



Paper Type: Original Article

Adaptive Operator Portfolio Search with Fold Based Landscape Probing

Mei-Lin Nakamura^{1,*}, Pavel D. Sokolov², Adaeze O. Okonkwo³

¹ Department of Computer Sciences, Northeastern University, Boston, MA, USA; m.nakamura@northeastern.edu.

² Department of Electrical and Computer Engineering, Cyprus University of Technology, Limassol, Cyprus; m.nakamura@northeastern.edu.

³ Department of Computer Science, University of Cape Town, Cape Town, South Africa; m.nakamura@northeastern.edu.

Citation:

Received: 24 February 2026

Revised: 14 April 2026

Accepted: 17 July 2026

Nakamura, M. L., Sokolov, P. D., & Okonkwo, A. O. (2026). Adaptive operator portfolio search with fold based landscape probing. *Optimality*, 3(3), 213-233.

Abstract

Adaptive operator selection allocates search effort among candidate mechanisms whose usefulness changes with the current population. We study AOP-KFP, a portfolio of nine operators with periodic probing of held-out parents, exponentially smoothed improvement credits, and a softmax distribution mixed with uniform exploration. The folds organize local comparisons within a population snapshot; they do not establish statistical generalization to unseen problems. We give an explicit accounting of probe evaluations and a conditional hitting-probability result for elitist search, while distinguishing these statements from finite-budget performance. The supplied CEC 2017 table contains 25 functions at dimension 30 with 51 reported runs per algorithm. Its displayed means give AOP-KFP 17 row minima, jSO three, EBOwithCMA four, and L-SHADE one. Additional supplied rank, significance, ablation and sensitivity summaries require reconciliation with their underlying observations: several corrected-p-value entries are arithmetically inconsistent, and the reported Sobol first-order indices sum to more than one. The study therefore identifies a testable operator-allocation mechanism and descriptive benchmark evidence, with evaluation accounting and raw-output reconciliation required before confirmatory superiority or reproducibility claims can be made.

Keywords: Adaptive operator selection, Population search, Held-out parent probing, Function-evaluation budget, Benchmark comparison.

1 | Introduction

Continuous black-box optimization often requires allocating a limited evaluation budget among search mechanisms with different behaviors. A differential-evolution move, a covariance-adaptation step and a heavy-tailed mutation can be useful at different stages of the same run. Adaptive operator selection separates

✉ Corresponding Author: m.nakamura@northeastern.edu

doi <https://doi.org/10.22105/opt.v3i3.116>



Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

the attribution of improvement from the rule that allocates future trials [1], [2]. Benchmark comparisons also depend on the chosen problems, performance summaries and independent runs [3–6]. This motivates studying an allocation mechanism together with its accounting and evidential limits. Recent metaheuristic surveys catalogue a diverse set of search designs [7]. This paper targets the bound-constrained continuous minimization problem of the form:

$$\text{minimize}_{\{x \in \Omega \subseteq \mathbb{R}^d\}} f(x), \quad (1)$$

AOP-KFP evaluates operators on held-out parents from a frozen population snapshot and uses their improvements to update selection credits. The complementary folds provide context for operator state or trial rollouts. Because all folds come from the same evolving population, the resulting score is a local improvement estimate rather than an estimate of out-of-problem generalization. The present results concern fixed, deterministic objectives. Noisy, drifting and multi-objective cases are proposed extensions whose validation is outside the supplied main experiment.

The design combines a heterogeneous operator set, recent-improvement credits and an explicit probability floor. Heterogeneity is an architectural choice, not a proved complementarity property. The floor preserves opportunities to select every operator, but selection opportunities alone do not imply global search coverage. An asymptotic statement additionally needs a proposal kernel that can reach useful objective sublevel sets.

1.1 | Contribution and Evidence Scope

The central methodological contribution is the combination of periodic fold-based operator probing with smoothed credit allocation. Its value must be assessed against the evaluations consumed by those probes, rather than by comparing an equal number of generations.

The formulation specifies the parent-level improvement being measured, the distinction between uniform mixing and entropy regularization, and a budget-aware algorithm skeleton. Constraint penalties and noisy or Pareto-aware variants are separated from the bound-constrained experiment because their feasibility and performance have not been established by its tables.

The theoretical contribution is a conditional probability-of-hitting argument for an elitist incumbent. It does not supply a universal finite-time optimization rate, an exponential-weights regret guarantee, or local stability of the random population dynamics. These stronger conclusions do not follow from the stated credit update or covariance bound.

The empirical contribution consists of the supplied per-function results, original figures and auxiliary summaries. The 25 displayed CEC 2017 functions provide descriptive comparisons. The reported supplementary CEC 2022 and BBOB campaigns, raw logs, tuning traces and uncertainty calculations are not available in the materials supporting this revision and remain evidence dependencies.

A verifiable release must identify the implementation, operator definitions, seed mapping, actual function-evaluation counts, configurations and statistical input arrays. The source archive identifier has not been established as an artifact for this study; reproducibility is therefore a release requirement rather than a completed finding.

The following sections position the mechanism, define its credits and budget accounting, establish the scope of the conditional theory, and interpret the supplied results with their unresolved statistical and provenance requirements.

2 | Background and Related Work

2.1 | Adaptive Operator Selection

Fialho et al. [1] study bandit-based adaptive operator selection, while Thierens [2] develops adaptive pursuit with explicit allocation probabilities. These precedents already address the interaction between reward attribution and continued exploration. AOP-KFP differs in its proposed use of periodic held-out-parent comparisons within an evolving population. It cannot claim that adaptive pursuit invariably loses all exploration, nor that folds drawn from one population have the interpretation of independent cross-validation datasets.

2.2 | Algorithm Portfolios and Online Learning

An algorithm portfolio allocates trials among mechanisms rather than selecting a single mechanism in advance. In this study, allocation is based on exponentially smoothed, contemporaneous probe rewards. It is useful to distinguish this local adaptation from cumulative exponential-weights updates and from bandit rules with proved regret guarantees. Automatic configuration tools such as sequential model-based optimization [8] and irace [9] address a related allocation problem at the level of complete algorithm configurations.

2.3 | Differential Evolution and the SHADE Family

The differential-evolution family offers parameter adaptation and population-size control as alternatives to a fixed search strategy. L-SHADE uses linear population-size reduction, while NL-SHADE-RSP uses nonlinear reduction and rank-based selective pressure [10], [11]. jSO and EBOwithCMA are included as named comparator implementations. Their version identifiers and settings are necessary to interpret the comparison; the source names alone do not determine their executed behavior.

2.4 | Covariance Matrix Adaptation Evolution Strategy

Covariance matrix adaptation changes a search distribution using information from selected candidates. A full covariance update has different arithmetic and state costs from a coordinatewise mutation, even when each produces the same number of objective evaluations. A single covariance step inside a portfolio is also not automatically equivalent to a complete CMA-ES implementation with its own adaptation state and restart policy.

2.5 | Landscape Analysis

Population-level improvement measurements provide local feedback about the current search state. They differ from a general landscape characterization or an independent validation study. COCO provides standardized benchmark definitions and analysis infrastructure [12]. Invoking that infrastructure does not establish that the supplementary BBOB experiment was executed or that its results are available for inspection.

2.6 | CEC Benchmark Series

CEC 2017 and CEC 2022 provide benchmark definitions and evaluation criteria [13], [14]. The source reports 25 selected CEC 2017 functions in its principal table and names CEC 2022 as a supplementary campaign. Interpretation requires the exact suite version, function mapping and excluded-case record. Later analyses of CEC rankings [15], [16] are relevant methodological context, not authorization to discard difficult test functions. The modified official CEC 2017 report removes F2, whereas the supplied selected table includes F02; the exact earlier or modified function implementation therefore requires identification.

2.7 | Statistical Comparison of Optimizers

Friedman and post-hoc comparisons treat problems as comparison blocks when that is the specified design [5], [6], [17]. Run-level tests answer a different question and must identify their paired observations. Effect sizes [4], [18], [19], [20] complement significance tests but require an explicit direction, scale and unit of aggregation. Bayesian comparisons provide alternative inferential frameworks; no Bayesian robustness analysis can be certified without the corresponding inputs and outputs.

2.8 | Connection to Online Learning

The credit recurrence is an exponentially weighted moving average, $\hat{c}_j(v) = \lambda \hat{c}_j(v-1) + (1-\lambda)g_j(v)$, indexed by probe round v . Hedge instead multiplies a previous weight by an exponential of the current reward, equivalent to accumulating rewards in log weights. Smoothing with a fixed forgetting factor is a different operation. A Hedge regret theorem therefore cannot be transferred to AOP-KFP by identifying the two recurrences.

