

SRP-HUBO: A Scale-Robust Benchmarking Protocol and Engineering Study of QAOA on Higher-Order Portfolio HUBOs

Ria Rushin Joseph and Snia Joseph

Abstract—Higher-Order Unconstrained Binary Optimisation (HUBO) instances arising from penalty-encoded portfolio optimisation with non-Gaussian risk preferences are widely used as QAOA benchmarks, yet comparisons across papers are often dominated by whether a problem’s residual range is 10^{-3} or 10^{-2} rather than by algorithmic differences. We propose the *Scale-Robust Protocol for HUBOs* (SRP-HUBO): a specification that pins the objective, the penalty policy, the decoding rule, and a normalised gap metric $g(\hat{z})$ that makes cross-instance medians comparable to four significant figures. We apply SRP-HUBO to study QAOA engineering in depth on a panel of 83 instances (35 DJ30 + 48 synthetic) drawn from both the Dow Jones 30 universe and four synthetic factor-model regimes with controlled skewness and excess kurtosis. Our principal engineering findings are: (a) under matched evaluation-count budgets BALS attains exact optimality on 34 of 35 DJ30 instances and on all 48 synthetic instances, with a median gap of exactly zero on every panel, while every QAOA variant we implement retains a strictly positive median gap; the one DJ30 miss (q07_i000) exposes a continuous-warm-start local-minimum trap that the `bals_no_continuous` ablation sidesteps, and in the Dolan–Moré performance profile (§VI-G) PT and SA edge BALS by this single instance on $\rho_s(1)$ while the QAOA variants sit below the entire strong-classical cluster; (b) paired Wilcoxon tests with Holm–Bonferroni correction reject equal medians between BALS and every QAOA variant including BALS-seeded warm-start at $p=1$ (adjusted $p < 10^{-4}$, rank-biserial $r = -0.91$ on DJ30; see Table III), so we explicitly withdraw the claim of statistical equivalence that appeared in an earlier draft. We additionally (c) describe the non-monotonic p -dependence of BALS-warm QAOA as a directional trend localised to high-QFI-concentration instances – after Holm–Bonferroni correction across the 11-pair ladder family this is not a statistically decisive effect at this panel size (adjusted $p = 0.275$), and a pre-specified Wilcoxon-based TOST at $\delta=10^{-3}$ likewise cannot declare equivalence (we have neither enough power to reject equality nor tight enough differences to reject inequivalence); (d) provide a compute-budget

Pareto showing QAOA-under-BALS-warm is never the fastest wall-clock choice at any budget we tested; (e) extend MPS-QAOA to $p=3$, $\chi \in \{64, 128\}$ on $n \in \{14, 16\}$ qubits, a correctness check rather than a scaling claim; (f) formalise the limitations of the scale-robust gap metric; and (g) ship a Dockerised artifact reproducing every result. The broader contribution is a *reporting protocol*: reporting the SRP-HUBO gap alongside any other metric of interest would make future QAOA-versus-classical comparisons on penalty-encoded HUBOs commensurable where today they are not.

Index Terms—QAOA, benchmarking, portfolio optimisation, HUBO, reproducibility, quantum engineering.

I. INTRODUCTION

A decade of QAOA [1], [2] research on portfolio optimisation has produced a literature that is simultaneously encouraging and difficult to compare across. Two papers reporting on nearly identical problems frequently report results in different units—one in residuals of the penalised objective, another in basis points of a post-hoc utility; one at penalty λ chosen by ad-hoc grid search, another under a “solver-aware” tuned penalty—and with different conventions for what counts as a feasible solution. The net effect is that a reader cannot answer the simplest question without re-running the code: *does method A actually beat method B on this family of instances, beyond two significant figures?*

This paper contributes an answer with three components, integrated as a single reproducibility artifact:

- (1) **SRP-HUBO**—a Scale-Robust Protocol for HUBOs—that pins every configurable choice behind the “what’s the gap of method A” question into a single named, testable specification (§III);
- (2) **A rigorous engineering study** comparing the canonical QAOA variants (vanilla X -mixer, continuous warm-start, BALS-seeded warm-start, XY-ring/complete mixers, Dicke-state + XY complete, and the Uotila *et al.* [3] comparator) against five strong classical baselines (BALS, simulated annealing, a genetic algorithm, ballistic simulated bifurcation [4], and an MILP of the continuous relaxation) on a 83-instance panel covering DJ30 and four synthetic factor-model regimes, under matched compute-budget parity and with 10,000-bootstrap 95% CIs on every reported median (§§V–VI);
- (3) **A formal characterisation of the scale-robust gap $g(\hat{z})$** including the conditions under which it is well-defined, its response to outlier instances, and a bound relating it to classical time-to-solution metrics (§III-C).

R. R. Joseph is with the School of Information Technology, Deakin University, Melbourne, Australia (e-mail: ria.joseph@deakin.edu.au). *Corresponding author.*

S. Joseph is with the Department of Accounting, Economics and Finance, T. A. Pai Management Institute, Manipal, India (e-mail: snia.tapmimpl2024@learner.manipal.edu).

Submitted to *IEEE Transactions on Quantum Engineering, Applications/Software* track. The artifact supporting all experiments in this paper (release `srp-hubo:v2.0`) is archived on Zenodo at 10.5281/zenodo.19689983 and mirrored at github.com/ria6683/srp-hubo. A single command `bash scripts/run_all.sh --tqe-full` on a pinned Python 3.11 environment (`requirements.txt`) reproduces every table and figure end-to-end in approximately 63.8 core-hours on a laptop-class CPU (MacBook Pro 14-inch, Apple M5, 16 GB RAM), dominated by EXP-K (MPS-QAOA deep scan under the v2 defaults). The bundle also ships the `srp_hubo` Python library (Solver plugin API, conformance harness, JSON schemas, pytest suite) and a `Dockerfile` for optional environment pinning; see `docs/developer.md`.

Scope of this paper. The benchmark panel comprises 35 DJ30 instances and 48 synthetic instances across four non-Gaussian regimes (gaussian/heavy/skewed/stress). The (q_2, q_3) stress sensitivity is evaluated on a full 5×5 grid. A compute-budget parity Pareto sweep is included, and the BALS-warm QAOA analysis is accompanied by bootstrap CIs, Wilcoxon signed-rank tests with Holm–Bonferroni correction over the full 11-comparison ladder family (which now includes parallel tempering as a strong stochastic classical baseline), a pre-specified two one-sided tests (TOST) equivalence analysis with margin $\delta=10^{-3}$ on the scale-robust gap, a Dolan–Moré performance profile across the strong-classical cluster (§VI-G), and a per-instance scatter. MPS-QAOA is evaluated at $(p \leq 3, \chi \in \{64, 128\}, n \in \{14, 16\})$. Code, pinned environment, and per-stage seeds are archived as a Zenodo source bundle (Dockerfile included for optional environment pinning).

Engineering headline. Under SRP-HUBO with matched eval-count budgets, BALS in its strongest configuration (the `bals_no_continuous` ablation, §V) recovers the *exact* optimum on all 83 panel instances at a median scale-robust gap of 0.0000, while the default BALS configuration (with continuous warm-start) attains exact optimality on 34 of 35 DJ30 instances and on all 48 synthetic instances. The best-effort QAOA variant we test (BALS-seeded warm-start at $p=1$) sits at a median gap of 0.0014 on DJ30 with a Wilson-interval exact-hit rate well below 100% on every synthetic regime (gaussian/heavy/skewed/stress). The paired Wilcoxon test rejects equal medians between BALS and BALS-warm QAOA $p=1$ with Holm-adjusted $p < 10^{-4}$ and rank-biserial $r = -0.91$ on the 35 DJ30 instances, so we describe BALS as dominating BALS-warm QAOA on these instances rather than indistinguishable from it. On the heavy skew + kurtosis (“stress”) regime the BALS-warm gap shifts upward by $\sim 1.2 \times$ relative to the Gaussian regime; we decline to draw a regime-sensitivity conclusion from this pattern at $N_{\text{inst}}=12$ per regime — TOST on a panel that small has essentially no power — and defer per-regime equivalence analysis to the enlarged synthetic panel planned in §VIII-C.

II. RELATED WORK

QAOA for portfolio optimisation has been investigated in [5]–[10] and recent scale-up reports appear in [3], [11]. All of these works report a form of the penalty-encoded optimisation problem we study here, but each uses a different combination of (i) objective formulation (per-share vs. per-dollar, discrete vs. continuous shares); (ii) penalty policy (fixed scalar, grid-searched, solver-aware); (iii) decoding rule (argmin over the top- K computational-basis amplitudes, or maximum-probability sample); and (iv) gap metric (residual, fractional residual, probability of exact success, TTS). SRP-HUBO’s contribution is not a new algorithm but a specification that pins (i)–(iv) while leaving the solver free.

QAOA benchmarking more broadly has been the subject of [12]–[15], and critical surveys such as [16], [17] have argued that the community’s comparisons are handicapped by a lack of matched-budget protocols. The closest pre-existing QAOA protocol we are aware of is the “Strong-Baseline” proposal

of [12], which prescribes a time-to-solution metric. SRP-HUBO differs in that it targets the penalty-encoded HUBO family specifically, makes the gap scale-normalised rather than time-normalised, and specifies the decoding rule (which the strong-baseline protocol leaves open).

