# Error Bars Change the Leaderboard: Handcrafted Features Are Not Beaten in Automated Alpha Mining

Anonymous authors
Paper under double-blind review

## Abstract

A benchmark can order methods it has no power to separate, and the ordering can depend on choices its table never reports. We measure how far both go on automated alpha mining, the search for formulas that predict the cross-section of stock returns, which has its own machine learning leaderboard. Miners built on genetic programming, reinforcement learning, GFlowNets and LLM agents are ranked by mean information coefficient (IC) on a fixed test period. Some report dispersion across search seeds, none reports uncertainty over time, and none tests whether the ordering between methods is significant. We re-run the released code of AlphaGen, AlphaSAGE and a genetic-programming baseline, at their published budgets and at a matched one, together with AlphaAgent's released formulas, the Alpha158 handcrafted library and the Kronos foundation model zero-shot, under one point-in-time protocol on CSI300 constituents. No miner is ever the unique survivor of a ninety percent model confidence set, at either horizon or on any window. At twenty days, the horizon they are trained for, a six-factor style model has the highest mean IC, 0.0768 against 0.0716 for its closest competitor, and every searcher run at its released budget falls outside the set. At one day, the label these benchmarks score on, the set retains a single handcrafted model, which kills one of our own pre-registered claims and we report the kill. The ordering it publishes then fails three separate checks. Only 16 of 66 pairs separate under familywise control, and among methods that sit next to each other in the ranking, 1 of 11 do. Switching to an information ratio or a cost-adjusted Sharpe drops the style model below the handcrafted library and below a mined pool, because its IC climbs with the forecast horizon while fitted signals peak and decay. Spending AlphaGen's released budget instead of a matched one buys +0.0047 training IC and loses -0.0080 out of sample. We also pre-registered an online false discovery layer over candidate streams; it works in simulation, abstains on real streams, and fails its own test, which we report as a negative result. Pools, candidate streams, pre-registrations and a 19-check verification suite are released.

## 1 Introduction

No automated alpha miner in this comparison is ever the single model a confidence set retains. At the horizon the miners are trained for, none of them beats a linear model built from six textbook style factors; at the one-day horizon their own benchmarks score, the model left standing is a handcrafted feature library. On CSI300 constituents over 969 trading days in 2021–2024, that style model has the highest mean information coefficient of every method we ran, and the ninety percent model confidence set (Hansen et al., 2011) contains it while excluding genetic programming, AlphaSAGE, a short-term reversal signal and random formulas drawn from the same grammar. Figure 1 and Table 1 are the result in one place.

The miners do find something. The claim here is about what their benchmarks can establish about which of them finds more. Papers in this line differ in what they report. AlphaGen



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

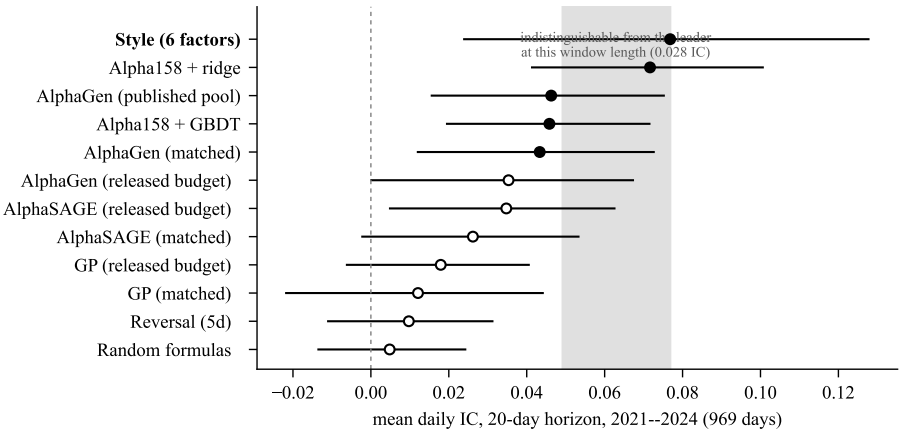



Figure 1: Twenty-day horizon, CSI300, 2021–2024. Mean daily IC with stationary-bootstrap confidence intervals. Filled markers are inside the ninety percent model confidence set, hollow markers are excluded. The shaded band is the median pairwise resolution at this window length, not a boundary of the confidence set: a method can fall outside the band and still be retained, because the set uses the full correlation structure between methods and the band uses a median.

gives means and standard deviations over ten runs and AlphaSAGE over five (Yu et al., 2023; Chen et al., 2026); the AlphaEval and AlphaBench evaluation frameworks report one number per method (Ding et al., 2026; Luo et al., 2026). None of the four attaches uncertainty over time to a mean IC, and none tests whether the gap between two methods is larger than sampling noise.

We therefore measured what the evaluation can resolve before interpreting its point estimates. A four-year daily window separates methods that are far apart and cannot separate methods that are close. Close is where every published leaderboard makes its claim. That finding also binds our own result: what the data support is not that the style model is the best of these methods, but that no method in this comparison is demonstrably better than it. None of the failure modes we report is specific to finance. Overlapping labels, a metric fixed by convention and a search budget reported as an implementation detail are common enough elsewhere; alpha mining is where all of them could be measured at once, because its labels are noisy enough that the gaps between methods sit near the resolution of the test.

