems

Quantum work on trial design has targeted optimization. Domingo and Johnson[17] apply annealing and QAOA to patient stratification and report efficiency and significance gains, Kowshik et al.[18] study QAOA for set balancing, and Lebl and Shakeel[19] use variational algorithms for covariate balancing. We found none that certify the output law of the sampler or connect it to randomization-test validity.

The fair-sampling literature shows why such certification is needed. Transverse-field and mixer-driven samplers are known to be biased among equal-cost states[20–22], and Grover-mixer QAOA gives equal amplitude to equal-cost states when the feasible uniform superposition can be prepared[23,24]. Our amplification control is a fixed-oracle instance of that uniform-over-marked-set construction, so its success is expected. The XY-mixer and penalty arms follow the constrained-mixer framework[25,26], and our negative QAOA results are consistent with the bias literature.

What is new here is the application framing: an exact accepted set, exact TV, and a bridge to p-value error in a trial-randomization setting.

3 Exact target and benchmark

The benchmark uses the pre-randomization covariates of the 16-county Colorado cluster-randomization example reported by Dickinson et al.[1], as distributed in cvcrand 0.1.1. The covariates are location, inciis, uptodateonimmunizations, hispanic and incomecat; categorical variables are coded with the alphabetically first level as reference. Each covariate column is standardized with its sample mean and standard deviation over the 16 counties. No trial outcomes are analyzed, and we do not use the simulated individual-level dataset that the package also distributes.

Let Ω8 be the set of C(16,8) = 12,870 allocations of 8 treated and 8 control counties. With standardized covariates Z, the l2 balance score of an allocation z with treated set T(z) is:

$$BL(z) = \sum_{j} \Big(\sum_{i \in T(z)} Z_{ij} \Big)^2 \tag{1}$$

For an 8-versus-8 allocation this equals 16 times the summed squared standardized treatment–control mean differences. That is an algebraic identity, not a quantity computed by cvcrand. Over Ω8 the score has minimum 1.161, mean 24 and maximum 116.66.

Exact accepted set. We compute BL(z) in exact rational arithmetic from the raw integer covariates and exact variances, and rank allocations in ascending order. Ties are broken by the canonical allocation index, the lexicographic order of the treated sets. The accepted set A holds the first m = round(12,870 × 0.1) = 1,287 allocations, and the target P* is uniform on A: P*(z) = 1/1,287 on A and 0 elsewhere. Exact arithmetic is deliberate, because floating-point summation can reorder nearly equal scores exactly where a tie sits.

The boundary tie. Exactly two allocations share the cutoff score, and they are complements: their treated sets partition the 16 counties. Because the standardized columns sum to zero, every allocation has the same score as its complement, so the tie is structural. The index rule decides membership at the boundary, and it is a stated definition, not a numerical accident. It breaks complement closure only at this pair, so county treatment probabilities are 643/1,287 or 644/1,287 instead of one half (largest deviation 1/2,574 ≈ 0.00039).

Sensitivity target. After the target was locked, we evaluated a complement-closed 1,288-state target with 644 complement pairs and exact one-half county marginals. This analysis is post-lock and is not the reported target. Under it the reported anchor laws move by at most 3.3×10^{-5} in raw TV (biased descent 0.958042 to 0.958075; original XY-QAOA headline 0.833403 to 0.833404).

Cross-check against cvcrand. We ran cvcrand 0.1.1 on the same frozen covariates with its documented settings (metric l2, cutoff 0.1, seed 12345). Allocation identifiers, design-matrix values and accepted-set membership agree, and balance scores agree within 2×10^{-14}. This agreement holds for this frozen instance only; we do not claim that cvcrand implements an explicit exact tie rule.

4 Certifying a sampler's output law

A sampler is described by its output distribution Q on the balanced domain. This section defines how Q is compared with the exact target P*, what a certificate means, and how the distance bounds p-value error.

4.1 Total variation and the lumped representation

We measure distance by total variation:

$$\mathrm{TV}(Q, P^*) = \tfrac{1}{2} \sum_{z \in \Omega_8} |Q(z) - P^*(z)| \tag{2}$$

Because P* has no mass outside A, the sum splits into the accepted states and everything else:

$$\mathrm{TV}(Q, P^*) = \tfrac{1}{2}\Big(\sum_{z \in A} |Q(z) - 1/m| + Q(\Omega_8 \setminus A)\Big) \tag{3}$$

Aggregating all outside-accepted states into one bin therefore preserves TV exactly. We report every distance on this 1,288-bin representation (1,287 accepted bins and one outside bin). When a sampler also emits strings of the wrong Hamming weight, we add their mass to the outside bin, and the same identity holds on the extended domain.

4.2 Certificate levels

Both thresholds were defined before any result was computed. A distribution is certified **close** when its exact TV is at most τ_close = 0.02, and **far** when it is at least τ_far = 0.10. Between the thresholds the result is inconclusive. Full-statevector probabilities carry no sampling error and are labelled exact, meaning direct evaluation from the ideal statevector to numerical precision. For sampled distributions, close means the upper endpoint of the TV confidence set is at most τ_close, and far means its lower endpoint is at least τ_far. A close verdict certifies that exact TV is at most τ_close at the stated confidence. The procedure never converts non-rejection into a claim of closeness. Identity testing against a known target[27,28] and closeness testing of two unknown distributions[29] are different tasks from TV interval estimation.

4.3 Finite-shot certification

For n IID draws over k bins and any radius ε, we use the distribution-free method-of-types inequality for the L1 deviation of an empirical distribution, as restated by Mardia et al.[30]:

$$\Pr(\|\hat{P}_n - P\|_1 \geq \varepsilon) \leq (2^k - 2)\, e^{-n\varepsilon^2/2} \tag{4}$$

The exponent uses the constant φ ≥ 2 of the cited lemma. With k = 1,288 bins and α = 0.05, this gives, under IID sampling, the conservative confidence set [max{0, $\hat{\mathrm{TV}}$ − h}, min{1, $\hat{\mathrm{TV}}$ + h}]

for TV, with h = ε/2. The minimum shot budgets are 179,154 for ε = 0.10 and 4,478,847 for ε = 0.02. Calibration used the upward-rounded tiers 179,200 and 4,480,000.

These budgets show the limits of finite-shot certification. Under the confirmatory budget the half-width is h = 0.01, so the simple worst-case sufficient condition for closeness ($\delta_{TV} + 2h \leq \tau_{close}$) covers only $\delta_{TV} = 0$. The 2^k factor is a union bound over all subsets and dominates the budget. Sharper plug-in, bias-corrected, bootstrap, asymptotic and test-inversion intervals are not validated here, and we give no coverage for them.

The bound proves coverage under IID sampling; the calibration shows how it behaves in practice and does not replace the proof. We applied the frozen verdict rules to synthetic distributions whose exact TV to P* is known in closed form: an exact target, leakage outside the accepted set, concentration on one accepted state, a structured within-set bias, a diffuse within-set bias, and a postselection distortion. Each family was evaluated on a grid of distances from 0 to 0.20 under both budgets, with 10,000 replicates per cell. Observed coverage was at or above nominal in every cell, and no verdict went the wrong way beyond Monte Carlo error.

- The screening budget (179,200 shots) never produced evidence-close, because its half-width exceeds τ_{close}. It produced evidence-far only at $\delta_{TV} = 0.20$.
- The confirmatory budget (4,480,000 shots) produced evidence-close at the exact target with probability 1.000 and at $\delta_{TV} = 0.005$ with probability 0.999. It never did so at $\delta_{TV} \geq 0.01$.

These figures describe the power of the frozen rules under these budgets only.

4.4 From TV to p-value error

Fix the observed outcomes Y_obs and the observed allocation Z_obs. Let T be a statistic and let $E = \{z : |T(z; Y_{obs})| \geq |T(Z_{obs}; Y_{obs})| - 10^{-10}\}$ be the two-sided extremeness event. Once the data are fixed, E is a fully defined subset of the domain and does not depend on which reference is used. For any distribution Q on the domain, the randomization p-value is Q(E).

Proposition 1. For any finite domain, any probability distributions Q and P, and any event E,

$$|Q(E) - P(E)| \leq \mathrm{TV}(Q, P) \tag{5}$$

Proof. Let D = Q − P, so ΣD = 0. Then $\Sigma D^{+} = \Sigma D^{-} = \tfrac{1}{2}\Sigma|D| = \mathrm{TV}(Q, P)$. Hence Q(E) − P(E) = Σ_E D ≤ Σ_E D^{+} ≤ $\Sigma D^{+} = \mathrm{TV}(Q, P)$. Exchanging Q and P gives the lower bound. ∎

This inequality is standard. We use it as an operational bridge from distributional distance to event-specific p-value error, not as a new theorem. Applied to a randomization test, it gives |p_Q − p_P*| ≤ TV(Q, P*) for every replicate and every reference. A close certificate therefore bounds the worst-case p-value error over all fully defined events at the certified level. A far certificate does not imply large error for every event: TV is a supremum over events, so a particular statistic can show a much smaller error than the bound.

