

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: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11499298/v1>

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 optimization algorithms are commonly evaluated through exact optimum recovery or objective-space approximation, although these criteria need not coincide with the quality of the resulting decision in its downstream operating environment. This paper develops an end-to-end evaluation framework that separates three notions of solver success: exact recovery of the mathematical optimum, near-optimality with respect to the optimization surrogate, and operational acceptability under stochastic downstream evaluation. We study the framework using a small binary inventory-control benchmark represented by a calibrated and held-out-validated QUBO surrogate and solved with exact enumeration, greedy search, simulated annealing, uniform feasible sampling, and ideal QAOA at depths $p = 1, 2, 3$.

The results show that mathematical near-optimality and downstream operational acceptability define distinct solution sets. For QAOA, the probability mass assigned to operationally acceptable solutions can differ materially from both exact-recovery and QUBO-near-optimal probability, producing substantial differences in the number of measurements required to observe an acceptable solution. For $n = 8$ at $p = 2$, the median number of shots required for 95% observation probability decreases from 392 under exact recovery to 84 under QUBO near-optimality and 31 under the operational criterion. However, operational success alone can be misleading when acceptable states are abundant. We therefore use a descriptive enrichment comparison against uniform feasible sampling. Statistically supported enrichment is observed only in selected regimes, while classical greedy and simulated-annealing baselines outperform QAOA on these small instances. The results therefore do not indicate quantum advantage; rather, they show that solver evaluation can change materially when success is defined in terms of downstream decision consequences rather than surrogate-objective recovery alone. The QAOA statevector implementation and selected frozen confirmatory cases were independently reproduced with Qiskit and Qiskit Aer to floating-point precision.

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

Combinatorial optimization methods are usually evaluated in the mathematical space in which the optimization problem is formulated. An exact solver can be assessed by whether it recovers an optimal solution, while approximate or heuristic solvers are commonly assessed through objective

gaps, approximation ratios, or the probability of recovering an optimum. This distinction is especially important for the Quantum Approximate Optimization Algorithm (QAOA), which produces a probability distribution over candidate bit strings rather than a single deterministic solution [1].

Existing QAOA benchmarking therefore commonly considers measures such as expected energy, approximation quality, and ground-state or optimal-solution probability [6, 7]. Recent systematic quantum-optimization benchmarking efforts likewise emphasize standardized optimization instances, classical reference methods, solution gaps, runtimes, and comparable measures of solver performance [11].

Exact recovery, however, is not the only possible definition of useful optimization performance. A measured solution can differ from the mathematical optimum while remaining close to it in objective value. This observation is already incorporated into QAOA benchmarking through expectation-based measures, approximation ratios, and the probability of obtaining a solution above a prescribed quality threshold [6, 7]. More recent work also connects approximate-solution criteria directly to the number of samples required to observe a target-quality solution [8]. Thus, distinguishing exact from mathematically approximate optimization is neither new nor specific to the present work.

A separate issue arises when the mathematical optimization problem is itself only a representation of a downstream decision problem. In practical decision pipelines, an optimization model may contain estimated coefficients, simplified constraints, surrogate objectives, or deterministic approximations to an uncertain environment. A solution that is near-optimal for this model is therefore not necessarily near-optimal when evaluated by the consequences of the resulting decision. This broader distinction is well established in decision-focused learning, which starts from the observation that quality measured at an intermediate predictive or optimization layer need not be a reliable proxy for downstream decision quality [14].

This paper investigates the corresponding evaluation problem for approximate quantum optimization. We consider a decision pipeline in which a stochastic operational problem is represented by a quadratic unconstrained binary optimization (QUBO) surrogate, the surrogate is approximately solved, and measured candidate solutions are subsequently deployed or evaluated in the original stochastic environment. QUBO and Ising formulations are mature representations of binary combinatorial optimization problems [3, 2, 4]; they are not quantum-specific constructs.

The pipeline creates three distinct levels at which success can be defined. A candidate solution may exactly recover the optimum of the QUBO surrogate; it may instead be mathematically near-optimal for that QUBO; or it may satisfy operational performance requirements when evaluated under stochastic downstream outcomes. The central question is not whether approximate solutions can be useful, but whether *mathematical near-optimality for the optimization surrogate contains the same decision-relevant information as downstream operational acceptability*.

The distinction matters particularly for probabilistic solvers. If a solver assigns probability distribution $P(x)$ over candidate solutions, exact-recovery success considers only probability mass assigned to the mathematical optimum. A near-optimal criterion instead aggregates probability over a neighborhood of sufficiently good surrogate solutions. An operational criterion can aggregate probability over a different set: solutions whose downstream expected performance and risk remain within prespecified tolerances.