Our contributions:

- A common-protocol comparison in which no automated miner is ever the unique survivor of the confidence set, at either horizon or on any window. At twenty days a six-factor style model has the highest mean IC and no miner beats it; at one day, including after every method's cutoff, the only retained models are variants of a handcrafted feature library.
- A pre-registered primary claim that fails. We committed to rejecting C1 if the style model left the one-day confidence set. It does, once the handcrafted baselines our own amendment required are included, and we report the kill rather than the version of the table that predates them.
- A resolution limit for this benchmark: which pairs a four-year daily window can separate, which it cannot, and how many years the ones it cannot would need.
- Three measurements of why the published ordering survives: a horizon mismatch inside released benchmark code, seed variation that can exceed time-sampling uncertainty, and rankings that do not persist across periods.
- To our knowledge, the first evaluation of these miners and of a financial foundation model on data postdating every method's cutoff.

- A pre-registered selection layer under online false discovery control that fails its prespecified test, which we retain as a negative result and analyze to identify the failure mechanism.

## 2 Protocol

Data. Daily bars for CSI300 constituents with point-in-time membership, 2010-01-01 to 2026-07-31, Qlib format. Halted days are dropped. Prices follow the Qlib convention in which each stock's first close is one.

Split. Train 2010–2019, validate 2020, test 2021–2024, and a further window 2025-01-01 to 2026-07-31. Every method uses the same split, rather than each paper's own.

Label. $\text{close}_{t+h}/\text{close}_t - 1$ for $h \in \{1, 20\}$. Miners are trained and scored at the same horizon, which is the comparison the released code does not make (Section 5). Signals use information up to day $t$ and the return starts at that same close, with no implementation lag. That is the convention in the work we compare against and it is applied identically to every method, so it moves the levels rather than the ordering.

Models. AlphaGen (Yu et al., 2023), AlphaSAGE (Chen et al., 2026) and the genetic-programming baseline from the AlphaGen repository are run from released code with the data loader pointed at our panel. We add AlphaAgent's released formulas and AlphaGen's published pool, the Alpha158 library (Yang et al., 2020) with a ridge and a gradient-boosted-tree model on top, the Kronos foundation model zero-shot (Shi et al., 2025b), a six-factor style model, a five-day reversal signal, and pools of random formulas from the miners' own grammar. The style factors are size proxied by traded amount, market beta, volatility, twelve-month momentum, one-month reversal and abnormal turnover, each rank-gaussianized daily, combined by a regression fit on the training period only.

What we changed in the competitors. Two deviations matter and neither is cosmetic. First, the combination weights of every pooled miner are refit by ridge on the training years rather than taken as released: AlphaGen's own mean-squared-error objective diverged on pools of tens of alphas, giving that pool a negative training IC, and AlphaSAGE's released pool stores absolute-IC placeholders rather than fitted weights. The scoring head of every competitor is therefore ours, and it was chosen with a competitor's numbers visible. Second, the planted-leakage null check failed as first written, because a real model scored above the weak planted signal; the control that the claims are gated on is a stronger planted model added after that failure. Both are recorded with dates and pre-fix numbers in the pre-registration amendments.

Statistics. We compute daily Pearson IC, with stationary-bootstrap confidence intervals using a mean block of 20 at the twenty-day horizon, matching the label overlap, and 10 at one day, which the pre-registration fixed and which is conservative for a nearly white series. Rankings are tested with the model confidence set at ninety percent. Two null checks are run before any claim is read: pools of random formulas must not separate from each other, and a planted leakage model must be the unique survivor.

## 3 The twenty-day leaderboard

Table 1 orders every model by mean daily IC over the test period. The style model leads at 0.0768 [0.0239, 0.1277], the handcrafted Alpha158 library with a ridge is the closest competitor, the two AlphaGen pools follow, and the confidence set keeps 5 of the twelve models. Random formulas drawn from the miners' own grammar score positively, which is what makes an interval necessary: a mean IC above zero is not by itself evidence of search working.

At one day, the label AlphaEval scores every method on, the ordering changes and the comparison stops being close. The Alpha158 library with gradient-boosted trees reaches

Table 1: Twenty-day horizon, CSI300, 2021–2024. Mean daily IC, stationary-bootstrap CI, and model confidence set membership at 90 percent. The style model is not beaten by any miner.

| Model | Mean IC | 95% CI | MCS $p$ | In set |
|---|---|---|---|---|
| Six-factor style model | +0.0768 | [+0.0239, +0.1277] | 1.000 | ✓ |
| Alpha158 + ridge | +0.0716 | [+0.0414, +0.1006] | 0.809 | ✓ |
| AlphaGen published pool | +0.0463 | [+0.0156, +0.0752] | 0.226 | ✓ |
| Alpha158 + GBDT | +0.0458 | [+0.0195, +0.0715] | 0.226 | ✓ |
| AlphaGen (matched budget) | +0.0433 | [+0.0121, +0.0726] | 0.226 | ✓ |
| AlphaGen (released budget) | +0.0353 | [+0.0001, +0.0673] | 0.049 | |
| AlphaSAGE (released budget) | +0.0348 | [+0.0049, +0.0625] | 0.031 | |
| AlphaSAGE (matched budget) | +0.0262 | [-0.0022, +0.0533] | 0.028 | |
| GP (released budget) | +0.0179 | [-0.0062, +0.0405] | 0.000 | |
| GP (matched budget) | +0.0121 | [-0.0217, +0.0441] | 0.001 | |
| 5-day reversal | +0.0097 | [-0.0110, +0.0312] | 0.028 | |
| Random formulas | +0.0048 | [-0.0135, +0.0242] | 0.001 | |

