

Backtests Are the Most Optimistic Estimates a Strategy Will Get

Of 212 published stock-return predictors, 208 enter the main comparison after exclusions. Historical returns are reconstructed for backtest, forward and post-publication periods. These are not verified live trading results.

SUBJECT

Backtests and forward validation of systematic equity strategies

EVIDENCE

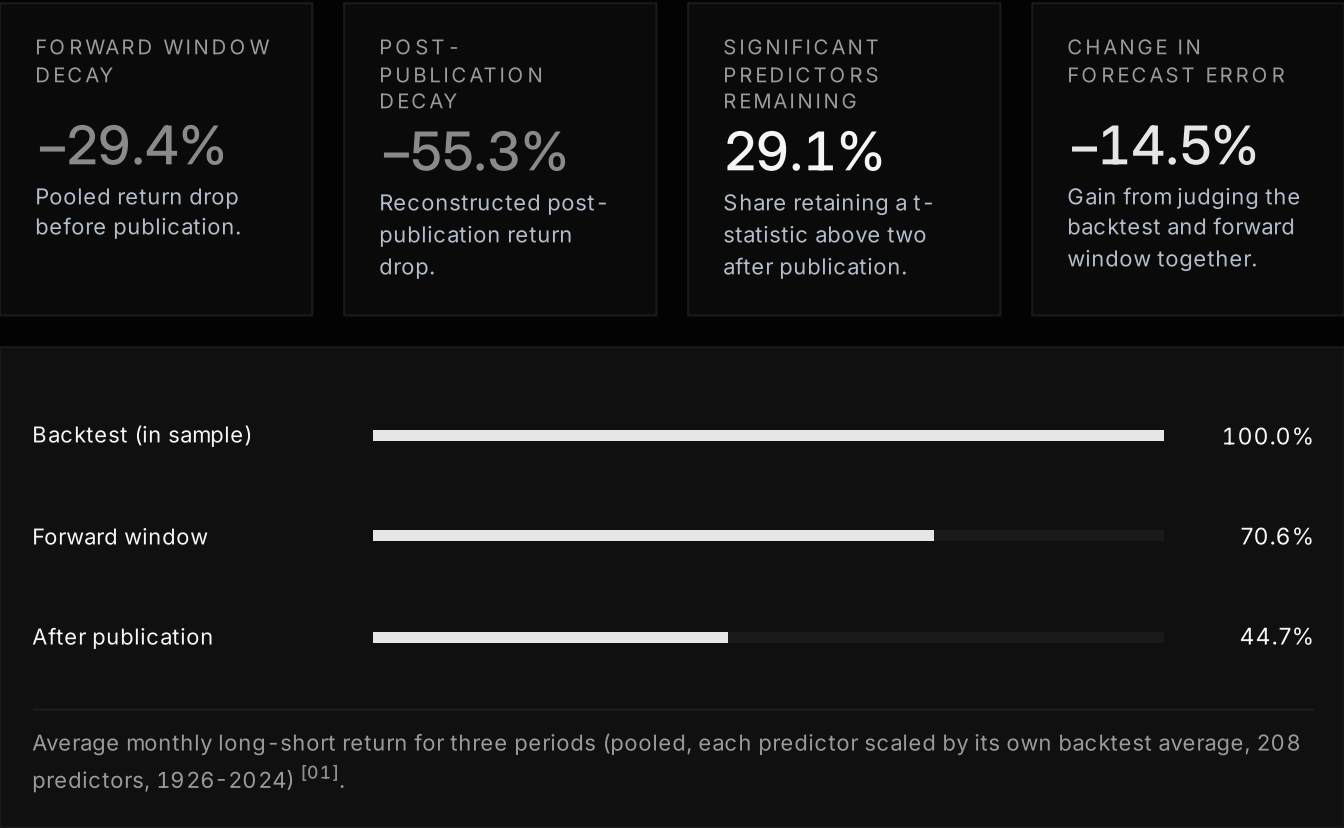
208 of 212 predictors, 1926-2024; factors to August 2026

DOCUMENT TYPE

Private research

Published predictors lost 29% of their backtest return in the forward window and 55% after publication