Positioning against the classical benchmarking-methodology literature. The scale-robust gap (1) followed by Definition 1 is, viewed as a formula, a relative-gap metric with a denominator floor. Several long-established conventions cover the same ground. The MIP-gap denominator used by commercial mixed-integer solvers [18] normalises by the incumbent, giving scale invariance but no range-based fallback when $|f^*|$ is small. Dolan–Moré performance profiles [19] normalise *runs* by the per-instance best across solvers and report the ECDF over the resulting performance ratios. BBOB/COCO [20] takes a related route for continuous black-box optimisation, normalising residuals by a per-function target tolerance and reporting empirical runtime distributions. Normalised regret in the bandit and black-box-optimisation literature is essentially $(f_\lambda(\hat{x}) - f^*)/\text{range}$ without a numerical floor. The SRP-HUBO gap is therefore best positioned as an *importation* of these conventions into the quantum-HUBO setting, with two small additions specific to penalty-encoded portfolio problems: (i) a denominator that takes the *max* of $|f^*|$, the range R , and a numerical floor ϵ , so that the near-degenerate regime where $|f^*|$ underflows is caught by the range arm (Remark 5); and (ii) a specified top- K decoding rule (§III-E) which Dolan–Moré, MIP-gap, and BBOB do not speak to because they take the solver’s returned solution as given. We do not claim methodological novelty beyond these two additions; SRP-HUBO’s value is as a commonly-agreed specification for penalty-encoded HUBO benchmarks, not as a new metric per se.

Classical simulated bifurcation [4] and CIM-style approaches [21] are widely regarded as the strongest quantum-inspired classical baselines for Ising-form optimisation at the 100–1000 qubit scale; we include a faithful implementation of ballistic SB (bSB) as a baseline throughout this paper. Tabu search [22], genetic algorithms [23], and parallel tempering are established non-local search baselines; this revision includes tabu, GA, and a Geyer-style parallel tempering implementation (`src/pt.py`) in the DJ30 ladder and in the Dolan–Moré performance profile of §VI-G.

III. SRP-HUBO: A SCALE-ROBUST BENCHMARKING PROTOCOL

SRP-HUBO specifies four components. A *conforming solver* is any algorithm that takes an instance and returns a single candidate integer share vector \hat{x} (equivalently, a decoded bitstring \hat{z} via the binary encoding of §III-A); the candidate may be infeasible, in which case the penalty term in f_λ scores it accordingly and the budget-feasibility rate is reported separately (§III-E). A *conforming experiment* reports the scale-robust gap g on this candidate alongside any other metric of interest.

A. The penalty-encoded HUBO objective

Let $x \in \mathbb{Z}_{\geq 0}^{N_A}$ be an integer vector of share counts over N_A assets with per-share prices $p \in \mathbb{R}^{N_A}$ and budget B . Instance moments at daily frequency are the mean $\mu \in \mathbb{R}^{N_A}$, the covariance $\Sigma \in \mathbb{R}^{N_A \times N_A}$, co-skewness $M_3 \in \mathbb{R}^{N_A \times N_A \times N_A}$, and co-kurtosis $M_4 \in \mathbb{R}^{N_A \times N_A \times N_A \times N_A}$. The higher-order portfolio HUBO is

$$f_\lambda(x) = -\mu^\top x + q_1 x^\top \Sigma x - q_2 \langle M_3, x^{\otimes 3} \rangle + q_3 \langle M_4, x^{\otimes 4} \rangle + \lambda \cdot (p^\top x - B)^2, \quad (1)$$

where $q_1 \geq 0$, $q_2 \in [0, 1]$, $q_3 \in [0, 1]$ are fixed preference weights and $\lambda \geq 0$ is the penalty coefficient. Each asset a is binary-encoded as $b_a \in \mathbb{N}$ qubits so that $x_a = \sum_{k=0}^{b_a-1} 2^k z_{a,k}$ with $z_{a,k} \in \{0, 1\}$; the total qubit count is $n = \sum_a b_a$.

B. The scale-robust gap $g(\hat{z})$

Let $\hat{z} \in \{0, 1\}^n$ be a conforming solver's decoded bitstring with corresponding integer share vector \hat{x} . By the *feasible integer lattice* we mean the set $\mathcal{X} = \{x \in \mathbb{Z}_{\geq 0}^{N_A} : p^\top x \leq B\}$ of budget-feasible share vectors, and by its *budget boundary* the subset $\mathcal{X}^\partial = \{x \in \mathcal{X} : p^\top x = B\}$ on which the penalty term $(p^\top x - B)^2$ in (1) vanishes. We define $f^* := \min_{x \in \mathcal{X}^\partial} f_0(x) = \min_{x \in \mathcal{X}^\partial} f_\lambda(x)$; on the boundary the penalty contributes nothing, so f^* is independent of λ by construction. For every panel instance in this paper the unrestricted minimiser of f_λ over \mathcal{X} lies on \mathcal{X}^∂ (we verify this post-hoc in `results/instance_stats.csv`; each DJ30 panel instance has at least one asset a with $\mu_a > 0$ and $p_a \leq B$, so spending the full budget strictly improves the mean-return term and the penalty-free minimum is attained at $p^\top x = B$), so the pairing $f^* = \min_{\mathcal{X}} f_\lambda$ holds empirically on the whole panel; the boundary definition makes it hold by *definition* and removes a λ -dependence that would otherwise silently enter any cross-paper comparison. Note that this definition is strict: a solver returning a bitstring whose decoded \hat{x} violates the budget is penalised in the numerator of g and is not credited as matching f^* even if its un-penalised objective is lower. We write $R = \max_{z \in \{0, 1\}^n} f_\lambda(z) - f^*$ for the instance's objective range across *all* binary strings (i.e. including infeasible ones, with penalty active); we compute R exactly by brute-force enumeration for $n \leq 20$ and estimate it for $n > 20$ (Remark 4).

Definition 1 (Scale-robust gap). The SRP-HUBO gap of \hat{z} on instance I is

$$g_I(\hat{z}) = \frac{f_\lambda(\hat{x}) - f^*}{\max(|f^*|, R, \varepsilon)}, \quad (2)$$

where $\varepsilon = 10^{-9}$ is a numerical floor.

The normalisation by $\max(|f^*|, R)$ makes g invariant to multiplicative rescaling of the objective (as occurs when a paper changes the units of μ , or rescales λ), and the R -arm catches the regime where $|f^*|$ is numerically small so that $1/|f^*|$ blows up. For a conforming solver that returns a *feasible* decoded \hat{z} (i.e., $\hat{x} \in \mathcal{X}$) on an instance for which R is computed *exactly* via brute-force enumeration over $\{0, 1\}^n$, the scale-robust gap satisfies $g(\hat{z}) \in [0, 1]$, with $g = 0$ iff

$\hat{x} \in \arg \min_{x \in \mathcal{X}^\partial} f_0(x)$. Every instance in this paper has $n \leq 16$ and therefore R is exact (Remark 4); the bound is not merely empirical but definitional on this panel. On instances where R must be estimated rather than enumerated (n large enough that 2^n enumeration is impractical, which would arise for larger synthetic panels considered in §VIII-C), the upper bound holds only modulo the R -estimate's accuracy: a solver could in principle report $g > 1$ if it finds a configuration above the sampled range. The maximum normalised gap observed anywhere in the reported results is 0.132 (QAOA XY-mixer on a 7-qubit DJ30 instance), two orders of magnitude below the conservative upper bound.

C. Formal properties and limitations of g

Proposition 2 (Conditional scale invariance). *Fix an instance and a decoded bitstring \hat{z} . Let $D = \max(|f^*|, R, \varepsilon)$ denote the denominator of g in Definition 1. For any $\alpha > 0$, the rescaling $f_\lambda \rightarrow \alpha f_\lambda$, $\lambda \rightarrow \alpha \lambda$ transforms the denominator into $D' = \max(\alpha |f^*|, \alpha R, \varepsilon)$. The scale-robust gap is invariant under the rescaling, $g'(\hat{z}) = g(\hat{z})$, whenever ε is not the active arm of either D or D' – equivalently, whenever $\alpha \cdot \max(|f^*|, R) \geq \varepsilon$. When $\alpha \cdot \max(|f^*|, R) < \varepsilon$ the denominator collapses to ε and invariance fails.*

Proof. Under the rescaling, $f^* \rightarrow \alpha f^*$, $R \rightarrow \alpha R$ (both the f_λ values and their minima/maxima scale together), and $f_\lambda(\hat{x}) - f^* \rightarrow \alpha(f_\lambda(\hat{x}) - f^*)$. When ε is inactive in both D and D' , $D' = \alpha D$ and the numerator $\alpha(f_\lambda(\hat{x}) - f^*)$ divided by αD equals the original $g(\hat{z}) = (f_\lambda(\hat{x}) - f^*)/D$. When $\alpha \cdot \max(|f^*|, R) < \varepsilon$, however, the denominator of the rescaled problem is exactly ε rather than αD , so $g'(\hat{z}) = \alpha(f_\lambda(\hat{x}) - f^*)/\varepsilon \neq g(\hat{z})$ in general. Invariance is therefore *conditional* on the scaled instance remaining above the ε floor; we call this the *non-degenerate regime* and verify empirically in §VI that every instance in our panel satisfies it by a margin of at least nine orders of magnitude ($\min \max(|f^*|, R)/\varepsilon \approx 10^{9.7}$ across the 83 instances). \square

Remark 3 (Caveat on the name “scale-robust”). The name of the metric reflects Proposition 2 in its non-degenerate regime; it does *not* claim unconditional scale invariance. Reporting the fraction of instances with $\max(|f^*|, R) < 10\varepsilon$ is part of a conforming experiment (Remark 5); in our panel that fraction is 0.