Consequently, two success probabilities can differ even when computed from the same quantum state. They also imply different finite-shot requirements because the probability of observing at

least one successful solution after N independent measurements is

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

where S is the probability mass assigned to the chosen success set. The number of measurements required to obtain at least one successful observation with probability $1 - \eta$ is therefore

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

Recent QAOA work treats measurement count as a practical resource and studies success through target approximation quality rather than exact optimum recovery alone [8].

A second issue is that a large operational-success probability is not necessarily evidence that an optimizer is concentrating probability on useful decisions. If the operationally acceptable set occupies a large fraction of the feasible search space, even uniform feasible sampling can achieve a substantial operational-success probability. We therefore compare solver probability mass with the prevalence of acceptable solutions under uniform feasible sampling. For an operationally acceptable set $\mathcal{O}$, we define the descriptive operational enrichment

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

This separates the question of whether operational evaluation assigns a different success probability from the question of whether an algorithm actually concentrates probability on operationally useful states.

We evaluate this framework on a small stochastic inventory-control benchmark. Stochastic inventory control has a long operations-research tradition [20, 21]. The application is deliberately used as a controlled decision environment rather than as evidence of practical quantum advantage. The underlying stochastic decision problem is approximated by a calibrated QUBO surrogate whose parameters are selected on calibration instances and frozen before held-out and confirmatory evaluation. We then compare exact enumeration, a greedy heuristic, simulated annealing [16], uniform feasible sampling, and ideal statevector QAOA at depths $p = 1, 2, 3$.

The inventory application is not itself a novelty claim. Quantum, classical, and hybrid optimization have already been investigated for inventory-management problems [22]. Here, inventory provides a small stochastic environment in which both the surrogate and downstream decision landscapes can be completely enumerated.

For each candidate solution we separately evaluate surrogate-objective quality and downstream stochastic performance using expected cost and Conditional Value-at-Risk (CVaR). CVaR is a standard tail-risk measure [12] and has also been used directly as a variational quantum optimization objective [13]. In this paper, by contrast, CVaR is not used to train QAOA; it is evaluated after a candidate decision is mapped back into the stochastic operating environment.

The experiments do not provide evidence of quantum advantage. Classical greedy search and simulated annealing perform substantially better than QAOA on these deliberately small instances, and shallow $p = 1$ QAOA can concentrate less operationally acceptable probability mass than uniform feasible sampling. These negative results are relevant to the evaluation question: an operational criterion does not mechanically improve the apparent performance of a solver.

The main contributions are:

1. an end-to-end distinction between exact surrogate recovery, surrogate near-optimality, and downstream stochastic operational acceptability;
2. an explicit set decomposition identifying probability mass that is mathematically near-optimal but operationally unacceptable, and vice versa;
3. a finite-shot interpretation of alternative success definitions without interpreting the resulting differences as quantum speedup;
4. a prevalence control based on uniform feasible sampling, summarized through operational enrichment; and
5. a frozen experimental pipeline independently cross-validated against Qiskit and Qiskit Aer [23].

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 formalizes the evaluation framework. Section 4 describes the stochastic benchmark and frozen experimental protocol. Section 5 presents the confirmatory 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:

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

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

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

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

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

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 $\epsilon_Q \geq 0$,

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

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

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

For tolerances ϵ_J and δ_C , define

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

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,

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

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

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

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

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

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

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

with output probability

$$P_Q(x) = |\langle x | \psi_p(\gamma, \beta) \rangle|^2.$$

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

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

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

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

The feasible set is

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

Demand is generated as

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

Operational loss is

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

Expected operational loss is estimated using

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

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

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

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

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

using NumPy's default linear quantile interpolation, and then

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

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

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

4.3 Frozen QUBO surrogate and calibration

For each item, define

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

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

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

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:

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

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

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

$$R_J \leq 0.05, \quad R_C \leq 0.05,$$

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

$$s_t = 1.2, \quad \kappa = 0.15, \quad \eta = 0, \quad s_c = 0.50.$$

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.

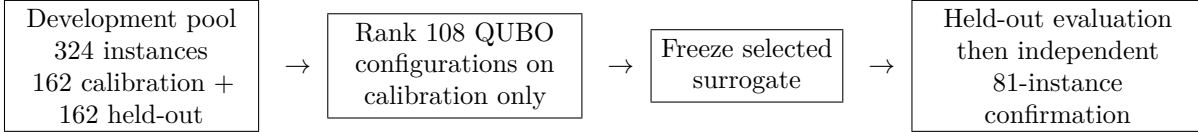


Figure 1: Calibration, freeze, and evaluation flow. Model selection used only the calibration split; held-out and confirmatory sets were used strictly after the surrogate was frozen.