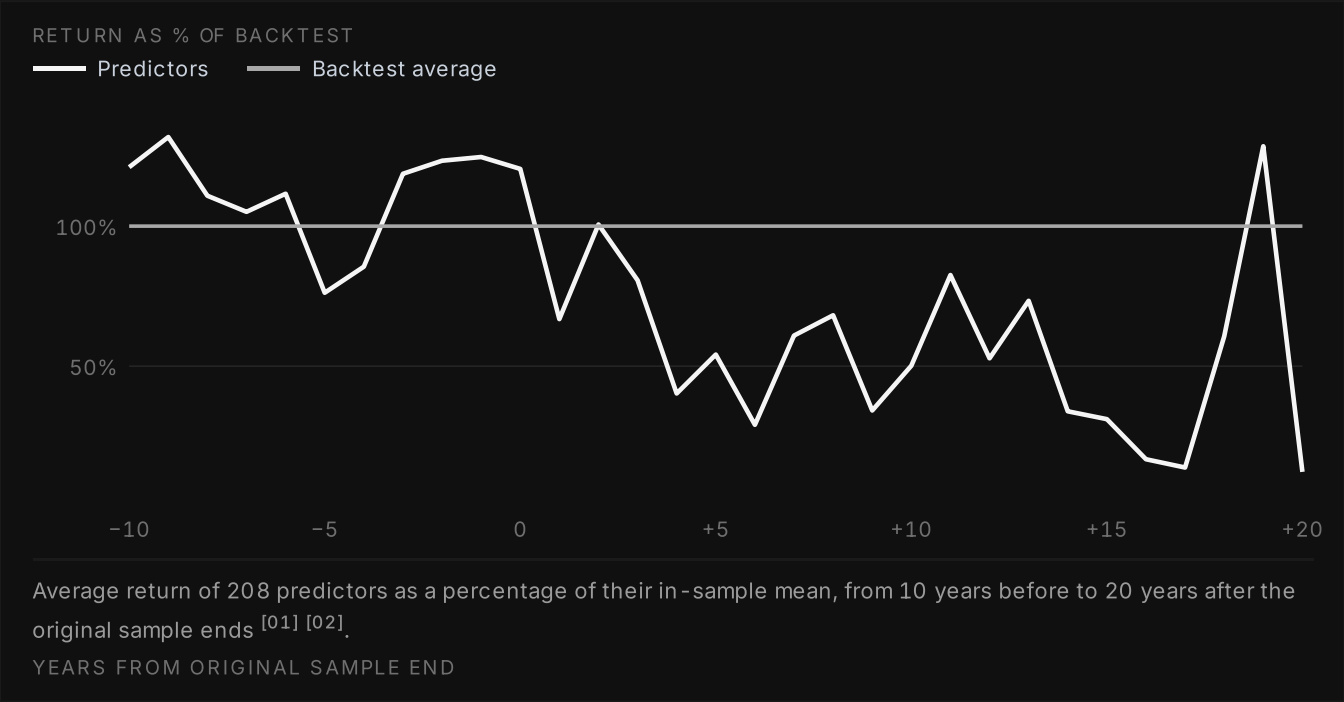
The backtest is the high- water mark of a strategy's evidence ^[04]. The decline starts before publication in these reconstructed returns. A forward window (sample end to publication, median 48 months) provides a useful check, although authors may have seen those years. Adding it to the backtest cut forecast error by about a seventh on average, a gain coming entirely from predictors published after 2006 (down 32.6%, with no improvement for older predictors) ^[01].



OBSERVATION	More than half of the decay happens before investors can trade on the paper, making it the closest available measure of backtest optimism and an upper bound ^[04] . Pooled returns dropped 29.4% in the forward window and 55.3% after publication. Simple averages show a steeper decline: 0.70% a month in sample fell to 0.42% in the forward window and 0.30% after publication, drops of 40.5% and 56.9% respectively ^[01] .
INTERPRETATION	The performance drop starts before publication. Strong simulated results offer no safety margin. The strongest backtests lost the most out of sample and ended up matching the middle of the pack after publication.
IMPLICATION	Do not discard the backtest mean to rely solely on the forward period, as it predicts reconstructed post - publication returns no better than the original sample (R - squared 0.17 versus 0.19). The forecasting advantage comes strictly from judging both periods together (0.26).
LIMITATIONS	These returns are rebuilt on today's data, and the forward window is short (median 48 months against 336 in sample). Authors may have seen these forward years while writing, making this an upper bound on true overfitting. Finally, the reliable post - publication decline is a US phenomenon that may not travel globally ^[10] .

