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.

Beyond Exact Recovery: Operational Success Probability and Sampling Cost for Approximate Quantum Optimization under Stochastic Uncertainty

Roberto Garrone

roberto.garrone@st.ouc.ac.cy

Open University of Cyprus https://orcid.org/0009-0005-7060-6774

Research Article

Keywords: Quantum Approximate Optimization Algorithm (QAOA,Quadratic Unconstrained Binary Optimization (QUBO),Quantum optimization,Quantum algorithm benchmarking

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11499298/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.

Beyond Exact Recovery: Operational Success Probability and Sampling Cost for Approximate Quantum Optimization under Stochastic Uncertainty

Roberto Garrone
roberto.garrone@st.ouc.ac.cy

Abstract

Approximate optimizers are usually judged by how well they recover a mathematical optimum or a near-optimal objective value. In many applications, however, the optimization model is only a surrogate for a downstream decision problem, so mathematical quality and practical decision quality need not coincide. We formalize this distinction through three success probabilities: exact surrogate recovery, surrogate near-optimality, and operational acceptability under stochastic evaluation. The framework is studied on a small binary inventory-control benchmark using a calibrated QUBO surrogate, exact enumeration, greedy search, simulated annealing, uniform feasible sampling, and ideal QAOA at depths $p = 1, 2, 3$.

The confirmatory results show that the near-optimal and operational acceptance sets are generally different. This changes finite-shot requirements substantially: for $n = 8, p = 2$, the median number of measurements needed for a 95% chance of observing a successful state falls from 392 under exact recovery to 84 under QUBO near-optimality and 31 under the operational criterion. Because operationally acceptable states can also be common, we compare QAOA with uniform feasible sampling; enrichment is supported only in selected regimes, and the classical baselines outperform QAOA on these small instances. A post-confirmatory IBM hardware study provides a complementary physical test. For $n = 4$, operational success can remain close to its ideal value even when the measured decision distribution is distorted. For $n = 6, p = 1$, hardware gives $S_N = 0.02475$ and $S_O = 0.01630$, showing directly that near-optimal surrogate solutions need not be operationally acceptable. At $n = 6, p = 2$, Fez and Kingston yield different distances from the ideal decision distribution (0.4475 and 0.5326) but similar operational-success probabilities (0.0185 and 0.0171). The study does not indicate quantum advantage; it shows that exact recovery, surrogate quality, operational success, and distributional fidelity answer different evaluation questions and should be reported jointly.

Keywords: Quantum Approximate Optimization Algorithm (QAOA); Quadratic Unconstrained Binary Optimization (QUBO); Quantum optimization; Stochastic optimization; Operational decision quality; Success probability; Finite-shot sampling complexity; Quantum algorithm benchmarking.

1 Introduction

An optimizer can miss the mathematical optimum and still return a decision that works well in practice. The reverse is also possible: a solution can be close to the optimum of a mathematical model yet perform poorly when it is evaluated in the uncertain environment that the model was meant to represent. This distinction is especially relevant for probabilistic solvers such as the Quantum

Approximate Optimization Algorithm (QAOA), which produces a distribution over candidate bit strings rather than a single deterministic answer [1].

Standard QAOA benchmarks therefore consider quantities such as expected energy, approximation quality, and ground-state or optimal-solution probability [6, 7]. Recent benchmarking work also emphasizes standardized instances, classical references, runtimes, and reproducible performance measures [11]. Approximate-solution criteria and their finite-shot consequences are likewise established topics: one can ask for the probability of drawing a solution above a chosen quality threshold rather than insisting on exact recovery [8].

The additional issue studied here appears when the optimization problem itself is only an intermediate model. A practical decision pipeline may contain estimated coefficients, simplified constraints, deterministic approximations, or surrogate objectives. In that setting, a solution that is near-optimal for the surrogate is not necessarily near-optimal for the downstream decision problem. This model-versus-decision distinction is closely related to decision-focused learning, where performance at an intermediate modeling stage need not predict downstream decision quality [14].

We study the corresponding question for approximate quantum optimization through the pipeline

$$\text{stochastic problem} \rightarrow \text{QUBO surrogate} \rightarrow \text{approximate solver} \rightarrow \text{stochastic evaluation}.$$

QUBO and Ising formulations are standard representations of binary optimization problems [3, 2, 4]. The question is not whether near-optimal solutions can be useful, but whether being near-optimal for the surrogate conveys the same decision-relevant information as being acceptable in the downstream environment.

We therefore distinguish three success events. Exact recovery asks whether the surrogate optimum is sampled. Near-optimal success allows any feasible solution within a prescribed QUBO gap. Operational success instead asks whether the sampled decision remains within prespecified expected-cost and risk tolerances when it is evaluated under stochastic demand. For a solver distribution $P(x)$, these events can contain different states and therefore different probability mass.

The distinction also has a direct sampling interpretation. If a chosen success set has probability mass S, then the probability of observing at least one successful state in N independent measurements is

$$P_{\geq 1}(N, S) = 1 - (1 - S)^N,$$

and the number of measurements needed for confidence $1 - \eta$ is

$$N_{1-\eta}(S) = \left\lceil \frac{\log \eta}{\log(1 - S)} \right\rceil.$$

Changing the definition of success can therefore change the apparent sampling burden even when the underlying solver distribution is unchanged.

Operational success must also be interpreted relative to how common acceptable decisions are. If a large share of the feasible space already satisfies the downstream criterion, a high S_O need not imply useful concentration by the optimizer. We therefore compare solver probability with uniform feasible sampling through the descriptive enrichment ratio

$$E_O = \frac{P_{\text{solver}}(\mathcal{O})}{P_{\text{uniform}}(\mathcal{O})}.$$

The empirical setting is a small stochastic inventory-control problem, chosen because all feasible decisions can be enumerated and evaluated exactly at the tested dimensions. The stochastic objective is represented by a calibrated QUBO surrogate whose parameters are selected on a calibration set, evaluated on a held-out set, and then frozen before an independent 81-instance confirmatory study. We compare exact enumeration, greedy search, simulated annealing [16], uniform feasible sampling, and ideal statevector QAOA at $p = 1, 2, 3$. Expected cost and CVaR are used to evaluate downstream performance; CVaR is a standard risk measure [12] and, unlike CVaR-based variational objectives [13], is applied here only after a candidate decision has been sampled.

The small instances are intentional. They make the complete surrogate and operational landscapes observable, while also providing classical baselines strong enough to prevent favorable interpretation of QAOA by construction. In fact, greedy search and simulated annealing outperform the tested QAOA configurations, and shallow QAOA can assign less operationally acceptable probability than uniform feasible sampling. The study therefore makes no claim of quantum advantage. A separate post-confirmatory hardware extension asks a narrower question: do the distinctions among exact recovery, near-optimality, operational success, and distributional fidelity remain visible after strict-QUBO encoding, routing, and physical sampling on IBM superconducting processors?

The main contributions are:

1. a joint evaluation of exact surrogate recovery, surrogate near-optimality, and downstream stochastic acceptability;
2. an explicit decomposition of probability mass that is near-optimal but operationally unacceptable, and vice versa;
3. a finite-shot interpretation of these alternative success definitions, complemented by a uniform-feasible prevalence baseline;
4. a frozen simulation pipeline independently cross-validated with Qiskit and Qiskit Aer [23]; and
5. a targeted physical-QPU study showing that decision-level exact recovery, near-optimality, operational success, and distributional fidelity can move differently across depth, encoded problem size, and backend/calibration conditions.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 formalizes the evaluation framework. Section 4 describes the benchmark and frozen protocol. Section 5 presents the confirmatory and hardware results. Section 6 discusses interpretation and limitations, and Section 7 concludes.

2 Related Work

2.1 QAOA and quantum optimization benchmarking

QAOA was introduced by Farhi, Goldstone, and Gutmann as a variational algorithm for approximate combinatorial optimization [1]. It alternates a problem-dependent cost Hamiltonian with a mixing Hamiltonian, with circuit depth governed by p.

Because the final state defines a probability distribution over computational-basis solutions, several quantities can be used to assess performance. Willsch et al. distinguish ground-state probability,

expected energy, and approximation-type quality measures, showing that conclusions depend strongly on the instance [6]. Larkin et al. further evaluate QAOA in terms of the approximation ratio of individual samples and the probability of sampling sufficiently good solutions [7].

More recent benchmarking efforts emphasize standardized problem formulations, classical references, reproducible protocols, and careful separation between algorithmic performance and claims of quantum advantage [11]. Our experiments follow this principle by evaluating QAOA alongside exact enumeration, greedy optimization, simulated annealing, and uniform feasible sampling.

Outer-loop parameter optimization is itself an important part of QAOA evaluation, especially as depth grows. Zhou et al. develop an efficient parameter-optimization procedure and structured initialization heuristics for higher-depth QAOA, showing that naive random initialization can become increasingly costly [9]. Hao et al. study an end-to-end protocol for obtaining high-quality QAOA parameters when the number of shots is limited, directly linking parameter-search design to practical resource constraints [10]. These considerations are relevant here because our cross-depth comparison fixes the number of parameter candidates per depth rather than attempting to equalize optimization quality across depths.

2.2 Approximate solutions and measurement requirements

Requiring recovery of a unique mathematical optimum can be unnecessarily restrictive for an approximate optimizer. Expectation values, approximation ratios, and target-quality sampling probabilities already assign value to solutions other than the exact optimum [6, 7].

Chernyavskiy et al. directly connect approximate QUBO solution quality to the number of QAOA measurements required to obtain a target-quality solution [8]. Accordingly, neither the use of a success set containing multiple solutions nor the transformation from success probability to a finite-shot observation requirement is claimed here as new.

