Fixing our own CMA-ES bug reversed the headline

A preregistered benchmark of four plain and six biologically labelled optimisers on shifted test functions, run once with a bug and once after the fix

Zain Dana Harper

Version 1.0, 9 November 2026. Finding F183 in the findings ledger. Honest null and self-correction (tier D).

Abstract

Optimisers named after slime moulds, wolves, bees and ants are often reported to beat standard methods. Our first preregistered benchmark seemed to agree: the biologically labelled group had the better mean rank, 5.04 against 6.20. Tracing a poor CMA-ES score showed a bug in our own code, and several other methods had a related fault. We kept that run, fixed the box handling for every method, wrote a new preregistration and ran again on seven test functions in 10 dimensions. After the fix, CMA-ES ranked first on the shifted functions with a mean rank of 2.00, and the best method with a biological label, ACOR, ranked second at 3.14. Moving the optimum away from the centre cost the slime mould method 12.1 and the grey wolf method 13.6 decades of median error, while ACOR, IWO and PSO lost 0.0006, 0.07 and 0.42 decades. Without random search the plain group's mean rank was 3.98 against 5.54 for the biologically labelled group; with random search counted as plain the groups were 5.45 and 5.54, a gap that sat inside the label-permutation range (p 0.45). The group contest does not settle much in either direction. The method-level results do.

1 What this means for a reader

When a benchmark says a new nature-inspired optimiser wins, check two things: whether the test functions have their best point at the centre of the search box, and whether the standard methods it beat were run correctly. Our own first run failed the second check, and the result looked like a win for the nature-inspired group. Two popular methods in that group did well only when the answer sat at the centre. A good result on centred test functions can come from a method that guesses the centre.

2 Introduction

A metaheuristic is a general recipe for searching a function for its minimum by sampling and updating candidate points. Sörensen [1] described a flood of "novel" metaheuristics built on metaphors of natural or man-made processes and argued that the trend threatens to lead the field away from scientific rigour. Camacho-Villalón, Dorigo and Stützle [2] took six widely used metaphor-based methods for continuous problems apart into components, the grey wolf optimiser among them, and reported that none of them proposes a new idea; one section of their paper is titled "The grey wolf optimizer is PSO".

Kudela [3] pointed to a defect in how such methods are tested. Several common test functions have their minimum at the centre of the feasible set, and a method whose update rule has a centre bias solves those functions easily. Comparisons between such a method and one without the bias on centred functions say little.

The no free lunch theorems of Wolpert and Macready [4] give the background for any optimiser comparison: averaged over all possible functions, no method beats another, so an advantage on one class is paid for on another.

This study tested biologically labelled methods against plain ones with the optimum moved off the centre. It became a study of our own error as well. The first run's headline came from a bug in our CMA-ES code. CMA-

ES, the covariance matrix adaptation evolution strategy, samples candidates from a normal distribution and adapts the distribution's centre, shape and step size. Auger and Hansen [5] described a version that restarts with a doubled population each time, known as IPOP. Ours restarts with the same population size.

3 Preregistration and design

The study ran twice, each time under its own preregistration. The first (here, the first run) was hash-locked on 31 October 2026 before any result of the run was seen; only smoke runs in 3 dimensions with small budgets had been made, to fix crashes, and no rankings were read from them. It named the methods, the functions, the measures, two controls, the forecasts and the kill conditions. The decision it informs was written in it: no slime or plant-labelled optimiser becomes a default search method in our tools unless it beats differential evolution and CMA-ES on the shifted functions.

The second (the rerun) was hash-locked on 31 October 2026 after the first run had been read in full. It records what went wrong, the fix, and new forecasts. Its own kill condition reads: if CMA-ES fails to bring the median error on the shifted ellipsoid below 1, it is still broken and the study stops. It also commits to reporting any biologically labelled method that beats the best plain one, whatever we expected.

Both preregistrations' SHA-256 hashes are in the reproduction section.

4 Data and provenance

No external data were used. All results come from our own implementations of ten optimisers and seven standard test functions, run on one desktop computer. The method update rules were written from their published descriptions with default settings and no tuning. Every number in this paper is read from the two runs' result files.