Samplers that emit invalid assignments. For the penalty output, the extended domain adds the wrong-weight strings, and the p-value is undefined until an extension rule is chosen for them. We bound it with two fully defined events: E_lo counts wrong-weight strings as non-extreme and E_hi counts them as extreme. The target assigns them zero probability, so |p_lo − p_P*| ≤ TV_ext and |p_hi − p_P*| ≤ TV_ext, with TV_ext = 0.949 the extended-domain TV. No extension is scientifically privileged, so choosing one would make a substantive assumption about invalid assignments that the data cannot settle.

Arithmetic audit. We checked both inequalities against all saved replicates, for both statistics and all eight cells, and found no failures. The global maximum observed difference is 0.998446, attained by the step-0 swap reference under both statistics in cells F3c and F3d, against a reference TV of 0.999223. Smaller per-reference maxima are 0.388 for the XY-QAOA reference under T1 and 0.543 for the step-10 swap marginal under T2. No difference exceeds its reference TV. The audit confirms the arithmetic of the simulation and does not strengthen the bound.

5 Samplers audited

Seven sampler laws are audited. Classical laws are exact by construction or by propagation of an exact matrix, and quantum laws are full-statevector probabilities from ideal simulation in Qiskit[31]. Local search uses SciPy[32].

Sampler	Construction	Output law
Direct uniform	Uniform draws from A	Exactly P* (TV 0)
Uncapped rejection	Draw uniform balanced allocations, accept those in A (acceptance 1/10)	Exactly P*
Swap chain	Symmetric proposal over 64 treated–untreated swaps, rejected outside A; one-time marginals at steps 0, 10, 41, 170, 1,000	Float64 propagation of an exact integer matrix; correlated, not IID
Biased descent	Steepest descent on the balance score from uniform starts; 54 local minima	Exact basin law; TV 137/143 ≈ 0.958
Amplification control	Uniform Dicke start, fixed accepted-set phase oracle, $t^* = 2$	Full statevector
XY-QAOA	Dicke start, ring XY mixer, depth $p \leq 3$, balance-score phase separator	Full statevector
Penalty-QAOA		$+\rangle$ start, RX mixer, penalty weight $\lambda = 48$, depth $p \leq 3$

5.1 Classical controls and the swap chain

Direct uniform sampling from A has TV zero by construction and is the positive control. Uncapped rejection sampling draws uniform balanced allocations and accepts those in A, so each attempt succeeds with probability 1/10 and its output is exactly uniform on A. A capped variant reports failures as a separate outcome, so its raw TV equals the failure mass and its conditional law is exactly uniform.

The swap chain moves one treated county for one untreated county, chosen uniformly among the $8 \times 8 = 64$ pairs, and stays in place when the proposal leaves A. The accepted-state graph is connected (one component of 1,287 states) and symmetric, and self-loop counts range from 41 to 57, so the chain is aperiodic. The integer transition matrix is exactly symmetric with row sums 64, so its stationary law is exactly uniform on A. Finite-time laws are float64 propagations of this matrix, and we never treat chain output as IID. The spectral gap is 0.022003. From the start state,

TV to uniform is 0.999 at step 0, 0.726 at step 10, 0.248 at step 41 (first below 0.25), 0.00985 at step 170 (first below 0.01), and 9.3×10^{-11} at step 1,000. The last value is a float64 number, not an exact zero. The inference study uses only these one-time marginals.

The deliberately biased control was defined before any result was seen. It starts from a uniform random allocation and runs steepest descent on the balance score over the same 64 swaps, stopping at a local minimum. Its exact output law is the basin distribution over all 12,870 starts. It places no mass outside A, but its outputs concentrate on 54 local minima, and its exact TV to P* is 137/143 ≈ 0.958.

5.2 Amplification control

The amplification control starts from the uniform superposition over the 12,870 balanced states, the uniform Dicke state D(16, 8). A fixed phase oracle multiplies the amplitude of each of the 1,287 accepted states by −1, followed by the reflection about the initial state, as in amplitude amplification[33]. The Grover angle is $\theta = \arcsin \sqrt{(1{,}287/12{,}870)} = 0.321751$. The iteration count $t^* = 2$ was fixed before the sweep and predicts acceptance $\sin^2(5\theta) = 0.99856$.

At $t^* = 2$ the raw accepted mass is 0.998560 and the full-statevector TV to P* is 0.001440, certified close as exact. The conditional law on A has numerically zero TV. Amplification preserves equal relative amplitudes among accepted states but does not remove the residual raw leakage. Raw TV at other iteration counts (0.90 at t = 0, 0.32 at t = 1, 0.40 at t = 3, 0.94 at t = 4) is diagnostic and did not change t*.

This arm is a fixed-oracle correctness control. Its oracle is a diagonal operator built from the classically enumerated accepted set, not a reversible membership circuit, so we report no resource estimate. Published Dicke-state constructions give gate-count and depth bounds[34–36], but we make no claim about the circuits used here.

5.3 QAOA arms

Two prospective QAOA arms were fixed in a protocol amendment before optimization. They follow the QAOA framework of Farhi et al.[37] and the constrained-mixer framework of Hadfield et al.[25]. The balance-energy family is intentionally not optimized to reproduce the uniform accepted-set target. It stress-tests whether a plausible balance-oriented sampling procedure preserves the required randomization law. Both arms use depth $p \in \{1, 2, 3\}$. The phase separator is a quadratic pseudo-Boolean cost in the 16 binary variables, compiled into single-qubit RZ and two-qubit RZZ rotations and normalized affinely to [0, 1] with constants computed by exact enumeration before any optimization.

XY arm. It is Hamming-weight-preserving[26] and starts from D(16, 8). The mixer is a fixed ordered sequence over the 16-cycle: all eight edges of matching E0, then all eight edges of matching E1, each edge applying RXX(β) then RYY(β). The sequence is the definition of the mixer; it is not claimed to equal continuous-time evolution under the summed ring Hamiltonian. Weight preservation was verified numerically at predetermined points and at every objective evaluation. Normalization uses BL_min = 1.161 and BL_max = 116.66 on the weight-8 subspace.

Penalty arm. It starts from $|+\rangle^{\otimes 16}$ and uses RX mixers. Its cost extends the balance score to all 2^{16} bitstrings:

$$BL_{all}(x) + \lambda\,(|x| - 8)^2 \tag{6}$$

The scale B_scale = 24 is the mean balance score over the 12,870 weight-8 strings, and the headline weight is $\lambda = 2 \cdot$ B_scale = 48. Two diagnostic weights, $0.5 \cdot$ B_scale = 12 and $8 \cdot$ B_scale = 192, were run with the same procedure and never enter selection. The normalization uses the minimum and maximum of the extended cost over all 2^{16} bitstrings; the maximum is 3,072, attained at the empty string. For every configuration the minimum cost over wrong-weight

strings exceeds the best weight-8 cost, so the penalty makes weight 8 the cost minimum. It does not make the output weight 8, because amplitudes evolve under the mixer and the cost is not the output distribution.

Optimization. Each arm used Nelder–Mead under box bounds (γ in $[0, \pi]$, β in $[0, \pi/2]$), a 300-evaluation budget per start, and 10 starts per depth from fixed seeds. We selected the run with the lowest final normalized energy, with ties broken by lower depth and seed. Selection never used TV. This is bounded local search, not global optimization. Across 120 runs, 88 stopped by stagnation and 32 exhausted the budget. A default-NumPy `isclose` boundary diagnostic flagged 90 of 120 runs, and a strict absolute re-audit disagrees on one run; we report these counts descriptively and do not read them as optimality or failure. We make no claim about other optimizers, a target-aligned ansatz, or QAOA in general; optimizer strategies are studied elsewhere[38].

6 Simulation design for randomization inference

The inference study asks what happens to a randomization test when its reference distribution is replaced by a sampler's law. The design follows the ADEMP structure for simulation studies[39].