The distinction introduced in this paper concerns *which set defines success* when the QUBO is only a surrogate. We distinguish a mathematical near-optimal set $\mathcal{Q}_\epsilon$ from an operational acceptance set $\mathcal{O}_{\epsilon,\delta}$ defined by a separate stochastic evaluator.

2.3 Risk-sensitive objectives and CVaR

CVaR was formalized as an optimization-oriented tail-risk measure by Rockafellar and Uryasev [12]. It has also been introduced directly into variational quantum optimization as an alternative objective aggregation function [13].

The present use is conceptually different. QAOA parameters are optimized with respect to the encoded surrogate energy. CVaR is evaluated only after candidate bit strings are mapped back into the stochastic operating environment:

$$\text{QUBO/QAOA objective} \longrightarrow \text{candidate decision} \longrightarrow \text{stochastic operational CVaR}.$$

This separation preserves the model-versus-deployment distinction studied in the paper.

2.4 Near-optimal solution spaces

The idea that multiple structurally distinct decisions can perform nearly as well as the mathematical optimum has a long history in operations research. Modeling to Generate Alternatives (MGA) explicitly explores near-optimal regions in order to identify alternatives that differ in properties not fully represented in a primary optimization objective [15].

Accordingly, we do not claim novelty for the mathematical near-optimal set

$$\mathcal{Q}_{\epsilon_Q} = \{x : Q(x) - Q^* \leq \epsilon_Q \Delta_Q\},$$

or for the existence of multiple near-optimal solutions. In the present study, $\mathcal{Q}_{\epsilon_Q}$ is a control condition used to test whether relaxing exact QUBO recovery is sufficient to capture downstream-acceptable decisions.

2.5 Surrogate objectives and downstream decision quality

The closest conceptual precedent comes from decision-focused learning. Wilder et al. argue that quality measured in an intermediate predictive or modeling layer can be misaligned with the quality of the downstream decision [14].

Our setting has an analogous structure but a different intermediate component:

$$\begin{aligned}\text{stochastic decision problem} &\to \text{QUBO surrogate} \to \text{approximate solver} \\ &\to \text{candidate solution} \to \text{stochastic evaluation}.\end{aligned}$$

The potential mismatch therefore lies between a surrogate optimization objective and an operational decision criterion.

2.6 Quantum optimization for inventory

Quantum and quantum-hybrid methods have already been explored for inventory-management problems. Fernandes et al. formulate an inventory problem for classical, quantum, and hybrid hardware [22]. The present benchmark differs in its stochastic binary replenishment setting, but inventory itself is not claimed as a new quantum-optimization domain.

Inventory is used because it provides a compact environment in which the complete end-to-end chain can be evaluated exactly for small instances. Exhaustive enumeration makes it possible to identify the QUBO optimum, the complete mathematical near-optimal set, the operational optimum, and the operationally acceptable set without relying on another approximate reference solver.

2.7 Positioning

The preceding literature establishes clear novelty boundaries. Multiple QAOA performance metrics are established; approximation-quality success and measurement requirements have been studied; near-optimal solution sets are longstanding; CVaR is mature; and intermediate-model quality versus downstream decision quality has substantial precedent.

The contribution of this paper lies instead in combining these distinctions into an end-to-end evaluation of a probabilistic approximate optimization pipeline in which the surrogate objective and stochastic deployment criterion are explicitly different.

For solver distribution $P_s(x)$, we separately evaluate

$$S_E = P_s(x = x_Q^*),$$

$$S_N(\epsilon) = P_s(x \in \mathcal{Q}_\epsilon),$$

and

$$S_O(\epsilon, \delta) = P_s(x \in \mathcal{O}_{\epsilon,\delta}).$$

Because S_O can be large simply when $\mathcal{O}$ occupies a large portion of the feasible space, we additionally compare it with uniform feasible sampling:

$$E_O = \frac{P_s(\mathcal{O}_{\epsilon,\delta})}{P_{\text{uniform}}(\mathcal{O}_{\epsilon,\delta})}.$$

The framework is therefore an evaluation layer for application-oriented approximate optimization, rather than a new QAOA variant, a new QUBO construction technique, a new risk measure, or an argument for quantum advantage.

3 Problem Formulation and Evaluation Framework

3.1 Decision pipeline and surrogate-based success

Consider a stochastic binary decision problem with decision vector $x \in \mathcal{X} = \{0,1\}^n$ and feasible subset $\mathcal{F} \subseteq \mathcal{X}$. For uncertainty realization $\omega \in \Omega$, let $L(x, \omega)$ denote operational loss, with expected operational objective

$$J(x) = \mathbb{E}_\omega[L(x, \omega)], \qquad x_J^* \in \arg\min_{x \in \mathcal{F}} J(x).$$

The stochastic function J is not optimized directly. Instead, the original problem is represented by the QUBO surrogate

$$Q(x) = c + \sum_{i=1}^{n} a_i x_i + \sum_{i<j} b_{ij} x_i x_j.$$

The resulting pipeline is

stochastic problem $\rightarrow$ QUBO surrogate $\rightarrow$ approximate solver
$\rightarrow$ candidate decision $\rightarrow$ stochastic evaluation.

This creates two separate approximation layers: surrogate error and solver error.

Let

$$\mathcal{X}_Q^* = \arg\min_{x \in \mathcal{F}} Q(x).$$

For solver distribution $P_s(x)$, exact-recovery probability is

$$S_E = \sum_{x \in \mathcal{X}_Q^*} P_s(x),$$

which reduces to $S_E = P_s(x_Q^*)$ for a unique optimum.

To define mathematical near-optimality, let

$$\Delta_Q = \max_{x \in \mathcal{F}} Q(x) - \min_{x \in \mathcal{F}} Q(x), \qquad Q^* = \min_{x \in \mathcal{F}} Q(x),$$

and

$$R_Q(x) = \frac{Q(x) - Q^*}{\Delta_Q}.$$

Feasible-range normalization is used instead of a multiplicative ratio such as $Q(x)/Q^*$ because QUBO energies are defined only up to an arbitrary additive constant and may have an optimum at or below zero. A multiplicative ratio can therefore change under an innocuous energy shift or become undefined when $Q^* = 0$. By contrast, for every positive affine transformation $Q'(x) = aQ(x) + b$, $a > 0$,

$$\frac{Q'(x) - Q'^*}{\max_{y \in \mathcal{F}} Q'(y) - \min_{y \in \mathcal{F}} Q'(y)} = R_Q(x).$$

Thus R_Q measures location within the feasible QUBO energy range while remaining invariant to additive shifts and positive rescaling. All generated confirmatory instances have $\Delta_Q > 0$.

For tolerance $\epsilon_Q \geq 0$,

$$\mathcal{Q}_{\epsilon_Q} = \{x \in \mathcal{F} : R_Q(x) \leq \epsilon_Q\}, \qquad S_N(\epsilon_Q) = \sum_{x \in \mathcal{Q}_{\epsilon_Q}} P_s(x).$$

3.2 Downstream operational success and set disagreement

Expected-cost regret relative to x_J^* is

$$R_J(x) = \frac{J(x) - J(x_J^*)}{J(x_J^*)},$$

assuming $J(x_J^*) > 0$. Let

$$C_\alpha(x) = \mathrm{CVaR}_\alpha[L(x,\omega)], \qquad R_C(x) = \frac{C_\alpha(x) - C_\alpha(x_J^*)}{C_\alpha(x_J^*)}.$$

For tolerances ϵ_J and δ_C, define

$$\mathcal{O}_{\epsilon_J,\delta_C} = \{x \in \mathcal{F} : R_J(x) \le \epsilon_J \wedge R_C(x) \le \delta_C\},$$

with operational-success probability

$$S_O(\epsilon_J, \delta_C) = \sum_{x \in \mathcal{O}_{\epsilon_J,\delta_C}} P_s(x).$$

To characterize disagreement between mathematical and operational acceptance, define

$$\begin{aligned} \mathcal{I} &= \mathcal{Q}_{\epsilon_Q} \cap \mathcal{O}_{\epsilon_J,\delta_C}, \\ \mathcal{Q}_{\text{only}} &= \mathcal{Q}_{\epsilon_Q} \setminus \mathcal{O}_{\epsilon_J,\delta_C}, \\ \mathcal{O}_{\text{only}} &= \mathcal{O}_{\epsilon_J,\delta_C} \setminus \mathcal{Q}_{\epsilon_Q}. \end{aligned}$$

Their solver probability masses are

$$\begin{aligned} S_I &= \sum_{x \in \mathcal{I}} P_s(x), \\ S_{Q\setminus O} &= \sum_{x \in \mathcal{Q}_{\text{only}}} P_s(x), \\ S_{O\setminus Q} &= \sum_{x \in \mathcal{O}_{\text{only}}} P_s(x). \end{aligned}$$

Therefore,

$$S_N = S_I + S_{Q\setminus O}, \qquad S_O = S_I + S_{O\setminus Q},$$

and

$$S_O - S_N = S_{O\setminus Q} - S_{Q\setminus O}.$$

3.3 Feasibility and finite-shot observation cost

Some approximate solvers can produce infeasible candidate solutions. This is relevant for QAOA with a conventional transverse-field X-mixer, which does not preserve a constrained feasible subspace; feasibility-preserving alternatives are discussed by Hadfield et al. [5]. Define feasible probability mass as

$$S_F = \sum_{x \in \mathcal{F}} P_s(x).$$

All primary success metrics are computed with respect to the full solver distribution rather than conditioning on feasibility. Infeasible probability mass therefore contributes zero to S_E, S_N, and S_O.