5 Methods

Methods compared

The plain group is random search, L-BFGS-B with restarts and a numeric gradient, differential evolution (rand/1/bin) and CMA-ES with restarts. The biologically labelled group is the slime mould algorithm (SMA), the grey wolf optimiser (GWO), particle swarm optimisation (PSO), artificial bee colony (ABC), ant colony optimisation for continuous domains (ACOR) and invasive weed optimisation (IWO). Every method gets the same budget of 20,000 function evaluations; a numeric gradient spends evaluations from that budget.

Functions and arms

The seven functions are the sphere, an ellipsoid with a condition number of one million, Rosenbrock, Rastrigin, Ackley, Griewank and Schaffer F7, each in 10 dimensions with a minimum value of 0. The search box allows every coordinate from -100 to 100. On the shifted arm each base function g is evaluated as

$$f(x) = g(R(x - s)) \tag{1}$$

where s is a random shift with every coordinate drawn from -70 to 70 and R a random rotation. Every method sees the same s and R within one repeat. On the unshifted arm s is zero and R is the identity, so the minimum sits at the centre of the box. Each method ran 15 repeats on each function on each arm.

Measures

For method m on function f, let M_{fm} be the median best error over the 15 repeats on the shifted arm and U_{fm} the same on the unshifted arm, each floored at 10^{-12}. Methods are ranked on each function by the base-10

logarithm of M_{fm} rounded to two decimals, with ties sharing their average rank. Call that rank ρ_{fm}. The mean rank is

$$r_m = \frac{1}{7} \sum_{f} \rho_{fm} \tag{2}$$

The change in error from moving the minimum, in decades, is

$$\delta_{fm} = \log_{10} M_{fm} - \log_{10} U_{fm} \tag{3}$$

and a method's shift penalty D_m is the median of δ_{fm} over the seven functions. The group gap is

$$G = r_{bio} - r_{plain} \tag{4}$$

where r_{bio} and r_{plain} are the averages of r_m over each group, so a positive gap means the plain group did better.

The bug and the fix

In the first run CMA-ES scored a median error of 14 million on the ellipsoid. Tracing it showed that the distribution's centre had left the box (its norm reached 498 against a box half-width of 100). Every sample was clipped to the same corner, the step size fell to about 10^{-8}, and the method never restarted. Differential evolution, ABC, ACOR and IWO also kept candidates outside the box in their internal state. These were errors in our code.

For the rerun every method keeps its candidates inside the box. CMA-ES adds a quadratic penalty for any part of a sample outside the box,

$$F = f + 10000(1 + a) \sum_{i} e_i e_i \tag{5}$$

where a is the absolute value of f and e_i is how far coordinate i lies outside the box, zero inside. It restarts from a random point when its step size falls below 10^{-11}. No method's parameters were tuned.

6 Controls and planted checks

Two planted checks guard the instrument. CMA-ES must bring the shifted sphere below 10^{-8}. A centre cheat, a dummy method that samples only points next to the centre, must solve the unshifted sphere (error below 0.001) and fail the shifted one (base-10 logarithm of its error above 3). The cheat shows that the unshifted arm can be won by guessing the centre and that the shifted arm cannot.

The label permutation control shuffles method names within each function's ranks 2,000 times and records the group gap each time. A real gap must clear the 95th percentile of that null distribution.

The no-free-lunch check runs every method on a test function in which each integer cell of the box gets an independent random value, with a budget of 5,000 evaluations and 30 repeats. No method can exploit such a function, so the methods should tie. Ties are tested with the Kruskal-Wallis test, with L-BFGS left out because it can stall on a flat cell, on the best value found and, in the rerun, on the best of the first 500 distinct cells visited.

In the rerun CMA-ES reached a median error of 1.6e-15 on the shifted sphere, written in the e-notation used in the tables. The centre cheat reached 7.2e-7 on the unshifted sphere and an error whose base-10 logarithm is 4.27 on the shifted sphere. Both planted checks held.