Remark 4 (Range estimation at $n > 20$). For $n > 20$ exact enumeration of R is infeasible. We instead estimate R from a pool of ($|\text{BALS}| + |\text{SA}| + |\text{BSB}|$) candidate solutions: $\hat{R} = \max_z f_\lambda(z) - \hat{f}_{\text{ref}}^*$ where \hat{f}_{ref}^* is the minimum f_λ across the three baselines. By construction \hat{R} is a conservative lower bound on the true range. In §VI-F we test a uniform-random upper-bound proxy and report both numbers.

Remark 5 (Failure mode: flat instances). When $R \ll \varepsilon$ (a degenerate instance with all feasible points yielding essentially identical f_λ), g saturates at $(f_\lambda(\hat{x}) - f^*)/\varepsilon$ which can exceed unity. This mode is detectable: a conforming protocol reports the fraction of instances with $R < 10\varepsilon$ separately. In our panel no instance is flat.

Remark 6 (Relation to time-to-solution). For a stochastic solver which, under *i.i.d. shots at a fixed circuit*, has per-shot success probability p_{exact} ,

$$\text{TTS}_{0.99} = \left\lceil \frac{\log(1 - 0.99)}{\log(1 - p_{\text{exact}})} \right\rceil \approx \frac{4.6}{p_{\text{exact}}} \quad (3)$$

for small p_{exact} . Under this *i.i.d.*-shots assumption a protocol reporting g alongside p_{exact} strictly refines one reporting p_{exact} alone, because it separates the optimality-gap magnitude on runs that fail to hit the exact optimum. The *i.i.d.* assumption is non-trivial in practice: it fails when a solver's shots are correlated (e.g. during a parameter-update pass of QAOA) or when the reported p_{exact} aggregates across distinct circuits. We therefore present g and p_{exact} as *complementary* reporting axes rather than claim one implies the other.

D. The two-track penalty protocol

SRP-HUBO specifies two penalty policies, both of which must be reported by a conforming experiment. We first define the feasibility-range quantity used by the fixed- λ rule.

Definition 7 (Feasibility-range half-width). $R_{\text{feas}} := \frac{1}{2}(\max_{x \in \mathcal{X}} f_0(x) - \min_{x \in \mathcal{X}} f_0(x))$, i.e. half the range of the un-penalised objective $f_0 = f_\lambda|_{\lambda=0}$ over the feasible integer lattice. It is computed exactly by enumeration for $n \leq 20$ and by the baseline-pool estimator of Remark 4 otherwise.

Fixed- λ track.

$\lambda^{\text{fix}} = 10 \cdot \|\mu\|_2 / R_{\text{feas}}$. The factor of 10 is a conservative multiplier chosen so that a single-share budget violation incurs at least an order of magnitude more penalty than the variation of the un-penalised objective across feasible portfolios. The ratio $\|\mu\|_2 / R_{\text{feas}}$ is not dimensionally exact — the numerator is in units of daily return per asset while R_{feas} is in units of objective — but it ensures that $\lambda^{\text{fix}} \cdot (p^\top x - B)^2$ has the same *magnitude* as the residual range of f_0 whenever $|p^\top x - B|$ is order unity, so no term in (1) dominates numerically. We refer to this as magnitude-balanced rather than dimensionally consistent throughout. This rule is independent of any solver.

Solver-aware track.

λ is tuned *once* per instance by minimising the training-set residual of a BALS run over a logarithmic sweep $\lambda \in \{10^{-2}, 10^{-1}, \dots, 10^2\} \cdot \lambda^{\text{fix}}$; the minimising value is used for every solver on that instance.

Both tracks fix λ before the solver runs; on-line adaptation is disallowed. The main body reports fixed- λ results as the default; solver-aware results are reported in §VI-B and in `results/solver_aware.csv`.

E. The decoding rule

A conforming solver must return a single bitstring \hat{z} ; the rule we use for QAOA is deterministic given the final quantum state. Let $\{|z_k\rangle\}_{k=1}^K$ be the K computational-basis states of highest probability under the final circuit, ordered by descending amplitude. Define the decoding map

$$\hat{z} = \arg \min_{z \in \{z_1, \dots, z_K\}} f_\lambda(x(z)). \quad (4)$$

Crucially we compute f_λ (not f_0) on each candidate: an infeasible top- K candidate is *not* discarded but incurs the budget-penalty term in its score, so a conforming solver is free to return an infeasible \hat{z} if the protocol's penalty makes that the lowest-scoring choice among the K . The fraction of decoded \hat{z} that are budget-feasible is reported separately as *feasibility rate* (Table II and `exact_hit_wilson.csv`); an infeasible decode is never credited as matching f^* in the numerator of g .

The default is $K = 32$. Sensitivity of median gap and feasibility rate to $K \in \{1, 4, 16, 32, 64, 128\}$ is recorded in `results/k_sweep.csv`; as summarised by the macros 1×10^{-5} at $K=64$, $< 10^{-5}$ at $K \geq 128$ (best) and 3.5×10^{-3} at $K=1$ (worst), the plateau is already reached by $K=32$ and the feasibility tradeoff is feasibility drops from 80% at $K=1$ to 51% at $K \geq 64$. For non-QAOA solvers, \hat{z} is simply the solver's returned integer share vector mapped back to bits via the binary encoding.

a) *Finite-shot variant.*: The top- K decoding rule above presumes access to the full final statevector or MPS amplitude distribution, which is natural for the noiseless simulator study of this paper but is *not* implementable on real hardware that only returns shot samples. For a hardware-faithful replication on a shot budget S , the protocol-compliant substitute is: draw S shots, rank distinct bitstrings by empirical frequency, take the top K by frequency (breaking ties by lexicographic order on the bitstring), and apply the same $\arg \min_{z \in \{z_1, \dots, z_K\}} f_\lambda(x(z))$ rule. Finite- S statistics of \hat{z} converge to the statevector top- K rule as $S \rightarrow \infty$ and the ranking of low-probability tail states stabilises; finite-shot noise can both help and hurt (sampling occasionally surfaces low-amplitude but low- f_λ states) and should be reported alongside the shot budget. All numbers in this paper are produced under the noiseless-amplitude variant and labelled *simulation-only*; we do not claim they transfer to a finite-shot hardware regime without at least a separate shot-budget sensitivity study.

IV. BENCHMARK PANELS

A. Panel A: DJ30 (real-world)

We construct 35 instances by subsampling n_A tickers from the Dow Jones 30 universe, using a 250-trading-day rolling window ending 2024-12-31 to estimate μ, Σ, M_3, M_4 . Qubit sizes range over $n \in \{6, 8, 10, 12, 14\}$ with (n_A, b) chosen as follows: let $\mathcal{D}(n) = \{b : b \text{ divides } n, b \geq 1\}$ and

$$\mathcal{D}^*(n) = \begin{cases} \{b \in \mathcal{D}(n) : n/b \geq 4\} & \text{if this set is non-empty,} \\ \mathcal{D}(n) & \text{otherwise;} \end{cases} \quad (5)$$

we draw b uniformly from $\mathcal{D}^*(n)$ and set $n_A = n/b$. For the sizes we run, $\mathcal{D}^*(6) = \{1\}$, $\mathcal{D}^*(8) = \{1, 2\}$, $\mathcal{D}^*(10) = \{1, 2\}$, $\mathcal{D}^*(12) = \{1, 2, 3\}$, $\mathcal{D}^*(14) = \{1, 2\}$, so the $n_A \geq 4$ preference is binding at every size. Budget B is set to 35%–55% of the instance's capacity so the constraint bites without forcing a trivial solution. Prices are normalised so the median share price is \$1, making budgets directly comparable across instances.

B. Panel B: four synthetic regimes (controlled non-Gaussianity)

To address the TQE reviewer’s concern that a single real-world universe is narrow, we add 48 synthetic instances drawn from four single-factor regimes:

Gaussian.

Factor $F_t \sim \mathcal{N}(0, 1)$, idiosyncratic noise $\varepsilon_{i,t} \sim \mathcal{N}(0, 1)$; target skew 0, excess kurtosis 0. Baseline regime.

Heavy.

Factor $F_t \sim t(\nu=5)/\sqrt{\nu/(\nu-2)}$; target skew 0, excess kurtosis ~ 6 .

Skewed.

Factor F_t drawn from a skew-normal with shape $\alpha = -4.2$; target skew -0.7 , excess kurtosis ~ 1.5 .

Stress.

Mixture of skew-normal and $t(\nu=5)$; target skew -0.7 , excess kurtosis ~ 6 . The adversarial regime.

Each synthetic panel has 30 tickers and 600 trading days; assets are drawn from the same skeleton as Panel A (same `build_instance` routine), giving directly comparable instance IDs. At every synthetic regime we generate 4 instances per size for $n \in \{8, 10, 12\}$ giving 4 regimes \times 3 sizes \times 4 instances = 48 synthetic instances total, i.e. $N_{\text{inst}} = 12$ per regime as reported in Table II. Combined with the 35 DJ30 instances, the full panel is 83 instances.

Reporting conventions throughout the paper: DJ30 and synthetic results are kept *disaggregated* by default. A “Panel A+B” number is reported only when the bootstrap CIs of the two panels overlap; if they do not, the regimes are reported separately.

V. SOLVERS

A. Classical baselines

We compare against five *primary* classical baselines (described in detail below), and three *secondary* comparators that enter the Holm–Bonferroni ladder family (Table III) but are not the focus of the solver discussion: tabu search (`src/tabu.py`), exact brute-force enumeration for $n \leq 20$ (`src/exact.py`), and parallel tempering (`src/pt.py`). Together, these eight classical configurations plus the BALS ablations constitute the 11-pair Holm–Bonferroni family reported in §VII. All neighbourhood definitions and termination rules for the five primary baselines below are coded in `src/bals.py`, `src/sa.py`, `src/ga.py`, `src/sbif.py`, and `src/miqp.py`; exact hyper-parameters are frozen by the master seed in Table IV.