Element	This study
Aim	Measure how replacing P* with an approximate law Q changes randomization-test type-I error and power
Data-generating mechanism	Observed Dickinson covariates with simulated Gaussian cluster means; eight cells; 50,000 replicates per cell
Estimands	Sharp-null rejection rate, power, and decision disagreement with P* at $\alpha = 0.05$
Methods	The reference laws of Section 5, applied to two statistics (T1, T2)
Performance measures	Rejection rates with Wilson 95% intervals; disagreement with P*; p-value difference against TV

6.1 Replicates and the primary experiment

Within each replicate we draw the realized allocation Z_obs from the exact target P*. We generate outcomes once from the cluster-level model below. We then compute a two-sided

randomization p-value under P* and the same p-value under each approximate reference, on the same outcomes and the same Z_obs. Drawing Z_obs from the approximate reference would mix two sources of error, so the primary validity experiment does not do it. The synthetic extension reports that variant separately as a sensitivity display.

6.2 Outcome model

The baseline cluster mean is:

$$\bar{Y}_i(0) = 0.5\,\tilde{x}_i + a_i + \bar{e}_i, \quad a_i \sim N(0, \rho), \quad \bar{e}_i \sim N(0, (1-\rho)/50) \tag{7}$$

Here x̃_i is the standardized observed inciis value for county i, a_i is a cluster effect, and ē_i is the mean of 50 independent individual errors. The individual-level residual has variance 1 and intra-cluster correlation ρ. The cluster-mean residual has variance ρ + (1 − ρ)/50, which is 0.069 (SD 0.263) at ρ = 0.05. Under an alternative, Ȳ_i(1) = Ȳ_i(0) + τ_i. The shifts τ ∈ {0.25, 0.50} are raw cluster-mean values, so τ = 0.50 is a half-SD shift relative to the individual-level SD of 1.

The eight cells use ρ = 0.05 for the base cells: a sharp null, constant shifts τ = 0.25 and τ = 0.50, and a heterogeneous shift τ_i = 0.50(1 + 0.5 x̃_i). Additional sharp-null and constant-shift cells vary ρ over {0.00, 0.20}. Each cell has 50,000 replicates and its own seed rule.

The cluster-level model is a deliberate simplification. Each cluster mean is Gaussian, and the design effect enters only through the intra-cluster correlation. We do not model unequal cluster sizes, individual-level covariates, or non-Gaussian outcomes. Those features would change the variance of the statistics, but they are not needed to isolate the effect of the reference distribution.

6.3 Statistics and references

Two statistics are computed on cluster means and are not studentized. T1 is the unadjusted difference in means between treated and control clusters. T2 is the coefficient on treatment in an

ordinary least-squares regression of the cluster means on treatment and $\tilde{x}$. Both are two-sided, a tie counts as extreme, and the decision threshold is $\alpha = 0.05$. The p-value under each reference sums its weights over allocations at least as extreme as the observed one, with absolute tolerance 10^{-10} for exact ties.

The reference distributions are the exact target; the two exact controls (direct uniform and uncapped rejection), which must reproduce the target exactly; swap-chain one-time marginals at steps 0, 10, 41, 170 and 1,000; the biased descent control; the amplification raw output at $t^* = 2$; and the two raw QAOA headline outputs. Conditional distributions appear only as labelled diagnostics.

For the penalty output, wrong-weight mass prevents a single balanced test statistic. We bound its p-value with the two extensions of Section 4.4 and report guaranteed and possible rejection rates: the rates at which the envelope lies wholly on one side of α, or touches the rejection region. The true rejection rate under any extension lies between them. Neither endpoint is the type-I error or the power of the penalty output.

6.4 Monte Carlo uncertainty and checks

Monte Carlo uncertainty is reported with Wilson 95% intervals. At 50,000 replicates, a rate near 0.05 has half-width about 0.2 percentage points and a rate near 0.5 has half-width about 0.4 points. Each replicate draws one Z_obs and one outcome vector, and the statistic vectors over all 12,870 allocations are computed once and shared by every reference. T2 uses the residualized representation with respect to the intercept and $\tilde{x}$, which is algebraically equal to the OLS coefficient; it agrees with direct least squares on fixed test outcomes to within 1.11×10^{-15}.

The exact controls were checked for exact equality of p-values with the target on a fixed set of synthetic outcomes, one per cell, in exact rational arithmetic. The amplification and QAOA

distributions were regenerated from their immutable parameter vectors, and their full-statevector TV values and certified flags agreed with the saved records to within 3.4×10^{-15}. The swap-chain marginals were regenerated from the exact integer matrix and reproduced the first crossings at steps 41 and 170.

7 Results

7.1 Certification of the sampler distributions

The amplification control is certified close on raw leakage, and both QAOA headline outputs are far from the target on the raw distribution. Table 1 gives the raw and postselected values. Raw distributions are the primary sampler outputs over all 1,288 bins. Conditional values renormalize the accepted bins and are postselected diagnostics that do not establish validity of the raw sampler.

[Table 1 about here]

The XY headline has accepted mass 0.168, and its conditional TV of 0.497 is still far. The penalty headline has accepted mass 0.051 and wrong-weight mass 0.520, so its raw law is not a valid 8-versus-8 randomization distribution. Its conditional TV of 0.0023 lies below the close threshold, but that value describes the accepted part of the output only and does not show the raw penalty sampler is valid. Figure 1 shows how fast the swap chain closes the distance for comparison.

[Figure 1 about here]

7.2 Downstream inference

Under the sharp null in the base cell, the exact target rejects at 4.98% under T1 and 5.05% under T2. Table 2 gives type-I rates and Table 3 gives decision disagreement with P* at $\tau = 0.50$. Brackets are Wilson 95% intervals from 50,000 replicates per cell, and penalty-QAOA entries

are bounds, not point rates. Both tables are descriptive only, with no regression or causal reading. Figure 3 plots these rates with their intervals.

[Figure 3 about here]

[Table 2 about here]

[Table 3 about here]

The swap-chain marginal at step 170 shows no disagreement with the target at the displayed Monte Carlo resolution. The step-41 marginal shows small disagreement (1.42% under T1 and 0.75% under T2), and the step-10 marginal, which still carries substantial distance, shows larger disagreement (10.92% and 5.87%). The biased control inflates T1 type-I error to 15.87% (interval 15.56–16.20%) and T2 to 6.50%, so its distortion is visible under both statistics but differs in size. In the cell with $\rho = 0.00$ the T1 rate rises to 28.53%.

The XY-QAOA headline has raw TV 0.833. Under T1 it disagrees with the target on 36.67% of replicates at $\tau = 0.50$, while its sharp-null T1 type-I rate is 0.20%, so it is conservative and distorted at once. Under T2 the same law disagrees on 2.87% of replicates and has a sharp-null type-I rate of 4.53% (Wilson 4.35–4.72%), closer to the exact-target rate than T1 is, but below the nominal 5% level. The two statistics respond very differently to one distribution, even though a fixed TV of 0.833 bounds the error of every event. The amplification control, with raw TV 0.0014, shows a T1 disagreement of 0.08%.

For the penalty output, the possible sharp-null rejection rate is 0.82% under T1 and 9.33% under T2, and the definite rejection rate is zero in both cases. At $\tau = 0.50$, guaranteed disagreement is 16.71% under T1 and possible disagreement is 68.21%; under T2 the bounds are 0.04% and 96.94%. These intervals are wide, and they are the honest description of a sampler whose wrong-weight mass prevents a single p-value.

The swap marginal at step 10 has smaller TV than the XY headline, yet under T2 it shows larger disagreement (5.87% against 2.87%). Figure 2 plots this directly. We therefore make no regression, slope, causal, or general TV-to-inference reading of these tables. Neither the biased-control nor the XY pattern generalizes beyond the simulated families and the benchmark.

[Figure 2 about here]

7.3 Synthetic multi-instance extension

The extension tests whether the workflow behaves consistently across constrained-randomization targets that differ in size, covariate strength and law. It uses eight synthetic instances with $J \in \{12, 16\}$ clusters, covariate-strength settings $\kappa_{cov} \in \{0.2, 0.6\}$, and two seeds per setting. Covariates mirror the anchor's structure (three continuous covariates on the anchor's integer ranges, a balanced binary covariate and a three-level categorical variable with dummy coding) but do not reproduce its joint distribution. Targets are complement-closed, so each cluster has treatment probability one half, and every accepted set has at least 100 allocations.

All eight instances carry the classical laws of Section 5, plus Boltzmann laws calibrated to expected normalized costs 0.02, 0.05 and 0.10. On four instances we also ran the XY-QAOA arm, with ten deterministic starts at each depth $p \in \{1, 2, 3\}$ and 300 evaluations per start. Its ideal-statevector law is far from the target in every case.