For success probability S, the probability of observing at least one successful solution in N independent measurements is

$$P_{\geq 1}(N; S) = 1 - (1 - S)^N.$$

For target confidence ρ,

$$N_\rho(S) = \left\lceil \frac{\log(1-\rho)}{\log(1-S)} \right\rceil,$$

and specifically

$$N_{95}(S) = \left\lceil \frac{\log(0.05)}{\log(1-S)} \right\rceil,$$

with $N_{95}(0) = \infty$ and $N_{95}(1) = 1$. These quantities isolate the sampling consequence of changing the definition of success; they are not end-to-end runtimes.

3.4 Operational prevalence and surrogate fidelity

Uniform feasible sampling is

$$P_U(x) = \begin{cases} 1/|\mathcal{F}|, & x \in \mathcal{F}, \\ 0, & x \notin \mathcal{F}. \end{cases}$$

Operational prevalence and enrichment are

$$S_O^U = \frac{|\mathcal{O}_{\epsilon_J, \delta_C}|}{|\mathcal{F}|}, \qquad E_O = \frac{S_O}{S_O^U}.$$

Thus $E_O > 1$ indicates concentration of operationally acceptable probability beyond its prevalence under uniform feasible sampling, whereas $E_O < 1$ indicates under-enrichment.

Surrogate fidelity addresses a different issue. If

$$x_Q^* \notin \mathcal{O}_{\epsilon_J, \delta_C},$$

then perfect recovery of the QUBO optimum still fails the downstream criterion. Across a collection of instances, define

$$F_M = \Pr\left[x_Q^* \in \mathcal{O}_{\epsilon_J, \delta_C}\right].$$

This distinguishes surrogate mismatch from solver behavior.

3.5 QAOA realization

For cost Hamiltonian H_C and standard mixer

$$H_M = \sum_{j=1}^{n} X_j,$$

the depth-p state is

$$|\psi_p(\boldsymbol{\gamma}, \boldsymbol{\beta})\rangle = \prod_{\ell=1}^{p} e^{-i\beta_\ell H_M} e^{-i\gamma_\ell H_C} |+\rangle^{\otimes n},$$

with output probability

$$P_Q(x) = \left|\langle x|\psi_p(\boldsymbol{\gamma}, \boldsymbol{\beta})\rangle\right|^2.$$

The framework evaluates the full distribution through

$$\begin{aligned} S_E &= \sum_{x \in \mathcal{X}_Q^*} P_Q(x), \\ S_N &= \sum_{x \in \mathcal{Q}_{\epsilon_Q}} P_Q(x), \\ S_O &= \sum_{x \in \mathcal{O}_{\epsilon_J, \delta_C}} P_Q(x). \end{aligned}$$

3.6 Primary evaluation questions

The confirmatory analysis asks how much mathematical near-optimality changes success relative to exact recovery, through $S_N - S_E$; whether $S_O - S_N$ is materially different from zero; whether $S_{O \setminus Q}$ and $S_{Q \setminus O}$ are non-negligible; how the alternative success definitions change

$$N_{95}^E, \qquad N_{95}^N, \qquad N_{95}^O;$$

and whether $E_O > 1$. No ordering such as $S_E < S_N < S_O$ is assumed a priori.

4 Experimental Design

4.1 Objective and stochastic inventory benchmark

The confirmatory experiment tests whether exact surrogate recovery, surrogate near-optimality, and downstream operational acceptability lead to materially different assessments of approximate optimization. The benchmark is intentionally small enough for complete binary-state enumeration,

allowing exact characterization of the surrogate optimum, mathematical near-optimal set, operational reference, and prevalence of acceptable solutions without introducing another approximate reference solver. The experiment is not designed to test quantum advantage.

For item i, the binary decision $x_i \in \{0,1\}$ determines whether an additional replenishment quantity is ordered, giving available stock

$$q_i(x) = b_i + d_i x_i,$$

where b_i is baseline stock and d_i is incremental replenishment. Instance parameters are generated as

$$\begin{aligned}
&\mu_i \sim U(8,20), && b_i = 0.70\mu_i,\\
&d_i = u_i\mu_i, \quad u_i \sim U(0.25,0.55), && c_i \sim U(0.3,1.2),\\
&h_i \sim U(0.7,1.3), && p_i = rh_i, \quad r \in \{1,3,10\},\\
&w_i \sim U(0.7,1.3).
\end{aligned}$$

The feasible set is

$$\mathcal{F} = \left\{ x \in \{0,1\}^n : \sum_i w_i x_i \le B \right\}, \qquad B = \tau \sum_i w_i, \quad \tau \in \{0.40, 0.65, 0.90\}.$$

Demand is generated as

$$D_i = \max(0, \tilde{D}_i), \qquad \tilde{D}_i \sim \mathcal{N}\Big(\mu_i, (cv\,\mu_i)^2\Big), \qquad cv \in \{0.10, 0.30, 0.50\}.$$

Operational loss is

$$L(x,D) = \sum_{i=1}^{n} \left[h_i \max(q_i(x) - D_i, 0) + p_i \max(D_i - q_i(x), 0) + c_i x_i \right].$$

Expected operational loss is estimated using

$$\hat{J}(x) = \frac{1}{M} \sum_{m=1}^{M} L(x, D^{(m)}), \qquad M = 1800.$$

The same demand realizations are used for every candidate within an instance. This common-random-numbers design is a standard variance-reduction strategy for comparing simulation alternatives [18, 19]. The expected-cost reference is

$$x_J^* = \arg\min_{x \in \mathcal{F}} \hat{J}(x).$$

4.2 Risk evaluation and operational criteria

For each decision x, let

$$L_m(x) = L(x, D^{(m)}), \qquad m = 1, \ldots, M.$$

Empirical CVaR is evaluated at $\alpha = 0.95$. The implementation first computes

$$\widehat{q}_\alpha(x) = \text{Quantile}_\alpha\{L_1(x), \ldots, L_M(x)\},$$

using NumPy's default linear quantile interpolation, and then

$$\widehat{\text{CVaR}}_\alpha(x) = \frac{\sum_{m=1}^{M} L_m(x)\, \mathbf{1}[L_m(x) \geq \widehat{q}_\alpha(x)]}{\sum_{m=1}^{M} \mathbf{1}[L_m(x) \geq \widehat{q}_\alpha(x)]}.$$

Observations equal to the quantile cutoff are included, so ties can make the empirical tail slightly larger than exactly $(1-\alpha)M$ observations. This definition matches the frozen confirmatory code.

Expected-cost regret and CVaR deviation are

$$R_J(x) = \frac{\hat{J}(x) - \hat{J}(x_J^*)}{\hat{J}(x_J^*)}, \qquad R_C(x) = \frac{\widehat{\text{CVaR}}_{0.95}(x) - \widehat{\text{CVaR}}_{0.95}(x_J^*)}{\widehat{\text{CVaR}}_{0.95}(x_J^*)}.$$

4.3 Frozen QUBO surrogate and calibration

For each item, define

$$r_i = \text{clip}\left(s_t \frac{\mu_i - b_i}{d_i}, 0, 1\right), \qquad a_i = \frac{(h_i + p_i)d_i}{\bar{d}}, \qquad \bar{d} = \frac{1}{n}\sum_i d_i,$$

and target weighted replenishment

$$B_t = \min\left(B, 0.85 \sum_i w_i\right).$$

The surrogate is

$$Q(x) = \sum_i a_i (x_i - r_i)^2 + \kappa \left(\sum_i w_i x_i - B_t\right)^2 + s_c \sum_i c_i x_i.$$

Using $x_i^2 = x_i$, this becomes

$$Q(x) = c_0 + \sum_i A_i x_i + \sum_{i<j} B_{ij} x_i x_j,$$

with

$$B_{ij} = 2\kappa w_i w_j, \qquad A_i = a_i(1 - 2r_i) + \kappa(w_i^2 - 2B_t w_i) + s_c c_i.$$

A separate development dataset contained 324 generated instances. Within each factorial cell, four instances were generated; the first two formed the calibration set and the remaining two the held-out set, yielding 162 instances in each split.

The calibration family contained 108 configurations:

$$s_t \in \{0.8, 1.0, 1.2\}, \qquad \kappa \in \{0.05, 0.15, 0.35, 0.70\},$$

$$\eta \in \{0, 0.04, 0.08\}, \qquad s_c \in \{0.25, 0.50, 1.00\}.$$

Here, η was a development-stage coefficient for an optional additional pairwise augmentation intended to test whether pairwise structure beyond that induced by the squared budget-target term improved downstream fidelity. The selected value was $\eta = 0$; hence this exploratory augmentation is absent from the frozen surrogate and affects no confirmatory result.

All 108 candidates were ranked using the 162 calibration instances only. The held-out set was not used for ranking, tie-breaking, or parameter adjustment. The primary selection criterion was the fraction of calibration instances for which the exact QUBO optimum satisfied

$$R_J \leq 0.05, \qquad R_C \leq 0.05,$$

with ties resolved by lower median expected-cost regret and then lower median CVaR deviation. The selected parameters were

$$\boxed{s_t = 1.2, \qquad \kappa = 0.15, \qquad \eta = 0, \qquad s_c = 0.50.}$$

After selection, these parameters were frozen and evaluated once on the held-out instances.

Table 1: Calibration and held-out validation of the selected QUBO surrogate under the primary 5%/5% operational criterion. Exact agreement denotes equality between the QUBO optimum and the stochastic expected-cost optimum.

Split	Operational acceptance	Exact agreement	Median cost regret	P90 cost regret	Median CVaR deviation
Calibration	90.12%	55.56%	0.00%	4.33%	0.00%
Held-out	85.19%	46.91%	0.39%	5.06%	0.00%

The held-out operational-acceptance rate was 4.94 percentage points below the calibration rate. Held-out evaluation assessed only the already selected surrogate and triggered no further tuning. The independent 81-instance confirmatory campaign was generated only after this freeze, and no QUBO parameter was retuned after observing held-out or confirmatory QAOA results.