The edge peaks in the last five backtest years and loses a third in the first five years after

The peak just before the sample ends is what selection looks like. Researchers stop the sample where the result is strongest. Sometimes the effect was simply strongest then ^[01] ^[02].



LAST FIVE SAMPLE YEARS	FIRST FIVE YEARS AFTER	YEARS SIX TO TEN AFTER
114.6%	68.4%	48.5%
Return relative to the full backtest average.	Average edge kept immediately after the sample.	Continued deterioration as the strategy ages.

INTERPRETATION	A strategy peaks right before its history ends. The subsequent drop exposes the cost of fitting rules to a known past. Live markets offer no such luxury.
LIMITATIONS	Four of the 212 predictors were excluded. Calendar years define the stage splits. Jacobs and Mueller find the decline reliably only in the US ^[10] .

The strongest backtests lost the most and ended level with the middle

Investors often treat a massive in- sample t- statistic as proof of a durable edge. The data tell a different story. Across 208 predictors, the in- sample t- statistic explains about 3% (R- squared 0.03) of the variation in later returns ^[01].

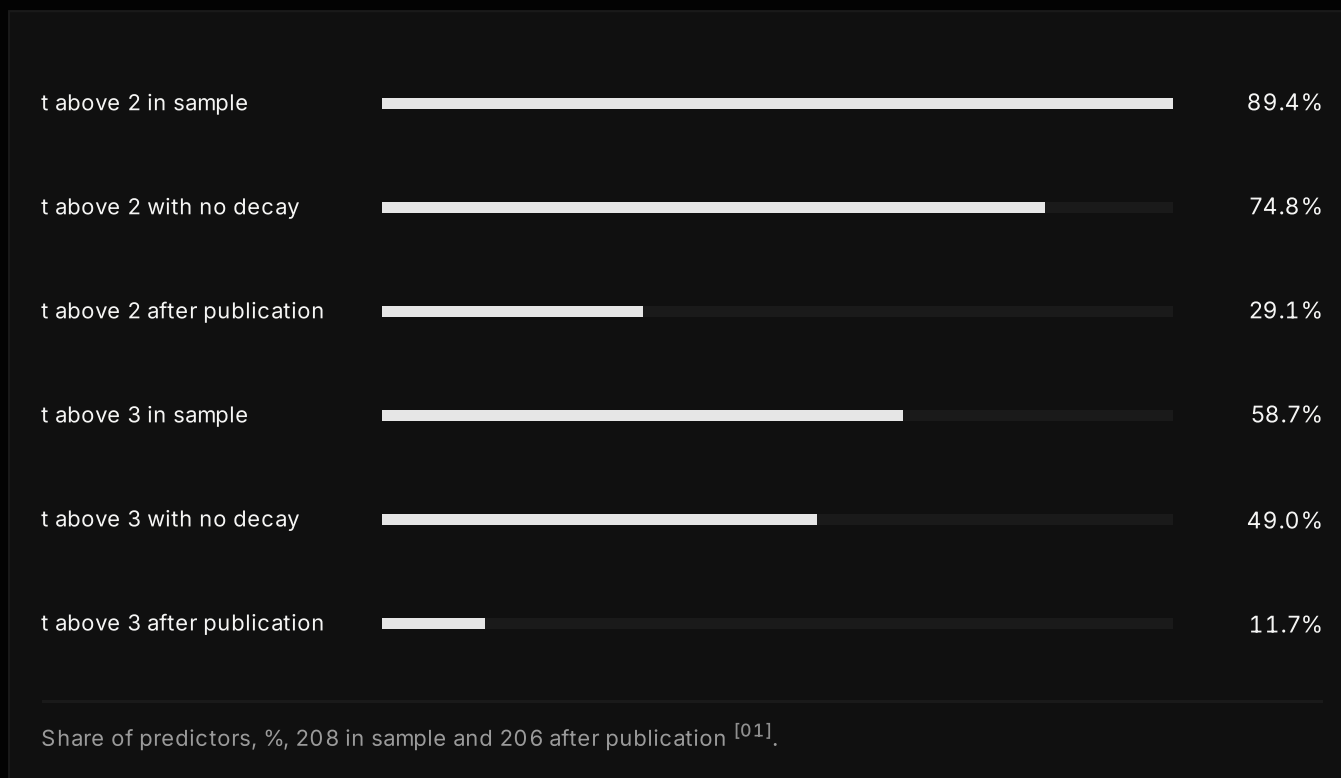


GROUP	IN - SAMPLE T RANGE	BACKTEST RETURN A MONTH AFTER PUBLICATION	A MONTH SHARE KEPT	
Weakest fifth	1.0–2.3	0.42%	–0.05%	–11.9%