[Table 4 about here]

These extension cells come from the original frozen run and were not re-run for this revision: inference used 10,000 replicates for the final primary cells and 5,000 for screening cells. The primary family is equal-size Gaussian cluster means ($n_i = 50$) at ICC 0.05 or 0.20, under the sharp null and a constant shift of 0.5. Binary equal-size, heterogeneous Gaussian and Gaussian unequal-size analyses are sensitivity-only, are inventoried in Supplementary Table S4, and do

not alter sampler selection or primary-cell designation. Power cells use the same replicate counts as their sharp-null counterparts, a rule fixed before outcomes were generated.

Under the sharp null, the fixed-event bound of Section 4.4 holds in all 176 primary cells examined, and it is the only bound applied to these results. Across laws with similar raw TV, sharp-null rejection rates fall on either side of the nominal level, so raw TV alone does not determine the rate. Table 5 gives target-design power under P* on the four XY instances; these are properties of the exact-target design and do not depend on any sampler. In every row T2 rejects more often than T1 at the same replicate count, the same pattern as in the anchor.

[Table 5 about here]

7.4 Independent replication and extension experiments

We rebuilt the full pipeline from the public Dickinson design data in a separate NumPy statevector implementation, with new random seeds, and re-ran every headline quantity. The exact target (1,287 of 12,870 allocations, tie set at indices 5,913 and 6,956, minimum score 1.1611, maximum 116.656) was reproduced exactly. The swap chain reached TV below 0.25 at step 41 and below 0.01 at step 170, matching the original. The biased descent control has 54 minima and TV 137/143. The fixed-oracle amplification control at $t^* = 2$ matched the original raw TV. The regenerated XY headline ($p = 3$) has raw TV 0.8339 and accepted mass 0.168, and the penalty headline ($\lambda = 48$, $p = 3$) has raw TV 0.949, accepted mass 0.051 and wrong-weight mass 0.520. These agree with the original values to about the second or third decimal, so the preprint's quantum-arm conclusions replicate. Of the 120 energy-minimization runs, 59 stagnated before the 300-evaluation budget, 61 exhausted it, and 93 finished with at least one angle at a bound.

Inference at 50,000 replicates. We raised the replicate count from 10,000 to 50,000 per cell (Tables 2 and 3). The earlier T2 sharp-null rate of 4.40% (4.38% in a 5,000-replicate pilot) was Monte Carlo noise, not real conservatism of the covariate-adjusted statistic: at 50,000 replicates the exact target rejects at 4.98% (T1) and 5.05% (T2), and we withdraw the earlier reading. The biased descent control inflates T1 type-I error to 15.87% (15.56–16.20). The regenerated XY headline gives 0.20% under T1 and 4.53% under T2. Swap-chain marginals at step 10 give 5.89% (T1) and 7.56% (T2) under the sharp null. Disagreement with the exact p-value at $\tau = 0.50$ is 36.67% (T1) and 2.87% (T2) for the XY headline, 14.70% and 8.63% for biased descent, 10.92% and 5.87% for swap step 10, and 1.42% and 0.75% for swap step 41. For the penalty envelope, guaranteed and possible disagreement are 16.71% and 68.21% under T1, and 0.04% and 96.94% under T2. The fixed-event bound $|p - p^*| \leq TV$ held for every law and every replicate checked.

Swap-chain start-state sensitivity. The swap chain's mixing depends on where it starts. Across start states, TV falls below 0.25 after 35 to 49 steps and below 0.01 after 110 to 194 steps, and the spectral gap is 0.0220. A single-start burn-in figure should therefore be reported as a range.

XY-QAOA ceiling. To test whether the XY failure is a training failure, we minimized raw TV directly (an oracle diagnostic that uses knowledge of the target) with 12 Nelder-Mead starts and 4,000 evaluations per depth. The best raw TV is 0.8846 at $p = 1$, 0.8360 at $p = 2$, 0.8294 at $p = 3$ and 0.8284 at $p = 4$, with accepted mass at most 0.172 and conditional TV near 0.49 throughout. The energy-selected headline (0.8339) is within 0.005 of this ceiling, so the XY negative result is structural: a Dicke-start XY mixer with the balance-score phase separator cannot weight the accepted allocations uniformly, however it is trained. Multistart Nelder-Mead gives an empirical

ceiling, not a proof of global optimality. Figure 4 plots these values with the energy-selected runs.

[Figure 4 about here]

Grover-mixer QAOA. Replacing the XY mixer with a Grover mixer on the weight-8 subspace, with the same balance-score phase separator, improves the result but stays far from the target. Energy-selected runs reach raw TV 0.8133, 0.7718 and 0.7124 at $p = 1$, 2 and 3, and TV-minimized ceilings reach 0.8078, 0.7471 and 0.6862, with accepted mass 0.314 and conditional TV 0.060 at $p = 3$. When the phase separator is replaced by the indicator of the accepted set (an oracle that already encodes the answer), the Grover mixer at $\gamma = \beta = \pi$ reproduces the amplification control. The amplification result is therefore an oracle result, not evidence that a balance-score QAOA samples the target.

[Table 6 about here]

8 Discussion

A sampler's distance from its target caps the worst-case error of randomization-based inference, but the realized error depends on the event being tested. The XY-QAOA law shows this directly: one distribution at raw TV 0.834 yields a T1 sharp-null rate of 0.20% and a T2 rate of 4.53%. A certificate should therefore be reported together with the downstream behavior of the statistics a trial would use. Disagreement under power and type-I error under the null are different quantities, and each should be reported.

Why audit instead of repairing validity. Valid-by-construction tests[8,10,11] are the right tool when a sampler has exchangeability structure or a known kernel. The samplers audited here have neither. A one-shot readout from a QAOA circuit or a descent heuristic gives a fixed law Q, and what matters is how far Q sits from the law the design promised. For that setting the audit gives a

number (TV) and an event-level consequence (the p-value bound), and it makes visible what a repair would otherwise hide.

Why audit at a size that can be enumerated. At 16 clusters the full space has 12,870 allocations, and trial guidance recommends enumeration below about 20 clusters, where the space has 184,756. We chose this size because the exact target is then known, which is what makes certification possible at all. The number of balanced allocations grows quickly: it is 155,117,520 at 30 clusters. An audit at enumerable sizes validates a sampler against truth before anyone relies on it where enumeration is impossible. We do not claim that the anchor trial needs a sampler, and we have not shown that conclusions at 16 clusters transfer to larger ones.

What the quantum arms show. The amplification control reaches TV 0.0014, which is expected: it is a fixed-oracle instance of uniform-over-marked-set amplification, and its oracle uses the classically enumerated accepted set. The QAOA arms stay far from the target, consistent with the fair-sampling literature on biased sampling among equal-cost states. At 10% acceptance, rejection sampling is already exact and needs about ten attempts per draw on average. These results are not a statement about quantum advantage, and we do not make one. Quantum-enhanced stratification has been proposed with claimed efficiency and significance gains[17]; that work addresses optimization of covariate stratification and does not certify randomization laws, and we do not evaluate its claims. A different, adaptive preparation model uses mid-circuit measurement, feedback and ancilla resources[36]; it is not a demonstrated replacement for the static, no-ancilla workflow used here.

Reporting standard. Three questions apply before an approximate sampler replaces an exact constrained target.

1. What is its exact distance from the target on the output space the randomization test uses?

2. Are invalid outputs, such as wrong-weight strings, excluded, penalized or counted, and under which rule?
3. Which test statistics will the trial use, given that the worst-case bound does not say how far a particular statistic moves?

Raw distributions should be certified on their own output space, with conditional or postselected quantities reported separately and labelled as such. A close certificate for a conditional distribution does not license the raw sampler, and a far certificate for a raw distribution does not establish the size of the error for any particular test. Both facts belong in the same report.

9 Limitations and scope

The findings describe these runs on this benchmark and do not establish general sampler behavior. The main limits are below.

Design and data

- The empirical input is one anchor with 16 clusters, plus eight small synthetic instances, four of which carry ideal-statevector XY-QAOA runs.
- Outcomes are simulated, and the cluster-level model and its variance components are our specification. Primary inference uses equal-size Gaussian outcomes; binary, heterogeneous and unequal-size analyses are sensitivity-only.
- We have not shown that results at 16 clusters transfer to larger cluster counts or to other covariate structures.

Quantum arms

- All main quantum results are ideal statevector simulations. At 16 qubits these are classically trivial, and they say nothing about noise robustness. The exploratory hardware runs in Appendices A and B are small and single-device.
- Each QAOA headline is one run per arm, chosen by energy under bounded local search (Nelder–Mead, 300 evaluations, 10 starts per depth). Most runs were flagged by the boundary diagnostic, which limits any reading of optimizer quality. The Grover-mixer variant of Section 7.4 is the only other ansatz tested; we did not test a target-aligned ansatz, stronger initializations, or other optimizers.