4.4 Confirmatory design and thresholds

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

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

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

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

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

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,

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

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,

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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.

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%

Thus,

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

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, median (S_E, S_N, S_O) is $(0.0133, 0.0278, 0.0535)$ for $p = 1$, $(0.0408, 0.0723, 0.1397)$ for $p = 2$, and $(0.0362, 0.0711, 0.1087)$ for $p = 3$. At $n = 8, p = 2$, the corresponding values are

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

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:

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]

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 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

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

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 $392 \rightarrow 84 \rightarrow 31$ 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 $n = 8$,

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

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

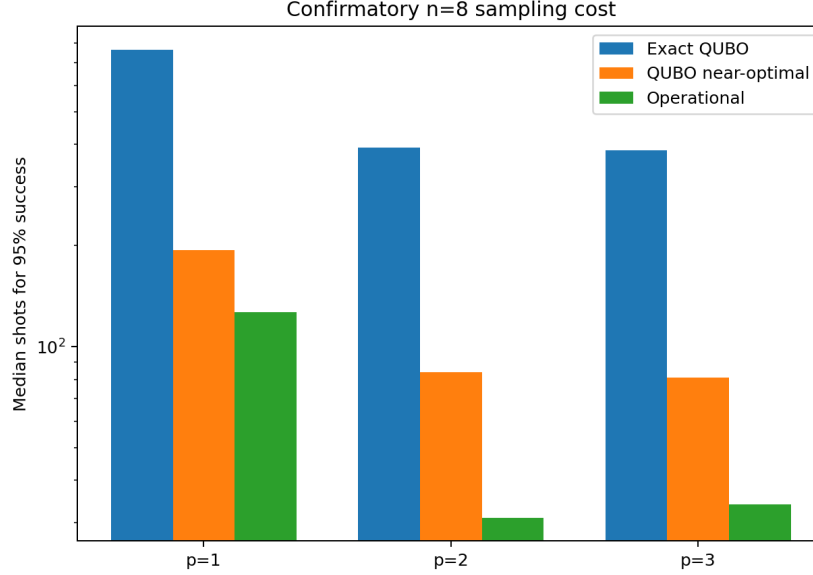


Figure 2: Median N_{95} for the confirmatory $n = 8$ instances under exact, QUBO-near, and operational success. The differences arise from the success definition applied to the same QAOA output distribution.

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

Threshold sensitivity shows the same dependence on the success definition:

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

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.

Solver	Exact S_E		QUBO-near S_N		Operational S_O	
	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

equal 1 because the deterministic greedy output is successful under both criteria on at least half of the instances.

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} .

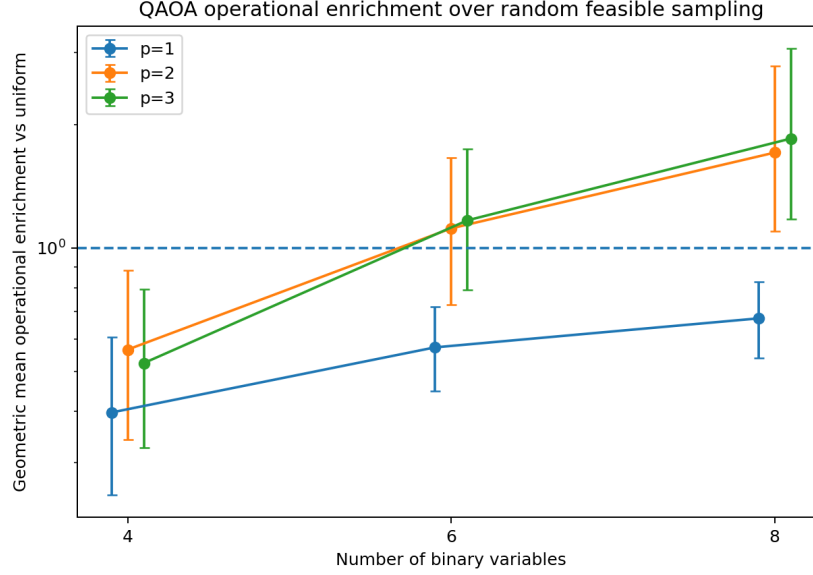


Figure 3: Geometric-mean operational enrichment relative to uniform feasible sampling. Values above 1 indicate concentration of operationally acceptable solutions beyond baseline prevalence; error bars show 95% bootstrap confidence intervals.

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

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

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 $392 \rightarrow 84 \rightarrow 31$ 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 confirmatory quantum results use ideal statevector simulation. No claim is made that S_O is more robust than S_E or S_N under physical quantum noise. An exploratory post-measurement bit-flip stress test conducted during development did not provide a defensible device-noise model and is therefore not used as evidence.