Second fifth	2.3–2.9	0.50%	0.34%	67.4%
Middle fifth	2.9–3.7	0.73%	0.43%	59.6%
Fourth fifth	3.7–5.3	0.79%	0.39%	49.3%
Strongest fifth	5.4–14.2	1.07%	0.40%	37.3%

INTERPRETATION	Beyond moderate significance, a higher t-statistic buys no extra post-publication return: the strongest fifth ended level with the middle (0.40 % against 0.43% a month), while only the weakest fifth lost money.
IMPLICATION	Stop paying a premium for paper significance. You should rank competing models by their out-of-sample decay instead.
LIMITATIONS	The weakest group turned negative partly because a t-statistic below 2.0 was never a strong claim. These are just averages. Those pooled numbers hide wide dispersion among the individual predictors inside each group.

Nine in ten predictors cleared $t = 2$ in the backtest, fewer than three in ten after publication

Most backtests clear traditional significance hurdles easily. Shorter track records after publication explain only a small fraction of the subsequent drop ^[01]. Harvey, Liu and Zhu have already argued that any new factor needs to clear a t -statistic of 3.0 ^[05].

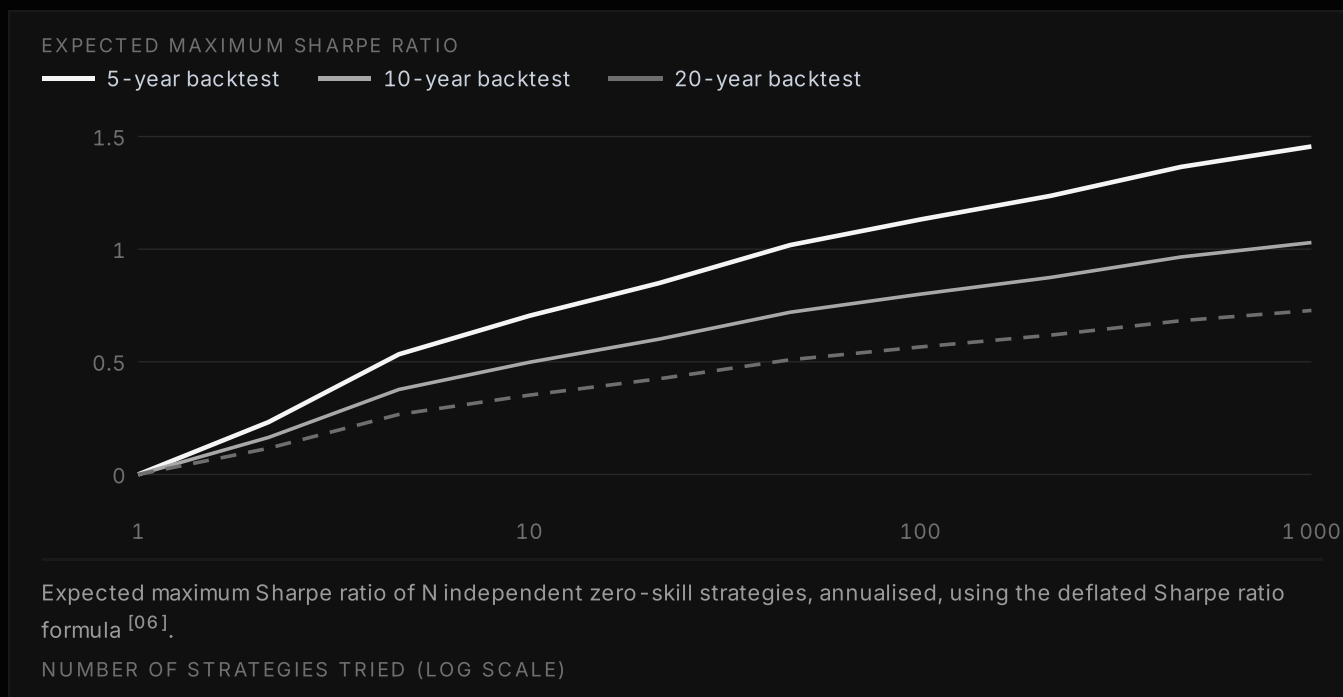


T - STATISTIC BAND	IN SAMPLE	AFTER PUBLICATION
below 0	0.0%	15.0%
0 to 1	0.5%	29.6%
1 to 2	10.1%	26.2%
2 to 3	30.8%	17.5%
above 3	58.7%	11.7%