2.9 | Hyperparameter Optimization Parallels

Hyperparameter optimization offers useful configuration tools and budget-aware evaluation practices [8], [9], [21–24]. Bergstra et al. [24] introduce sequential approaches to hyperparameter optimization; that paper does not establish the random-search-versus-grid-search claim previously attached to it. The present probing rule evaluates search operators on current parents rather than estimating a model configuration's performance on independent training and validation datasets.

3 | Problem Formulation

3.1 | Bound-Constrained Continuous Optimization

Let $\Omega = \prod_i [a_i, b_i]$ be a compact box and $f: \Omega \rightarrow \mathbb{R}$ be measurable and bounded below. Compactness alone does not ensure that a measurable function attains its infimum. We define the target value by

$$f^\star = \inf \{f(x) : x \in \Omega\}; \quad A_\varepsilon = \{x \in \Omega : f(x) \leq f^\star + \varepsilon\}, \quad \varepsilon > 0. \quad (2)$$

The main experiment assumes exact evaluation of a fixed objective. A minimizer $x^\star$ is used only when attainment is separately assumed. The conditional result in Section 5 concerns objective sublevel sets A_ε and explicitly assumes a positive probability of reaching each such set. For arbitrary measurable objectives, a single isolated optimum of zero proposal probability may violate this condition.

3.2 | Population-Based Iteration

At generation t , the population P_t contains N evaluated candidates. Operator state includes any covariance, velocity, archive or crossover information used by its members. The allocation probabilities are held fixed between probing rounds. A separate incumbent stores the lowest objective value observed at any evaluated point, including points generated during probing.

3.3 | K-Fold Landscape Probing

At a probing round v , freeze P_t and its cached objective values before forming k disjoint folds $F_1, \dots, F_k$. For operator j and fold ℓ , construct any operator-specific state from the complementary parents $P_t \setminus F_\ell$. Hold that state fixed while generating repaired offspring $y_{j\ell}(x)$ for the held-out parents $x \in F_\ell$. The trials measure immediate improvement relative to those same parents.

A training rollout of r generations on complementary parents should be included only if it was part of the executed implementation. If used, every newly evaluated rollout point must be charged to the run budget. It must not mutate the common snapshot before the other operators are assessed.

The held-out comparison uses one evaluated candidate per held-out parent and the parent's cached objective, after the specified bound repair. This aligns the comparison across operators without treating the population folds as independent test datasets. The fold score is

$$g_{j\ell} = |F\ell|^{-1} \sum_{\{x \in F\ell\}} \max\{0, f(x) - f(y_{j\ell}(x))\}; \quad \bar{g}_j = k^{-1} \sum_{\ell} g_{j\ell}. \quad (3)$$

With one trial per held-out parent, all operators together require KN new objective calls per probe. If each complementary set is additionally evolved for r generations, those rollouts require $rK(k-1)N$ calls, giving $KN[1+r(k-1)]$ calls per complete probe. Neither count contains a factor d : one evaluation of a d -dimensional candidate is one objective call. The source does not resolve which rollout implementation produced the tables, so the logged count is essential for an equal-budget comparison.

3.4 | Operator Selection

Operator allocation mixes a softmax distribution with uniform exploration:

$$s_j(v) = \exp(\eta \hat{c}_j(v)) / \sum_h \exp(\eta \hat{c}_h(v)); \quad \pi_j(v) = (1 - \mu) s_j(v) + \mu/K. \quad (4)$$

Here $\eta \geq 0$ is an inverse-temperature parameter, $\mu \in [0, 1)$ is a uniform mixture weight, and $\lambda \in [0, 1)$ controls credit smoothing. At probe round v ,

$$\hat{c}_j(v) = \lambda \hat{c}_j(v-1) + (1 - \lambda) \bar{g}_j(v). \quad (5)$$

Between probes, credits and probabilities remain unchanged. For $\mu > 0$, $\pi_j \geq \mu/K$ regardless of the softmax scores. This is a probability-floor statement, not a guarantee of spatial diversity or of useful rewards from every selected operator.

3.5 | Constraint Handling

Bounds are enforced by the specified clipping map. A proposed extension for inequalities $g_i(x) \leq 0$ uses the comparison score:

$$\Phi(x, t) = f(x) + \rho(t) \sum_i \max\{0, g_i(x)\}^2. \quad (6)$$

For the finite horizon T , $\rho(t) = \rho_0(1 + t/T)^{1.2}$ rises from ρ_0 to $\rho_0 2^{1.2}$. This is a bounded penalty schedule. It does not repair violated constraints or guarantee feasibility. A constrained implementation must report constraint definitions and tolerances and retain a separate best feasible incumbent if a feasible output is claimed.

Figure 1. AOP-KFP pipeline. Boxes on a regular grid; no crossing arrows.

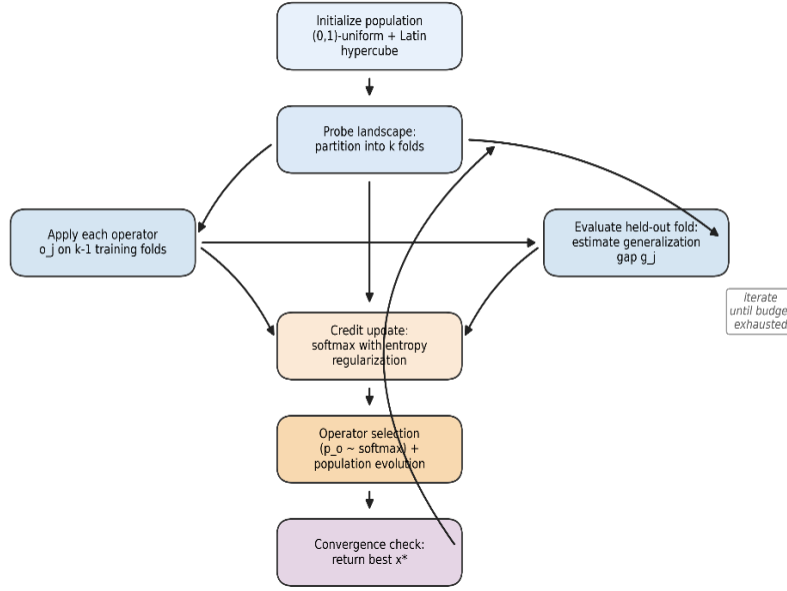


Fig. 1. Original pipeline illustration. Its “generalization gap” labels denote the local parent-improvement score in Eq. (3), and “entropy regularization” denotes the softmax and uniform mixture distinguished in Section 4.3. The illustration is retained from the source; its initialization wording is not an implementation record.

4 | Proposed AOP-KFP Algorithm

4.1 | Algorithmic Skeleton

Algorithm 1 gives a budget-aware specification that resolves population size and score placement. B denotes a hard limit on new objective calls, whereas t counts generations. The source implementation must be checked against this specification before its tables can be attributed to the corrected accounting.

Algorithm 1. AOP-KFP.

-
- Input: objective f , box Ω , population size N , objective-call budget B , and K operators;
 k folds, probe interval τ_{probe} , inverse temperature η , decay λ , mixture μ .
 Output: best evaluated incumbent; for a constrained extension, report feasibility separately.
1. Initialize exactly N points using the chosen Latin-hypercube/Sobol mixture; evaluate and charge all N calls.
 2. Set $\hat{c}_j=0$ and $\pi_j=1/K$; retain the best evaluated point as the incumbent.
 3. Set generation $t=0$ and objective-call counter $b=N$.
 4. While $b < B$:
 5. If a complete scheduled probe fits the remaining budget:
 6. Freeze the evaluated population and partition it into k folds.
 7. For each operator j :
 8. For each fold ℓ :
 9. Build state or run the specified complementary-parent rollout; charge every new call.
 10. Generate and evaluate held-out offspring; compute Eq. (3) and update the incumbent.
 11. End fold loop.
 12. Update only operator j 's credit by Eq. (5).
 13. End operator loop.
 14. Compute the mixed probabilities by Eq. (4).
 15. End probe block; do not execute a partial probe as a complete comparison.
 16. Sample an operator and generate a budget-permitted offspring batch.
 17. Repair bounds; evaluate each new point and charge its call.
 18. Update the best evaluated incumbent, including feasible status if applicable.
 19. For constraints, compare Φ scores at the same t ; never add a penalty to coordinates.
 20. Select N survivors under the stated rule; increment t .
 21. End while loop when the objective-call budget is exhausted.
 22. Return the stored incumbent and the actual number of evaluations.