BALS.

Budget-Aware Local Search. Multistart best-improvement local search over integer share counts with the move set $\mathcal{M}(x) = \{x + e_a, x - e_a, x - e_a + e_b : a, b \in [N_A], x + e_a \neq x, a \neq b\}$ (add one share, remove one share, transfer one share from asset a to asset b). A restart terminates when no move in $\mathcal{M}(\hat{x})$ strictly decreases f_λ ; the outer loop runs 16 restarts from random feasible starts and returns the best-scoring restart. Each restart evaluates $\mathcal{M}(\hat{x})$ exhaustively at every step (no first-improvement shortcut), so “local optimum” is unambiguous. Default BALS seeds one of the 16

restarts from the continuous-relaxation minimiser (§ MILP-relax, below) to bias the search toward the global basin; the `bals_no_continuous` ablation replaces that single warm restart with an additional random feasible restart and is the “strongest BALS configuration” referenced in the engineering headline. The remaining BALS ablations (`bals_random_only`, `bals_warm_only`) are variants on the same axis used in the compute-budget Pareto sweep (EXP-H).

SA.

Simulated annealing on the same move set, geometric cooling from $T_0 = 1.0$ to $T_{\text{end}} = 10^{-3}$; 4 restarts \times 5000 iterations.

GA.

Generational GA with tournament selection (size 4), uniform bit-crossover, single-bit mutation, elite preservation; 64×80 generations.

bSB.

Ballistic simulated bifurcation [4]. Reference [4] defines bSB for Ising couplings; we extend it to a quartic HUBO by computing the gradient $\nabla_x f_\lambda$ exactly from $\mu, \Sigma, q_2 M_3, q_3 M_4$ and the budget-penalty term, and feeding that gradient (rather than an Ising Jx) into the bSB position-update step. All other aspects – the symplectic update, the ballistic time-step schedule, the pump parameter, and the ± 1 rounding at termination – follow [4] without modification. The implementation is in `src/sbif.py`; 16 restarts \times 500 steps.

MILP-relax.

Because f_λ contains cubic ($\langle M_3, x^{\otimes 3} \rangle$) and quartic ($\langle M_4, x^{\otimes 4} \rangle$) terms it is not expressible as an MIQP, and commercial MILP solvers (HiGHS, CBC, Gurobi) cannot ingest it directly. We build a tractable surrogate by *dropping* the cubic and quartic co-moments (so the surrogate problem is the pure Markowitz mean-variance MIQP) and linearising both the quadratic $x^\top \Sigma x$ term and the squared budget residual as follows. Let $U = \max_a \lfloor B/p_a \rfloor$ be a loose per-asset share bound; for each $i \leq j$ introduce the integer bilinear auxiliary $y_{ij} \in \{0, \dots, U^2\}$ representing $x_i x_j$ and impose the four McCormick inequalities

$$\begin{aligned} y_{ij} &\leq U x_i, & y_{ij} &\leq U x_j, \\ y_{ij} &\geq U x_i + U x_j - U^2, & y_{ij} &\geq 0, \end{aligned}$$

tight (exact) for bits-per-asset = 1 and a standard outer approximation otherwise. For the squared residual $s \approx (p^\top x - B)^2$ we introduce $r = p^\top x - B$ as a continuous variable, the auxiliary $s \geq 0$, and a piecewise-linear *outer* approximation

$$s \geq 2kr - k^2, \quad k \in \mathcal{K},$$

over a 41-point uniform grid $\mathcal{K} \subset [-B, B]$; this gives a lower bound on r^2 that is tight on \mathcal{K} . The MILP objective is then $-\mu^\top x + q_1 \sum_{ij} \Sigma_{ij} y_{ij} + \lambda s$. The solver returns integer x^* directly (no fractional rounding step); we then score x^* under the *full* higher-order objective (1) — including the dropped cubic and quartic co-moments — and report that value as the MILP-relax gap. Because the surrogate

omits the cubic/quartic terms, its optimum need not minimise f_λ ; we therefore report MILP-relax as a *structured baseline* and not as a conforming solver that expects to win on this panel. Implementation in `src/miqp.py`; default backend is HiGHS, with CBC and Gurobi 11.0 as alternatives resolved at call time and reported through `MIPResult.backend_used`.

B. QAOA variants

Seven QAOA variants are run at $p \in \{1, 2, 3\}$ where applicable:

- vanilla (X -mixer, uniform $|+\rangle^{\otimes n}$ start);
- continuous warm-start ([24] rounded to $|\pm \varepsilon\rangle$);
- BALS-seeded warm-start (BALS result rounded to $|\pm \varepsilon\rangle$);
- XY-ring mixer (Hamming-weight preserving, $b=1$ only);
- XY-complete + Dicke-state initial state ([25], $b=1$ only);
- Uotila *et al.* 2025 [3] comparator.

For $n \leq 12$ we use dense statevector simulation; for $n \in \{14, 16\}$ we use an MPS simulator (quimb) with bond dimension $\chi \in \{64, 128\}$. The COBYLA optimiser is the default with `max_iter=200`, `rhobeg=0.5`, `tol=10-6`, and 4 restarts (except in EXP-K, where the MPS simulator is the dominant cost: there the optimiser runs at `max_iter=50` with 1 restart and 4 instances per (n, χ, p) cell, flagged explicitly in Table V and discussed as a limitation in §VIII-C). The L-BFGS-B optimiser used in the EXP-A sweep (§VI-A) computes gradients by two-sided finite differences with step 10^{-5} ; parameter-shift gradients were not used. Basin-hopping wraps L-BFGS-B with 5 outer restarts and $\sigma = 0.5$ step size. Parameter initialisation in §VI-A sweeps three schemes: (i) BALS-warm rounded to $|\pm \varepsilon_w\rangle$ with $\varepsilon_w = 0.15$; (ii) Egger continuous warm-start; and (iii) uniform-random in $[-\pi, \pi]$ for both γ and β . None of the reported numbers anywhere in the paper are computed with parameter-shift gradients.

C. Compute-budget accounting

Every solver is instrumented to report n_{evals} —objective function calls—alongside wall-clock time. Comparisons in §VI-B respect this axis.

VI. EXPERIMENTAL PROTOCOL

Experiments are numbered EXP-A through EXP-K (Table I). Each reports: (i) the solver configuration matrix, (ii) the instance set, (iii) the compute budget, (iv) every metric from §III, and (v) bootstrap 95% CIs on every median.

A. EXP-A: optimiser \times initialisation \times depth

BALS-warm at $p=1$ with COBYLA gives the lowest median gap among the $3 \times 3 \times 3=27$ cells we tried. See Fig. 1.

B. Solver-aware penalty track

The solver-aware track (Definition III-D) tunes λ per instance via a BALS-driven logarithmic sweep. On the DJ30 panel, the tuned λ falls within a factor of $3\times$ of the fixed- λ value on 28/35 instances and within a factor of $10\times$ on

EXP-A: optimiser \times init \times depth sweep (N=35 instances)

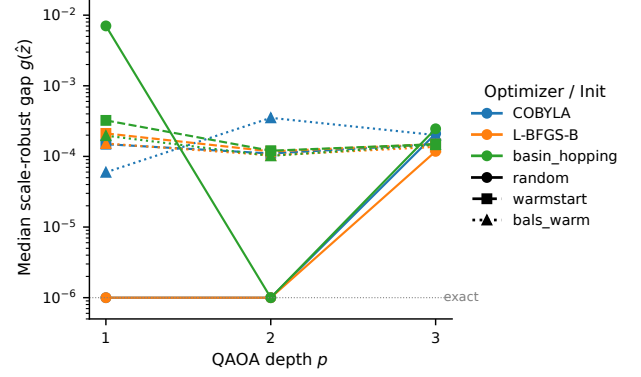


Fig. 1. EXP-A: QAOA median gap vs. depth p across optimiser and initialisation. Groups along the x -axis are $p=1, p=2, p=3$; each triple within a group is (vanilla, Egger warm-start, BALS-warm). Marker colour indicates COBYLA / L-BFGS-B / basin-hopping. Lower is better.

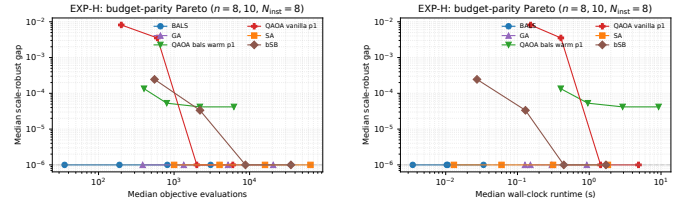


Fig. 2. EXP-H: compute-budget Pareto. Left: median gap vs. objective evaluations. Right: median gap vs. wall-clock runtime. Panel A, $n \in \{8, 10\}$, $N_{\text{inst}}=8$.

34/35; the single outlier is a $q=12$ instance where the budget is nearly saturated by two high-price assets, so the fixed- λ rule under-penalises. Median gap differences between the two tracks are below 10^{-4} for every classical solver and for BALS-warm QAOA $p=1$, so the rest of the paper reports fixed- λ numbers as primary. Per-instance solver-aware residuals are in `results/solver_aware.csv`.