OBSERVATION	Backtested strategies project high confidence, with 89.4% clearing a t -statistic of 2 and 58.7% clearing $t = 3$. Reality sets in after publication when those figures collapse to 29.1% and 11.7% ^[01] . Fully 15.0% of predictors recorded a negative average return after publication ^[01] .
INTERPRETATION	The original results still contain information: across 206 predictors the in-sample mean explains an R -squared of 0.19 of post-publication returns ^[01] . Yet the collapse highlights that historical t -statistics dramatically overstate the actual signal. Initial certainty rarely survives contact with unseen data.
LIMITATIONS	The no-decay baseline assumes the return-to-risk ratio stays completely constant. Post-publication t -statistics are lower partly because arbitrage took the return, not only because the original authors overfit the data ^[04] .

The best of 100 useless ten-year backtests shows a Sharpe ratio of 0.8 on average

Every variant a quantitative team tries acts as a lottery ticket. The winning ticket's reported Sharpe ratio demands harsh judgment against the sheer number of tickets bought. Selection makes the best of many zero-skill results look like skill; it does not isolate forecasting ability ^[06].



TEN-YEAR BACKTESTS

0.80

Expected outcome from one hundred independent zero-skill trials.

FIVE-YEAR, 1,000 TRIALS

1.46

Average top result when searching across one thousand variations.

TWENTY-YEAR BACKTESTS

0.57

Apparent performance generated entirely by luck across one hundred attempts.

INTERPRETATION

Finding a good backtest tells you nothing if you ignore how many bad ones hit the cutting room floor. A high Sharpe ratio often measures the computing power spent searching for it, rather than any structural edge in the market. The shorter the sample period, the easier the noise masquerades as alpha.

IMPLICATION

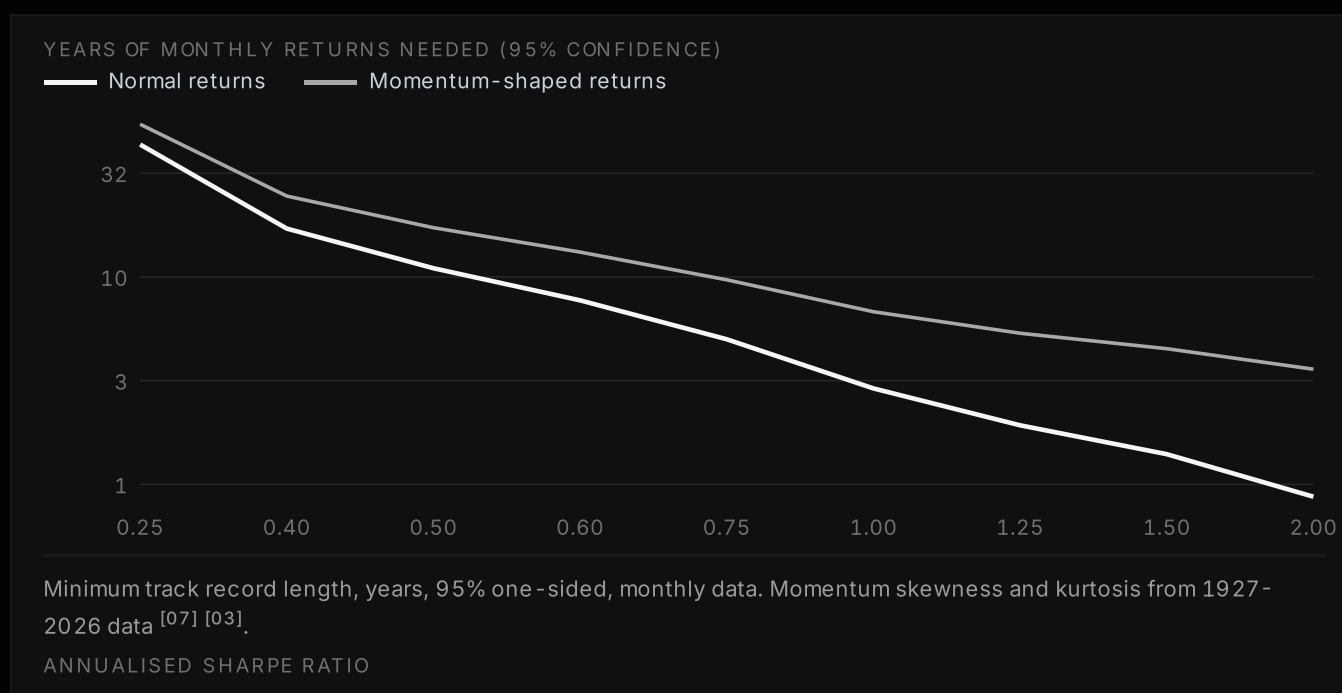
Investment committees should discount presented performance by the estimated number of discarded variations. Demand a substantially longer track record when the underlying strategy relies on aggressive computational searches.