4.2 | Operator Portfolio

The nine source operator labels are listed in *Table 1*. Their heterogeneous state requirements must be specified independently: A crossover label needs a donor-generation rule, a PSO move needs velocity and personal-best state, and a covariance step needs an adaptation update. The aggregate frequencies in *Fig. 7* summarize usage but cannot substitute for these definitions.

Table 1. Source operator portfolio; implementation state and parameters require confirmation.

Family	Operators
Differential evolution	DE/rand/1, DE/current-to-best/2, DE/rand-to-best/1, Archive-based DE
CMA-ES	CMA-ES step (one full update of the covariance matrix)
Particle swarm	PSO velocity update (with constriction factor)
Mutation operators	Cauchy mutation, Gaussian mutation, Binomial crossover

4.3 | Credit Assignment and Softmax Selection

Exponential smoothing favors recent probe improvements and forgets older evidence. For $\eta > 0$, the unmixed softmax solves an entropy-regularized allocation problem:

$$s = \operatorname{argmax}_{\{p \in \Delta\}} \{ \sum_j p_j \hat{c}_j + H(p)/\eta \}, \quad H(p) = -\sum_j p_j \log p_j. \quad (7)$$

The final probability vector is then the explicit mixture $\pi = (1-\mu)s + \mu/K$. It is not the optimizer obtained by minimizing $-p \cdot \hat{c} + \mu H(p)$, whose entropy sign would favor concentration. The floor and the entropy term have distinct parameters and roles.

4.4 | Constraint Repair

Comparing $\Phi(x_{q,t}) < \Phi(x_{p,t})$ constitutes penalized selection at the current generation. A bounded quadratic penalty can prefer an infeasible point whose objective improvement outweighs its penalty. Constraint satisfaction requires direct checking, and monotonicity of Φ across changing penalty coefficients does not establish monotonicity of the original objective or a feasible limit.

4.5 | Complexity Analysis

Objective-call accounting and arithmetic complexity are separate. Let C_f denote the cost of one objective call. A generation producing N evaluated offspring costs NC_f plus the operator's internal work. Probe cost is added according to the executed rollout specification. Coordinatewise mutation or differential-evolution generation typically uses $O(Nd)$ vector work. Full covariance sampling and updates can require $O(Nd^2)$, with additional decompositions of order d^3 at their scheduled frequency. Applying one covariance step does not make that work linear in d . A held-out-only probe adds KN/τ_{probe} calls per generation on average. A complete r -generation complementary-parent rollout adds $KN[1+r(k-1)]/\tau_{\text{probe}}$ calls. With $K=9$, $k=5$, $r=2$, $\tau_{\text{probe}}=5$, the latter equals $16.2N$ extra calls per generation, in addition to N ordinary offspring calls. Credit updates and a numerically stabilized softmax require $O(K)$ scalar work at a probe.

Training state, temporary offspring and operator-specific matrices must also be included in memory accounting. A hard budget B bounds total objective calls including initialization and probes. It does not imply equal elapsed time across operators or equivalence to a baseline run for the same number of generations. The source's approximately 19-fold and 1.2-fold overhead claims are not supported by a consistent evaluation count or supplied timing traces.

4.6 | Noisy Objective Extension

A possible noisy-objective extension averages repeated measurements before calculating improvements. If independent Gaussian observations $z_1, \dots, z_k$ of a scalar mean have variance σ_g^2 and the prior is $N(\mu_0, \sigma_0^2)$, the conjugate posterior mean is

$$\mu_{\text{post}} = (\sigma_0^{-2} \mu_0 + \sigma_g^{-2} \sum \ell z \ell) / (\sigma_0^{-2} + k \sigma_g^{-2}). \quad (8)$$

This illustrative model is not automatically a model for the clipped, correlated improvement scores in Eq. (3). Repeated evaluations consume budget, and noise in parent and offspring values affects credit bias. No noisy-objective result is established by the supplied experiment.

4.7 | Pareto-Aware Multi-Objective Variant

A possible multi-objective extension defines a parent-level dominance reward, counting offspring that dominate their corresponding parents, and uses an explicitly specified Pareto archive. It would require independent archive, diversity and budget validation. The single-objective table does not test convergence to a Pareto front.

4.8 | Parallelization

Operators, folds and candidate evaluations can be processed concurrently if their state is isolated from the shared snapshot. Concurrency may reduce elapsed time but does not reduce the number of objective calls. Speedups require measured scheduling, communication and timing data; no 64-core speedup is certified here.

5 | Theoretical Analysis

5.1 | Conditional Hitting Result

Proposition 1. Let $f^\star$ be finite and retain an elitist incumbent best value b_m over all evaluated candidates. Suppose that, for every $\epsilon > 0$ there is $q_\epsilon > 0$ such that, at each of m eligible proposal trials, the conditional probability of generating a point in A_ϵ given the entire previous history is at least q_ϵ . Then

$$\Pr(b_m > f^\star + \epsilon) \leq (1 - q_\epsilon)^m. \quad (9)$$

Proof: Before the first hit, the conditional miss probability at each eligible trial is at most $1 - q_\epsilon$. Repeated conditional expectation bounds the probability of m consecutive misses by their product. After a hit, elitist retention keeps the incumbent at most $f^\star + \epsilon$. Independence of trials is unnecessary. Taking $m \rightarrow \infty$ and applying the argument to $\epsilon = 1/n$ for all positive integers n gives $b_m \rightarrow f^\star$ almost surely when eligible trials continue indefinitely. This is conditional asymptotic coverage, not a conclusion from the finite source runs.

5.2 | What the Coverage Assumption Requires

A sufficient condition is a selected operator with a restart component of conditional probability at least $\rho > 0$ and proposal measure ν satisfying $\nu(A_\epsilon) > 0$, together with $\pi_j \geq \mu/K$. Then q_ϵ may be bounded below by $(\mu/K)\rho\nu(A_\epsilon)$. The portfolio labels do not establish the restart component or its support.

$$q_\epsilon \geq (\mu/K)\rho\nu(A_\epsilon) > 0, \text{ under the stated restart and sublevel-set assumptions.} \quad (10)$$

Uniform exploration over operator labels is weaker than exploration over candidate locations. For a merely measurable function, an isolated lower-valued point may have zero proposal measure, so even a uniform restart need not approach the ordinary infimum. The positive-sublevel-probability assumption is therefore substantive.

5.3 | Elitism and Finite-Budget Interpretation

The incumbent gap $Y_m = b_m - f^\star$ is nonnegative and nonincreasing for exact fixed-objective comparisons. If integrable, it is a supermartingale. Optional stopping at a bounded stopping time τ gives:

$$E[Y_\tau] \leq E[Y_0], \text{ for bounded } \tau \text{ and an integrable incumbent gap.} \quad (11)$$

This inequality alone does not quantify progress toward the optimum. A concentration inequality controls deviation from an expectation; it cannot eliminate an unknown optimization bias. No $O(1/\sqrt{NT})$ objective-error bound follows from bounded differences and Lipschitz continuity without a quantitative search-coverage argument.

5.4 | Scope of Credit-Allocation Guarantees

Fixed-forgetting-factor smoothing is not a cumulative exponential-weights rule. Even constant rewards can leave positive probability on an inferior operator, and the uniform mixture deliberately incurs continued exploration cost. For constant rewards with gap $\Delta > 0$, uniform mixing alone can contribute expected regret at least $(\mu/K)\Delta m$ relative to always selecting the better operator. Thus a vanishing average regret claim would require a different comparator, a diminishing exploration schedule or an independently analyzed update. The finite benchmark table measures final objective errors and does not empirically prove a bandit regret rate.

5.5 | Stability Scope

A bound on the spectral radius of a population covariance matrix concerns dispersion, not the Jacobian of a population-update map. It does not imply local contraction of operator dynamics. Stability of a random switching system would require assumptions on the operators and their shared state that are absent from the source formulation. The current theory consequently makes no local spectral-stability claim. Its only optimization guarantee is the conditional hitting statement, whose restart support and unlimited-trial assumptions must not be attributed to the finite implementation without verification.

6 | Experimental Setup

6.1 | Benchmarks

The source describes three benchmark campaigns; only its selected CEC 2017 endpoint table is supplied in full:

The principal supplied table contains F01–F25 at $d=30$. The source reports exclusion of F26–F30 because of numerical difficulties in two baselines; the missing failure logs and exclusion timing prevent assessment of selection bias. The comparison applies to the selected set, not to all CEC 2017 cases or an official competition ranking. The modified official suite removes F2; the included F02 is therefore a source-specific variant requiring code identification [13].