- The amplification oracle is a fixed diagonal operator built from the enumerated accepted set. We report no resource estimate, and no computational-effort-matched classical-versus-quantum comparison was performed.
- The penalty output is not a valid 8-versus-8 randomization distribution, so its inference is bounded rather than point-estimated.

Certification

- The finite-shot sets are conservative. Under IID sampling and the confirmatory budget, the worst-case sufficient condition for closeness covers only the exact target, although calibration observed evidence-close at $\delta_{TV} = 0.005$. Sharper intervals are not validated here.
- Far detection, closeness certification and event-specific safety bounds are different tasks. Only the lumped interval of Section 4.3 supports closeness; the hardware appendix gives far-type event lower bounds and no closeness claim.
- The p-value bound of Section 4.4 is a standard inequality. Its role is as a link in the audit, not as a result.

Records and registration

- The protocol and the synthetic extension were internally time-ordered and hash-recorded before execution. They were not externally registered, and no registry identifier exists.
- One synthetic replicate run was interrupted. An audit supports its saved stream outputs, but its source hash cannot be recovered.
- The hardware analysis used replicate-specific Clopper–Pearson bounds, a documented post-protocol exploratory deviation.
- Assignment from Q with reference P* is not a valid design for the constrained target, and internal coherence under the sharp null is not evidence that the design was restored.

Several claims are outside scope: quantum advantage or speedup, noise robustness, scalable oracle cost, clinical benefit, and general QAOA performance. Future work would vary the outcome models, cluster count and covariate structure, and would add noise models and a matched classical search comparison. Each of those is a separate study with its own pre-specification, hash recording and freezing before execution.

The replication in Section 7.4 used a NumPy implementation rather than Qiskit, different random seeds from the original run, and did not repeat the cvcrand cross-check. The original code and data bundle are not yet public, so the replication is independent but not yet verifiable by readers.

10 Declarations and reproducibility

Reproducibility. All numerical results are generated from frozen scripts, pinned environments and saved machine-readable artifacts. The protocol, its amendments and the exact seed rules are recorded in the project files and the prepared release bundle. Hashed manifests record SHA-256 hashes for the protocol, code, preconditions, aggregate tables, per-replicate results, theorem audit and test record; the manifest families are listed in Supplementary Table S5. A clean regeneration reproduced the 44 synthetic-extension streams that predate an interrupted run bit-for-bit. Source hashes were not recorded at that launch, so its code state cannot be proven by hash.

Independent replication. The main benchmark results in Sections 6, 7.1, 7.2 and 7.4 and the abstract come from a separate NumPy statevector re-implementation with new seeds and 50,000 inference replicates per cell. The synthetic extension (Section 7.3, Tables 4 and 5) and Appendix A come from the original frozen run and were not re-run. Differences in the T2 sharp-null rate between runs (4.40% at 10,000 replicates against 5.05% at 50,000) are attributed to Monte Carlo error.

Registration. The protocols and the synthetic extension were internally time-ordered and hash-recorded before execution. They were not externally registered, and no registry identifier exists. We use "pre-specified" only for items fixed in these records before the relevant execution, and we state each amendment or deviation in the text.

Hardware records. The hardware submission record exists in two forms. The original contains account-identifying metadata and is private. The public manifest is a redacted release that records the hash of the original.

Availability. All code and data needed to regenerate Sections 6, 7.1, 7.2 and 7.4 and Appendix B are provided in a release bundle (34 files, SHA-256 manifest included, MIT license for code) that includes the Dickinson county-level covariates, the saved QAOA runs, the transpiled hardware circuits and the raw hardware counts. A clean re-run of the 50,000-replicate inference from the bundle reproduced the saved results exactly. The bundle is public at https://github.com/Hammazb/cluster-randomization-sampler-audit; a permanent archive DOI [to be inserted after Zenodo release] will accompany the accepted version. The synthetic extension of Section 7.3 and Appendix A come from the original frozen run and are described in the preprint's records.

Data accessibility statement. The data that support the findings of this study are openly available in the GitHub repository https://github.com/Hammazb/cluster-randomization-sampler-audit (a Zenodo DOI will be added on release). The covariates analyzed are from Dickinson LM, et al.[1], as distributed in the cvcrand R package.

Data. The only empirical input is the pre-randomization county-level covariates of the Dickinson et al.[1] example as distributed in cvcrand 0.1.1. No trial outcomes and no individual-level data are analyzed.

Ethics statement. This study used only simulation and the pre-randomization, county-level covariates of the Dickinson et al.[1] example, which are public aggregate data distributed with the cvcrand package. No human participants, individual-level data or trial outcomes were recruited, accessed or analyzed, so institutional review board approval and informed consent were not

required. The quantum-hardware runs executed a small synthetic circuit and involved no patient data.

Author contributions. Mohamed Faisal Sindhi and Muhammad Hammaz conceived the study. M.H. implemented the simulations and hardware runs and performed the analysis. Both authors wrote and revised the manuscript and approved the final version.

Funding. The authors received no specific funding for this work. IBM Quantum hardware time was provided under IBM's open plan.

Competing interests. The authors declare no competing interests.

Use of AI tools. Anthropic's Claude was used to help re-implement the simulations, re-run the analyses, generate the figures and draft and revise text. The authors reviewed the code, results and manuscript and take responsibility for the content.

Appendix A: Exploratory hardware run

This appendix reports one exploratory hardware campaign on one IBM device, ibm_marrakesh. It is not a certification study. Four frozen circuit types were used: two readout baselines (all-zero and all-one preparations), the Dicke preparation D1 for c = 2, and the complete QAOA circuit Q1 with the frozen p = 1, seed-0 parameters. Each circuit ran as two separate 2,048-shot jobs, eight jobs in total, and all completed. Raw counts were analyzed without mitigation, readout correction or postselection.

The frozen protocol specified the lower bound through a pooled upper Wilson 95% bound on the event proportion. The reported bounds instead use replicate-specific one-sided exact Clopper–Pearson bounds with Bonferroni allocation: four tests (two replicates by two events), familywise level 0.05, per-test level 0.0125. This replicate-specific analysis is a documented post-protocol exploratory deviation and is not a preregistered analysis.

For Q1, the event lower bounds on TV(Q, P*) are 0.9794 and 0.9835 for the accepted event, and 0.8572 and 0.8546 for the weight-8 event, for replicates 1 and 2. They use event probabilities only, so they are not full-distribution TV estimates and not closeness certificates. Observed accepted mass for Q1 was 0.0137 and 0.0103, against an ideal value of 0.1155. For D1 the observed weight-8 mass was 0.1416 and 0.1294 (pooled 0.1355), against an ideal value of 1. A readout-only diagnostic that models independent bit flips predicts a D1 weight-8 mass near 0.70, so readout error alone does not explain the observed loss. The evidence is consistent with noisy execution of the deep compiled Dicke preparation on this device. It does not uniquely attribute the loss to state preparation, routing, calibration drift, readout or the QAOA gates. The exact submitted Q1 transpilation could not be reproduced after calibration changes.

[Table A1 about here]

Appendix B: Hardware noise scaling on an 8-cluster instance

This appendix reports a new exploratory run on IBM ibm_fez (Heron r2, 156 qubits) on 21 October 2026, submitted through the IBM Quantum Platform Composer on the open plan. It is not a certification study and it does not repeat Appendix A. Its purpose is narrower: to measure how far device noise moves the XY-QAOA output away from the ideal circuit as depth grows, on a size where the ideal output is known exactly.

Instance. We took the first eight Dickinson clusters, dropped design columns that are constant in those eight, and standardized the rest. There are 70 balanced 4-versus-4 allocations. The accepted set is the 8 best-balanced allocations (4 complement pairs), which is the top 10% rounded up and made complement-closed; there is no tie at the boundary. Eight qubits encode the allocation. The circuit starts from the Néel state $|10101010\rangle$ (weight 4) and applies p = 1, 2 or 3 layers of a balance-score phase separator and a ring XY mixer. Angles were trained by Nelder–Mead on a

noiseless simulator. This is a shallow stand-in for the Dicke start used in the main text, so the ideal outputs at p = 1 and p = 2 stay near the start state and are far from the target; the experiment measures noise degradation, not sampler quality.

Runs. One job per depth, 4,096 shots each, on ibm_fez, with no error mitigation, readout correction or postselection (jobs dbcl6m7i0auc73dv6mf0, dbcl8us24mrc73cjsjrg and dbcl9ogc77vs7386h320). Each used about 3 s of device time. Circuits were pre-transpiled to the native gate set; the two-qubit gate counts below are from our all-to-all transpilation and exclude the routing the platform added, which we did not record.