LIMITATIONS

Tested variations are rarely completely independent. The effective number of trials is smaller than the raw count, and estimating that true effective number remains the hardest part of the evaluation.

A Sharpe ratio of 0.5 needs 11 years of evidence, 17 if it crashes like momentum

A six-month paper-trading period says almost nothing about a modest strategy ^[07]. Sample size dictates confidence. The shape of returns matters too: momentum-like skew stretches the requirement from 11 to 17 years ^[07] ^[03]. Any strategy carrying crash risk requires a massive observation window just to rule out luck.



NORMAL RETURNS AT 0.5 SHARPE

11.0y

Years required to reject zero skill at 95% confidence.

MOMENTUM SHAPE AT 0.5 SHARPE

17.3 y

Years required when returns carry fat left tails.

MOMENTUM'S LONG-RUN SHARPE (0.45)

20.2 y

Years required to reject zero skill at momentum's own long-run Sharpe ratio.

INTERPRETATION

A standard two-year forward test offers less security than the industry assumes. With normal returns, a two-year record at 95% confidence is enough only for a true Sharpe ratio of about 1.25 or higher. Reaching that 1.25 mark takes 1.9 years. A 1.0 Sharpe needs 2.9 years.

IMPLICATION

Price the shape of returns into any evaluation of a live track record. Short histories offer zero protection against strategies with left-tail risk.

LIMITATIONS

This test only rejects zero skill, leaving the actual backtest level unconfirmed. High-frequency strategies run many independent bets and can shorten the time requirement. They never bypass it entirely.

Dropping micro caps and weighting by size removes a quarter to a third of the backtest before any cost

Statistical overfitting only explains part of the performance gap. Academic backtests routinely assume equal weights across universes thick with tiny, illiquid equities that no institutional fund could efficiently trade ^[01]. The problem worsens with complexity, as bank - promoted alternative beta strategies suffer massive degradation when moved from paper to production ^[08].



EXCLUDING MICRO CAPS

-26.2%

Return impact of dropping stocks below the NYSE 20th size percentile.

VALUE WEIGHTING

-34.6%

Return impact of weighting by market capitalization rather than equally.