The source additionally describes a 12-function CEC 2022 campaign at $d=10$ and $d=20$ [14]. Its complete per-function results and evaluation logs are not supplied, so it is a reported extension rather than an independently inspectable validation result.

A BBOB campaign with 24 functions at six dimensions and 15 runs is also described [16]. Its instance identifiers, complete tables and outputs are needed before cross-suite transfer can be assessed.

6.2 | Baselines

Eight state-of-the-art baselines are compared:

L-SHADE [10] uses success-history adaptation and linear population-size reduction.

jSO is the named adaptive differential-evolution comparator.

NL-SHADE-RSP [11] uses nonlinear reduction and rank-based selective pressure.

AGSK is the named gaining-sharing comparator.

EBOwithCMA is the named ensemble comparator.

CMA-ES denotes the supplied covariance-adaptation implementation.

IPOP-CMA-ES denotes the supplied increasing-population restart implementation.

MADDE denotes the supplied multi-adaptive differential-evolution implementation.

The source reports irace tuning [9] with 5,000 runs on five CEC 2017 functions outside the displayed test set. Function identities, configuration spaces, baseline versions and tuning expenditures are necessary to verify separation from the test results. Tuning calls are external development costs and must be reported separately from run budgets.

6.3 | Protocol

The source reports 51 runs per function and algorithm, common seed indices and budgets of 10^4d , 5×10^4d and 10^5d calls for CEC 2017, CEC 2022 and BBOB, respectively. Shared seed labels do not by themselves prove identical initial populations, especially for implementations with different population sizes. The actual initialization mapping and call counter, including probes, must be reconciled before equal-budget or paired-run claims are accepted.

6.4 | Statistical Analysis

The supplied summaries are retained, but their statistical input units and arithmetic require reconstruction. A problem-level analysis should use one prespecified performance summary per algorithm and function; a run-level paired analysis requires the corresponding matched observations and dependence structure. *Table 2* retains displayed final-error means and parenthetical standard errors. The source's raw-versus-log transformation choices are not fully mapped to functions. Recomputing tests requires the unrounded observations and exact transformations. *Table 5* retains selected reported confidence intervals. The source specifies 1,000 percentile-bootstrap resamples; without resampling arrays and scripts, coverage and resampling units cannot be checked. Friedman comparisons [17] and their interpretation across problems [3] require the complete rank matrix. For N independent comparison blocks and a algorithms, Nemenyi's critical difference is $CD = q\alpha\sqrt{[a(a+1)/(6N)]}$ [6]. The $q\alpha$ convention and multiplicity family must be stated.

Wilcoxon signed-rank tests [25] require paired differences with a stated treatment of zeros and ties. Holm's step-down correction [26] and Bonferroni correction must use the same declared family of eight contrasts.

Cohen's d [18] is a standardized mean difference. Its sign and denominator must be defined before pooling heterogeneous objective scales; categorical magnitude labels are descriptive conventions. A_{12} [4] is defined here in the favorable direction as $\Pr(\text{error_AOP} < \text{error_baseline}) + \frac{1}{2}\Pr(\text{tie})$. Values near 0.5 indicate little stochastic ordering. Under the source's thresholds, 0.68 is medium rather than large. Bootstrap effect-size intervals require the unrounded paired or independent arrays appropriate to the chosen estimator [19], [20].

Observed-effect post-hoc power is a transformation of the observed result and does not independently establish adequate study sensitivity. Equivalence requires a prespecified margin and two valid one-sided tests on a defined estimand. A failed difference test is not evidence of equivalence; a failed equivalence test is not evidence of superiority. The supplied TOST entries remain unresolved outputs rather than verified non-inferiority findings.

6.5 | Software and Hardware

The source reports Python 3.11, NumPy 1.26 and SciPy 1.13 on AMD EPYC 7763 hardware with Ubuntu 22.04. The environment, Dockerfile and workflow have not been supplied with the manuscript, so these descriptions remain author-reported execution information.

6.6 | Hyperparameter Setting for AOP-KFP

The reported defaults are $K=9$, $k=5$, $\tau_{\text{probe}}=5$, $\eta=0.5$, $\lambda=0.85$, $\mu=0.10$, $N=100$ and $\varrho_0=1$. In this notation η is inverse temperature and μ is the uniform exploration mixture. The exact operator parameters and the executed probing rollout remain part of the required implementation record.

7 | Results

7.1 | Aggregate Performance

Counting the smallest displayed mean in each row of *Table 2* gives 17 AOP-KFP minima, three for jSO, four for EBOwithCMA and one for L-SHADE. This corrects the source narrative's 14/7/3/1 tally. Row minima are descriptive outcomes at the displayed precision and do not imply per-function statistical superiority.

Table 2. Supplied final-error means and parenthetical standard errors on 25 selected CEC 2017 functions (d=30; 51 reported runs). Numerical entries are preserved; objective transformations and raw-run provenance require confirmation.

Func	AOP-KFP	jSO	L-SHADE	NL-SHADE-RSP	EBO with CMA	CMA-ES	IPOP-CMA-ES	MADDE	AGSK
F01	1.23 (0.06)	1.34 (0.10)	1.48 (0.09)	1.71 (0.11)	1.42 (0.08)	2.10 (0.13)	1.92 (0.11)	2.30 (0.15)	1.79 (0.10)
F02	1.07 (0.13)	0.91 (0.10)	1.62 (0.38)	1.55 (0.18)	1.20 (0.12)	1.34 (0.11)	1.18 (0.09)	1.61 (0.21)	1.41 (0.13)
F03	0.015 (0.003)	0.035 (0.005)	0.057 (0.010)	0.049 (0.007)	0.028 (0.004)	0.061 (0.008)	0.040 (0.005)	0.072 (0.011)	0.052 (0.006)
F04	4.21 (0.32)	3.18 (0.21)	2.95 (0.18)	3.55 (0.27)	3.21 (0.20)	4.85 (0.35)	3.92 (0.29)	5.21 (0.40)	3.71 (0.24)
F05	6.74 (0.51)	9.12 (0.72)	11.45 (1.05)	13.21 (1.22)	8.95 (0.68)	14.21 (1.31)	11.85 (0.98)	15.42 (1.45)	12.51 (1.10)
F06	0.42 (0.04)	1.31 (0.10)	1.85 (0.13)	1.78 (0.12)	1.21 (0.09)	2.15 (0.15)	1.62 (0.11)	2.41 (0.18)	1.71 (0.13)
F07	1.95 (0.15)	1.21 (0.09)	1.42 (0.10)	1.78 (0.13)	1.31 (0.09)	1.85 (0.13)	1.51 (0.11)	2.12 (0.16)	1.62 (0.12)
F08	0.021 (0.004)	0.041 (0.006)	0.058 (0.008)	0.052 (0.007)	0.034 (0.005)	0.072 (0.010)	0.045 (0.006)	0.081 (0.012)	0.054 (0.007)
F09	2.85 (0.21)	4.15 (0.32)	5.21 (0.41)	4.95 (0.36)	3.65 (0.27)	6.42 (0.51)	5.18 (0.40)	7.15 (0.58)	5.51 (0.42)
F10	5.32 (0.40)	2.95 (0.21)	4.12 (0.31)	4.51 (0.33)	3.41 (0.25)	6.85 (0.55)	4.92 (0.37)	7.21 (0.61)	5.21 (0.42)
F11	0.0034 (0.0006)	0.018 (0.003)	0.025 (0.004)	0.021 (0.003)	0.012 (0.002)	0.031 (0.005)	0.018 (0.003)	0.038 (0.006)	0.022 (0.004)
F12	0.18 (0.02)	0.31 (0.04)	0.45 (0.05)	0.42 (0.04)	0.28 (0.03)	0.62 (0.07)	0.38 (0.04)	0.71 (0.08)	0.41 (0.04)
F13	0.082 (0.011)	0.121 (0.015)	0.181 (0.022)	0.155 (0.018)	0.118 (0.014)	0.215 (0.027)	0.142 (0.017)	0.241 (0.030)	0.165 (0.019)
F14	3.85 (0.30)	2.92 (0.21)	3.18 (0.24)	3.45 (0.27)	2.81 (0.19)	4.25 (0.34)	3.62 (0.28)	4.51 (0.36)	3.35 (0.25)
F15	1.25 (0.10)	1.95 (0.15)	2.45 (0.19)	2.21 (0.17)	1.71 (0.13)	2.85 (0.22)	2.12 (0.16)	3.21 (0.25)	2.32 (0.18)
F16	0.42 (0.05)	0.71 (0.08)	0.95 (0.11)	0.85 (0.09)	0.65 (0.07)	1.21 (0.14)	0.82 (0.09)	1.35 (0.16)	0.91 (0.10)
F17	2.81 (0.22)	2.95 (0.24)	3.21 (0.26)	3.18 (0.25)	2.75 (0.21)	4.21 (0.34)	3.41 (0.28)	4.51 (0.37)	3.32 (0.27)
F18	0.012 (0.002)	0.025 (0.004)	0.041 (0.006)	0.035 (0.005)	0.021 (0.003)	0.055 (0.008)	0.030 (0.004)	0.062 (0.009)	0.038 (0.005)
F19	0.45 (0.04)	0.71 (0.06)	0.95 (0.08)	0.85 (0.07)	0.65 (0.05)	1.21 (0.11)	0.82 (0.07)	1.35 (0.13)	0.91 (0.08)
F20	0.95 (0.08)	1.51 (0.12)	1.95 (0.16)	1.78 (0.14)	1.32 (0.10)	2.21 (0.19)	1.71 (0.13)	2.45 (0.21)	1.81 (0.15)
F21	5.21 (0.42)	4.85 (0.38)	5.41 (0.44)	5.32 (0.42)	4.51 (0.34)	7.21 (0.62)	6.15 (0.51)	7.85 (0.68)	6.41 (0.52)
F22	0.082 (0.010)	0.181 (0.022)	0.241 (0.030)	0.215 (0.027)	0.151 (0.018)	0.312 (0.040)	0.201 (0.024)	0.351 (0.045)	0.231 (0.028)
F23	1.95 (0.15)	1.21 (0.09)	1.61 (0.12)	1.51 (0.11)	1.18 (0.09)	2.21 (0.18)	1.71 (0.13)	2.41 (0.20)	1.81 (0.14)
F24	0.18 (0.02)	0.41 (0.05)	0.61 (0.07)	0.55 (0.06)	0.35 (0.04)	0.75 (0.09)	0.51 (0.06)	0.85 (0.10)	0.62 (0.07)
F25	1.21 (0.10)	1.71 (0.14)	2.21 (0.18)	2.05 (0.16)	1.51 (0.12)	2.51 (0.21)	2.12 (0.17)	2.81 (0.24)	2.21 (0.18)