Table 2: One-day horizon, the label AlphaEval scores every method on, CSI300 2021–2024.

| Model | Mean IC | 95% CI | MCS $p$ | In set |
|---|---|---|---|---|
| Alpha158 + GBDT | +0.0511 | [+0.0423, +0.0607] | 1.000 | ✓ |
| Alpha158 + ridge | +0.0339 | [+0.0251, +0.0432] | 0.000 | |
| AlphaGen (released budget) | +0.0249 | [+0.0164, +0.0337] | 0.000 | |
| AlphaAgent released formulas | +0.0189 | [+0.0101, +0.0279] | 0.000 | |
| GP (matched budget) | +0.0138 | [+0.0065, +0.0214] | 0.000 | |
| AlphaGen published pool | +0.0116 | [+0.0037, +0.0194] | 0.000 | |
| Six-factor style model | +0.0064 | [-0.0043, +0.0169] | 0.000 | |
| 5-day reversal | -0.0007 | [-0.0126, +0.0117] | 0.000 | |
| Random formulas | -0.0028 | [-0.0088, +0.0031] | 0.000 | |

0.0511 against 0.0249 for the best miner and 0.0064 for the style model, and the confidence set retains it alone (Table 2). The style model is excluded at $p = 0.0002$, and so is every miner. Our pre-registration names this outcome and its consequence: claim C1 was to be rejected if the style model left the one-day set, and amendment 5, which added these baselines, committed us to reporting a handcrafted or gradient-boosted model that dominates. Both conditions are met, so C1 fails at one day. It survives at twenty days, which is the horizon the miners are trained for and the one the title refers to.

## 4 What a four-year window can resolve

A confidence interval on a single method does not answer the leaderboard's comparative question. The leaderboard asks whether method $A$ beats method $B$, and two methods scored on the same days have correlated daily ICs, so the paired difference is estimated far more precisely than either mean. We therefore resample the paired daily difference of every pair with the stationary bootstrap at the label's block length, and record how often the published ordering reverses.

The coarseness is specific. At twenty days, 39 of 66 pairs separate at five percent uncorrected, so the window does distinguish a mined pool from random formulas. But testing sixty-six pairs at once creates sixty-six chances to be impressed by noise. Under Holm control at five percent familywise the count falls to 16 of 66, and among pairs adjacent in the ranking it falls from 3 of 11 to 1. The typical detectable gap at this window length is 0.0276 against a median neighbouring gap of 0.0052, so the detectable gap is more than five times the median observed gap. Published tables are decided in the region the test cannot see.

Table 3: What a test window of this length can resolve. Every pair is compared on its own paired daily IC difference, so the correlation between two methods is used rather than ignored. These are simultaneous comparisons, so the Holm columns are the honest ones: distant pairs separate and neighbours on the leaderboard never do.

| | | Typical | All pairs separating | | Adjacent pairs | |
|---|---|---|---|---|---|---|
| Horizon | Window | detectable gap | at 5% | Holm | at 5% | Holm |
| Raw IC | | | | | | |
| 1-day | 2021–2024 | 0.0134 | 26 of 36 | 22 of 36 | 2 of 8 | 1 of 8 |
| 1-day | 2025–2026 | 0.0236 | 33 of 55 | 21 of 55 | 2 of 10 | 0 of 10 |
| 20-day | 2021–2024 | 0.0276 | 39 of 66 | 16 of 66 | 3 of 11 | 1 of 11 |
| 20-day | 2025–2026 | 0.0460 | 20 of 66 | 4 of 66 | 0 of 11 | 0 of 11 |
| After neutralization | | | | | | |
| 1-day | 2021–2024 | 0.0088 | 28 of 36 | 28 of 36 | 3 of 8 | 3 of 8 |
| 1-day | 2025–2026 | 0.0125 | 34 of 55 | 20 of 55 | 2 of 10 | 0 of 10 |
| 20-day | 2021–2024 | 0.0154 | 29 of 66 | 10 of 66 | 1 of 11 | 0 of 11 |
| 20-day | 2025–2026 | 0.0276 | 19 of 66 | 4 of 66 | 0 of 11 | 0 of 11 |