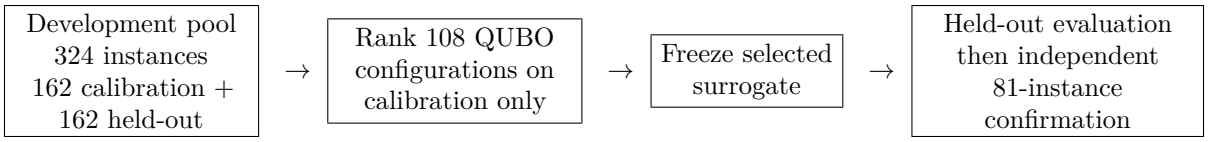


Figure 1: Calibration, freeze, and evaluation flow. Model selection used only the calibration split; held-out and confirmatory sets were used strictly after the surrogate was frozen.

4.4 Confirmatory design and thresholds

The final confirmatory dataset is independent of the calibration and held-out datasets. The factorial design varies

$$n \in \{4, 6, 8\}, \qquad cv \in \{0.10, 0.30, 0.50\},$$

$$p/h \in \{1, 3, 10\}, \qquad \tau \in \{0.40, 0.65, 0.90\},$$

giving $3^4 = 81$ confirmatory instances. The frozen random seed is 20260927.

The primary criterion uses

$$\epsilon_Q = \epsilon_J = \delta_C = 0.05.$$

Sensitivity analysis uses

$$\epsilon, \delta \in \{0.01, 0.02, 0.05, 0.10\}, \qquad \epsilon_Q = \epsilon_J = \epsilon,$$

producing 16 threshold combinations per solver distribution.

For $n \leq 8$, at most 256 binary states exist, so every state is enumerated to determine feasibility, x_Q^*, normalized QUBO gaps, x_J^*, the complete mathematical near-optimal and operational acceptance sets, and their prevalence under uniform feasible sampling.

4.5 Classical and uniform-feasible baselines

The greedy solver optimizes the frozen QUBO over feasible states. Starting from the all-zero state, it activates the feasible inactive bit producing the largest immediate decrease in $Q(x)$ and stops when no feasible single-bit activation improves the objective.

Simulated annealing [16] also optimizes the frozen QUBO over feasible states using a single-bit-flip neighborhood. Each run starts from a uniformly selected feasible state, with infeasible proposals rejected. For n variables,

$$N_{\text{steps}} = \max(300, 80n), \qquad T_0 = 0.25\Delta_Q, \qquad T_f = 0.0025\Delta_Q,$$

with geometric cooling. Forty independent runs are performed per instance.

Uniform feasible sampling assigns

$$P_U(x) = \frac{1}{|\mathcal{F}|}$$

for every feasible state. Because $\mathcal{F}$ is exactly known,

$$S_E^U = \frac{|\mathcal{X}_Q^*|}{|\mathcal{F}|}, \qquad S_N^U = \frac{|\mathcal{Q}_{\epsilon_Q}|}{|\mathcal{F}|}, \qquad S_O^U = \frac{|\mathcal{O}_{\epsilon_J,\delta_C}|}{|\mathcal{F}|}.$$

4.6 QAOA cost, circuit, and parameter search

QAOA uses the frozen QUBO as its mathematical optimization objective. Because the standard X-mixer does not preserve feasibility, infeasible states receive the additional diagonal penalty

$$v(x) = \max\left(0, \sum_i w_i x_i - B\right), \qquad \sigma_Q = \mathrm{sd}[Q(x) : x \in \mathcal{F}],$$

giving simulated energy

$$E(x) = Q(x) + 4\sigma_Q v(x)^2.$$

The energies are then linearly normalized to $[0, 1]$.

The mathematical surrogate and the Hamiltonian used in the statevector experiment must be distinguished. The surrogate $Q(x)$ is a genuine quadratic binary model and admits the standard 2-local Ising representation under $x_i = (1 - z_i)/2$ [2, 4]. By contrast,

$$4\sigma_Q \max\left(0, \sum_i w_i x_i - B\right)^2$$

is a general diagonal function over computational-basis states. Because of the hinge at zero, it is not generally a quadratic pseudo-Boolean polynomial and should not be identified with an additional 2-local Ising term. The simulated cost Hamiltonian is therefore

$$H_C = H_Q + H_{\mathrm{viol}},$$

where H_Q is the 2-local Ising image of the frozen QUBO and H_{viol} is diagonal with eigenvalues equal to the violation penalty. The statevector implementation applies the complete diagonal energy exactly, penalizing infeasible probability mass without mischaracterizing the full simulated objective as a strict QUBO.

The confirmatory experiments use

$$H_M = \sum_j X_j, \qquad p \in \{1, 2, 3\}, \qquad |+\rangle^{\otimes n}$$

as mixer, depth set, and initial state. For every instance and depth, exactly 100 parameter candidates are evaluated, with

$$\gamma_\ell \sim U(0, 2\pi), \qquad \beta_\ell \sim U(0, \pi/2).$$

The candidate minimizing expected normalized penalized energy is retained.

Because every depth receives the same number of candidate evaluations while parameter dimension grows as $2p$, cross-depth comparisons characterize performance under a fixed parameter-evaluation budget rather than equal parameter-optimization quality. This distinction matters because parameter-search difficulty is known to increase with depth and because practical QAOA performance can depend strongly on the design of the outer-loop optimization procedure [9, 10].

4.7 Sampling, enrichment, and statistical analysis

Primary QAOA probabilities are obtained exactly from the statevector. Finite-shot consequences are derived analytically as

$$N_{95}(S) = \left\lceil \frac{\log(0.05)}{\log(1 - S)} \right\rceil.$$

For every confirmatory instance and depth, operational enrichment is

$$E_O = \frac{S_O^{QAOA}}{S_O^U}.$$

Because enrichment ratios are skewed, aggregation uses the geometric mean; inference is performed on $\log E_O$ and transformed back to the original ratio scale. Paired raw differences $S_O^{QAOA} - S_O^U$ are also reported.

Problem instances are the independent experimental units. At the primary 5%/5% criterion, paired contrasts include

$$S_O - S_N, \qquad S_N - S_E, \qquad S_{O\setminus Q}, \qquad S_{Q\setminus O}.$$

Nonparametric paired bootstrap confidence intervals are obtained by resampling complete instances with replacement [17]; the primary set-comparison analysis uses 4000 bootstrap resamples and the enrichment analysis 5000.

4.8 Independent validation and evidence separation

The custom NumPy statevector implementation was independently checked against Qiskit and Qiskit Aer [23]. Synthetic validation covered $n = 4, 6, 8$ and depths up to $p = 3$.

A second validation used actual frozen confirmatory cases. To avoid outcome-dependent selection, the central factorial cell

$$cv = 0.30, \qquad p/h = 3, \qquad \tau = 0.65$$

was selected for each problem size, corresponding to confirmatory instances 14, 41, and 68; each was evaluated at $p = 1, 2, 3$. The validation environment was Python 3.9.19, Qiskit 2.2.3, and Qiskit Aer 0.17.2.

Exploratory development included initial optimization-error analysis, aligned-versus-surrogate comparisons, operational-equivalence exploration, preliminary classical experiments, initial $p = 1$ QAOA, finite-shot pilots, penalty sensitivity, a simplified post-measurement bit-flip stress test, initial QUBO construction, and QUBO calibration. These experiments informed the final protocol but are not pooled with the confirmatory results. The simplified bit-flip stress test did not support a defensible claim of noise robustness and is not used as evidence for such a claim.

4.9 Post-confirmatory physical-QPU extension

The physical-QPU study was added only after the simulation experiment and its independent implementation checks had been frozen. Its purpose is to test whether the same evaluation distinctions remain useful once the problem is encoded as an implementable circuit and sampled on real hardware. The hardware results are therefore reported as a separate post-confirmatory extension and do not alter the thresholds, success sets, or classical comparisons used in the 81-instance study.

The extension uses confirmatory instances 14 and 41, the $n = 4$ and $n = 6$ members of the central factorial cell $cv = 0.30$, $p/h = 3$, and $\tau = 0.65$. The semantic simulator can apply the positive-part feasibility penalty directly as a diagonal function, but that penalty is not generally a strict 2-local QUBO. For hardware, the budget inequality is therefore rewritten with binary slack variables:

$$\sum_i \widetilde{w}_i x_i + \sum_k 2^k s_k = \widetilde{B},$$

with quadratic penalty

$$\lambda \left(\sum_i \widetilde{w}_i x_i + \sum_k 2^k s_k - \widetilde{B} \right)^2 .$$

The integerization scale is the smallest positive integer a for which $\widetilde{w}_i = \mathrm{round}(a w_i)$ and $\widetilde{B} = \mathrm{round}(aB)$ preserve the feasible/infeasible classification of all 2^n decision strings. Writing

$$Q^{\mathrm{all}}_{\min} = \min_x Q(x), \qquad Q^{\mathcal{F}}_{\min} = \min_{x \in \mathcal{F}} Q(x), \qquad \Delta^{\mathcal{F}}_Q = \max_{x \in \mathcal{F}} Q(x) - \min_{x \in \mathcal{F}} Q(x),$$

the equality penalty is fixed as

$$\lambda = Q^{\mathcal{F}}_{\min} - Q^{\mathrm{all}}_{\min} + 0.10\, \Delta^{\mathcal{F}}_Q + 10^{-9}.$$

This makes any nonzero integer residual more expensive than the largest possible advantage of an infeasible decision-QUBO state over the best feasible state, with a 10% feasible-span margin. Instance 14 uses scale 2, weights $(2, 2, 2, 3)$, budget 5, three slack qubits, and $\lambda = 4.910824$; instance 41 uses scale 8, weights $(10, 6, 6, 9, 9, 7)$, budget 30, five slack qubits, and $\lambda = 4.778769$. The resulting encodings contain seven and 11 logical qubits. Exhaustive checks confirm energy reconstruction, a unique zero-residual slack representation for each feasible decision, and projection of the encoded ground state to the original frozen QUBO optimum.