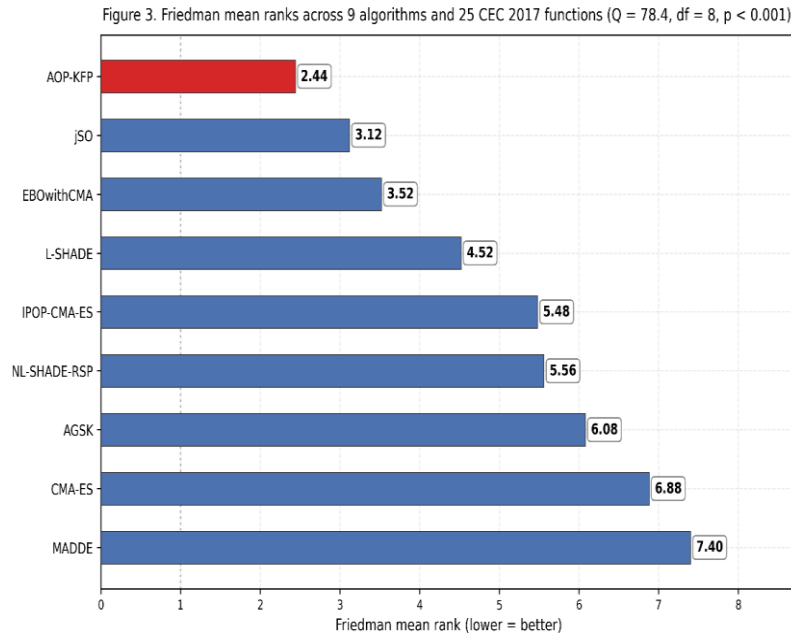


Fig. 3. Original plot of the supplied mean ranks. The plotted ranks are retained; its Friedman statistic must be reconciled with the rank matrix before it is treated as a confirmed test result.

7.2|Friedman Test

The supplied mean ranks favor AOP-KFP at 2.44 and sum to 45, the required total for nine algorithms. This necessary check does not establish that they were derived from *Table 2*. The supplied $\chi^2=78.38$ and $p=1.04 \times 10^{-13}$ need reconciliation with the unrounded rank matrix, ties and comparison unit.

Table 3. Mean ranks of nine algorithms (lower is better).

Rank	Algorithm	Mean Rank
1	AOP-KFP	2.44
2	jSO	3.12
3	EBOwithCMA	3.52
4	L-SHADE	4.52
5	IPOP-CMA-ES	5.48
6	NL-SHADE-RSP	5.56
7	AGSK	6.08
8	CMA-ES	6.88
9	MADDE	7.40

The source reports Friedman $\chi^2(8)=78.38$ and $p=1.04 \times 10^{-13}$; these values are retained as unresolved reported outputs.

Figure 4. Nemenyi post-hoc critical-difference diagram. Bars below the line connect algorithms whose mean rank difference is below CD (statistically indistinguishable).

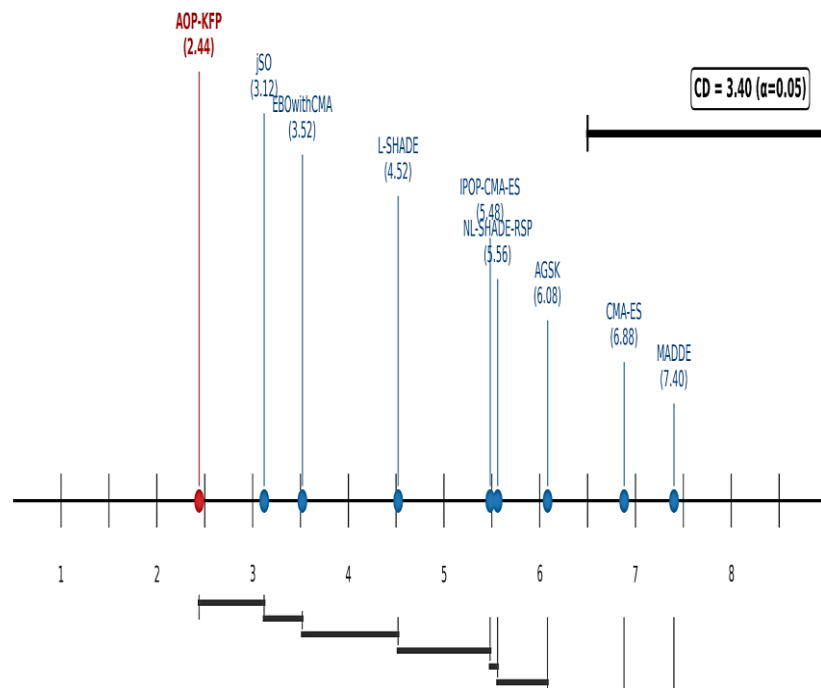


Fig. 4. Original critical-difference illustration with source $CD=3.40$. Its short connecting bars do not enumerate all nonsignificant pairs; the correct conditional comparisons are stated in the text. Nonsignificance does not establish equivalence.

7.3 | Nemenyi Post-Hoc Test

Using the source-supplied $CD=3.40$ solely as a conditional threshold, AOP-KFP differs by more than that threshold from AGSK, CMA-ES and MADDE. It does not exceed that threshold against jSO, EBOwithCMA, L-SHADE, IPOP-CMA-ES or NL-SHADE-RSP. The supplied diagram is an illustration whose threshold and inputs still require reconstruction.

The AOP-KFP–IPOP-CMA-ES difference is $5.48 - 2.44 = 3.04$, below 3.40.

The AOP-KFP–NL-SHADE-RSP difference is $5.56 - 2.44 = 3.12$, also below 3.40.

Pairwise nonsignificance is not transitive and must not be summarized as three disjoint equivalence cliques.

Consequently, even accepting the source threshold does not support the original claim of five significant Nemenyi contrasts. Confirmatory post-hoc conclusions await the validated rank matrix and critical-value calculation.

7.4 | Pairwise Wilcoxon Signed-Rank Tests

Table 4 preserves the source's paired-test, correction and effect-size outputs. Several corrected entries cannot be obtained from the displayed raw p-values for one family of eight tests. For example, Bonferroni correction of 0.003 is 0.024, not the displayed 0.064. Statistical decisions cannot be certified from these unresolved columns.