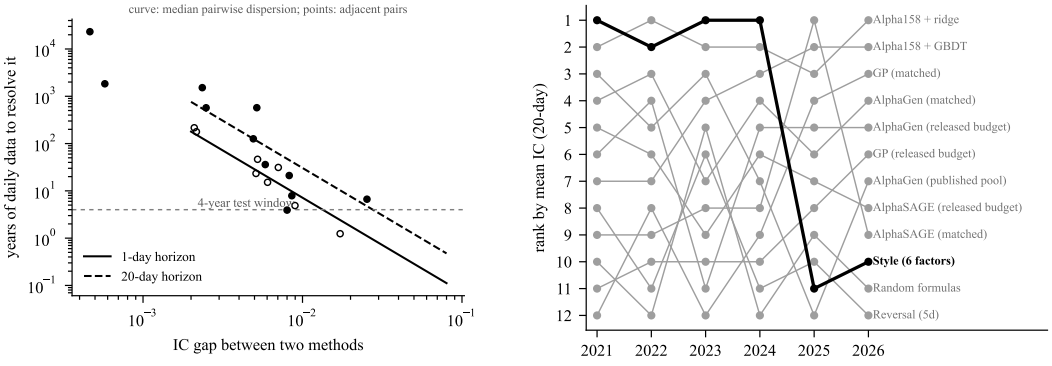



Figure 2: Left: years of daily data needed to resolve a gap of a given size. The curve uses the median pairwise dispersion; the points are the adjacent pairs themselves, which carry their own standard errors and therefore sit off it. Right: each method's rank by mean IC, year by year. The style model is the heavy line.

How much data would close it? Extrapolating the observed dispersion, the median neighbouring pair needs 126 years of daily data. That number should be read as an order of magnitude and nothing finer: it is a squared reciprocal of a gap estimated with a signal-to-noise ratio below one, and resampling the same window puts it anywhere from 5 to 24241 years. It also assumes the gap is a constant, which Section 5 shows it is not. Both caveats point the same way, so we state it as a lower bound on the difficulty rather than a forecast: decades at least, under an assumption that flatters the benchmark.

The estimand matters as much as the window. Removing style and statistical-factor exposure halves the typical detectable gap, from 0.0276 to 0.0154, because the common component that dominates raw daily IC is also what makes it volatile. A benchmark that reported residual rather than raw IC would see more with the same data.

## 5 Why the ordering survives

The horizon in the code is not the horizon in the table. In the released AlphaEval repository, three of the eight searchers define a twenty-day training target, AlphaGen, AlphaQCM and AlphaForge, while the evaluation code scores every method on one-day returns; the genetic-programming baseline trains on one-day returns and AlphaAgent's released formulas are scored on them. Our direct estimate of the effect is weaker than the mismatch itself. Training the same genetic-programming searcher at twenty days and scoring it at one moves

Table 4: After removing style and Marchenko–Pastur statistical factor exposure day by day, twenty-day horizon, 2021–2024. The ordering changes and the set tightens: what survives is the handcrafted feature library, not a mined pool.

| Model | Neutralized mean IC | MCS $p$ | In set |
|---|---|---|---|
| Alpha158 + ridge | +0.0312 | 1.000 | ✓ |
| Alpha158 + GBDT | +0.0255 | 0.274 | ✓ |
| AlphaGen (matched budget) | +0.0144 | 0.174 | ✓ |
| AlphaGen published pool | +0.0143 | 0.174 | ✓ |
| AlphaGen (released budget) | +0.0141 | 0.174 | ✓ |
| AlphaSAGE (released budget) | +0.0136 | 0.174 | ✓ |
| GP (matched budget) | +0.0099 | 0.166 | ✓ |
| AlphaSAGE (matched budget) | +0.0097 | 0.115 | ✓ |
| GP (released budget) | +0.0073 | 0.083 | |
| 5-day reversal | +0.0066 | 0.174 | ✓ |
| Random formulas | +0.0005 | 0.060 | |
| Six-factor style model | -0.0093 | 0.060 | |

mean IC by 0.0123 across 3 seeds at an identical budget, but the seeds run from -0.0128 to 0.0268 and 1 of 3 has the opposite sign, and the only other searcher we could test this way moves 0.0006 on one seed with budget confounded with horizon. We report the mismatch as a fact about the code, verified at file and line, and the cost as not established at the number of seeds we could afford.

Seed variation can exceed time-sampling uncertainty. For genetic programming at one day, the standard deviation of test IC across search seeds is 0.0221, against a time-sampling standard error of 0.0050. The two sources are different questions: one asks whether this training run was lucky, the other whether this test period was. Where a paper reports seed dispersion it is reporting the first; where it reports one run, as the evaluation frameworks do, it is reporting neither. Table 6 in the appendix puts both columns side by side, with the seed counts, which are small.

The exclusions do not depend on averaging over seeds. Families with several runs enter the comparison as a seed average, which is less noisy than any single run and therefore easier to exclude. Redoing the twenty-day set with one randomly chosen run per family, over 5 draws, leaves the conclusion intact: 4 models are retained in no draw at all, while 4 are retained in every draw, including the style model and both handcrafted baselines. The one membership that moves is AlphaGen at the matched budget, retained in 4 of 5.

Rankings persist inside the test period and break after it. Within-period stability overstates out-of-period stability. At twenty days, the mean Kendall rank correlation between two test years is 0.53 (range 0.24 to 0.79); between a test year and the 2025–2026 period it is 0.07 (range -0.21 to 0.33). Against a null standard deviation of 0.22 at this number of models, no single correlation is significant, so the supported statement is directional: the ordering reproduces from year to year inside the reported window, but not outside it. The right panel of Figure 2 shows the consequence. The style model ranks first in each of 2021–2024, then 11 and 10 of 12 in 2025 and 2026, where its mean IC turns negative, -0.0323 and -0.0011. Alpha158 with a ridge is the exception: it never ranks worse than 3, ranks 1 in both later years, and stays positive throughout, +0.0950 in 2026.