To avoid compressing the decision-relevant phase differences by the much larger penalty range, the encoded phase energy is shifted by the global strict-QUBO minimum and scaled by the feasible decision-QUBO span,

$$E_{\text{phase}}(z) = \frac{E_{\text{strict}}(z) - E_{\min}^{\text{strict}}}{\Delta_Q^{\mathcal{F}}}.$$

The corresponding phase scales are 8.590866 for instance 14 and 12.437754 for instance 41. QAOA parameters are then selected before any QPU outcome is observed, using the same 100 uniformly sampled parameter vectors per depth as in the simulation study. These frozen strict-encoding distributions provide the ideal hardware comparators.

Circuit compilation is likewise selected without reference to measured outcomes. Fez, Kingston, and Marrakesh are each transpiled at optimization level 3 with SABRE layout and routing and deterministic seeds $20261027, \ldots, 20261031$. For each case, the preferred seed minimizes two-qubit operations, then ISA depth, circuit size, and seed value. Because the three backends produced identical routing totals, the final backend choice uses a descriptive calibration score on the actual mapped circuit,

$$C_{\text{cal}} = \prod_{g \in \text{CZ}} (1 - e_g) \prod_{q \in \mathcal{M}} (1 - r_q),$$

where e_g is the reported error for an executed CZ occurrence and r_q is the readout error for measured qubit q. Higher C_{cal} ranks first, followed by lower mean CZ error, shorter queue, and backend name. The score is only a hardware-selection heuristic, not an estimate of circuit fidelity. Exact transpiled QPY artifacts are hashed and reloaded for execution without retranspilation.

For instance 14, Kingston ranked best for all three depths. The reviewed circuits contain 69, 151, and 241 CZ operations at ISA depths 157, 316, and 473, and each is sampled 5000 times without dynamical decoupling or gate/measurement twirling. A preceding 1000-shot $p = 1$ infrastructure pilot is retained only as a repeatability check. For instance 41, Kingston ranked best at $p = 1$; the 209-CZ, depth-374 circuit is sampled 20000 times. At $p = 2$, the reviewed circuit contains 488 CZ operations at depth 601. Fez ranked best in the contemporaneous calibration screen, so the same frozen $p = 2$ experiment is run with 20000 shots on both Fez and Kingston, preserving a within-Kingston depth comparison and adding a calibration-favored cross-backend comparison.

Measured encoded strings are projected back to the original decision variables before evaluation. Hardware results report decision-level exact recovery $S_{E,\text{dec}}$, QUBO-near probability S_N, operational probability S_O, N_{95}, and total-variation distance from the strict-encoding ideal decision distribution. Thus $S_{E,\text{dec}}$ means recovery of the optimal decision bits, not necessarily recovery of the complete decision-plus-slack encoded ground state.

5 Results

5.1 Surrogate fidelity and primary QAOA success

Under the primary operational criterion $R_J \leq 0.05$ and $R_C \leq 0.05$, the exact QUBO optimum belongs to the operational acceptance set in $70/81 = 86.4\%$ of confirmatory instances.

Thus,

Table 2: Operational acceptability of the exact QUBO optimum under the primary 5%/5% criterion.

n	QUBO optimum operationally acceptable
4	81.5%
6	88.9%
8	88.9%

$$x = x_Q^* \not\Rightarrow x \in \mathcal{O}_{0.05,0.05}.$$

Table 3 reports median QAOA probability mass under the exact, QUBO-near, and operational success definitions.

Table 3: Median QAOA probability mass at the primary 5%/5% criterion.

Depth	n	S_E	S_N	S_O
1	4	0.0621	0.0700	0.0750
1	6	0.0154	0.0307	0.0500
1	8	0.0039	0.0154	0.0234
2	4	0.1000	0.1081	0.1748
2	6	0.0297	0.0592	0.1114
2	8	0.0076	0.0353	0.0927
3	4	0.0803	0.0937	0.1291
3	6	0.0289	0.0670	0.1003
3	8	0.0078	0.0366	0.0864

Across all 81 instances, the median (S_E, S_N, S_O) triplets are

$$\begin{aligned} p = 1 &: (0.0133, 0.0278, 0.0535), \\ p = 2 &: (0.0408, 0.0723, 0.1397), \\ p = 3 &: (0.0362, 0.0711, 0.1087). \end{aligned}$$

At $n = 8, p = 2$, the corresponding values are

$$S_E = 0.0076, \qquad S_N = 0.0353, \qquad S_O = 0.0927.$$

Thus, only 0.76% of the statevector mass lies on the exact surrogate optimum, whereas 9.27% lies on decisions satisfying the downstream cost-and-risk criterion.

5.2 Near-optimality, operational acceptance, and set disagreement

Operational acceptance is not equivalent to QUBO near-optimality. The mean paired differences are:

All three confidence intervals exclude zero. The gain from relaxing exact recovery to mathematical near-optimality is $0.0182\,[0.0132, 0.0238]$ for $p = 1$, $0.0480\,[0.0361, 0.0607]$ for $p = 2$, and

Table 4: Paired bootstrap contrasts at the primary criterion.

Depth	Mean $S_O - S_N$	95% bootstrap CI
$p = 1$	0.0851	[0.0462, 0.1297]
$p = 2$	0.0770	[0.0406, 0.1186]
$p = 3$	0.0837	[0.0462, 0.1246]

$0.0483\,[0.0356, 0.0618]$ for $p = 3$. The distributions are heterogeneous: median $S_O - S_N$ is approximately 0.0095, 0.0079, and 0.0101, respectively.

The set decomposition confirms that the two acceptance sets are non-nested. Mean operational-only probability mass is $0.0929\,[0.0548, 0.1369]$ for $p = 1$, $0.0989\,[0.0657, 0.1373]$ for $p = 2$, and $0.1045\,[0.0693, 0.1451]$ for $p = 3$. Mean QUBO-near-only mass is $0.0077\,[0.0040, 0.0124]$, $0.0219\,[0.0127, 0.0322]$, and $0.0208\,[0.0120, 0.0306]$, respectively. Hence both

$$\mathcal{Q}_{0.05} \setminus \mathcal{O}_{0.05,0.05} \qquad \text{and} \qquad \mathcal{O}_{0.05,0.05} \setminus \mathcal{Q}_{0.05}$$

receive nonzero probability mass.

5.3 Sampling consequences, depth, and threshold sensitivity

For $n = 8$, median N_{95} values are:

Table 5: Median number of measurements for 95% probability of observing at least one successful $n = 8$ state.

Depth	N_{95}^E	N_{95}^N	N_{95}^O
$p = 1$	766	194	127
$p = 2$	392	84	31
$p = 3$	384	81	34

At $p = 2$, the sequence $392 \to 84 \to 31$ implies a ratio $392/31 \approx 12.6$. This is not a quantum speedup; it is the sampling consequence of applying different success definitions to the same underlying output distribution.

Figure 2 shows that the operational criterion substantially lowers the sampling burden relative to exact surrogate recovery, especially at $p = 2$ and $p = 3$.

Depth does not produce monotonic improvement under the fixed parameter-search budget. For $n = 8$,

$$S_O(p = 1) = 0.0234, \qquad S_O(p = 2) = 0.0927, \qquad S_O(p = 3) = 0.0864,$$

so increasing from $p = 2$ to $p = 3$ reduces the operational metric.

Threshold sensitivity shows the same dependence on the success definition:

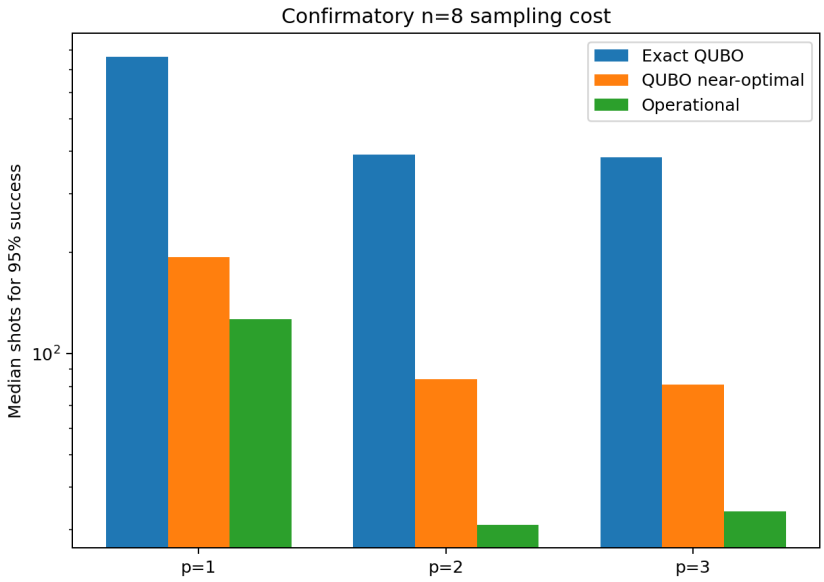


Figure 2: Median N_{95} for the confirmatory $n = 8$ instances under exact, QUBO-near, and operational success. The differences arise from the success definition applied to the same QAOA output distribution.

Table 6: Threshold sensitivity for $p = 2$.