7 Results

Table 1 gives every method's mean rank in both runs and its shift penalty in the rerun, and Figure 1 draws the ranks. After the fix, CMA-ES ranked first on the shifted functions with a mean rank of 2.00, and the best method

with a biological label, ACOR, ranked second at 3.14. The fix moved CMA-ES from 4.86 to 2.00 and PSO, the first run's leader, from 2.79 to 4.57.

Table 1. Mean rank over seven functions in 10 dimensions, 15 repeats each (1 is best), and the shift penalty in decades. The first run carries the bug.

Method	Group	First run, shifted	Rerun, shifted	Rerun, centred	Shift penalty
CMA-ES	plain	4.86	2.00	3.57	0
ACOR (ant colony)	labelled	4.14	3.14	4.36	0.0006
IWO (invasive weed)	labelled	6.29	4.43	7.86	0.07
L-BFGS	plain	3.64	4.57	5.93	0.05
PSO (particle swarm)	labelled	2.79	4.57	5.50	0.42
Differential evolution	plain	6.43	5.36	7.71	0.08
ABC (bee colony)	labelled	4.93	5.86	5.86	0.70
SMA (slime mould)	labelled	4.86	6.36	2.07	12.06
GWO (grey wolf)	labelled	7.21	8.86	2.43	13.56
Random search	plain	9.86	9.86	9.71	0.02

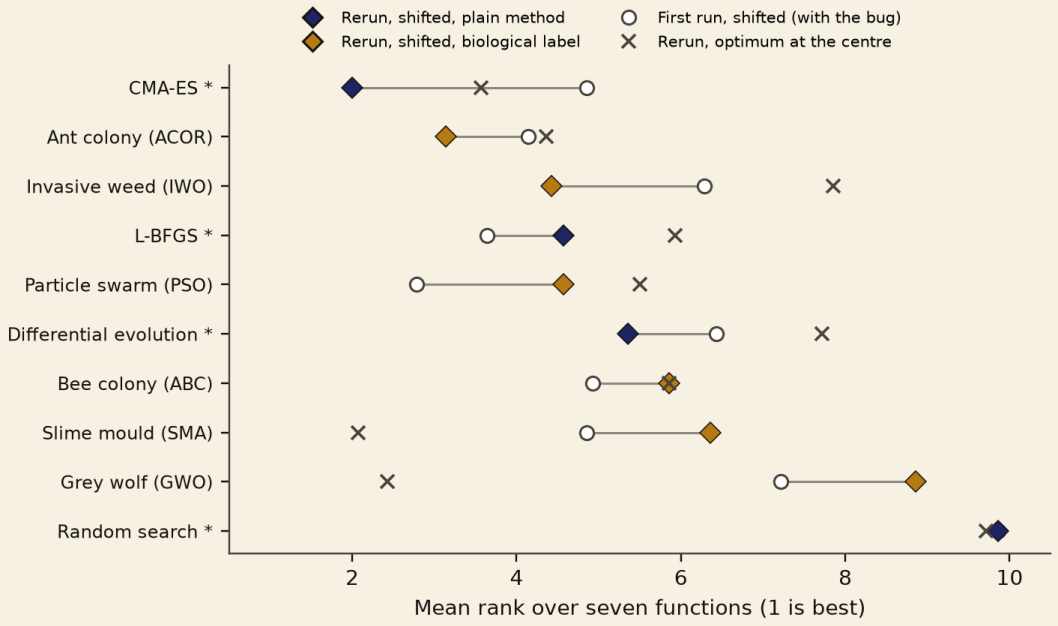


Figure 1. Mean rank on the shifted arm in the first run (circles) and the rerun (diamonds), with the rerun's rank when the optimum sits at the centre (crosses). Plain methods carry an asterisk. Drawn from the two runs' result files.

With the minimum at the centre, SMA ranked first (2.07) and GWO second (2.43). Figure 2 shows what moving the minimum did to each method. Moving the optimum away from the centre cost the slime mould method 12.1 and the grey wolf method 13.6 decades of median error, while ACOR, IWO and PSO lost 0.0006, 0.07 and 0.42 decades. The centre bias described by Kudela [3] shows in two of the six labelled methods and not in the other four.