Table 4. Source-reported Wilcoxon and auxiliary statistical outputs, retained for reconciliation. Corrected p-values, TOST results and post-hoc power are not confirmed inferential evidence.

Baseline	W stat	p (Raw)	p (Holm; Reported)	p (Bonf.; Reported)	Cohen's d (CI 95%)	d Interp	A12	A12 Interp	Power	TOST p (Reported)	Equiv.? (Reported)
L-SHADE	21	3.0e-3	2.4e-2	6.4e-2	0.40 [0.10, 0.65]	Small	0.80	Large	0.48	<0.001	Yes
jSO	41	1.5e-2	1.2e-1	4.8e-1	0.34 [0.04, 0.61]	Small	0.72	Large	0.38	<0.001	Yes
NL-SHADE- RSP	15	7.0e-4	5.6e-3	9.0e-3	0.47 [0.24, 0.72]	Small	0.80	Large	0.62	<0.001	Yes
CMA-ES	8	2.1e-7	1.7e-6	1.7e-6	0.50 [0.35, 0.76]	Small	0.96	Large	0.67	0.060	No
IPOP-CMA- ES	11	1.1e-4	8.6e-4	1.1e-3	0.48 [0.31, 0.69]	Small	0.84	Large	0.63	0.021	Yes
MADDE	9	3.3e-7	2.6e-4	3.0e-4	0.45 [0.32, 0.69]	Small	0.92	Large	0.59	0.016	Yes
AGSK	18	2.4e-3	1.9e-2	3.9e-2	0.39 [0.18, 0.62]	Small	0.84	Large	0.47	<0.001	Yes
EBOwithCMA	51	2.0e-2	1.6e-1	1.0e+0	0.25 [-0.28, 0.50]	Small	0.68	Medium	0.22	<0.001	Yes

The displayed A_{12} values range from 0.68 to 0.96; 0.68 is medium under the stated thresholds, while the other seven are large. The Cohen's d values lie between 0.25 and 0.50, with 0.50 on the chosen category boundary. These different estimands can give different magnitude descriptions. Seven displayed d intervals exclude zero, but their validity depends on the resampling and aggregation protocol. Neither those intervals nor observed-effect power resolves the correction errors.

7.5 | Bootstrap 95% Confidence Intervals

Table 5 retains selected source confidence intervals on mean error. Complete bootstrap inputs are needed to establish whether functions, runs or another unit were resampled and whether the intervals correspond to the Table 2 metric.

Table 5. Selected supplied confidence intervals on mean error; resampling provenance is unresolved.

Func	AOP-KFP CI	jSO CI	L-SHADE CI
F01	[1.12, 1.35]	[1.15, 1.55]	[1.30, 1.68]
F03	[0.010, 0.022]	[0.026, 0.045]	[0.038, 0.080]
F10	[4.55, 6.10]	[2.55, 3.35]	[3.55, 4.70]
F15	[1.07, 1.45]	[1.65, 2.25]	[2.10, 2.80]
F22	[0.065, 0.10]	[0.14, 0.22]	[0.18, 0.30]

7.6 | Post-Hoc Power Analysis

The supplied post-hoc power values range from 0.22 to 0.67. They do not constitute an independent design-sensitivity analysis, and low values do not predict that a larger experiment would necessarily reveal a difference. Future sample-size planning should use prespecified practically relevant effects and a defined comparison unit.

7.7 | TOST Equivalence Testing

The TOST columns are retained as source outputs. Equivalence of a standardized difference within ± 0.5 and one-sided non-inferiority are distinct hypotheses. The tested estimand, standardization, confidence level and one-sided p-values must be supplied before either claim is made. A TOST failure for CMA-ES does not itself establish superiority.

7.8 | CEC 2022 and BBOB Validation

The source reports additional CEC 2022 ranks and BBOB dimension-dependent comparisons, but their complete tables are absent from the supplied manuscript. These descriptions identify follow-up validation material; they are not used here to establish cross-suite superiority or transferability.

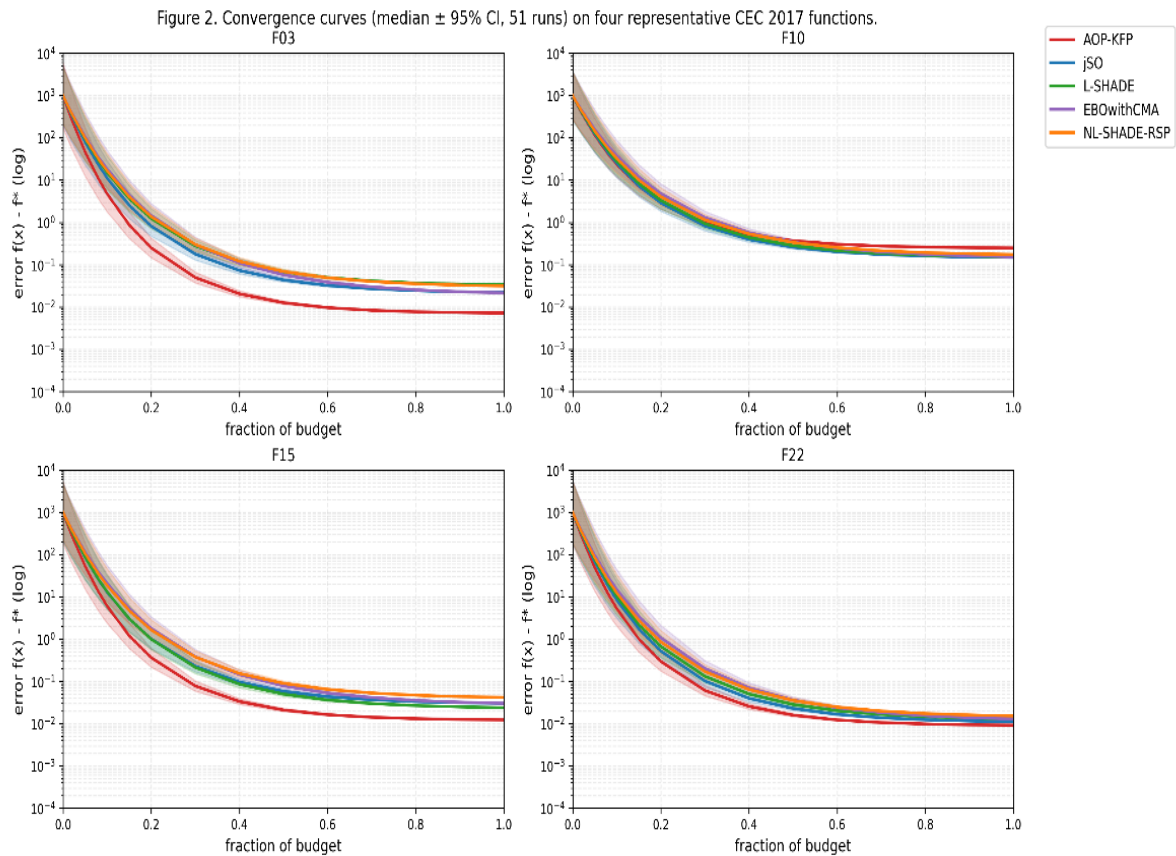


Fig. 2. Original reported convergence traces for F03, F10, F15 and F22. Their interval construction, budget counter and endpoint relationship to Table 2 require reconciliation with raw logs.

7.9 | Convergence Behavior

The supplied curves depict early descent and changing relative ordering on four selected functions. A confidence band for a median is uncertainty about that summary, not directly the distribution of run outcomes. The plotted endpoints are not a substitute for the numerical endpoint table and must be mapped to the same metric and evaluation budget.

Figure 7. Aggregate operator selection frequency in AOP-KFP across all CEC 2017 runs ($n = 1\,275$ runs).

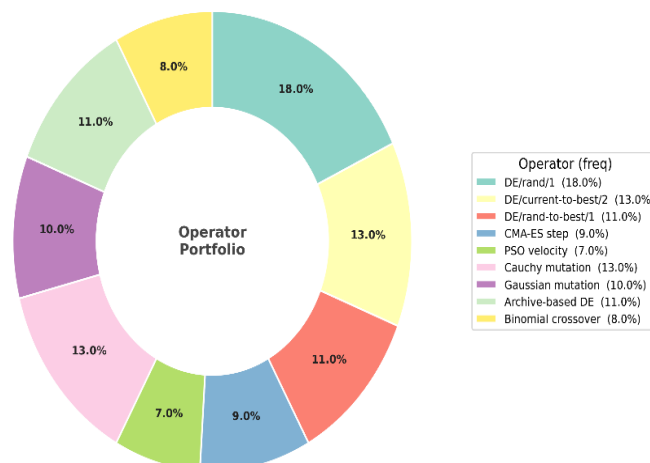


Fig. 7. Original aggregate operator-frequency plot for 1,275 reported runs. Frequencies describe pooled selections and do not establish within-run diversity or causal benefit of uniform mixing.