Tolerance	S_E	S_N	S_O	N_{95}^E	N_{95}^N	N_{95}^O
1%	0.0408	0.0488	0.0523	72	60	56
2%	0.0408	0.0517	0.0699	72	57	42
5%	0.0408	0.0723	0.1397	72	40	20
10%	0.0408	0.1081	0.2256	72	27	12

At $p = 3$ under the strict 1%/1% criterion, $S_N = 0.0522$ and $S_O = 0.0508$. The experiment therefore does not support a universal ordering $S_E < S_N < S_O$.

5.4 Classical and uniform-feasible baselines

To avoid comparing QAOA medians with classical means, Table 7 reports both statistics for every solver over the same 81 confirmatory instances at the primary 5%/5% criterion.

The like-for-like comparison confirms that both classical optimization baselines substantially outperform the tested QAOA configurations on these small instances under both mean and median aggregation. The uniform-feasible row also shows why operational probability requires a prevalence control: its mean and median S_O both exceed the corresponding QUBO-near values despite the absence of optimization.

Operational evaluation also does not mechanically improve apparent solver quality. For greedy search, the mean values satisfy $S_N = 0.9877 > S_O = 0.8519$, while the medians of both quantities equal 1 because the deterministic greedy output is successful under both criteria on at least half of the instances.

Table 7: Unified solver comparison at the primary criterion. Each QAOA row summarizes the full statevector success probabilities across 81 instances; simulated annealing uses its empirical 40-run distribution per instance; greedy is deterministic; and uniform feasible probabilities are computed exactly.

	Exact S_E		QUBO-near S_N		Operational S_O	
Solver	Mean	Median	Mean	Median	Mean	Median
QAOA $p = 1$	0.0231	0.0133	0.0412	0.0278	0.1263	0.0535
QAOA $p = 2$	0.0541	0.0408	0.1021	0.0723	0.1791	0.1397
QAOA $p = 3$	0.0565	0.0362	0.1048	0.0711	0.1885	0.1087
Greedy	0.9506	1.0000	0.9877	1.0000	0.8519	1.0000
Simulated annealing	0.6877	0.7000	0.7929	0.8500	0.7744	0.9500
Uniform feasible	0.0513	0.0213	0.0786	0.0455	0.2154	0.0909

5.5 Operational enrichment

Geometric-mean operational enrichment is:

Table 8: Geometric-mean operational enrichment E_O with 95% bootstrap intervals.

Depth	$n = 4$	$n = 6$	$n = 8$
$p = 1$	0.40 [0.25, 0.61]	0.57 [0.45, 0.72]	0.67 [0.54, 0.83]
$p = 2$	0.57 [0.34, 0.88]	1.12 [0.73, 1.66]	1.71 [1.10, 2.78]
$p = 3$	0.52 [0.33, 0.79]	1.17 [0.79, 1.75]	1.85 [1.18, 3.07]

Across all 81 instances, $E_O(p = 2) = 1.026\,[0.774, 1.362]$ and $E_O(p = 3) = 1.042\,[0.793, 1.367]$, so aggregate enrichment is not statistically established across the complete factorial design. Paired raw differences are likewise inconclusive in the $n = 8$ enriched regimes: $0.0074\,[-0.0489, 0.0567]$ for $p = 2$ and $0.0193\,[-0.0319, 0.0723]$ for $p = 3$. The ratio and raw-difference analyses therefore answer different questions and should be interpreted jointly.

Figure 3 shows that the enrichment crossover occurs only in selected regimes: $p = 2$ and $p = 3$ exceed unity for the larger problem sizes, whereas $p = 1$ remains below 1 throughout.

5.6 Independent implementation validation

The custom NumPy statevector simulator was independently validated against Qiskit and Qiskit Aer [23]. Across the nine actual frozen confirmatory validation cases, the maximum discrepancies were 1.67×10^{-16} for the full probability distribution, 2.22×10^{-16} for Qiskit/Aer success metrics, 1.25×10^{-16} relative to the frozen confirmatory outputs, and 2.66×10^{-15} for normalization. All are several orders of magnitude below the predefined tolerance 10^{-10}.

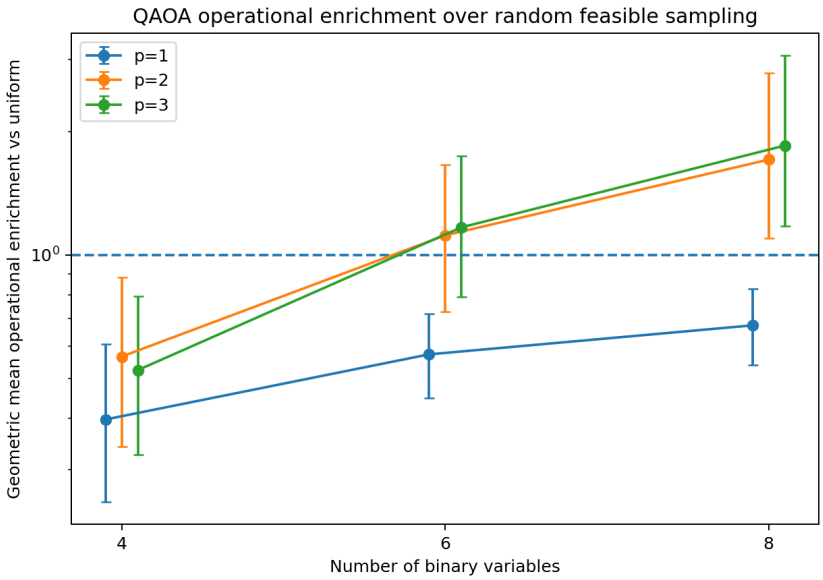


Figure 3: Geometric-mean operational enrichment relative to uniform feasible sampling. Values above 1 indicate concentration of operationally acceptable solutions beyond baseline prevalence; error bars show 95% bootstrap confidence intervals.

5.7 Physical-QPU validation: depth, scale, and backend/calibration-condition sensitivity

The hardware experiments show most clearly why task-level success and distributional fidelity should be reported separately. Table 9 gives the Kingston depth series for instance 14. For this instance only, the QUBO-near and operational acceptance sets coincide, so $S_N = S_O$.

Table 9: Strict-encoding ideal and IBM Kingston hardware results for frozen instance 14. Hardware values use 5000 shots per depth. TV_{dec} is total-variation distance between measured and ideal decision-projected distributions.

p	CZ	ISA depth	$S_{E,\mathrm{dec}}^{\mathrm{ideal}}$	$S_{E,\mathrm{dec}}^{\mathrm{hw}}$	$S_N^{\mathrm{ideal}} = S_O^{\mathrm{ideal}}$	$S_N^{\mathrm{hw}} = S_O^{\mathrm{hw}}$	TV_{dec}	$N_{95,O}^{\mathrm{ideal}}$	$N_{95,O}^{\mathrm{hw}}$
1	69	157	0.0099	0.1118	0.0202	0.1438	0.5087	147	20
2	151	316	0.0589	0.0532	0.1040	0.1048	0.1347	28	28
3	241	473	0.0392	0.0468	0.0736	0.0992	0.1945	40	29

At $p = 2$, the hardware operational-success probability is 0.1048, almost identical to the ideal value 0.10396; its Wilson 95% interval, $[0.0966, 0.1136]$, contains the ideal value. Yet the measured decision distribution is not the ideal distribution: $TV_{\mathrm{dec}} = 0.1347$. The practical success rate is therefore preserved much more closely than the full distribution.

The other depths show why this agreement cannot be taken for granted. At $p = 3$, hardware gives $S_O = 0.0992$ versus 0.07363 ideally, with $TV_{\mathrm{dec}} = 0.1945$. At $p = 1$, the difference is much larger: S_O rises from 0.02024 to 0.1438 while $TV_{\mathrm{dec}} = 0.5087$. A separate 1000-shot pilot is directionally consistent with this pattern ($S_{E,\mathrm{dec}} = 0.1280$, $S_N = S_O = 0.1620$, $TV_{\mathrm{dec}} = 0.5285$). The high $p = 1$ operational probability is therefore not evidence of better optimization; it is consistent with

hardware imperfections redistributing probability mass toward states that happen to satisfy the operational criterion. The corresponding ideal/hardware N_{95} values are 147/20, 28/28, and 40/29 for $p = 1, 2, 3$.

Instance 41 provides a different test because its QUBO-near and operational sets do not coincide. Table 10 reports the $n = 6$ extension.

Table 10: Physical-QPU results for frozen instance 41 ($n = 6$, 11 encoded qubits). Kingston provides a within-backend depth comparison; Fez provides a calibration-favored cross-backend comparison at $p = 2$. All rows use 20000 shots and no dynamical decoupling or twirling.

Backend	p	CZ	ISA depth	$S^{\mathrm{hw}}_{E,\mathrm{dec}}$	S^{hw}_{N}	S^{hw}_{O}	S^{ideal}_{O}	TV_{dec}	$N^{\mathrm{hw}}_{95,O}$	$N^{\mathrm{ideal}}_{95,O}$
Kingston	1	209	374	0.00650	0.02475	0.01630	0.00325	0.28633	183	920
Kingston	2	488	601	0.00150	0.01810	0.01710	0.00651	0.53261	174	459
Fez	2	488	601	0.00590	0.02655	0.01850	0.00651	0.44749	161	459

The $p = 1$ Kingston run gives $S_N = 0.02475$ and $S_O = 0.01630$, providing direct physical evidence that surrogate near-optimality and operational acceptability are not the same event. Since

$$S_N - S_O = S_{Q\setminus O} - S_{O\setminus Q} = 0.00845,$$

the QUBO-near-only mass must be at least 0.00845. Thus at least 34.1% of the near-optimal probability mass is operationally unacceptable. At the same time, S_O is about five times the strict-encoding ideal value and $TV_{\mathrm{dec}} = 0.28633$, again showing that a higher task-level success probability does not imply closer reproduction of the ideal optimizer distribution.