## 6 What survives neutralization

A style model can lead on raw IC because raw IC rewards exposure to well-known factors. The fair follow-up is to strip that exposure from every model and re-run the comparison, so we residualize each signal day by day on the six style factors and on Marchenko–Pastur statistical factors, then score again.

Table 5: The same signals under a yardstick a desk would use. IR is the information ratio of the daily IC series and $t$ its Newey–West statistic. The book is dollar-neutral and rank-weighted, held for twenty days so a twentieth turns over each day, and net figures charge 20 basis points on one-way turnover. Ranking by mean IC is not the same as ranking by what it earns.

| Model | Mean IC | IR | NW $t$ | Turnover | Gross SR | Net SR |
|---|---|---|---|---|---|---|
| Six-factor style model | +0.0768 | +0.342 | +2.70 | 1.8% | +1.03 | +0.91 |
| Alpha158 + ridge | +0.0716 | +0.476 | +4.33 | 2.5% | +1.67 | +1.42 |
| ALPHAAGENT TRAINED H1 | +0.0507 | +0.315 | +3.57 | 2.9% | +1.15 | +0.86 |
| AlphaGen published pool | +0.0463 | +0.352 | +2.73 | 2.1% | +1.21 | +0.99 |
| Alpha158 + GBDT | +0.0458 | +0.350 | +3.23 | 2.8% | +1.15 | +0.84 |
| AlphaGen (matched budget) | +0.0433 | +0.300 | +2.47 | 2.8% | +0.97 | +0.68 |
| AlphaGen (released budget) | +0.0353 | +0.238 | +1.83 | 2.6% | +0.66 | +0.41 |
| GP (released budget) | +0.0267 | +0.196 | +1.48 | 2.6% | +0.74 | +0.44 |
| GP (matched budget) | +0.0173 | +0.121 | +0.91 | 2.5% | +0.41 | +0.16 |
| 5-day reversal | +0.0097 | +0.055 | +0.87 | 3.3% | +0.26 | -0.12 |
| Random formulas | +0.0048 | +0.035 | +0.27 | 1.5% | +0.32 | +0.15 |

The ordering changes and the style model's advantage disappears entirely. What is left of it is -0.0093, which is what removing its own factors should do, and it is one of only 3 models the set excludes. The Alpha158 library with a ridge leads at 0.0312, the mined pools retain a little, 0.0141 for AlphaGen, and the set keeps 9 of 12 models because at this resolution it cannot separate what is left.

An earlier version of this comparison, run before the released-budget arms existed, retained a single model and appeared to exclude every mined pool. Adding those arms dissolved the exclusion. What is stable here is the ordering, not the set. Alpha158 with a ridge leads the neutralized comparison under every roster and at both conventional levels, and the style model is excluded under both. The raw comparison is not fragile this way, keeping the same 5 models at 0.05 and 0.10.

Read with Table 1, neither leader survives the other's test. The style model leads on raw IC and the handcrafted library leads on residual IC, and no searched pool leads on either.

## 7 After every cutoff

Kronos is scored only after its stated cutoff. Its paper states that the pre-training data extends up to June 2024 and that its own test period therefore begins in July 2024 (Shi et al., 2025b). It reaches 0.0117 [0.0010, 0.0226] at one day, positive and below the mined pools on the same days.

One pre-registered null check fails on the twenty-day 2025–2026 window: only 7 of 10 random pools are retained there. We retained the prespecified test and investigated the failure. Demeaning the random columns makes the null exactly true while preserving their autocorrelation and cross-correlation, and on resamples of that construction the confidence set retains all ten with probability 0.72, so the observed failure is an ordinary event and not evidence of a miscalibrated set. The underlying assumption also fails: random formulas from one grammar do not share an expected IC, and on this window their means run from -0.0512 to 0.0236. We therefore read N1 as a conservative power diagnostic, report the window, and flag it as the one place a pre-registered gate did not pass.

## 8 The ordering changes under a different yardstick

Mean IC scores a book re-formed every day at no cost, which is not how a twenty-day signal is held. We rebuilt each signal into a dollar-neutral, rank-weighted book held for twenty days, so a twentieth of it turns over daily, and scored the same models on what they earn rather than on what they correlate with.

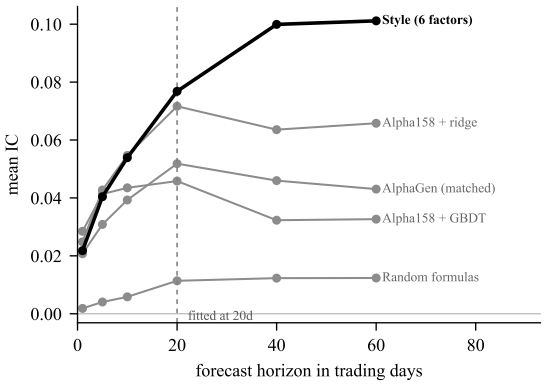



Figure 3: Mean IC against the forecast horizon, for signals built at twenty days. A fitted or mined signal peaks near the horizon it was built for and falls away. The style model does not peak over the evaluated horizons.