[Table B1 about here]

Fully depolarized output would put 70/256 = 0.273 of its mass on weight-4 strings, and the binomial standard error of each weight-4 figure is about 0.008. At p = 1 the device keeps 53% of its mass in the correct weight sector and 44% on the correct string. By p = 3 the weight-4 mass (0.294) is within about three standard errors of the fully depolarized value, and the output is hardly distinguishable from noise. The hardware TV to the target is 0.96 or higher at every depth, and at p = 3 it is worse than the ideal circuit's own 0.872. Figure 5 shows the weight-4 mass by depth.

[Figure 5 about here]

What this does and does not show. It shows that, on one current device, the circuits needed for this task lose their structure before they reach useful depth, which supports treating raw hardware output as unfit for randomization without certification. It is one device, one day, one job per depth, and the platform's routing is unrecorded. It does not show how any mitigation or newer hardware would behave, and it says nothing about type-I error or power. Raw counts are saved in the project files.

References

1. Dickinson LM, Beaty B, Fox C, Pace W, Dickinson WP, Emsermann C, Kempe A. Pragmatic cluster randomized trials using covariate constrained randomization: a method for practice-based research networks (PBRNs). J Am Board Fam Med. 2015;28(5):663–672. doi:10.3122/jabfm.2015.05.150001

2. Yu H, Li F, Gallis JA, Turner EL. cvcrand: a package for covariate-constrained randomization and the clustered permutation test for cluster randomized trials. R J. 2019;11(2):191–204. doi:10.32614/RJ-2019-027

3. Morgan KL, Rubin DB. Rerandomization to improve covariate balance in experiments. Ann Stat. 2012;40(2):1263–1282. doi:10.1214/12-AOS1008

4. Li X, Ding P, Rubin DB. Asymptotic theory of rerandomization in treatment–control experiments. Proc Natl Acad Sci USA. 2018;115(37):9157–9162. doi:10.1073/pnas.1808191115

5. Wang B, Li F. On flexible covariate adjustment under covariate-constrained randomization. Clin Trials. 2026; online 4 March 2026. doi:10.1177/17407745261423479

6. Goldstein R, Jerzak CT, Kamat A, Zhu FW. fastrerandomize: fast rerandomization using accelerated computing. arXiv:2501.07642; 2025 (v3, January 2026).

7. Lu J, Zhang T, Zhu K. Fast rerandomization for balancing covariates in randomized experiments: a Metropolis–Hastings framework. arXiv:2602.07613; 2026. Preprint.

8. Besag J, Clifford P. Generalized Monte Carlo significance tests. Biometrika. 1989;76(4):633–642.

9. Phipson B, Smyth GK. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. Stat Appl Genet Mol Biol. 2010;9(1):Article 39. doi:10.2202/1544-6115.1585

10. Howes A. Markov chain Monte Carlo significance tests. arXiv:2310.04924; 2023.

11. Ramdas A, Barber RF, Candès EJ, Tibshirani RJ. Permutation tests using arbitrary permutation distributions. Sankhya A. 2023;85(2):1156–1177. arXiv:2204.13581.

12. NIH Pragmatic Trials Collaboratory. Rethinking Clinical Trials Living Textbook: covariate-constrained randomization. https://rethinkingclinicaltrials.org/?p=28747 [accessed 2026-10-19]

13. Zhu K, Liu H. Pair-switching rerandomization. Biometrics. 2023;79(3):2127–2142. doi:10.1111/biom.13712

14. Lu J, Liu D, Lin Z, Wang X. Fast rerandomization via the BRAIN. arXiv:2312.17230; 2023 (v2, 2025).

15. Toulis P. Asymptotic validity and finite-sample properties of approximate randomization tests. arXiv:1908.04218; 2019.

16. Mehta CR, Patel NR, Wei LJ. Constructing exact significance tests with restricted randomization rules. Biometrika. 1988;75(2):295–302.

17. Domingo L, Johnson C. Quantum-enhanced optimization for patient stratification in clinical trials. arXiv:2601.11413; 2026. Preprint.

18. Kowshik N, Manna S, Pal SP. Quantum approximate and quantum walk optimization approaches to set balancing. arXiv:2509.07200; 2025.

19. Lebl J, Shakeel A. Variational quantum algorithms for Euclidean discrepancy and covariate-balancing. arXiv:2103.09090; 2021.

20. Mandrà S, Zhu Z, Katzgraber HG. Exponentially biased ground-state sampling of quantum annealing machines with transverse-field driving Hamiltonians. Phys Rev Lett. 2017;118:070502. arXiv:1606.07146.

21. Könz MS, et al. Uncertain fate of fair sampling in quantum annealing. Phys Rev A. 2019;100:030303. arXiv:1806.06081.

22. Ide Y, Tanaka K. Unfair sampling of quantum annealing in weighted graph bipartitioning problems. arXiv:2604.11449; 2026. Preprint.

23. Bärtschi A, Eidenbenz S. Grover mixers for QAOA: shifting complexity from mixer design to state preparation. arXiv:2006.00354; 2020.

24. Pelofske E. Biased degenerate ground-state sampling of small Ising models with converged QAOA. Phys Rev E. 2025;111:054103. arXiv:2411.05294.

25. Hadfield S, Wang Z, O'Gorman B, Rieffel EG, Venturelli D, Biswas R. From the quantum approximate optimization algorithm to a quantum alternating operator ansatz. Algorithms. 2019;12(2):34. doi:10.3390/a12020034

26. Wang Z, Rubin NC, Dominy JM, Rieffel EG. XY mixers: analytical and numerical results for the quantum alternating operator ansatz. Phys Rev A. 2020;101:012320.

27. Valiant G, Valiant P. An automatic inequality prover and instance optimal identity testing. SIAM J Comput. 2017;46(1):429–455. doi:10.1137/151002526

28. Paninski L. A coincidence-based test for uniformity given very sparsely sampled discrete data. IEEE Trans Inf Theory. 2008;54(10):4750–4755. doi:10.1109/TIT.2008.928987

29. Chan S-O, Diakonikolas I, Valiant G, Valiant P. Optimal algorithms for testing closeness of discrete distributions. In: Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA); 2014:1193–1203. doi:10.1137/1.9781611973402.88

30. Mardia J, Jiao J, Tánczos E, Nowak RD, Weissman T. Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types. Information and Inference. 2020;9(4):813–850. arXiv:1809.06522.

31. Javadi-Abhari A, Treinish M, Krsulich K, Wood CJ, Lishman J, Gacon J, et al. Quantum computing with Qiskit. arXiv:2405.08810; 2024.

32. Virtanen P, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–272. doi:10.1038/s41592-019-0686-2

33. Brassard G, Høyer P, Mosca M, Tapp A. Quantum amplitude amplification and estimation. In: Quantum Computation and Quantum Information, Contemporary Mathematics vol. 305; 2002:53–74.

34. Bärtschi A, Eidenbenz S. Deterministic preparation of Dicke states. In: Fundamentals of Computation Theory (FCT 2019), Lecture Notes in Computer Science. Springer; 2019:126–139. doi:10.1007/978-3-030-25027-0_9

35. Bärtschi A, Eidenbenz S. Short-depth circuits for Dicke state preparation. In: IEEE International Conference on Quantum Computing and Engineering (QCE); 2022:87–96. doi:10.1109/QCE53715.2022.00027

36. Yu J, Muleady SR, Wang Y-X, Schine N, Gorshkov AV, Childs AM. Efficient preparation of Dicke states. Phys Rev Lett. 2026;136:030601. doi:10.1103/9gjk-rgql. arXiv:2411.03428.

37. Farhi E, Goldstone J, Gutmann S. A quantum approximate optimization algorithm. arXiv:1411.4028; 2014.

38. Zhou L, Wang S-T, Choi S, Pichler H, Lukin MD. Quantum approximate optimization algorithm: performance, mechanism, and implementation on near-term devices. Phys Rev X. 2020;10(2):021067.

39. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–2102.

Table 1. Selected quantum-arm results from full ideal statevectors. The three masses sum to one.

Sampler (selected)	**Raw TV to P***	**Certified (≤ 0.02)**	**Conditional TV (postselected)**	**P(A)**	**P(Ω8 \ A)**	**P(wrong weight)**
Amplification, t* = 2	0.001440	Yes	6.98×10^{-17}	0.998560	0.001440	0
XY-QAOA, p = 3, seed 9	0.833922	No (far)	0.497427	0.167845	0.832155	0.000000
Penalty-QAOA, p = 3, seed 6	0.949081	No (far)	0.002334	0.050919	0.429113	0.519968

Table 2. Sharp-null type-I rejection rates at α = 0.05, base cell.