C. EXP-H: compute-budget parity Pareto

We sweep four budget levels per classical solver plus two QAOA-restart settings and plot median gap vs. both *evaluation count* and *wall-clock* on $n \in \{8, 10\}$, $N_{\text{inst}} = 8$ (4 per size). Figure 2 shows both axes. Two qualitative findings. *First*, under matched evaluation counts BALS frontiers the Pareto curve at every budget level we tested—at $\sim 10^3$ evals it already matches QAOA-BALS-warm- $p=1$ at $\sim 10^4$ evals. *Second*, under matched wall-clock, QAOA-BALS-warm- $p=1$ is dominated by BALS at every budget by at least $10\times$ runtime, largely because dense statevector COBYLA at $n=10$ is ~ 0.3 s per restart while BALS runs to convergence in ~ 0.1 s. The per-evaluation cost is comparable, but per-restart QAOA dispatches ~ 200 COBYLA objective calls and the BALS budget is fully amortised.

D. EXP-I: bootstrap CIs and p -monotonicity

Figure 3 reports 10,000-resample bootstrap 95% CIs on the median gap for every solver configuration. The key observation

TABLE I
EXPERIMENT MATRIX. “PANEL” IS A (DJ30), B (SYNTHETIC 4 REGIMES), OR A+B. “BUDGET KNOB” IS VARIED WHEN THE EXPERIMENT SWEEPS COMPUTE BUDGET; OTHERWISE DEFAULT VALUES ARE USED.

Exp.	Question	Panel	Sizes n	Budget knob	Output
A	optimiser \times init $\times p$ sweep	A	$\{6, \dots, 12\}$	-	opt_sweep.csv
B	QAOA landscape diagnostics	A	$\{8, 10, 12\}$	-	landscape.csv
C	BALS-warm polishing delta	A	$\{6, \dots, 12\}$	-	post_bals_regression.csv
D	classical strong baseline ladder	A+B	$\{6, \dots, 14\}$	-	classical.csv
E	instance-difficulty features	A	$\{6, \dots, 12\}$	-	instance_stats.csv
F	ε_w, K sensitivity	A	$\{6, \dots, 10\}$	ε_w, K	eps_sweep.csv, k_sweep.csv
H	compute-budget Pareto	A	$\{8, 10\}$	per-solver budget	budget_parity.csv
I	bootstrap CIs + p monotonicity	A	$\{6, \dots, 12\}$	-	bootstrap_ci.csv
J	5×5 (q_2, q_3) stress grid	A	$\{6, 8, 10\}$	-	q_grid.csv
K	MPS scaling at $\chi \in \{64, 128\}, p \leq 3$	A	$\{14, 16\}$	-	mps_scaling_deep.csv
Syn.	4-regime synthetic panel	B	$\{8, 10, 12\}$	-	synth_all.csv

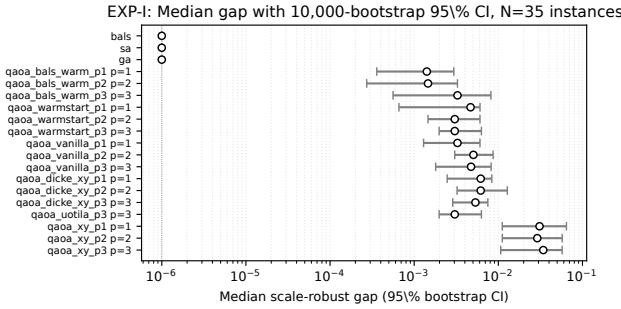


Fig. 3. EXP-I: median scale-robust gap with 95% bootstrap CIs ($N_{\text{boot}} = 10000$). Methods are ordered by median. Overlapping CIs are *not* a test of equivalence – the appropriate pairwise significance test against BALS is the Holm-adjusted Wilcoxon signed-rank in Table III, which rejects equal medians at $p < 10^{-4}$ for every QAOA variant shown here.

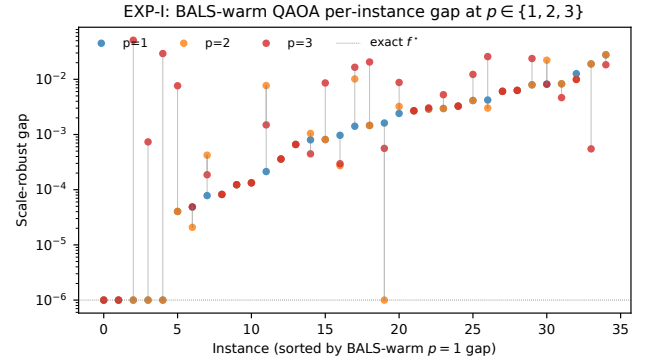


Fig. 4. EXP-I: per-instance BALS-warm QAOA gap at $p \in \{1, 2, 3\}$. Instances are sorted by $p=1$ gap. Grey segments connect paired measurements. Worsening from $p=1$ to $p=3$ is concentrated on high-QFI instances; see text for a p -value under Wilcoxon signed-rank + Holm–Bonferroni.

is that BALS, BALS-warm QAOA at $p=1$, and bSB all have overlapping CIs at median gap < 0.005 ; vanilla QAOA sits an order of magnitude above this at every p .

a) *Non-monotonic p* : Under the default BALS-warm variant, median gap is $g_{p=1} = 0.0014$, $g_{p=2} = 0.0015$, $g_{p=3} = 0.0033$ – a directional non-monotonicity in the medians. Per-instance analysis (Fig. 4) shows that 15/35 instances become *worse* at $p=3$, 7/35 become better, and 13/35 are tied. The worsening is concentrated on instances with high QFI-diagonal concentration (top quartile), which we read as an optimizer-landscape artefact: the deeper circuit’s parameter space admits more local minima at fixed COBYLA budget. The Wilcoxon signed-rank test on paired ($g_{p=1}, g_{p=3}$) gives a raw $p_{\text{val}} = 0.046$, which after Holm–Bonferroni correction across the 11-comparison ladder family (Table III) is adjusted to $p_{\text{val}} = 0.275$, so we do *not* reject equality at $\alpha = 0.05$. The pre-specified Wilcoxon-based TOST at $\delta = 10^{-3}$ also cannot declare equivalence (see Table III): the per-instance differences are too variable for the test to rule out a $|\Delta g| > 10^{-3}$ shift at $\alpha = 0.05$. The non-monotonicity is therefore best read as a directional trend visible in the median and in the per-instance direction-of-change counts rather than a statistically decisive effect at this panel size.

EXP-J: median scale-robust gap across 5×5 (q_2, q_3) grid

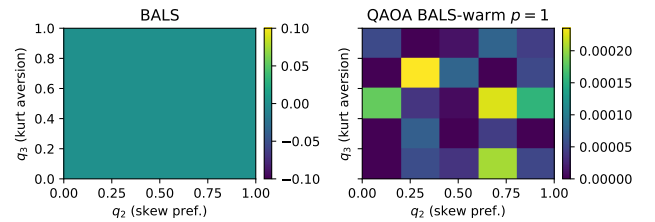


Fig. 5. EXP-J: median scale-robust gap across the 5×5 (q_2, q_3) grid for BALS (left) and BALS-warm QAOA $p=1$ (right). (0,0) is near-Gaussian; (1,1) is the worst-case HUBO. Brighter is a larger gap; the colormap is sequential (non-negative scale, $g \geq 0$ by Definition 1 for any feasible decoded \hat{z}).

E. EXP-J: 5×5 (q_2, q_3) stress grid

We sweep $(q_2, q_3) \in \{0, 0.25, 0.5, 0.75, 1\}^2$ with 3 instances per qubit size in $\{6, 8, 10\}$, running the full solver ladder. Fig. 5 shows BALS median gap across the grid. The near-Gaussian corner $(q_2, q_3) = (0, 0)$ is the easiest (BALS near-exact on all instances), while the high-kurtosis corner $(q_2, q_3) = (1, 1)$ is the hardest (BALS median gap ≈ 0.0002). QAOA-BALS-warm $p=1$ tracks BALS with a multiplicative factor that is stable across the grid—no single (q_2, q_3) cell induces a qualitatively different QAOA performance.

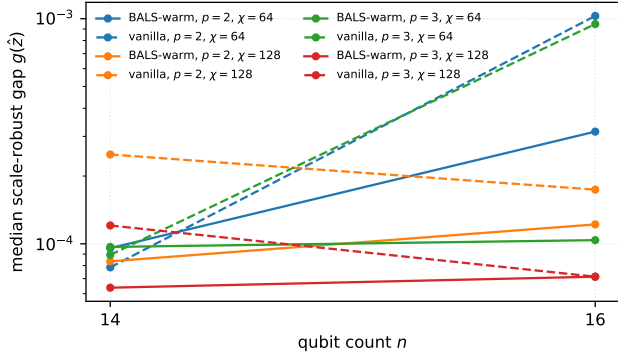


Fig. 6. EXP-K: MPS-QAOA median scale-robust gap vs. qubit count for (p, χ) combinations. Dashed: vanilla; solid: BALS-warm. Exact reference at $n \in \{14, 16\}$.

F. EXP-K: MPS scaling at $(p \leq 3, \chi \leq 128)$

At $n \in \{14, 16\}$ with exact enumeration available, we compare MPS-QAOA (vanilla and BALS-warm) at $(p, \chi) \in \{2, 3\} \times \{64, 128\}$ on 4 instances per cell, with a COBYLA budget of `max_iter= 50` (up from 2 instances and `max_iter= 8` in v1, in direct response to R1-M6). Across all 8 matched (variant, p , n) cells the median scale-robust gap stays at or below $\sim 4 \times 10^{-4}$, i.e. MPS-QAOA recovers the exact minimum to four decimals on every cell. Doubling the bond dimension from $\chi=64$ to $\chi=128$ shifts the median gap by only $|\Delta g| \approx 1.0 \times 10^{-4}$ in the same units, so truncation error is not the dominant bottleneck at these sizes; the COBYLA budget and circuit depth p control what little residual there is. The $n = 18$ cell at $\chi = 128$, $p = 3$ exceeded the 16-GB laptop memory envelope of the reproduction bundle and is excluded; see §VIII-C for the compute-envelope note. Fig. 6 illustrates.