LIVE SHARPE DETERIORATION

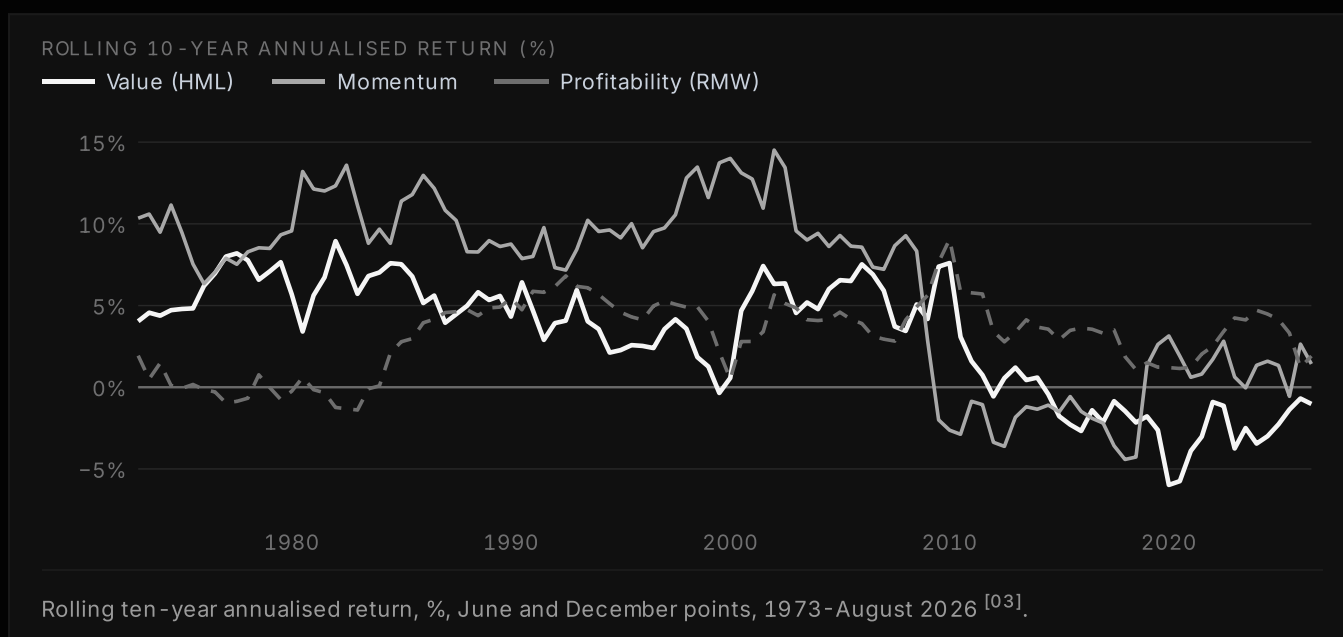
-73%

Median drop for bank-promoted alternative beta strategies compared to their backtests ^[08].

OBSERVATION	Data from 888 trading algorithms reveals the backtest Sharpe ratio predicts out of sample performance with an R-squared below 0.025 ^[09] . A simulated track record explains practically nothing about future success.
INTERPRETATION	Paper portfolios ignore market friction and capital capacity. Adjusting the rules to reflect realistic institutional mandates destroys a massive portion of the theoretical returns. You are pricing in execution constraints that the original authors simply ignored.
LIMITATIONS	These adjustments capture only structural portfolio rules, as transaction costs are not deducted anywhere in the primary dataset. Trading costs would further reduce these reconstructed returns. The alternative beta data also suffers from self selection, reflecting only the strategies banks chose to actively promote.

Every classic factor earned less after publication. Value lost money for 14 years.

Even the best-known factors kept only a fraction of their backtest returns after decades of scrutiny ^[03]. Value earned -5.7% a year from 2007 to 2020, then surged to +7.9% a year from 2021 through August 2026 ^[03]. A forward test that covers one regime can mislead you in either direction.



FACTOR	PUBLISHED	ORIGINAL SAMPLE	IN SAMPLE, % A YEAR	AFTER PUBLICATION, % A YEAR	WORST DECLINE AFTER
Value (HML)	1980	1962-1976	7.0%	2.7%	-57.8%
Size (SMB)	1981	1926-1975	1.7%	-0.4%	-54.9%
Momentum	1993	1964-1989	9.3%	2.9%	-57.8%
Profitability (RMW)	2006	1977-2003	4.2%	2.8%	-25.9%
Investment (CMA)	2008	1968-2003	5.4%	0.0%	-27.6%

INTERPRETATION A backtest provides an optimistic baseline that needs a haircut. You cannot dismiss it entirely. Across 206 predictors the in-sample mean still explains 0.19 of the variation in post-publication returns ^[01].

LIMITATIONS French's factor definitions differ from the original papers. Returns are before costs.

Adding a four - year forward window cut forecast error by 14.5%, all of it among papers published after 2006

The forward window alone offers no more predictive power than the original backtest. Averaging the two produces a clear improvement ^[01]. R-squared rises from 0.19 to 0.26, and a flat or negative forward window is a strong warning (19.3% of the backtest kept, against 51.1%).



CHANGE IN FORECAST ERROR

-14.5%

Drop in mean absolute forecast error when using the combined average.

NEGATIVE OR FLAT FORWARD WINDOW

19.3%

Share of the backtest return kept after publication by the 55 flagged predictors.