The leaderboard reorders. The style model has the highest mean IC and ranks fourth by information ratio among 11 models, 0.342 against 0.476 for Alpha158 with a ridge, because its daily IC is the most volatile in the comparison. Net of twenty basis points on turnover it earns a Sharpe ratio of 0.91 against 1.42, and 1 mined pool overtakes it on both measures. Mean IC is a choice of yardstick that changes the answer, and no paper in this line reports which yardstick its ranking depends on.

Figure 3 shows why, and it is the mechanism behind the neutralization result. The style model's IC rises monotonically with the forecast horizon, from 0.0218 at one day to 0.1012 at sixty, and has not turned over where we stop looking. Alpha158 peaks at the horizon it was fitted for, 0.0716 at twenty days falling to 0.0658 at sixty, and the mined pools behave the same way. A style model is not a competing forecast that happens to win at twenty days; it is an exposure whose accumulated premium grows with the holding period. At twenty days the benchmark is comparing an accumulating exposure against decaying alpha and calling the result a ranking.

## 9 A selection layer that fails

Every miner evaluates thousands of formulas and reports ten to fifty, which is a multiple-testing problem its scoring rule ignores. We pre-registered OFA, a selection layer for any miner: it tests the candidate stream on the validation year after removing style and statistical-factor exposure, and admits a formula only while SAFFRON (Ramdas et al., 2018) allows it.

In simulation with drifting factor returns, OFA holds the false discovery proportion at 0.087 against a 0.10 target, where the same online rule applied to raw validation IC reaches 0.911. On real candidate streams it fails. It admitted nothing in 2 of 5 runs at one day and 3 of 25 at twenty days, and its neutralized test IC is -0.0066 against selecting the top formulas by the same orthogonalized statistic without the error budget.

The mechanism is visible in its admission counts. An online FDR rule spends an error budget that is replenished by discoveries, so a stream that is mostly null starves it. One validation year does not provide the evidence the budget demands, and the layer abstains.

## 10 What would make these benchmarks identifiable

Six changes follow from the measurements above, and none of them needs new data. Report the time-sampling interval on every mean IC, not only dispersion across seeds, since the two answer different questions and only the first speaks to the test period. Test rankings with a procedure built for many models at once, such as the model confidence set, rather than

comparing point estimates. Train and score at the same horizon, and say which one the table used. Report how many pairs the window can separate, corrected for the number of comparisons made, so that a table's silent precision is visible; here that is 1 of 11 among neighbours. The last two we only found by being wrong first. Report residual rather than raw IC, which halves the detectable gap on this data. And report turnover and a risk-adjusted figure beside the mean, because the ordering is not invariant to that choice (Section 8).

## 11 What we would test next

Each of the four experiments below includes a falsifiable prediction.

Where the budget turns. We ran two budgets and found that AlphaGen's published one costs -0.0080 in out-of-sample IC while the other two searchers gain. Two points cannot locate a turning point. A sweep across six budgets at five seeds, recording training and test IC at each, would show whether every searcher has an interior optimum and whether the published budget sits past it. If training IC rises monotonically while test IC peaks early for all three, then the budget a paper reports is a hyperparameter tuned against the wrong quantity.

A long-only book. Section 8 holds a dollar-neutral book, which the borrow constraints on A-shares make optimistic. Rebuilding it long-only against the index would change turnover and could reorder the net Sharpe column, and the style model, which has the lowest turnover in the comparison, would gain most from that change. We expect the handcrafted library to keep its lead, and the expectation is cheap to falsify.

One market or many. One market cannot separate a claim about alpha mining from a claim about A-share factor premia. Running the same protocol on CSI500 and on a US universe would separate them. If the style model loses its twenty-day lead outside China while the handcrafted library keeps its neutralized lead, the first finding narrows to one market and the second survives.

The mechanism behind OFA's failure. We argue the layer abstains because a single validation year cannot fund its error budget. That is a mechanism claim, and a rolling multi-year validation window tests it, because admissions should rise under one. If they do not, our explanation is wrong and the layer is broken for some other reason.

## 12 What we cannot settle

Our own ordering is subject to the limit in Section 4. The defensible statement is the exclusion: genetic programming, AlphaSAGE, reversal and random formulas fall outside the confidence set at twenty days, and the style model does not. That the style model is first is not established. After neutralization the set is sharper and the conclusion changes hands, though those exclusions hold at ten percent and not at five: what survives both levels is that the handcrafted library leads, not that the mined pools are ruled out. Under a risk-adjusted or cost-aware yardstick the style model is not first at twenty days either.

The style model is weak in absolute terms at one day, where a handcrafted library beats it and every miner outright, so nothing on that window supports style factors as forecasts. Our size factor is a traded-amount proxy and the data contain no industry codes, so this is not a full risk model. We show what released benchmark code does, not what produced any published table. Learned miners are run at a matched budget of 5,000 candidate evaluations. The cap binds less than its ratio to a step budget suggests: uncapped runs at this horizon reached only 4246 to 11189 distinct formulas, because the searchers re-propose expressions they have already seen. AlphaQCM and AlphaForge exceeded the available compute budget and are absent. Verification found that AlphaGen's own float32 normalization moves the daily IC of its published pool by as much as 0.0042 on a single day against float64. That

is a pointwise maximum, not a bound on the precision of a mean, but it is a reminder that formulas with extreme intermediate magnitudes are not numerically portable between implementations.