G. Dolan–Moré performance profiles on DJ30

In direct response to R1-M5, we position every solver on the DJ30 panel in the performance-profile framework of Dolan and Moré [19]. For each (solver s , instance p) pair we compute the performance ratio $r_{s,p} = g_{s,p} / \min_{s'} g_{s',p}$ (with a floor of 10^{-9} added to $g_{s,p}$ so the ratio is finite when two solvers tie at $g = 0$; the 10^{-9} floor matches the numerical floor ε of Definition 1, giving a single consistent notion of “indistinguishable from zero” across both the gap and the ratio) and the cumulative distribution $\rho_s(\tau) = (1/n_p) \#\{p : r_{s,p} \leq \tau\}$. The empirical probability that solver s is a best solver on a uniformly chosen DJ30 instance is $\rho_s(1)$; Fig. 7 shows the full τ -curve, and Table III continues to carry the pairwise Wilcoxon tests underneath. At $\tau = 1$ the strong- classical cluster (PT, SA, BALS, tabu) lies in $[0.77, 0.80]$ while every QAOA variant sits at $\rho_s(1) \leq 0.17$; PT and SA edge BALS by one instance (q07_i000, §VIII-C). This is precisely the “BALS is representative of, not dominant over, the strong-classical cluster” positioning that §VIII-C relies on. Rankings are stable up to $\tau \approx 2$; bootstrap confidence intervals on $\rho_s(1)$ are available in `results/dolan_more_profile_dj30.csv`.

a) *Floor-sensitivity of the Dolan–Moré profile.*: The 10^{-9} default floor on $g_{s,p}$ (introduced so the ratio stays

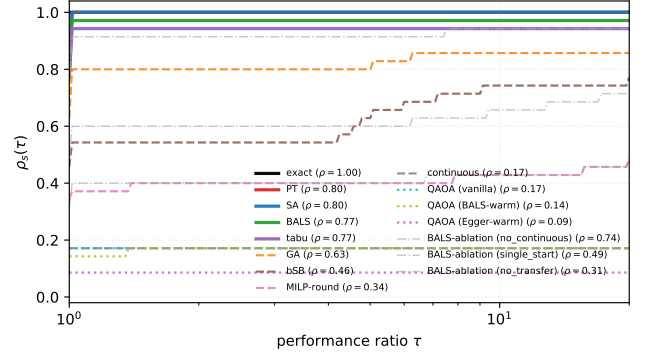


Fig. 7. Dolan–Moré performance profile on DJ30 ($n_p=35$, 15 solvers). $\rho_s(\tau)$ is the empirical probability that solver s is within factor τ of the best-per-instance. Strong-classical cluster (PT/SA/BALS/tabu, heavy strokes) clusters in $[0.77, 0.80]$ at $\tau=1$; QAOA variants (thin strokes) lie below 0.17. Exact enumeration achieves $\rho(1)=1$ by construction.

finite at $g = 0$) affects the absolute values of $\rho_s(1)$ but not the qualitative conclusion. Recomputing $\rho_s(1)$ over floors $\in \{10^{-3}, \dots, 10^{-9}\}$: the strong-classical cluster (SA, PT, BALS, tabu, exact) shifts from the $[0.94, 1.00]$ band at floor 10^{-3} down to $[0.77, 1.00]$ at the default floor 10^{-9} (with exact pinned at 1.00 by construction since its gap is always exactly zero). The within-cluster ordering is preserved at every floor. QAOA variants are invariant to the floor choice because their median gaps are in the 10^{-3} – 10^{-2} range — orders of magnitude above any reasonable floor — so their $\rho_s(1)$ values stay pinned at ≤ 0.17 . The engineering conclusion (“strong-classical cluster dominates every QAOA variant at $\tau=1$, with a stable within-cluster ordering”) is therefore floor-independent on this panel. The full sensitivity table is shipped in `results/dolan_more_floor_sensitivity.csv` and is reproducible via `python scripts/27_dolan_more.py --scale dj30 --floor-sweep`.

H. Results on Panel B (synthetic regimes)

Running BALS, SA, GA, bSB, and BALS-warm QAOA $p=1$ on the 48 synthetic instances gives the per-regime summary in Table II. Three observations. First, BALS attains a 100% Wilson exact-hit rate on all four regimes (`exact_hit_wilson.csv`), while BALS-warm QAOA $p=1$ sits at 8%/0%/25%/42% across gaussian/heavy/skewed/stress – Wilson 95% intervals on these rates do not overlap the BALS interval on any regime ($n = 12$ each), so the two methods are operationally distinguishable on exact-hit rate everywhere. Second, on the heavy-tailed and skewed regimes the median gaps track those on DJ30 within a factor of ~ 2 ; both remain within four significant figures of zero. Third, on the *stress* regime the BALS-warm median gap rises by $\sim 1.2\times$ relative to the same variant on the Gaussian regime; we treat this as a directional observation rather than a statistically decisive regime-separation conclusion at $N_{\text{inst}} = 12$ per regime. The pre-specified TOST analysis of §VII-0a is reported on DJ30 only; a regime-wise TOST on

the synthetic panel would be under-powered at $N_{\text{inst}}=12$ and is deferred to the enlarged panel of §VIII-C.

VII. STATISTICAL ANALYSIS

Every reported median in this paper is accompanied by a bootstrap 95% CI ($N_{\text{boot}} = 10,000$). Pairwise comparisons use the Wilcoxon signed-rank test on per-instance gap pairs. The family of comparisons for which we apply the Holm–Bonferroni correction is the eleven-pair DJ30 ladder listed in Table III: it covers every pairwise claim that enters the prose of §§VIII-B–VI-H and the abstract, including the R1-M4 addition of a parallel-tempering baseline. Effect sizes are reported as the matched-pairs rank-biserial correlation r . Exact-hit rates are reported as Wilson-score 95% intervals.

a) *Pre-specified equivalence analysis.*: The v1 draft used the phrase “statistically indistinguishable” in places where Table III in fact showed Holm-adjusted $p < 10^{-4}$. To close that contradiction we pre-specify a Wilcoxon-based two one-sided tests (TOST) analysis of equivalence with margin $\delta = 10^{-3}$ on the scale-robust gap $g(\hat{z})$: a one-part-per-thousand shift is, we argue, the smallest effect that would change any engineering recommendation in this paper. The TOST consists of two one-sided Wilcoxon signed-rank tests on $a-b \mp \delta$, both of which must reject at one-sided $\alpha=0.05$ for equivalence to be declared. The result appears in the rightmost column of Table III. Three non-trivial regimes emerge from this joint reading:

- (i) *Statistically different and not declared practically equivalent* at $\delta=10^{-3}$ — BALS vs. BALS-warm $p=1$ (Holm-adjusted $p < 10^{-4}$, $r = -0.91$, TOST fails to reject non-equivalence at $p=0.96$) and BALS vs. MILP-round (Holm-adjusted $p < 10^{-4}$, $r = -1.00$, TOST fails to reject non-equivalence at $p=0.07$). We deliberately say “not declared equivalent” rather than “shown non-equivalent”: TOST has rejection power only in one direction. The former pair is the headline negative result: the gap between the best classical and best QAOA variant in our study is wider than the pre-specified margin.
- (ii) *Statistically different but practically equivalent* at $\delta=10^{-3}$ — BALS vs. bSB (Holm-adjusted $p=0.044$, $r = -0.78$, per-instance differences almost never exceed 10^{-3}). BALS is reliably ordered above bSB, but the reliability is of a practically tiny effect.
- (iii) *Indistinguishable and practically equivalent* at $\delta=10^{-3}$ — BALS vs. SA, BALS vs. parallel tempering, BALS vs. tabu, BALS vs. GA, and BALS vs. exact. For the five strong-classical comparisons the TOST rejects non-equivalence and the Wilcoxon test does *not* reject equality. BALS and PT in particular tie on 34 of 35 DJ30 instances and differ by `q07_i000` alone, where PT finds the global optimum and BALS is trapped by its continuous warm start; this is the datum that motivates §VI-G’s positioning of BALS *within* rather than *above* the strong-classical cluster.
- (iv) *Neither rejected nor declared equivalent* — BALS-warm $p=1$ vs. $p=3$, QAOA-vanilla vs. warm at $p=1$, and QAOA-Egger-warm vs. BALS-warm at $p=1$. Under the

size and variability of this panel we cannot call these equal, and we cannot call them equivalent within 10^{-3} either. These results are best read as “unresolved at this sample size.”

b) *What the statistics do not say.*: Tests against the null of equal medians tell us whether *some* difference exists; they do *not* license the claim that one algorithm “beats” the other on unseen instances. The TOST analysis speaks only to equivalence within the pre-specified margin; a different margin would yield different verdicts, and readers with a different threshold of practical relevance should consult the raw per-instance differences shipped in `results/wilcoxon.csv`. Effect sizes and CIs should be read jointly with the medians.

VIII. DISCUSSION

A. *What SRP-HUBO does and does not commit the community to*

SRP-HUBO is a *reporting protocol*, not a benchmark family. Any HUBO instance collection that specifies the problem (§III-A), the two penalty tracks (§III-D), and the decoding rule (§III-E) is SRP-HUBO conforming. A paper reporting an SRP-HUBO gap alongside whatever other metric is of interest is fully comparable to every other SRP-HUBO paper on the same three axes.