Independent Qiskit/Aer reproduction establishes implementation correctness, including basis indexing, cost phases, mixer evolution, layer ordering, probability distributions, and success-mass aggregation, but it does not establish hardware efficiency. In particular, only the frozen surrogate $Q(x)$ maps directly to the 2-local Ising form

$$H_Q = C + \sum_i h_i Z_i + \sum_{i < j} J_{ij} Z_i Z_j.$$

The complete simulated cost Hamiltonian also contains H_{viol} , the diagonal representation of the positive-part squared feasibility penalty. The semantic Qiskit validation implements the complete diagonal cost unitary directly, which is appropriate for checking circuit semantics but should not be interpreted as a hardware-efficient decomposition.

Physical-QPU implementation would therefore require an additional encoding choice outside the scope of the present simulation study. One option is to compile the general diagonal penalty directly, potentially introducing higher-locality or multi-controlled phase operations. A second is to reformulate the weighted inequality as a strict QUBO by choosing a finite-precision representation

of the real-valued weights and budget and introducing binary slack variables with an equality penalty. A third is to replace the standard X -mixer with a feasibility-preserving alternating-operator construction so that infeasible states are not generated [5]. These alternatives impose different qubit, circuit-depth, and approximation costs, and the present paper does not claim that any one is optimal for hardware execution.

Accordingly, the paper reports no physical-QPU result and makes no claim regarding execution time, connectivity, calibration, error mitigation, or reproducibility across hardware calibration periods. Likewise, N_{95} is not time-to-solution: it measures only the number of independent measurements required after the underlying output distribution has been prepared.

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

$$S_E, \quad S_N, \quad S_O, \quad E_O,$$

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, explicit gate-level noise, and physical-QPU execution. The framework could also be extended from binary acceptable/unacceptable sets to utility-weighted downstream criteria.

7 Conclusion

This paper examined how approximate optimization should be evaluated when the mathematical problem supplied to a solver is only a surrogate for a downstream stochastic decision problem. The study distinguished three notions of success: S_E , the probability of recovering the exact surrogate optimum; S_N , the probability of obtaining a mathematically near-optimal surrogate solution; and

S_O , the probability of obtaining a decision that satisfies downstream operational cost-and-risk criteria. The confirmatory experiments show that these quantities are not interchangeable.

The mathematical near-optimal set $\mathcal{Q}_{\epsilon_Q}$ and downstream operational acceptance set $\mathcal{O}_{\epsilon_J, \delta_C}$ are empirically distinct. QAOA probability mass occurs in both

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

The distinction also has measurable sampling consequences. In the $n = 8, p = 2$ confirmatory regime, median N_{95} decreases from 392 under exact recovery to 84 under mathematical near-optimality and 31 under operational acceptance. These differences are success-definition effects, not quantum speedups.

Operational probability should also not be interpreted in isolation. If acceptable decisions are common in the feasible state space, uninformed sampling can already achieve substantial S_O . The enrichment statistic

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

distinguishes absolute operational success from concentration beyond baseline prevalence. The classical baselines provide an equally important boundary: greedy search and simulated annealing outperform the tested QAOA configurations on these deliberately small instances. The study therefore provides no evidence of quantum advantage.

The contribution is methodological. When an optimization model is embedded inside

stochastic problem $\rightarrow$ surrogate model $\rightarrow$ approximate solver $\rightarrow$ deployment evaluation,

solver quality can be characterized at more than one level. Exact mathematical recovery asks whether the surrogate optimum was found; near-optimality asks whether the solver remained close to that optimum; operational evaluation asks whether the resulting decision remains acceptable in the environment for which the optimization model was constructed. These questions are related, but they are not equivalent.

For application-oriented approximate optimization, a more complete evaluation should therefore report mathematical recovery, mathematical approximation quality, downstream acceptability, and the prevalence of acceptable solutions under an appropriate baseline. Future work should determine whether the same distinctions persist for larger optimization problems, other application domains, explicit circuit-level noise, and physical quantum processors.

Code and Data Availability

The 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 code is distributed under the MIT License. The complete reproducibility package is archived on Zenodo [24] and is available at [doi:10.5281/zenodo.22817579](https://doi.org/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](https://arxiv.org/abs/1411.4028), 2014. [doi:10.48550/arXiv.1411.4028](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://arxiv.org/abs/1811.11538), 2019. [doi:10.48550/arXiv.1811.11538](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://arxiv.org/abs/2006.04831), 2020. [doi:10.48550/arXiv.2006.04831](https://doi.org/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](https://arxiv.org/abs/2509.19035), 2025. [doi:10.48550/arXiv.2509.19035](https://doi.org/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](https://doi.org/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. 023240, 2025. [doi:10.1103/24gg-7p8z](https://doi.org/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.