## 13 Related work

Alpha mining and its benchmarks. Formulaic alphas have been searched with genetic programming, reinforcement learning over an expression grammar (Yu et al., 2023), generative flow networks (Chen et al., 2026), generator–predictor combinations (Shi et al., 2025a) and LLM agents (Tang et al., 2025). AlphaEval (Ding et al., 2026) and AlphaBench (Luo et al., 2026) widen the set of metrics used to compare them. Neither attaches time-sampling uncertainty to a mean IC, tests whether two methods differ by more than noise, or reports what its window can resolve; AlphaEval's own significance analysis relates one metric to another rather than one method to another. We do not propose another searcher. We run three of these from released code under one protocol and ask what their shared evaluation can support.

Comparing many models at once. Testing a whole leaderboard is the data-snooping problem (White, 2000), with familywise control (Romano & Wolf, 2005) and the model confidence set (Hansen et al., 2011) as standard answers, and the stationary bootstrap (Politis & Romano, 1994) as the resampling scheme under dependence. Finance has its own literature on the same issue: inflated $t$-statistics across published factors (Harvey et al., 2016) and publication bias in the anomaly record (Chen & Zimmermann, 2022). These tools are standard in econometrics and absent from the alpha-mining leaderboard; applying them is most of what we do. Online false discovery control (Ramdas et al., 2018; Javanmard & Montanari, 2018; Ramdas et al., 2017) motivates the selection layer of Section 9, which fails on real candidate streams.

## Reproducibility statement

Pre-registrations with every amendment, all mined pools, complete candidate streams, the evaluation and analysis code, and a verification suite of 19 checks are in the supplementary material. An independently written NumPy evaluator reproduces the published pool's mean IC to within 0.00015. The suite includes truncation tests for every model that carries a conclusion here: rebuilding the style factors, the factor-neutralization loadings and all 158 handcrafted features from a panel cut off at the evaluation day reproduces that day exactly, so none of them can be reading forward. Recomputation through our own evaluator agrees to $5 \times 10^{-7}$, which is the storage precision of the released series rather than a measure of independent agreement. The pre-registrations were written before the runs they govern and every amendment is dated and retained, but they carry no external timestamp, so they should be read as a disclosed record rather than as a registered protocol.

## AI use statement

AI-based language tools were employed exclusively to enhance linguistic quality, including clarity, grammar, and overall readability. These tools were not used to generate scientific ideas, analyses, results, or interpretations. The authors assume full responsibility for the accuracy, originality, and integrity of the work presented in this manuscript.

## References

Andrew Y. Chen and Tom Zimmermann. Publication bias in asset pricing research. arXiv preprint arXiv:2209.13623, 2022.

Binqi Chen, Hongjun Ding, Ning Shen, Jinsheng Huang, Taian Guo, Luchen Liu, and Ming Zhang. AlphaSAGE: Structure-aware alpha mining via GFlowNets for robust exploration. In International Conference on Learning Representations, 2026.

Hongjun Ding, Binqi Chen, Jinsheng Huang, Taian Guo, Zhengyang Mao, Guoyi Shao, Lutong Zou, Luchen Liu, and Ming Zhang. AlphaEval: A comprehensive and efficient evaluation framework for formula alpha mining. In Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2026.

Peter R. Hansen, Asger Lunde, and James M. Nason. The model confidence set. Econometrica, 79(2):453–497, 2011. doi: 10.3982/ECTA5771.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. ... and the cross-section of expected returns. The Review of Financial Studies, 29(1):5–68, 2016. doi: 10.1093/rfs/hhv059.

Adel Javanmard and Andrea Montanari. Online rules for control of false discovery rate and false discovery exceedance. The Annals of Statistics, 46(2):526–554, 2018. doi: 10.1214/17-AOS1559.

Haochen Luo, Ho Tin Ko, Jiandong Chen, David Sun, Yuan Zhang, and Chen Liu. AlphaBench: Benchmarking large language models in formulaic alpha factor mining. In International Conference on Learning Representations, 2026.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. Journal of the American Statistical Association, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.

Aaditya Ramdas, Fanny Yang, Martin J. Wainwright, and Michael I. Jordan. Online control of the false discovery rate with decaying memory. In Advances in Neural Information Processing Systems, 2017.

Aaditya Ramdas, Tijana Zrnic, Martin J. Wainwright, and Michael I. Jordan. SAFFRON: An adaptive algorithm for online control of the false discovery rate. In International Conference on Machine Learning, 2018.

Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. Econometrica, 73(4):1237–1282, 2005. doi: 10.1111/j.1468-0262.2005.00615.x.

Hao Shi, Weili Song, Xinting Zhang, Jiahe Shi, Cuicui Luo, Xiang Ao, Hamid Arian, and Luis Seco. AlphaForge: A framework to mine and dynamically combine formulaic alpha factors. In Proceedings of the AAAI Conference on Artificial Intelligence, 2025a.

Yu Shi, Zongliang Fu, Shuo Chen, Bohan Zhao, Wei Xu, Changshui Zhang, and Jian Li. Kronos: A foundation model for the language of financial markets. arXiv preprint arXiv:2508.02739, 2025b.