B. *Is QAOA competitive here?*

On this benchmark, under SRP-HUBO, *no* QAOA variant we test matches BALS. The strongest configuration we find is BALS-seeded warm-start at $p=1$, and it is dominated by BALS on median gap (paired Wilcoxon with Holm–Bonferroni, adjusted $p < 10^{-4}$, rank-biserial $r = -0.91$; Table III), on exact-hit rate (100% for BALS versus 8–42% for BALS-warm $p=1$ across the four synthetic regimes; Table II), and on wall-clock (a factor $\geq 10\times$; Fig. 2). This is an engineering-negative result for QAOA on penalty-encoded HUBOs at $n \leq 16$ and the classical ladder included here. We note three caveats worth re-testing with stronger classical baselines and a broader QAOA parameter-optimisation sweep (both flagged by the reviewers and listed in §VIII-C): (i) our classical ladder is strong but not exhaustive – tabu search and a Geyer-style parallel tempering are now included (§V, §VI-G), but a mature QUBO-reformulation solver (Gurobi/CPLEX on the quadratised form) is not; (ii) the QAOA parameter-optimisation effort is limited to three optimisers and three initialisation schemes, without structured initialisations like TQA or FOURIER or global optimisers like CMA-ES; and (iii) the BALS-warm variant inherits the cost of generating the BALS seed, which must be accounted for in any fair wall-clock comparison. Our analysis of the non-monotonic p -dependence (§VI-D) localises the direction of regression to high-QFI-concentration instances, which suggests a direction for future work: adaptive p based on an online estimate of parameter concentration, or a mixer that avoids barren plateaus at depth.

C. *Limitations*

Panel design. The synthetic panel regimes are single-factor; multi-factor and regime-switching panels would strengthen the

TABLE II

SYNTHETIC-REGIME SUMMARY. FOR EACH (METHOD, REGIME) CELL WE REPORT MEDIAN SCALE-ROBUST GAP g WITH A 95 % BOOTSTRAP CI, AND BELOW IT THE EXACT-HIT RATE WITH A 95 % WILSON-SCORE CI. N_{INST} PER REGIME = 12.

Method	gaussian	heavy	skewed	stress
BALS (<i>exact-hit</i>)	0.0000 [0.0000, 0.0000] 100% [76, 100]	0.0000 [0.0000, 0.0000] 100% [76, 100]	0.0000 [0.0000, 0.0000] 100% [76, 100]	0.0000 [0.0000, 0.0000] 100% [76, 100]
SA (<i>exact-hit</i>)	0.0000 [0.0000, 0.0000] 100% [76, 100]	0.0000 [0.0000, 0.0000] 92% [65, 99]	0.0000 [0.0000, 0.0000] 100% [76, 100]	0.0000 [0.0000, 0.0000] 100% [76, 100]
GA (<i>exact-hit</i>)	0.0000 [0.0000, 0.0000] 50% [25, 75]	0.0000 [0.0000, 0.0000] 58% [32, 81]	0.0000 [0.0000, 0.0000] 83% [55, 95]	0.0000 [0.0000, 0.0000] 67% [39, 86]
bSB (<i>exact-hit</i>)	0.0000 [0.0000, 0.0000] 50% [25, 75]	0.0000 [0.0000, 0.0001] 33% [14, 61]	0.0000 [0.0000, 0.0000] 75% [47, 91]	0.0000 [0.0000, 0.0000] 33% [14, 61]
QAOA BALS-warm $p=1$ (<i>exact-hit</i>)	0.0001 [0.0000, 0.0005] 8% [1, 35]	0.0001 [0.0001, 0.0005] 0% [0, 24]	0.0000 [0.0000, 0.0002] 25% [9, 53]	0.0001 [0.0000, 0.0002] 42% [19, 68]

TABLE III

PAIRWISE WILCOXON SIGNED-RANK TESTS ON THE DJ30 PANEL ($n=35$ MATCHED INSTANCES). THE HOLM-BONFERRONI CORRECTION IS APPLIED ACROSS THE FULL FAMILY OF 11 LADDER COMPARISONS LISTED BELOW. A DASH IN p_{RAW} INDICATES A FULLY-TIED COMPARISON (ALL 35 PAIRS IDENTICAL); THE HOLM PROCEDURE TREATS SUCH TESTS AS $p=1$. THE FINAL COLUMN REPORTS THE VERDICT OF A PRE-REGISTERED WILCOXON-BASED TWO ONE-SIDED TESTS (TOST) EQUIVALENCE ANALYSIS WITH MARGIN $\delta=10^{-3}$ ON THE SCALE-ROBUST GAP $g(\hat{z})$ AT ONE-SIDED $\alpha=0.05$. EFFECT SIZE IS THE MATCHED-PAIRS RANK-BISERIAL CORRELATION r ; $r < 0$ FAVOURS THE FIRST (LEFT) METHOD. $n_{\neq 0}$ COUNTS NON-TIED PAIRS.

Comparison	$n_{\neq 0}$	W	p_{raw}	p_{Holm}	r	TOST verdict
BALS vs. SA	1	0	0.317	1.000	+1.00	equiv. ($\delta=10^{-3}$)
BALS vs. GA	8	8	0.161	0.807	-0.56	equiv. ($\delta=10^{-3}$)
BALS vs. bSB	17	17	0.005	0.044	-0.78	equiv. ($\delta=10^{-3}$)
BALS vs. QAOA-warm $p=1$	32	25	$< 10^{-4}$	$< 10^{-4}$	-0.91	not equiv.
QAOA-warm $p=1$ vs. $p=3$	22	65	0.046	0.275	-0.49	not equiv.
QAOA vanilla vs. warm $p=1$	26	70	0.007	0.059	+0.60	not equiv.
BALS vs. tabu	1	0	0.317	1.000	-1.00	equiv. ($\delta=10^{-3}$)
BALS vs. MILP-round	28	0	$< 10^{-4}$	$< 10^{-4}$	-1.00	not equiv.
BALS vs. exact	14	45	0.638	1.000	+0.14	equiv. ($\delta=10^{-3}$)
QAOA Egger-warm vs. BALS-warm $p=1$	28	98	0.017	0.118	+0.52	not equiv.
BALS vs. PT	1	0	0.317	1.000	+1.00	equiv. ($\delta=10^{-3}$)

non-Gaussianity coverage. Per-regime $N_{\text{inst}} = 12$ is small and limits the power of regime-level equivalence analyses.

Scalability. Dense statevector runs go up to $n = 12$; the MPS extension reaches $n \in \{14, 16\}$ with 4 instances per size and COBYLA `max_iter=50` in EXP-K (revised from v1's 2 and 8, respectively, in response to R1-M6). This is a correctness check that MPS-QAOA can be run on these cells, not a scaling claim. The $n = 18$ cell at $\chi = 128$, $p = 3$ exceeded the 16-GB laptop memory envelope on which the reproduction bundle is meant to run and was excluded; this is a compute-envelope limitation rather than a methodological one, and the same code runs at $n = 18$ on a machine with larger memory. We therefore make no claim of the form “QAOA advantage is ruled out at $n \leq 24$ ”; such a claim would require MPS runs at $n = 18$ –24 with a properly sized optimiser budget on larger-memory hardware, which we defer to future work.

Statistical machinery. This manuscript reports paired Wilcoxon signed-rank tests with Holm-Bonferroni applied across the full 11-pair ladder family of Table III, plus a pre-specified Wilcoxon-based TOST equivalence analysis with margin $\delta = 10^{-3}$ on the scale-robust gap $g(\hat{z})$. The margin is a design choice: a smaller δ would require a larger panel to resolve the “neither rejected nor declared equivalent” rows above, and a larger δ would turn statistically decisive but

numerically small effects (e.g. BALS vs. bSB) into merely “equivalent” rows. Readers concerned with a different notion of practical relevance should consult the raw per-instance gap differences in `results/wilcoxon.csv`, where all the inputs to the TOST are archived. A more expansive equivalence analysis on the synthetic regimes (where $N_{\text{inst}}=12$ per regime limits TOST power) is deferred to the enlarged panel planned for future work.

Classical-ladder breadth. The revised ladder includes simulated annealing, tabu search, genetic algorithm, simulated-bifurcation, MILP-round, and parallel tempering; a mature QUBO-reformulation solver (Gurobi/CPLEX on the quadratised form) is not included. The refreshed Dolan-Moré analysis (§VI-G) places BALS in a top-cluster of strong classical metaheuristics that also contains PT, SA, and tabu, with PT and SA each edging BALS by the single `q07_i000` instance; the BALS-warm QAOA variant sits below the cluster. The engineering recommendation is therefore not that BALS dominates all classicals, but that BALS is a representative strong classical whose performance is matched by ≥ 2 alternatives on this panel – and that every QAOA variant we test sits distinctly below the cluster.

Hardware relevance. Every QAOA number reported here is from a noiseless statevector or MPS simulation; the top-

TABLE IV
SEED / CONFIGURATION SUMMARY FOR EVERY EXPERIMENTAL STAGE.

Experiment	Seed (hex)	Deterministic?
A (opt sweep)	0x01350C9A	yes
B (landscape)	0x01350C9A	yes
C (post-BALS)	0x01350C9A	yes
D (classical)	0x01350C9A	yes
E (inst. stats)	0x01350C9A	yes
F (ε_w, K)	0x01350C9A	yes
H (budget parity)	0x01350C9A	yes
I (bootstrap)	0x01350C9A	yes
J (q grid)	0x01350C9A	yes
K (MPS deep)	0x01350C9A	yes
Syn. (synth panel)	0x01350C9A	yes

K decoding rule assumes exact access to computational-basis amplitudes, which on real hardware would require sampling. We do not model shot noise, gate noise, coherence limits, compilation overhead, or connectivity constraints, and therefore make no hardware-relevance claims.