Table 2 gives the group contest. Without random search the plain group's mean rank was 3.98 against 5.54 for the biologically labelled group; with random search counted as plain the groups were 5.45 and 5.54, a gap that sat inside the label-permutation range (p 0.45). The preregistered label-permutation control was computed only for the gap with random search included. The gap of 1.56 without it has no control of its own.

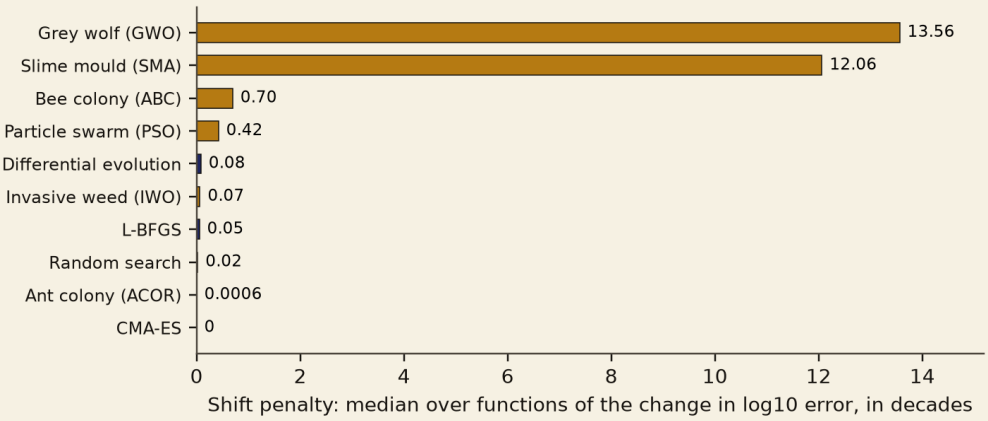


Figure 2. Shift penalty per method in the rerun: the median over seven functions of the growth in median error, in decades, when the minimum leaves the centre for a random shifted and rotated position.

Table 2. Group mean ranks on the shifted arm and the label-permutation control, 2,000 draws per run. The interval is the 5th to 95th percentile of the permuted gaps.

Comparison	Plain	Labelled	Gap	Permuted interval, p
First run, random counted as plain	6.20	5.04	-1.16	-1.19 to 1.19, p 0.95
Rerun, random counted as plain	5.45	5.54	0.09	-1.22 to 1.16, p 0.45
Rerun, random left out	3.98	5.54	1.56	not computed

In the first run the labelled group's lead of 1.16 ranks came partly from our bug and partly from counting random search as plain, and it sat inside its own label-permutation range. The first run's preregistration had a kill condition we did not apply when we first read it: if the no-free-lunch check fails to tie, the ranking is void. That check failed (p 0.0015), so by its own rule the first run's ranking was void. In the rerun the methods tied on that function, at p 0.13 on the best value and p 0.74 on the best of the first 500 distinct cells. The rerun's preregistration traced the first run's failure to converged methods revisiting the same cells, which the theorem does not count; in the rerun IWO visited a median of 2,870 distinct cells in 5,000 evaluations, against 5,000 for random search and ACOR.

Table 3 sets the rerun's forecasts against its outcomes.

Table 3. Forecasts in the rerun's preregistration against what the rerun gave.

Quantity	Forecast	Outcome
CMA-ES mean rank	at most 3	2.00
PSO among the top three	yes	no, 4.57
Plain mean rank, random left out	3.8	3.98
Labelled mean rank	5.0	5.54
Best plain ahead of best labelled	at least 0.3 ranks	1.14 ranks
SMA and GWO shift penalty	above 3 decades	12.06 and 13.56
No-free-lunch tie on distinct cells	p above 0.01	p 0.74
L-BFGS in the top two on smooth functions	yes	not scored

8 Discussion

The first run is a record of how a benchmark goes wrong in the direction of a popular story. CMA-ES was the strongest plain method we had, and our implementation of it was broken in a way that did not crash: it returned numbers, just bad ones. The labelled group won that run, and the error showed only because one score, 14 million on the ellipsoid, looked wrong. The rerun's preregistration was written after reading the first run, so the

rerun is not blind to it; what keeps it honest is that its forecasts and kill condition were fixed before it ran.