The two Kingston depths also behave differently. The observed TV_{dec} rises from 0.28633 at $p = 1$ to 0.53261 at $p = 2$, while S_O changes only from 0.01630 to 0.01710. Their shot-level Wilson intervals, $[0.01464, 0.01815]$ and $[0.01539, 0.01899]$, overlap substantially. These intervals describe finite-shot uncertainty within each execution; they are not independent replications of a processor state or calibration condition.

Finally, the Fez and Kingston $p = 2$ runs show that backend/calibration conditions can alter the measured distribution more than the downstream success rate. Fez has lower TV_{dec} (0.44749 versus 0.53261), higher decision-level exact recovery (0.00590 versus 0.00150), and higher S_N (0.02655 versus 0.01810). In contrast, S_O is similar: 0.01850 on Fez and 0.01710 on Kingston, with overlapping shot-level Wilson intervals. Because backend, mapping, calibration state, and execution time vary together, this is a descriptive sensitivity comparison rather than an isolated causal processor effect.

Taken together, the hardware results reproduce the paper's central distinction in several forms. Exact decision recovery, surrogate near-optimality, operational success, and distance from the ideal distribution do not move together. Which metric is reported can therefore change the apparent conclusion of a hardware benchmark, even when every metric is computed from the same set of measured bit strings.

6 Discussion and Limitations

6.1 Interpretation of mathematical and operational success

The central result is not that approximate solutions are preferable to exact solutions, nor that quantum optimization outperforms classical optimization. The result is narrower: when the mathematical optimization model is an intermediate surrogate for a downstream stochastic decision problem, different notions of solver success can classify the same output distribution differently.

The confirmatory results show that

$$\mathcal{Q}_{\epsilon_Q} \neq \mathcal{O}_{\epsilon_J,\delta_C},$$

with probability mass occurring in both

$$\mathcal{Q}_{\epsilon_Q} \setminus \mathcal{O}_{\epsilon_J,\delta_C} \qquad \text{and} \qquad \mathcal{O}_{\epsilon_J,\delta_C} \setminus \mathcal{Q}_{\epsilon_Q}.$$

The distinction is therefore not reducible to replacing exact recovery with a looser approximation-ratio criterion. The QUBO criterion asks whether a candidate remains sufficiently close to the mathematical surrogate optimum, whereas the operational criterion asks whether the same decision remains within specified tolerances after stochastic evaluation. A solution in $\mathcal{Q} \setminus \mathcal{O}$ is mathematically close but operationally unacceptable; a solution in $\mathcal{O} \setminus \mathcal{Q}$ falls outside the selected mathematical near-optimality region while remaining acceptable in deployment. This non-nested structure is the key empirical distinction.

The $n = 8, p = 2$ result $392 \to 84 \to 31$ similarly does not imply that the circuit becomes faster when operational evaluation is used: the circuit and output distribution remain unchanged. The values instead answer three different questions—how many measurements are required to observe the exact surrogate optimum, a near-optimal surrogate state, or any downstream-acceptable decision.

6.2 Operational success, baselines, and quantum performance

Operational probability should not be interpreted independently of baseline prevalence. Uniform feasible sampling shows that S_O alone can exaggerate optimizer quality when acceptable states are abundant. The enrichment statistic

$$E_O = \frac{P_Q(\mathcal{O})}{P_U(\mathcal{O})}$$

therefore complements absolute operational success. At $p = 1$, QAOA is under-enriched relative to uniform feasible sampling for every tested size. At $n = 8$, $p = 2$ and $p = 3$ show multiplicative enrichment, although paired absolute probability differences remain uncertain.

The classical baselines also show that downstream evaluation is not constructed to mechanically improve QAOA scores. For greedy search, $S_N > S_O$, while at the strict 1%/1% criterion with $p = 3$, $S_O < S_N$. Thus, the relationship among exact, near-optimal, and operational success depends on the solver distribution and acceptance criterion.

The study provides no evidence of quantum advantage. Greedy search and simulated annealing outperform the tested QAOA configurations, as expected for problems containing at most eight binary variables, for which complete enumeration is trivial. The small problem dimension is a methodological choice that makes the complete surrogate and operational landscapes observable rather than an attempt to establish computational superiority.

6.3 Depth, tolerances, and surrogate alignment

Every QAOA depth receives exactly 100 random parameter candidates. Since the parameter-space dimension is $2p$, increasing depth while holding this budget fixed makes the search increasingly sparse. This should be interpreted in light of the broader QAOA literature, where higher-depth performance depends materially on parameter-optimization strategy, initialization, and shot budget [9, 10]. The experiment therefore compares depths under equal parameter-evaluation budgets, not equal parameter-optimization quality.

The 5%/5% operational criterion is a controlled experimental convention rather than a universal standard. Sensitivity analysis shows that the mathematical-versus-operational distinction becomes small under very strict thresholds and larger under broader tolerances. In practical applications, such tolerances should be determined from domain requirements rather than selected to maximize measured solver performance.

The QUBO optimum satisfies the primary operational criterion in 86.4% of confirmatory instances, indicating strong but imperfect surrogate alignment. A perfectly aligned surrogate could make the distinction between S_N and S_O nearly redundant, whereas a very poor surrogate would make solver evaluation uninterpretable. The calibrated benchmark intentionally occupies an intermediate regime.

Expected cost defines the operational reference, while CVaR serves as a secondary risk guardrail rather than an independently optimized target. CVaR itself is not novel [12, 13]; other domains could replace it with service-level, reliability, fairness, emissions, latency, or other deployment-level criteria.

6.4 Statistical and external-validity limitations

Expected cost and CVaR are estimated from 1800 common demand realizations per instance, so the operational acceptance sets are empirical approximations. States close to an acceptance boundary could change classification under a substantially larger simulation budget, and future work could quantify uncertainty in operational-set membership explicitly.

The inventory environment is synthetic. Its parameter ranges and Gaussian demand model provide controlled variation but are not a validated representation of a specific supply chain. Numerical values of S_O, E_O, and N_{95} should therefore not be generalized directly to industrial systems; the transferable object is the evaluation framework.

At $n \leq 8$, complete enumeration remains possible and classical heuristics are highly effective. Larger systems would require estimation of $\mathcal{O}$, $\mathcal{Q}$, and uniform-set prevalence rather than exact enumeration.

6.5 Simulation, validation, and hardware limitations

The hardware evidence is intentionally narrow. The 81-instance confirmatory analysis remains an ideal-statevector study, while the physical-QPU extension covers two frozen instances, two processors, and a small set of depths. It shows that the evaluation metrics can separate under physical execution, but it does not estimate average behavior across problem instances, devices, or calibration periods.

The hardware representation is also more expensive than the semantic simulation model. The simulator can apply the positive-part feasibility penalty directly; the QPU implementation replaces it with finite-precision integer weights and binary slack variables so that the Hamiltonian is a strict 2-local QUBO/Ising model. This raises the selected $n = 4$ case from four decision bits to seven logical qubits and the $n = 6$ case to 11 logical qubits. Routing adds a second cost: the executed $n = 4$ circuits require 69, 151, and 241 CZ operations, while the executed $n = 6$ circuits require 209 and 488 CZ operations at $p = 1$ and $p = 2$. The heavier preflighted $n = 6, p = 3$ and $n = 8$ circuits were not executed, so the present results should not be read as a scaling claim.

The reported hardware distributions are deliberately unmitigated. Dynamical decoupling and gate or measurement twirling are disabled, which keeps the comparison close to the frozen strict-encoding ideal distribution but means that the measured probabilities combine algorithmic structure with gate errors, stochastic noise, readout error, and calibration-specific behavior. This is visible in the data: the $n = 4, p = 2$ circuit is closer to ideal than $p = 1$ despite using more than twice as many CZ operations, and Fez is closer than Kingston for $n = 6, p = 2$ while their operational-success probabilities remain similar. Two-qubit gate count and the calibration heuristic are therefore useful engineering diagnostics, not sufficient predictors of task-level output quality. Replication at later calibration epochs and on additional instances is needed before attributing these patterns to a stable hardware mechanism.

Finally, N_{95} is a sampling-count measure, not a wall-clock time-to-solution. Queueing, compilation, classical parameter search, calibration drift, and repeated execution are separate resources and are not included in N_{95}.

6.6 Scope of the evidence and generalization

Several exploratory formulations were rejected or modified during development, including an aligned-objective formulation that reduced to a tautological relationship, an initial near-optimal-set framing that overlapped with established MGA literature [15], and a simplified noise model that did not support the proposed noise claim. The final inferential claims rely on the frozen 81-instance confirmatory experiment. The study was not formally preregistered, so replication on independently designed benchmarks remains important.

The distinction developed here is not inherently quantum. Any approximate solver producing repeated candidate decisions can be evaluated through exact recovery, surrogate near-optimality, and downstream operational success. QAOA provides a particularly clear probabilistic case because its full output distribution and finite-shot consequences can be analyzed explicitly.

For application-oriented approximate-optimization benchmarks, it may therefore be useful to report the four quantities

$$S_E, \qquad S_N, \qquad S_O, \qquad E_O,$$

which respectively characterize exact surrogate recovery, mathematical approximation, downstream acceptability, and concentration relative to baseline prevalence. Operational evaluation should complement, rather than replace, conventional optimization metrics whenever the mathematical model is an intermediate representation rather than the ultimate decision objective.