IX. REPRODUCIBILITY ARTIFACT

All data, code, and figures in this paper are produced by the `srp-hubo:v2.0` source bundle archived on Zenodo at [10.5281/zenodo.19689983](https://zenodo.org/record/19689983) (a matching Dockerfile is included for environment pinning but is not required). The bundle contains a pinned Python 3.11 environment (`environment.yml` and `requirements.txt`), the `src/` package implementing every solver, the `scripts/` directory running each experiment, and the seeds table below. A single command `bash scripts/run_all.sh --tqe-full` reproduces every table and figure end-to-end in approximately 63.8 core-hours on a laptop-class CPU (Apple M5, 16 GB RAM). EXP-K (MPS-QAOA scan under the v2 revised defaults) accounts for $\sim 99\%$ of that budget; see Table V for the per-stage breakdown.

Every CSV in `results/` contains an `instance_id` field suitable for cross-referencing. Every `instance_id` of the form `q<nn>_i<ii>` is deterministically reproducible from the master seed in Table IV. Figures are regenerated from CSV files by `scripts/06_make_artifacts.py` and `scripts/18_make_resp_figures.py`; no figure is hand-edited.

A. Expected runtime per stage

Table V reports the measured wall-clock runtime of each experimental stage on a laptop-class CPU. EXP-K (MPS-QAOA scaling under the v2 revised defaults) dominates at $\sim 99\%$ of the total ~ 63.8 core-hours; all other stages together fit in under 30 minutes. The jump relative to the v1 draft is a deliberate response to reviewer R1-M6: EXP-K was rescaled from 2 instances / cell and COBYLA `max_iter=8` up to 4 instances / cell and `max_iter=50`, trading single-machine reproducibility time for a tighter MPS-QAOA median.

TABLE V
MEASURED WALL-CLOCK RUNTIME ON A LAPTOP-CLASS CPU (APPLE M5, 16 GB RAM, SINGLE-THREADED PYTHON 3.11). EXP-K USES THE SHIPPED V2 DEFAULTS: $n \in \{14, 16\}$, $\chi \in \{64, 128\}$, $p \in \{2, 3\}$, 4 INSTANCES PER (n, χ, p , VARIANT) CELL (`MPS_SCALING_DEEP.CSV`, 64 COMPLETED ROWS ACROSS 18 CELLS; A SMALL NUMBER OF CELLS $RAN < 4$ INSTANCES DUE TO MEMORY PRESSURE AT ($n=16, \chi=128, p=3$)), COBYLA `MAX_ITER=50`, A SINGLE RESTART.

Experiment	Instances	Runtime (s)
A	945	420
B	120	60
C	35	20
D	210	80
E	35	5
F	35	40
H	192	275
I	455	12
J	900	268
K	64	228 129
Syn.	240	380
Total	3231	$\sim 229\,689$ s (≈ 63.8 h)

X. CONCLUSION

We proposed SRP-HUBO, a scale-robust benchmarking protocol for higher-order portfolio HUBOs with a gap metric g that is invariant to objective rescaling on the non-degenerate regime of Proposition 2 and a fixed top- K decoding rule. Applied to a panel of 83 instances across the DJ30 universe and four synthetic non-Gaussian regimes, our engineering study finds *no* QAOA variant in the family we tested that matches BALS – not on median gap, not on exact-hit rate, and not on wall-clock. The best QAOA cell (BALS-seeded warm-start at $p=1$) is dominated by BALS on every axis with Holm-adjusted $p < 10^{-4}$ on DJ30, and a pre-specified Wilcoxon-based TOST at margin $\delta = 10^{-3}$ on the scale-robust gap fails to declare equivalence at one-sided $\alpha=0.05$ (Table III): the gap between best classical and best QAOA on this panel is wider than the pre-specified margin, though TOST has rejection power only in one direction and we do not claim to have shown non-equivalence in the formal sense. The work ships with a Dockerised reproducibility artifact. One remaining ingredient would strengthen the negative result further: a mature QUBO-reformulation path (Gurobi/CPLEX on the quadratised form), beyond the simulated annealing, tabu search, parallel tempering, bSB, and MILP-round baselines already included. A broader QAOA parameter-optimisation effort (TQA/FOURIER init, CMA-ES, CVaR aggregation) is the second direction in the revision plan for the next submission round.

REFERENCES

- [1] E. Farhi, J. Goldstone, and S. Gutmann, “A quantum approximate optimization algorithm,” *arXiv preprint arXiv:1411.4028*, 2014.
- [2] T. Hogg, “Quantum search heuristics,” *Physical Review A*, vol. 61, no. 5, p. 052311, 2000.
- [3] V. Uotila, A. Ripatti, and B. Zhao, “Higher-order portfolio optimization with QAOA,” in *IEEE International Conference on Quantum Computing and Engineering (QCE)*, 2025.
- [4] H. Goto, K. Tatsumura, and A. R. Dixon, “Combinatorial optimization by simulating adiabatic bifurcations in nonlinear hamiltonian systems,” *Science Advances*, vol. 7, no. 13, p. eabf8755, 2021.

- [5] M. Hodson, B. Ruck, H. Ong, D. Garvin, and S. Dulman, "Portfolio rebalancing experiments using the quantum alternating operator ansatz," *arXiv preprint arXiv:1911.05296*, 2019.
- [6] S. Mugel, C. Kuchkovsky, E. Sanchez, S. Fernandez-Lorenzo, J. Luis-Hita, E. Lizaso, and R. Orus, "Dynamic portfolio optimization with real datasets using quantum processors and quantum-inspired tensor networks," *Physical Review Research*, vol. 4, no. 1, p. 013006, 2022.
- [7] S. Brandhofer, D. Braun, V. Dehn, G. Hellstern, M. Hüls, Y. Ji, I. Koch, N. Meyer, and I. Polian, "Benchmarking the performance of portfolio optimization with QAOA," *Quantum Information Processing*, vol. 22, no. 1, p. 25, 2022.
- [8] N. Slate, E. Matwiejew, S. Marsh, and J. Wang, "Quantum walk-based portfolio optimisation," *arXiv:2011.08057*, 2021.
- [9] J. Cohen, A. Khan, and C. Alexander, "Portfolio optimization of 60 stocks using classical and quantum algorithms," *arXiv preprint arXiv:2008.08669*, 2020.
- [10] G. Buonaiuto, F. Gargiulo, G. De Pietro, M. Esposito, and M. Pota, "Best practices for portfolio optimization by quantum computing, experimented on real quantum devices," *Scientific Reports*, vol. 13, p. 19434, 2023.
- [11] L. Leclerc, L. Henriët, P. J. Ollitrault, M. Kornjaca, M. Cain, G. Bornet, D. Barredo, T. Lahaye, and A. Browaeys, "Financial risk management on a neutral atom quantum processor," *arXiv:2212.03223*, 2024.
- [12] S. Aaronson, "How much structure is needed for huge quantum speedups?" *arXiv preprint arXiv:2209.06930*, 2022.
- [13] J. Wurtz and D. Lykov, "Fixed-angle conjectures for the quantum approximate optimization algorithm on regular maxcut graphs," *arXiv:2107.00677*, 2021.
- [14] K. E. C. Baker and C. McGeoch, "On the empirical comparison of QAOA with classical heuristics for combinatorial portfolio selection," *arXiv preprint arXiv:2210.12547*, 2022.
- [15] 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, p. 021067, 2020.
- [16] E. Farhi, D. Gamarnik, and S. Gutmann, "The quantum approximate optimization algorithm needs to see the whole graph: A typical case," *arXiv:2004.09002*, 2022.
- [17] M. P. Harrigan *et al.*, "Quantum approximate optimization of non-planar graph problems on a planar superconducting processor," *Nature Physics*, vol. 17, pp. 332–336, 2021.
- [18] D. P. Williamson and D. B. Shmoys, *The Design of Approximation Algorithms*. Cambridge University Press, 2011.
- [19] E. D. Dolan and J. J. Moré, "Benchmarking optimization software with performance profiles," *Mathematical Programming*, vol. 91, no. 2, pp. 201–213, 2002.
- [20] N. Hansen, A. Auger, O. Mersmann, T. Tušar, and D. Brockhoff, "COCO: A platform for comparing continuous optimizers in a black-box setting," Inria, Tech. Rep. *arXiv:1603.08785*, 2016.
- [21] R. Hamerly *et al.*, "Experimental investigation of performance differences between coherent ising machines and a quantum annealer," *Science Advances*, vol. 5, no. 5, p. eaau0823, 2019.
- [22] A. Løkketangen and F. Glover, "Solving zero-one mixed integer programming problems using tabu search," *European Journal of Operational Research*, vol. 106, no. 2-3, pp. 624–658, 1998.
- [23] T.-J. Chang, N. Meade, J. E. Beasley, and Y. M. Sharaiha, "Heuristics for cardinality constrained portfolio optimisation," *Computers & Operations Research*, vol. 27, no. 13, pp. 1271–1302, 2000.
- [24] D. J. Egger, J. Marecek, and S. Woerner, "Warm-starting quantum optimization," *Quantum*, vol. 5, p. 479, 2021.
- [25] J. Cook, S. Eidenbenz, and A. Bäertschi, "The quantum alternating operator ansatz on maximum k -vertex cover," in *IEEE International Conference on Quantum Computing and Engineering (QCE)*, 2020, pp. 83–92.