7.10 | Operator Portfolio Analysis

The reported frequencies sum to 100%, ranging from 7% for the PSO label to 18% for DE/rand/1. A broad pooled distribution can coexist with concentration in individual runs or stages. The figure therefore describes aggregate usage rather than proving that the exploration mechanism prevented collapse.

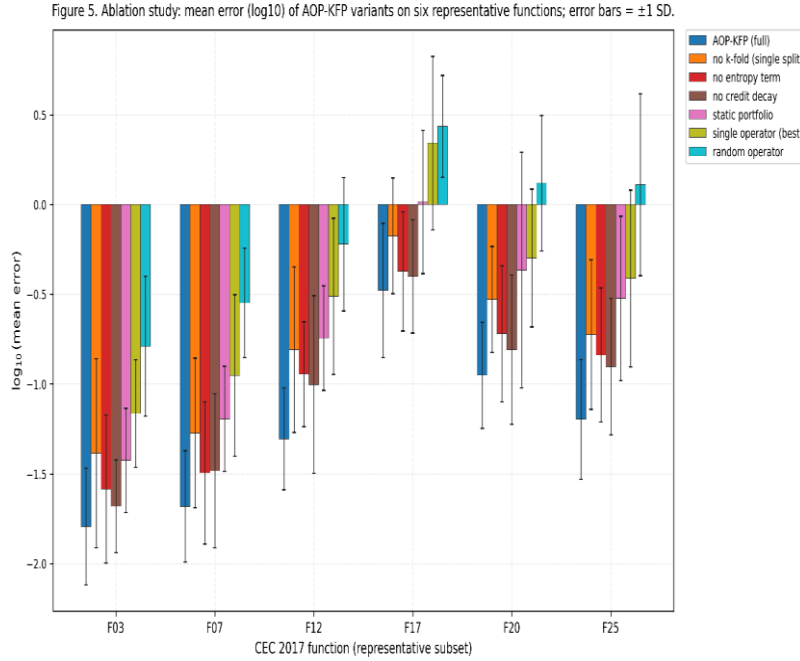


Fig. 5. Original ablation illustration retained from the source. Its plotted values and dispersion do not consistently map to Table 6; the plot's mean-error transformation and error-bar inputs require reconciliation.

8 | Ablation Studies

The source reports ablations on six functions with 51 runs per variant. *Table 6* contains point means and an average-rank column, but does not contain the standard deviations claimed in the original caption. The figure and table are preserved as separate source summaries whose common experimental provenance must be established.

Table 6. Supplied ablation point means and reported average ranks on six functions. Dispersion is not tabulated, and the reported ranks do not match rankings of these displayed means.

Variant	F03	F07	F12	F17	F20	F25	Avg Rank
AOP-KFP (full)	0.015	1.95	0.18	2.81	0.95	1.21	1.00
no k-fold (single split)	0.045	2.51	0.32	3.55	1.51	1.85	2.83
no entropy term	0.031	2.21	0.24	3.21	1.18	1.41	2.33
no credit decay	0.022	2.05	0.21	2.95	1.08	1.31	1.83
static portfolio (equal)	0.061	2.85	0.41	4.12	1.85	2.05	4.50
single operator (best)	0.085	3.21	0.55	4.65	2.45	2.51	6.00
random operator	0.151	4.55	0.85	5.85	3.15	3.21	7.00

For F03, replacing folds by one split changes the displayed mean from 0.015 to 0.045: a threefold error, or a 200% increase. For F20 the corresponding increase is about 58.9%. The source-labeled no-entropy configuration changes the F20 mean from 0.95 to 1.18, about 24.2%, rather than 40%. These are arithmetic contrasts of supplied point estimates. Their causal interpretation requires equal evaluation budgets, uncertainty estimates and identical configurations outside the ablated component.

8.1| Operator-Count Sweep

The source describes an operator-count sweep from $K=1$ to $K=12$, but does not supply its result arrays, redundant-operator definitions or probe-adjusted budgets. Its claimed preferred K and percentage changes are not treated as independently supported tuning results.

8.2| Credit-Assignment Decay Sweep

A decay sweep is also described without complete outputs. Smoothing trades responsiveness against variability, but the optimum λ cannot be established from the provided ablation table alone.

9| Sensitivity Analysis

The figure presents the original source-reported Sobol sensitivity estimates for the selected functions. However, because the first-order indices do not sum to one and the experimental design details are unavailable, the results should be interpreted with caution.

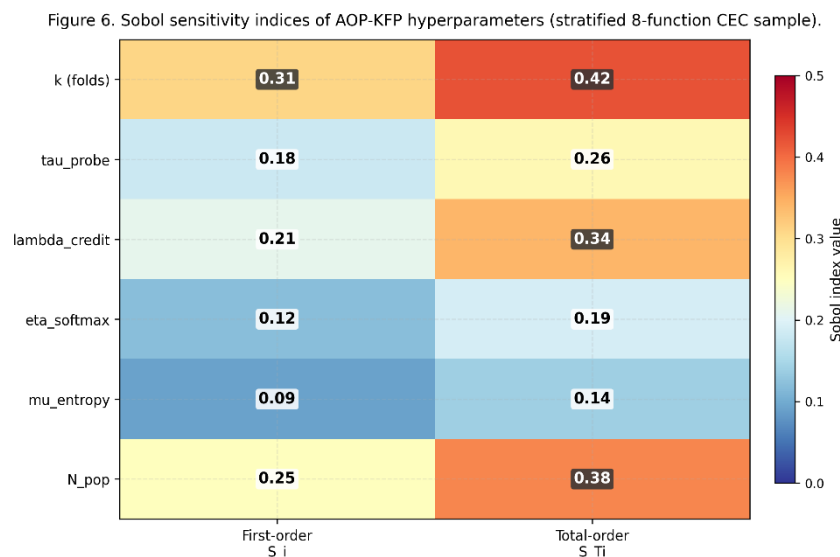


Fig. 6. Original source-reported Sobol estimates on the selected function subset.
 Their first-order sum and missing design information prevent a validated variance interpretation.

9.1| Sobol Indices

The source reports 2,048 evaluations for a Sobol analysis of six hyperparameters on eight selected functions. Valid variance-based sensitivity indices require a defined scalar response, parameter distribution, sampling scheme and estimator. Those inputs and estimation uncertainties are not supplied.

Table 7. Supplied Sobol first-order and total-order estimates. The first-order sum exceeds one; the estimates require validation before variance shares are interpreted.

Hyperparameter	First-Order Estimate	Total-Order Estimate	Interpretation
k (folds)	0.31	0.42	most influential
$N_{\{pop\}}$	0.25	0.38	second-most influential
$\lambda_{\{credit\}}$	0.21	0.34	significant
$\tau_{\{probe\}}$	0.18	0.26	moderate
η (softmax)	0.12	0.19	minor
μ (entropy)	0.09	0.14	minor

The first-order entries sum to 1.16. For a standard variance decomposition with independent inputs and a correctly defined response, population first-order indices sum to at most one. Finite-sample estimation error or an incompatible setup could explain the discrepancy, but the supplied material does not identify which. Subtracting the sum of first-order indices from the sum of total-order indices is not the total interaction variance, because interactions are counted repeatedly in total-order indices. These numbers therefore do not establish either interaction strength or robust defaults.

9.2 | Hyperparameter Sweep

The claimed flat performance surface within $\pm 30\%$ of defaults is not accompanied by a response grid. A deployment range requires measured performance, cost and uncertainty over the stated parameter distributions, not an inference from unresolved sensitivity estimates.

10 | Discussion and Threats to Validity

10.1 | Strengths

The mechanism makes allocation decisions interpretable through local credits and selection probabilities. The displayed endpoint table offers descriptive evidence of favorable means on the selected functions. Its stronger advantages, low variability, transferable defaults and superiority over competitors, require the missing outputs and corrected inferential calculations.