Future work should examine a second combinatorial decision family, larger dimensions, sampled rather than exhaustive operational-set characterization, replication across independent hardware calibration epochs, and mitigation-aware physical-QPU experiments. The present $n = 6$ extension shows that the separation between distributional fidelity and operational success can persist beyond the seven-qubit encoding, but additional independently selected instances are required before making a scaling claim. The framework could also be extended from binary acceptable/unacceptable sets to utility-weighted downstream criteria.

7 Conclusion

This paper asks a simple evaluation question: when an optimizer solves a surrogate model, does mathematical success mean that the resulting decision is also good in the environment for which the model was built? The answer in this benchmark is no. Exact recovery S_E, surrogate near-optimality S_N, and operational success S_O describe related but distinct events.

The confirmatory simulation makes the distinction explicit. The QUBO-near set and the downstream operational set are non-nested, so probability mass can lie in either $\mathcal{Q} \setminus \mathcal{O}$ or $\mathcal{O} \setminus \mathcal{Q}$. This changes the finite-shot interpretation of the same QAOA distribution: for $n = 8, p = 2$, the median N_{95} is 392 for exact recovery, 84 for QUBO near-optimality, and 31 for operational acceptability. These are different success definitions, not quantum speedups. Uniform feasible sampling further shows that a high operational probability must be interpreted relative to the prevalence of acceptable states, while greedy search and simulated annealing establish that the tested QAOA configurations do not provide quantum advantage on these small instances.

The physical-QPU extension reaches the same methodological conclusion from a different direction. For $n = 4$, the operational-success probability can remain close to its ideal value even when the full decision distribution is measurably distorted. For $n = 6$, hardware produces $S_N > S_O$, so a nontrivial share of near-optimal surrogate outcomes is operationally unacceptable. At $n = 6, p = 2$, Fez and Kingston yield different decision distributions and different exact/near-optimal probabilities while producing similar operational-success probabilities. These observations show that task-level success, surrogate quality, decision-level exact recovery, and distributional fidelity need not move together on real hardware.

For application-oriented approximate optimization, no single metric is therefore sufficient. Exact recovery describes fidelity to the surrogate optimum; near-optimality describes mathematical approximation quality; operational success describes downstream acceptability; and enrichment indicates whether that acceptability exceeds its baseline prevalence. When physical hardware is involved, distributional fidelity should be reported alongside them. The framework is not specific to QAOA and can be applied whenever an approximate solver sits between a surrogate optimization model and a downstream decision environment. Future work should test the same distinctions on

independently selected problem families, larger instances, later calibration epochs, and mitigation-aware hardware executions.

Code and Data Availability

The archived reproducibility package contains the confirmatory simulation code, synchronized classical and random baselines, enrichment analysis, frozen QAOA validation cases, independent Qiskit/Aer validation scripts and outputs, confirmatory result tables, bootstrap results, figures, exploratory-development artifacts retained for provenance, a manifest, and SHA-256 checksums.

The physical-QPU extension uses a separate hardware-execution package containing the frozen strict-QUBO cases, encoding validations, Qiskit circuit checks, routing and calibration-selection scripts, reviewed QPY artifacts, submission/retrieval utilities, and hardware-analysis code. The hardware package and raw physical-QPU outputs will be incorporated into the next archived revision of the reproducibility record.

The code is distributed under the MIT License. The currently archived simulation package is available on Zenodo [24] at `doi:10.5281/zenodo.22817579`.

Declaration of Generative AI Use

Generative artificial intelligence tools, including ChatGPT (OpenAI), were used to assist with the manuscript drafting, scientific and linguistic editing, LaTeX preparation, and the development, debugging, and documentation of computational code. The author directed the development and refinement of the research methodology, made the final methodological and scientific decisions, and assumes full responsibility for the accuracy, integrity, reproducibility, and conclusions of the work.

References

[1] E. Farhi, J. Goldstone, and S. Gutmann, "A Quantum Approximate Optimization Algorithm," arXiv:1411.4028, 2014. `doi:10.48550/arXiv.1411.4028`.

[2] A. Lucas, "Ising formulations of many NP problems," *Frontiers in Physics*, vol. 2, art. 5, 2014. `doi:10.3389/fphy.2014.00005`.

[3] G. Kochenberger, J.-K. Hao, F. Glover, M. Lewis, Z. Lü, H. Wang, and Y. Wang, "The unconstrained binary quadratic programming problem: a survey," *Journal of Combinatorial Optimization*, vol. 28, no. 1, pp. 58–81, 2014. `doi:10.1007/s10878-014-9734-0`.

[4] F. Glover, G. Kochenberger, and Y. Du, "A Tutorial on Formulating and Using QUBO Models," arXiv:1811.11538, 2019. `doi:10.48550/arXiv.1811.11538`.

[5] S. Hadfield, Z. Wang, B. O'Gorman, E. G. Rieffel, D. Venturelli, and R. Biswas, "From the Quantum Approximate Optimization Algorithm to a Quantum Alternating Operator Ansatz," *Algorithms*, vol. 12, no. 2, art. 34, 2019. `doi:10.3390/a12020034`.

[6] M. Willsch, D. Willsch, F. Jin, H. De Raedt, and K. Michielsen, "Benchmarking the quantum approximate optimization algorithm," *Quantum Information Processing*, vol. 19, art. 197, 2020. doi:10.1007/s11128-020-02692-8.

[7] J. Larkin, M. Jonsson, D. Justice, and G. G. Guerreschi, "Evaluation of QAOA based on the approximation ratio of individual samples," arXiv:2006.04831, 2020. doi:10.48550/arXiv.2006.04831.

[8] A. Yu. Chernyavskiy, D. A. Kulikov, B. I. Bantysh, Yu. I. Bogdanov, A. K. Fedorov, and E. O. Kiktenko, "Improving QAOA to find approximate QUBO solutions in $O(1)$ shots," arXiv:2509.19035, 2025. doi:10.48550/arXiv.2509.19035.

[9] L. Zhou, S.-T. Wang, S. Choi, H. Pichler, and M. D. Lukin, "Quantum Approximate Optimization Algorithm: Performance, Mechanism, and Implementation on Near-Term Devices," *Physical Review X*, vol. 10, no. 2, art. 021067, 2020. doi:10.1103/PhysRevX.10.021067.

[10] T. Hao, Z. He, R. Shaydulin, J. Larson, and M. Pistoia, "End-to-end protocol for high-quality quantum approximate optimization algorithm parameters with few shots," *Physical Review Research*, vol. 7, art. 033179, 2025. doi:10.1103/24gg-7p8z.

[11] T. Koch, D. E. Bernal Neira, Y. Chen, G. Cortiana, D. J. Egger, R. Heese, N. N. Hegade, et al., "The Quantum Optimization Benchmarking Library," *Nature Computational Science*, vol. 6, pp. 653–671, 2026. doi:10.1038/s43588-026-00991-1.

[12] R. T. Rockafellar and S. Uryasev, "Optimization of Conditional Value-at-Risk," *The Journal of Risk*, vol. 2, no. 3, pp. 21–41, 2000. doi:10.21314/JOR.2000.038.

[13] P. K. Barkoutsos, G. Nannicini, A. Robert, I. Tavernelli, and S. Woerner, "Improving Variational Quantum Optimization using CVaR," *Quantum*, vol. 4, art. 256, 2020. doi:10.22331/q-2020-04-20-256.

[14] B. Wilder, B. Dilkina, and M. Tambe, "Melding the Data-Decisions Pipeline: Decision-Focused Learning for Combinatorial Optimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 1658–1665, 2019. doi:10.1609/aaai.v33i01.33011658.

[15] E. D. Brill, Jr., S.-Y. Chang, and L. D. Hopkins, "Modeling to Generate Alternatives: The HSJ Approach and an Illustration Using a Problem in Land Use Planning," *Management Science*, vol. 28, no. 3, pp. 221–235, 1982. doi:10.1287/mnsc.28.3.221.

[16] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983. doi:10.1126/science.220.4598.671.

[17] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979. doi:10.1214/aos/1176344552.

[18] P. Glasserman and D. D. Yao, "Some Guidelines and Guarantees for Common Random Numbers," *Management Science*, vol. 38, no. 6, pp. 884–908, 1992. doi:10.1287/mnsc.38.6.884.

[19] R. D. Wright and T. E. Ramsay, Jr., "On the Effectiveness of Common Random Numbers," *Management Science*, vol. 25, no. 7, pp. 649–656, 1979. doi:10.1287/mnsc.25.7.649.

[20] K. J. Arrow, T. Harris, and J. Marschak, "Optimal Inventory Policy," *Econometrica*, vol. 19, no. 3, pp. 250–272, 1951. doi:10.2307/1906813.

[21] G. Hadley and T. M. Whitin, *Analysis of Inventory Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1963.

[22] G. P. L. M. Fernandes, M. S. Fonseca, A. G. Valério, A. C. Ricardo, N. A. C. Carpio, P. C. C. Bezerra, and C. J. Villas-Boas, "Optimization Algorithm for Inventory Management on Classical, Quantum and Quantum-Hybrid Hardware," arXiv:2411.11756, 2024. doi:10.48550/arXiv.2411.11756.

[23] A. Javadi-Abhari, M. Treinish, K. Krsulich, C. J. Wood, J. Lishman, J. Gacon, S. Martiel, P. D. Nation, L. S. Bishop, A. W. Cross, B. R. Johnson, and J. M. Gambetta, "Quantum computing with Qiskit," arXiv:2405.08810, 2024. doi:10.48550/arXiv.2405.08810.

[24] R. Garrone, *Beyond Exact Recovery: Code and Reproducibility Package for Operational Evaluation of Approximate Quantum Optimization*, Zenodo, 2026, computer software. doi:10.5281/zenodo.22817579.