The method-level results are the ones that hold up. CMA-ES ranked first once it was run correctly. SMA and GWO ranked first and second on centred functions and near the bottom on shifted ones, which is the pattern Kudela [3] describes for methods with a centre-biased update step. ACOR, IWO and PSO lost almost nothing when the minimum moved, so the centre-bias charge does not reach every method with a biological label. Camacho-Villalón and colleagues [2] argue from the structure of the update rules that GWO is a form of PSO. In our runs the two behave differently once the minimum moves: GWO loses 13.6 decades and PSO 0.42. A component that pulls candidates toward the centre would explain the gap, and we did not test which component it is.

The group result is weak. With random search counted as plain, the two groups' mean ranks differ by 0.09, well inside the label-permutation range. Leaving random search out gives a gap of 1.56, but that comparison had no control. Ranking methods by group label says less than ranking them one at a time.

9 Limitations

- Seven functions, 10 dimensions, one budget of 20,000 evaluations and 15 repeats. Other functions, dimensions or budgets could reorder the methods.
- Every method used default parameters. Tuning could change the ranking.
- The shifted arm moves and rotates the function at once, so the shift penalty mixes the two effects. A rotation-only arm was not run.
- Errors are floored at 10^{-12} before ranking, so methods that reach that level tie, and a shift penalty measured from a centred error of 0 counts from 10^{-12}. The 12 and 13.6 decades for SMA and GWO are bounded by that floor.
- The methods were written from published descriptions by us, and the first run shows how such code can be wrong without crashing. Other bugs may remain.
- PSO's biological label is thin: it is a velocity rule with a metaphor, and it predates the recent flood of such methods.
- Scoring was done by our own code and checked by the same analyst who wrote it.

10 Verified

Verified: on 9 November 2026 every mean rank, group mean, gap, shift penalty, planted-check value, label-permutation mean, percentile and p value, and both no-free-lunch p values in this paper were re-read from the two runs' result files. The rerun's values match the findings ledger. The function count, dimension, budget, repeats, box, shift range and rank rules were read from the study's code and build files. Both preregistrations' SHA-256 hashes were recomputed from the locked files and match the hashes recorded in the result files. The first run left two result files with identical analysis values; this paper cites the later one. Both figures were drawn by a script that reads only the two result files. The study was not re-run for this paper. The literature in the introduction was found by searches run on 9 November 2026 and read at the level of abstracts and publisher pages; the full texts of the cited papers were not read.

11 Corrections

- 31 October 2026: the first run's headline, that the labelled group beat the plain group, came partly from a bug in our CMA-ES code and in the box handling of four other methods. The run is kept as it ran, and the rerun replaces its ranking.
- 9 November 2026: the first run's preregistration voids its ranking when the no-free-lunch check fails to tie, and it failed. Our ledger entry did not say so; this paper does.

- 9 November 2026: our ledger and write-up did not report that the group gap with random search sat inside the label-permutation range in both runs, nor that the rerun's forecast of PSO in the top three failed. This paper reports both.
- 9 November 2026: first published version.

12 Reproduction

- Implement the ten methods with default parameters and keep every candidate inside the box from -100 to 100. Give CMA-ES the penalty of eq. 5 and a restart when its step size falls below 10^{-11}.
- Run each method on the seven functions in 10 dimensions with 20,000 evaluations and 15 repeats, once with the minimum at the centre and once shifted and rotated as in eq. 1, the same shift and rotation for every method within a repeat.
- Rank methods per function by the median best error, floored at 10^{-12} and taken as a base-10 logarithm rounded to two decimals, and average the ranks, eq. 2. Compute eqs. 3 and 4.
- Permute method labels within each function 2,000 times for the null distribution of the gap.
- Rerun preregistration SHA-256: dbf3b21010c50808e984aa422ad15639f3fd93af28409e44eb4ca34c8cc53a54. Plan SHA-256: 234b69bfb9702c86b356344553e6e8de24169a843d994fdac870bbc0575bef39. Run 20261031T085324Z, seed 20261054.
- First run preregistration SHA-256: 43edb85abe56495fcd70d726ef7b6f224917033ac10fefbd56fa9d1f2aa559c0. Plan SHA-256: e6fab72152c93c516de1f03d02999c1bcd691360bbef5123781902484f1059ba. Run 20261031T083319Z, seed 20261041.
- Cost: the rerun's timing record gives 1,179 seconds for the analysis on one desktop CPU.