Reference	Raw TV	Type-I, T1 (%)	Type-I, T2 (%)
Exact target P*	0.000000	4.98 [4.79–5.18]	5.05 [4.86–5.25]
Swap, t = 10	0.725991	5.89 [5.69–6.10]	7.56 [7.33–7.79]
Swap, t = 41	0.248146	4.93 [4.74–5.12]	5.02 [4.83–5.21]
Swap, t = 170	0.009850	4.98 [4.79–5.18]	5.05 [4.86–5.25]
Biased descent	0.958042	15.87 [15.56–16.20]	6.50 [6.29–6.72]
Amplification, t* = 2	0.001440	4.87 [4.68–5.06]	5.05 [4.86–5.25]
XY-QAOA raw	0.833922	0.20 [0.16–0.24]	4.53 [4.35–4.72]
Penalty-QAOA raw (bounds)	0.949081	[0.00, 0.82]	[0.00, 9.33]

Table 3. Decision disagreement with P at $\alpha = 0.05$, constant shift $\tau = 0.50$, base cell.*

Reference	**Raw TV**	**Disagreement, T1 (%)**	**Disagreement, T2 (%)**
Exact target P*	0.000000	0 (by construction)	0 (by construction)
Swap, t = 10	0.725991	10.92 [10.65–11.20]	5.87 [5.67–6.08]
Swap, t = 41	0.248146	1.42 [1.32–1.52]	0.75 [0.68–0.83]
Swap, t = 170	0.009850	0.00 [0.00–0.01]	0.00 [0.00–0.01]
Biased descent	0.958042	14.70 [14.39–15.01]	8.63 [8.39–8.88]
Amplification, t* = 2	0.001440	0.08 [0.06–0.10]	0.00 [0.00–0.01]
XY-QAOA raw	0.833922	36.67 [36.24–37.09]	2.87 [2.73–3.02]
Penalty-QAOA raw (bounds)	0.949081	[16.71, 68.21]	[0.04, 96.94]

Table 4. XY-QAOA on four synthetic instances (ideal statevector). Status is far when raw TV ≥ 0.10.

Instance	κ_cov	J	Accepted set size	Raw TV	Status
S12-k02-a	0.2	12	100	0.592	far
S12-k06-a	0.6	12	100	0.613	far
S16-k02-a	0.2	16	1,288	0.595	far
S16-k06-a	0.6	16	1,288	0.722	far

Table 5. Sharp-alternative rejection counts out of 10,000 under P for both assignment and reference, constant shift $\tau = 0.5$, equal-size Gaussian outcomes.*

Instance	J	Accepted set size	ICC 0.05, T1	ICC 0.05, T2	ICC 0.20, T1	ICC 0.20, T2
S12-k02-a	12	100	3,385	6,878	2,029	2,784
S12-k06-a	12	100	4,109	7,090	2,345	2,924
S16-k02-a	16	1,288	6,451	9,249	3,845	4,841
S16-k06-a	16	1,288	6,710	9,265	4,031	4,918

Table 6. Extension experiments (ideal statevector, raw TV over all 1,288 bins).

Arm	**p = 1**	**p = 2**	**p = 3**	**p = 4**
XY-QAOA, TV-minimized ceiling	0.8846	0.8360	0.8294	0.8284
Grover-mixer QAOA, energy-selected	0.8133	0.7718	0.7124	not run
Grover-mixer QAOA, TV-minimized ceiling	0.8078	0.7471	0.6862	not run

Table A1. Implementation-error ladder: what each layer measured, supports and does not support.

Layer	Measured	Supported	Unsupported
Mathematical target P*	Exact rational construction of the 1,287-allocation accepted set and the uniform target	P* is the intended uniform law on the accepted set	Anything about a physical sampler
Ideal full-statevector sampler	Ideal D1 weight-8 mass; ideal Q1 accepted mass 0.1155	The logical circuits implement the intended ideal laws	Any device claim; any noise attribution
Compiled gate-level circuit	Ideal simulation of transpiled D1 (TV 6×10^{-14}); current-pipeline Q1 reconstruction (1×10^{-13})	Transpilation preserves the ideal D1 law; decoder and measurement map are consistent	Direct re-establishment of the submitted Q1 transpilation
Raw hardware counts	Eight integrity-checked jobs; D1 weight-8 mass about 0.13–0.14; Q1 accepted mass about 0.01–0.014; readout rates	These raw observations for these jobs; readout alone does not explain the D1 weight loss	A device-wide distribution; full-distribution closeness; causal attribution
Downstream inference	Not measured on hardware (no hardware p-values, rejection rates or power)	Nothing new	Any claim that hardware sampling changes type-I error or power

Table B1. Hardware against the ideal circuit output (4,096 shots per depth).

Depth	CZ gates (pre-routing)	Weight-4 mass, hardware	Mass on start state, hardware / ideal	TV, hardware vs ideal	TV, hardware vs target	TV, ideal vs target
p = 1	56	0.533	0.440 / 1.000	0.560	0.986	1.000
p = 2	88	0.353	0.033 / 0.329	0.841	0.971	0.999
p = 3	213	0.294	0.011 / 0.530	0.886	0.968	0.872

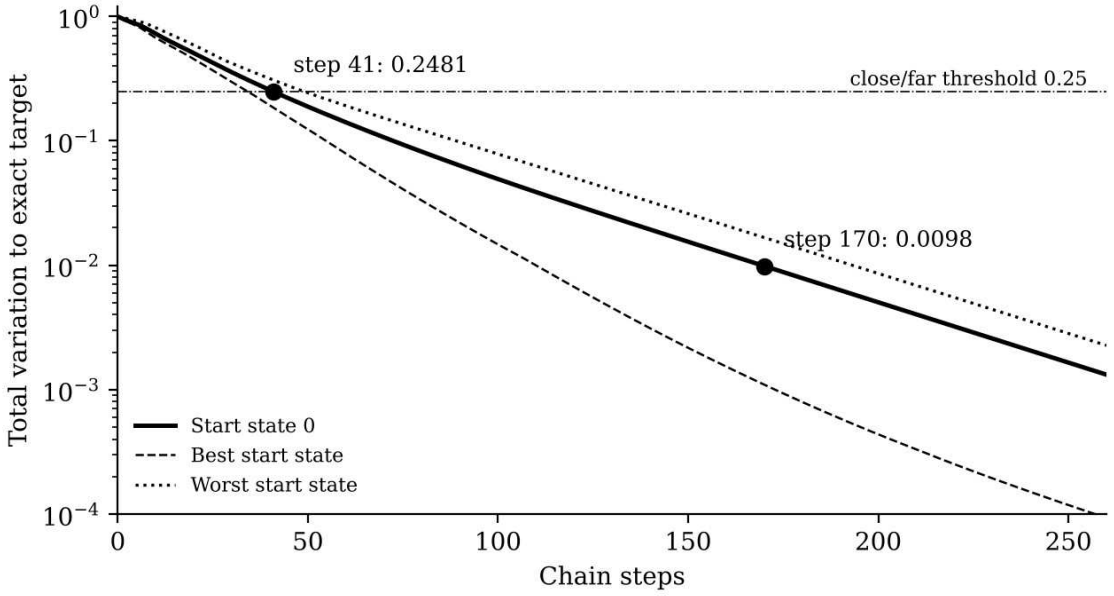


Figure 1. Swap-chain convergence to the exact target, computed from the 1,287 x 1,287 transition matrix.