Ziyi Tang, Zechuan Chen, Jiarui Yang, Jiayao Mai, Yongsen Zheng, Keze Wang, Jinrui Chen, and Liang Lin. AlphaAgent: LLM-driven alpha mining with regularized exploration to counteract alpha decay. arXiv preprint arXiv:2502.16789, 2025.

Halbert White. A reality check for data snooping. Econometrica, 68(5):1097–1126, 2000. doi: 10.1111/1468-0262.00152.

Xiao Yang, Weiqing Liu, Dong Zhou, Jiang Bian, and Tie-Yan Liu. Qlib: An AI-oriented quantitative investment platform. arXiv preprint arXiv:2009.11189, 2020.

Shuo Yu, Hongyan Xue, Xiang Ao, Feiyang Pan, Jia He, Dandan Tu, and Qing He. Generating synergistic formulaic alpha collections via reinforcement learning. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023.

## A Terms used here

The benchmark's vocabulary comes from finance. This appendix is for readers who do not work in it; nothing in the paper depends on knowing more than the following.

Table 6: Two sources of uncertainty in the same number. The seed column asks whether this training run was lucky; the time column asks whether this test period was. The papers we compare against report at most the first, and none of them reports the second.

| | $h = 1$ | | | $h = 20$ | | |
|---|---|---|---|---|---|---|
| Method | Seeds | Across seeds | Over time | Seeds | Across seeds | Over time |
| AlphaGen | 2 | 0.0051 | 0.0045 | 3 | 0.0087 | 0.0161 |
| Genetic programming | 3 | 0.0221 | 0.0050 | 5 | 0.0191 | 0.0175 |
| AlphaSAGE | – | – | – | 2 | 0.0031 | 0.0145 |
| Random formulas | 10 | 0.0026 | 0.0036 | 10 | 0.0108 | 0.0118 |

- Cross-section. The set of stocks on a single day. Every model scores all of them, and only the ordering within that day is used, never a comparison across days.
- Information coefficient (IC). The Pearson correlation, on one day, between a model's scores and the returns that follow. One number per day, so a four-year test window is a series of 969 of them. Mean IC is the leaderboard's metric. RankIC is the same with Spearman correlation.
- Horizon $h$. How far ahead the label looks, in trading days. A signal built for $h = 20$ is predicting three-to-four weeks out, so consecutive labels overlap, which is why the bootstrap here uses blocks.
- Information ratio. Mean IC divided by the standard deviation of the daily IC series, so it is monotone in the $t$-statistic. A model can lead on mean IC and trail on this if its daily IC is volatile.
- Style factors. Long-known return predictors. The six used here are size, market beta, volatility, twelve-month momentum, one-month reversal and abnormal turnover. They play the role strong known features play elsewhere: a method that cannot beat them has not shown it found anything new.
- Neutralization. Residualizing a signal on those factors day by day, which asks what the signal predicts beyond what the known features already explain.
- Dollar-neutral book. A portfolio holding equal value long and short, so its return depends on the ordering a model produces rather than on the market's direction.
- Turnover. The fraction of the portfolio replaced per day. It sets the trading cost, which is why a metric that ignores it can prefer a model that loses money.
- Formulaic alpha, and a pool. A formula over price and volume primitives; a pool is a weighted combination of several, which is what these methods output.

# B Supporting tables

Table 7: After every cutoff: 2024H2–2026, the window Kronos can be scored on, one-day horizon.

| Model | Mean IC | 95% CI | MCS $p$ | In set |
|---|---|---|---|---|
| Alpha158 + GBDT | +0.0550 | [+0.0418, +0.0681] | 1.000 | ✓ |
| Alpha158 + ridge | +0.0470 | [+0.0333, +0.0596] | 0.148 | ✓ |
| AlphaGen (released budget) | +0.0328 | [+0.0188, +0.0461] | 0.015 | |
| GP (matched budget) | +0.0245 | [+0.0131, +0.0363] | 0.015 | |
| AlphaAgent released formulas | +0.0206 | [+0.0050, +0.0356] | 0.015 | |
| Kronos (zero-shot) | +0.0117 | [+0.0010, +0.0226] | 0.015 | |
| Kronos (normalized signal) | +0.0074 | [-0.0022, +0.0174] | 0.015 | |
| Random formulas | +0.0038 | [-0.0037, +0.0108] | 0.000 | |
| 5-day reversal | +0.0026 | [-0.0154, +0.0207] | 0.015 | |
| Six-factor style model | +0.0024 | [-0.0120, +0.0162] | 0.015 | |
| AlphaGen published pool | -0.0017 | [-0.0172, +0.0142] | 0.015 | |

Table 8: OFA against the selection rules it must beat, on identical candidates. Neutralized mean IC on 2021–2024, averaged over runs.

| Selection rule | $h = 1$ | $h = 20$ | Members admitted |
|---|---|---|---|
| Miner's own pool | +0.0205 | +0.0115 | |
| Top-$k$ by train IC | +0.0210 | +0.0083 | |
| Top-$k$ by validation IC | +0.0149 | +0.0055 | |
| Top-$k$ by orthogonalized $t$ | +0.0212 | +0.0076 | |
| Online FDR on raw validation IC | +0.0051 | +0.0027 | |
| OFA | +0.0145 | +0.0016 | 0–7 at $h = 1$, 0–5 at $h = 20$ |