13 Appendix A. The preregistered verdicts

Each verdict is a set of rules written before the run. A verdict passes only when every rule holds. Values are read from the result files.

Table 4. Preregistered verdicts of the rerun and the rule that decided each one.

Verdict	Deciding rule	Value	Holds
CMA-ES solves the sphere	shifted sphere error below 1e-8	1.6e-15	yes
CMA-ES solves the ellipsoid	shifted ellipsoid median error below 1	2.4e-15	yes
Cheat caught	unshifted below 0.001 and shifted log10 above 3	7.2e-7; 4.27	yes
No free lunch, raw	Kruskal-Wallis p above 0.01	0.13	yes
No free lunch, distinct	Kruskal-Wallis p above 0.01	0.74	yes
Plain beats labelled	gap with random left out above 0	1.56	yes
Best plain beats best labelled	best labelled rank minus best plain rank above 0	1.14	yes
No labelled method dominates	wins over every plain method on at most 2 of 7 functions	1	yes
SMA centre bias	shift penalty above 3 decades	12.06	yes
GWO centre bias	shift penalty above 3 decades	13.56	yes
Plain methods unbiased	CMA-ES and differential evolution penalty below 0.5	0; 0.08	yes

14 Glossary

Each term is defined where it is first used in the text. The same definitions, with the field each comes from and what each excludes, are collected here.

- **metaheuristic**. A general recipe for searching for the minimum of a function by sampling and updating a set of candidate points, used when the function gives no gradient or has many local minima. Field: optimisation. Excludes exact methods that come with a proof of finding the minimum.

Table 5. Preregistered verdicts of the first run, with the bug.

Verdict	Deciding rule	Value	Holds
CMA-ES solves the sphere	shifted sphere error below 1e-8	3.9e-15	yes
Cheat caught	unshifted below 0.001 and shifted log10 above 3	6.2e-7; 4.18	yes
No free lunch	Kruskal-Wallis p above 0.01	0.0015	no
Plain beats labelled	gap with random counted as plain above 0	-1.16	no
Best plain beats best labelled	best labelled rank minus best plain rank above 0	-0.86	no
No labelled method dominates	wins over every plain method on at most 2 of 7 functions	2	yes
SMA centre bias	shift penalty above 3 decades	12.06	yes
GWO centre bias	shift penalty above 3 decades	13.49	yes
Plain methods unbiased	CMA-ES and differential evolution penalty below 0.5	0.45; 0.26	yes