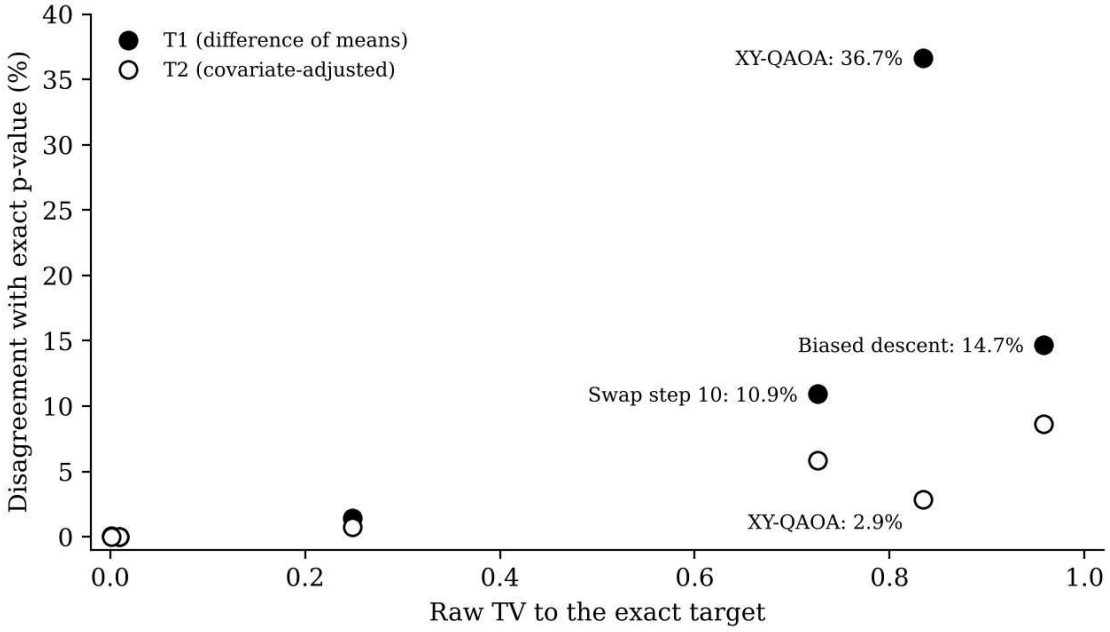


Figure 2. Raw TV against decision disagreement, base cell, R = 50,000 per law. Penalty-QAOA is omitted because its p-value is only bounded.

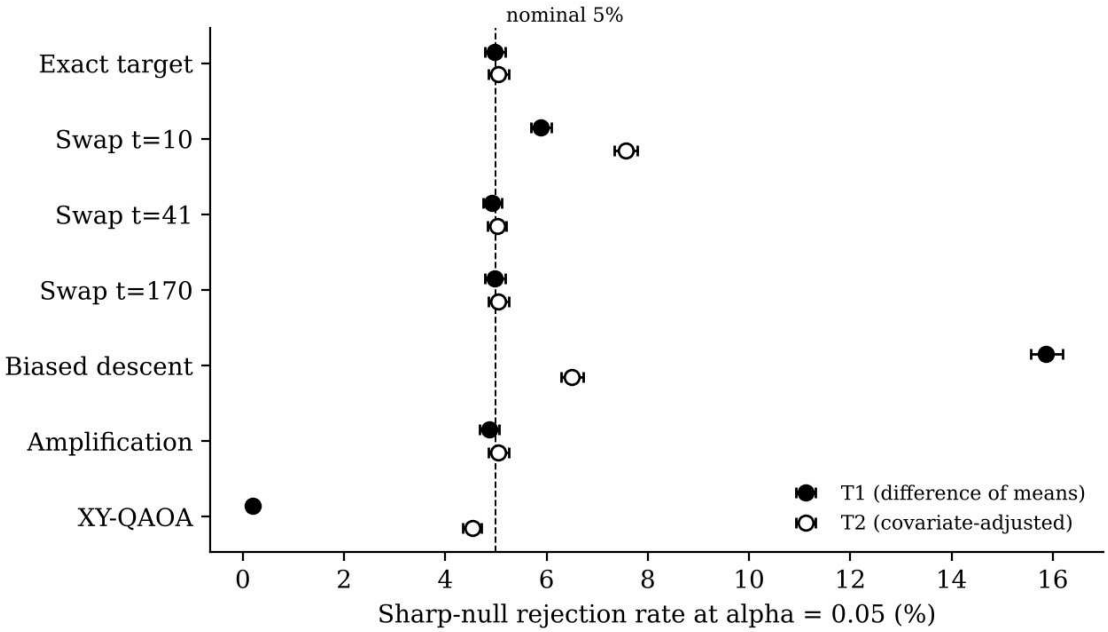


Figure 3. Sharp-null rejection rates (alpha = 0.05) under each sampler law, base cell, 50,000 replicates per law. Bars are Wilson 95% intervals; the dashed line is the nominal 5% level.

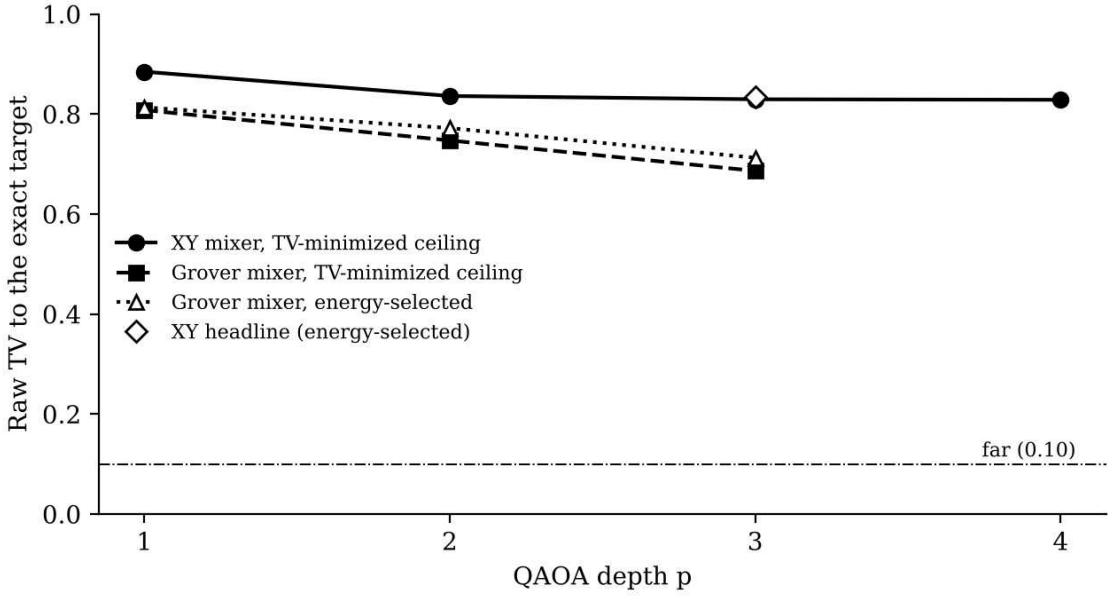


Figure 4. Raw TV to the exact target by QAOA depth for the XY and Grover mixers. Ceilings come from direct TV minimization (an oracle diagnostic); the diamond is the energy-selected XY headline.

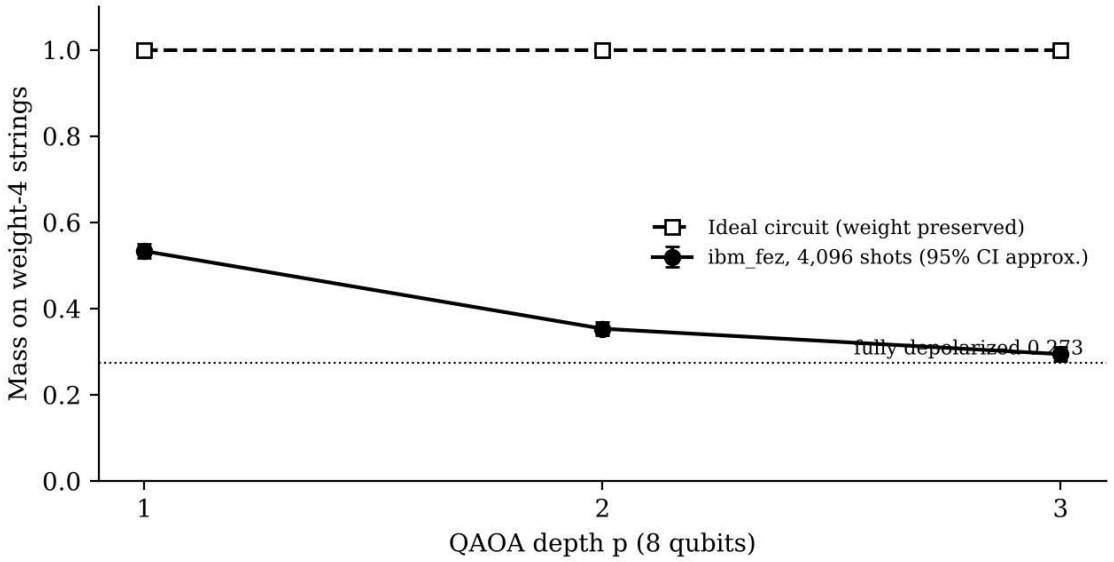


Figure 5. Weight-4 mass of the 8-qubit circuit output on ibm_fez against the ideal circuit (4,096 shots per depth, no error mitigation).