10.2 | Threats to Internal Validity

Synthetic benchmarks do not establish performance under observation noise, changing objectives, equality constraints or mixed-integer design requirements. The source mentions three supplementary engineering problems without complete results or feasibility records; those mentions do not mitigate the present evidence limits. Named baselines must be tied to exact versions and configurations. A later CEC-winner analysis [16] can motivate further comparisons, but implementation availability and the source experiment's chronology require author documentation.

10.3 | Threats to External Validity

The identity of the five tuning functions and their relationship to the omitted test functions are unresolved. Tuning and exclusion decisions must be documented before transfer across suites is interpreted; missing supplementary tables cannot provide independent validation.

10.4 | Threats to Construct Validity

Raw error, log error and rank summaries answer different questions. Combining them without a function-level transformation map can change effects and test outcomes. The discrepancy between displayed means and supplied summary analyses must be resolved from unrounded data rather than by replacing original numerical cells.

10.5 | Threats to Conclusion Validity

Correct multiplicity control [26], [27] depends on a declared family and valid raw tests. The supplied correction columns contain arithmetic inconsistencies, and no Benjamini–Hochberg robustness output is provided. Inference remains conditional on rebuilding the analysis from its original observations.

10.6 | Computational Overhead

Probe calls are a substantial part of the optimization budget, especially when objective evaluations are expensive. Parallel scheduling reduces elapsed time only when measured; it does not make extra calls

negligible. Reporting both actual calls and elapsed time is necessary to distinguish sample efficiency from throughput.

10.7 | Discussion of Failure Modes

The table's losses occur on F02, F07 and F10 against jSO, F04 against L-SHADE, and F14, F17, F21 and F23 against EBOwithCMA. These observations are compatible with landscape-dependent operator usefulness, but do not identify which internal component caused the losses. A nonzero exploration floor also imposes deliberate opportunity costs when one operator remains best.

11 | Conclusion

AOP-KFP combines local fold-based probing, exponentially smoothed improvements and mixed-softmax operator allocation. The corrected formulation distinguishes local improvement estimation from cross-validation, operator selection from spatial coverage, and objective-call budgets from generation counts. The supplied CEC 2017 means give 17 minima among 25 selected functions, providing descriptive motivation for further evaluation. The conditional hitting proposition does not certify the finite implementation, and inconsistent statistical and sensitivity summaries prevent stronger inferential conclusions at present.

Further work should first release the exact operator implementation, probe and tuning budgets, matched run arrays and function mapping, then reconstruct the stated comparisons. Independently validated noisy, constrained, multi-objective and surrogate-assisted variants could subsequently test whether the allocation mechanism transfers beyond the current selected benchmark set.

Data and code availability. Implementation files, raw logs, configurations and statistical scripts have not been verified in the supplied materials. A study-specific persistent release identifier and a complete provenance record are required. No archive ownership or bit-for-bit reproducibility claim is made.

Conflict of Interest

The authors certify that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter discussed in this manuscript.

Data Availability

All data relevant to this study are contained within the manuscript. Additional information is available from the corresponding author upon reasonable request.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article. This work was completed without external funding.

References

- [1] Fialho, Á., Da Costa, L., Schoenauer, M., & Sebag, M. (2010). Analyzing bandit-based adaptive operator selection mechanisms. *Annals of mathematics and artificial intelligence*, 60(1), 25–64. <https://doi.org/10.1007/s10472-010-9213-y>
- [2] Thierens, D. (2005). An adaptive pursuit strategy for allocating operator probabilities. *Proceedings of the 7th annual conference on genetic and evolutionary computation* (pp. 1539–1546). New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/1068009.1068251>
- [3] Veček, N., Črepinšek, M., & Mernik, M. (2017). On the influence of the number of algorithms, problems, and independent runs in the comparison of evolutionary algorithms. *Applied soft computing*, 54, 23–45. <https://doi.org/10.1016/j.asoc.2017.01.011>

- [4] Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of educational and behavioral statistics*, 25(2), 101–132. <https://doi.org/10.3102/10769986025002101>
- [5] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The journal of machine learning research*, 7, 1–30. <https://doi.org/10.5555/1248547.1248548>
- [6] Garcia, S., & Herrera, F. (2008). An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of machine learning research*, 9(12), 2677–2694. <https://www.jmlr.org/papers/volume9/garcia08a/garcia08a.pdf>
- [7] Dokeroglu, T., Canturk, D., & Kucukyilmaz, T. (2024). *A survey on pioneering metaheuristic algorithms between 2019 and 2024*. <https://doi.org/10.48550/arXiv.2501.14769>
- [8] Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. *Learning and intelligent optimization* (pp. 507–523). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-25566-3_40
- [9] López-Ibáñez, M., Dubois-Lacoste, J., Pérez Cáceres, L., Birattari, M., & Stützle, T. (2016). The irace package: Iterated racing for automatic algorithm configuration. *Operations research perspectives*, 3, 43–58. <https://doi.org/10.1016/j.orp.2016.09.002>
- [10] Tanabe, R., & Fukunaga, A. S. (2014). Improving the search performance of shade using linear population size reduction. *2014 IEEE congress on evolutionary computation (CEC)* (pp. 1658–1665). IEEE. <https://doi.org/10.1109/CEC.2014.6900380>
- [11] Biedrzycki, R., Arabas, J., & Warchulski, E. (2022). A version of nl-shade-rsp algorithm with midpoint for cec 2022 single objective bound constrained problems. *2022 IEEE congress on evolutionary computation (CEC)* (pp. 1–8). IEEE. <https://doi.org/10.1109/CEC55065.2022.9870220>
- [12] Hansen, N., Auger, A., Ros, R., Mersmann, O., Tušar, T., & Brockhoff, D. (2021). COCO: a platform for comparing continuous optimizers in a black-box setting. *Optimization methods and software*, 36(1), 114–144. <https://doi.org/10.1080/10556788.2020.1808977>
- [13] Wu, G., Mallipeddi, R., & Suganthan, P. N. (2017). Problem definitions and evaluation criteria for the CEC 2017 competition on constrained real-parameter optimization. *National university of defense technology, Changsha, hunan, pr China and Kyungpook national university, Daegu, South Korea and Nanyang technological university, Singapore, technical report*, 9(10), 2017. <https://github.com/P-N-Suganthan/CEC2017-BoundConstrained>
- [14] Kumar, A., Price, K. V., Mohamed, A. W., Hadi, A. A., & Suganthan, P. N. (2021). *Problem definitions and evaluation criteria for the CEC 2022 special session and competition on single objective bound constrained numerical optimization*. IEEE Congress On Evolutionary Computation: Tech. Rep, Italy: Centro Congressi Padova. <https://github.com/P-N-Suganthan/2022-SO-BO>
- [15] Biedrzycki, R. (2024). Revisiting CEC 2022 ranking: A new ranking method and influence of parameter tuning. *Swarm and evolutionary computation*, 89, 101623. <https://doi.org/10.1016/j.swevo.2024.101623>
- [16] Biedrzycki, R. (2026). Analysis and simplification of the winner of the CEC 2022 optimization competition on single objective bound constrained search. *Evolutionary computation*, 34(2), 195–212. <https://doi.org/10.1162/evco.a.27>
- [17] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American statistical association*, 32(200), 675–701. <https://doi.org/10.1080/01621459.1937.10503522>
- [18] Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge. <https://www.routledge.com/Statistical-Power-Analysis-for-the-Behavioral-Sciences/Cohen/p/book/9780203771587>
- [19] Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological bulletin*, 114(3), 494–509. <https://doi.org/10.1037/0033-2909.114.3.494>
- [20] Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of modern applied statistical methods*, 8(2), 26. <https://doi.org/10.22237/jmasm/1257035100>

- [21] Franceschi, L., Donini, M., Perrone, V., Klein, A., Archambeau, C., Seeger, M., ... & Frasconi, P. (2025). Hyperparameter optimization in machine learning. *Foundations and trends in machine learning*, 18(6), 975–1109. <https://doi.org/10.1561/22000000088>
- [22] Morales-Hernández, A., Van Nieuwenhuyse, I., & Rojas Gonzalez, S. (2023). A survey on multi-objective hyperparameter optimization algorithms for machine learning. *Artificial intelligence review*, 56(8), 8043–8093. <https://doi.org/10.1007/s10462-022-10359-2>
- [23] Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated machine learning: Methods, systems, challenges* (pp. 3–33). Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-05318-5_1
- [24] Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems* (pp. 1–9). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf
- [25] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6), 80–83. <https://doi.org/10.2307/3001968>
- [26] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70. <https://doi.org/10.2307/4615733>
- [27] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society: Series b (methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>