- **test function**. A formula with a known minimum, chosen to be easy or hard in a known way, on which optimisers are compared by how close they get. Field: optimisation benchmarking. Excludes real engineering problems, whose minimum is unknown.
- **centre bias**. A tendency of a method's update rule to pull candidates toward the centre of the search box, which helps on any function whose minimum happens to sit there. Field: evolutionary computation benchmarking, after Kudela. Excludes good performance that survives when the minimum is moved.
- **no free lunch**. The result that, averaged over all possible functions, every search method that never revisits a point does equally well, so any advantage on one class of functions is paid for on another. Field: optimisation theory, after Wolpert and Macready. Excludes claims about any one class of functions, such as the test functions here.
- **biologically labelled method**. An optimiser whose published description is framed by an animal, plant or other organism, here slime mould, grey wolf, particle swarm, bee colony, ant colony and invasive weed. Field: our usage for this study. Excludes any claim that the method works the way the organism does.
- **CMA-ES**. Covariance matrix adaptation evolution strategy: it samples candidates from a multivariate normal distribution and adapts the distribution's centre, shape and step size from the best samples. Field: evolutionary computation, after Hansen and colleagues. Excludes the larger-population restart schedule known as IPOP, which our version did not use.
- **preregistration**. A dated written plan for a study, fixed and hash-locked before any result is seen: the question, the data, the statistic, the controls, the forecasts and the kill conditions. Field: research method (metascience). Excludes a plan revised after looking at the outcome, and a protocol written only for the record after the run.
- **control**. A version of the data with the structure under test removed (for example shuffled or pooled), scored the same way as the real data, so the real score has something to beat. Field: statistics and experimental design. Excludes a comparison with a different dataset or with a published number.
- **kill condition**. A rule written in the preregistration that, when met, decides against the claim with no room for judgement after the run. Field: research method (our usage, after Popper's falsification). Excludes reasons found after the run for setting a result aside.
- **search box**. The allowed region, here every coordinate between -100 and 100; a method may only propose and keep points inside it. Field: optimisation. Excludes constraints other than simple bounds on each coordinate.
- **shifted arm**. The version of each test function whose minimum is moved to a random point with every coordinate between -70 and 70 and whose axes are turned by a random rotation, the same for every method within one repeat. Field: our usage for this study. Excludes the unshifted arm, with the minimum at the centre and the axes aligned.
- **mean rank**. Each method's rank on each function by median final error, averaged over the seven functions;

1 is best and 10 is worst. Field: benchmarking statistics. Excludes how large the differences in error are.

- **decade**. A factor of ten; a change of one decade in error is a change of one in its base-10 logarithm. Field: measurement. Excludes a difference on a linear scale.
- **shift penalty**. For one method, the median over the seven functions of how many decades its median error grows when the function moves from the unshifted arm to the shifted arm. Field: our usage for this study. Excludes effects of the rotation alone, which the design does not separate from the shift.
- **planted check**. Data built so that the right answer is known, run through the same analysis before the real data, to show that the analysis can see the effect when it is there. Field: research method; called a positive control in experimental science. Excludes any check made on the real data, which cannot show what the analysis misses.
- **centre cheat**. A dummy method that only samples points within a tiny distance of the centre of the box, so it solves any function whose minimum sits there and nothing else. Field: our usage for this study. Excludes any real optimiser.
- **label permutation**. Shuffling which method name goes with which rank within each function, 2,000 times, to see how large a gap between the two groups' mean ranks arises by chance. Field: statistics (permutation test). Excludes uncertainty from the random repeats of each run.
- **null distribution**. The spread of scores a control gives over many random draws; its 95th percentile is the bar a real score must clear. Field: statistics. Excludes a theoretical distribution assumed rather than drawn from the data.
- **Kruskal-Wallis test**. A rank-based test of whether several groups of numbers come from the same distribution. Field: statistics. Excludes a test of which group is better.
- **void**. The verdict of a run whose planted check failed or whose sample fell below its stated minimum; it says nothing for or against the claim. Field: research method (our usage). Excludes a null result, which comes from a working instrument.

15 References

[1] Sörensen, K. (2015). Metaheuristics: the metaphor exposed. International Transactions in Operational Research 22(1), 3-18. doi:10.1111/itor.12001

[2] Camacho-Villalón, C. L., Dorigo, M. and Stützle, T. (2023). Exposing the grey wolf, moth-flame, whale, firefly, bat, and antlion algorithms: six misleading optimization techniques inspired by bestial metaphors. International Transactions in Operational Research 30(6), 2945-2971. doi:10.1111/itor.13176

[3] Kudela, J. (2022). A critical problem in benchmarking and analysis of evolutionary computation methods. Nature Machine Intelligence 4, 1238-1245. doi:10.1038/s42256-022-00579-0

[4] Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation 1(1), 67-82. doi:10.1109/4235.585893

[5] Auger, A. and Hansen, N. (2005). A restart CMA evolution strategy with increasing population size. Proceedings of the IEEE Congress on Evolutionary Computation 2005, 1769-1776. doi:10.1109/CEC.2005.1554902