POSITIVE FORWARD WINDOW

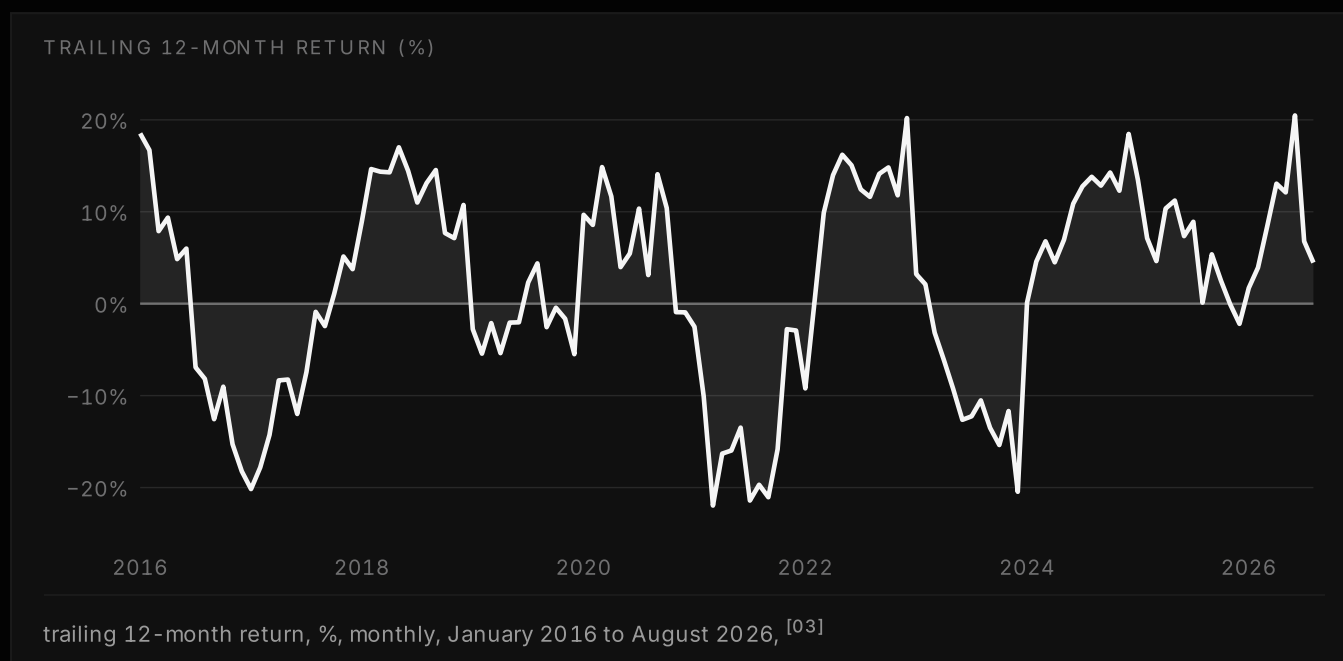
51.1%

Share of the backtest return kept after publication by predictors that earned a profit.

OBSERVATION	We measure forecast error as the mean absolute error of the post-publication monthly mean in percentage points a month. Using a fixed 50/50 average confirms the gain beats mere shrinkage: the best hindsight shrinkage leaves the error essentially unchanged at 0.49, while the forward window drops it to 0.42.
INTERPRETATION	Backtests simply measure what fit historical noise. The forward window tests whether the exact same rules survived their first contact with unseen data, stripping away the benefit of hindsight.
IMPLICATION	Freeze your rules to evaluate the backtest and forward window together. Treat a flat or negative forward window as a red flag that needs an explanation, not a reason to discard the strategy outright.
LIMITATIONS	The forward window was never a clean blind test, and an R-squared of 0.26 leaves three quarters of the variation unexplained. The error fell 32.6% for 81 predictors published after 2006 but 0.0% for 125 older ones, while 17 of the 55 flagged predictors still kept half their backtest.

Momentum gave back 17.2% in July and August 2026, a two-month loss seen in 1.3% of windows since 1927

A regime turn is when a backtest calibrated on the previous regime is found out. The gap between the trailing year in June and in August shows how fast. French's momentum factor recorded exactly this shift ^[03].



MOMENTUM DECLINE

-17.2%

Combined loss for the factor in July and August 2026.

HISTORICAL RANK

15 / 1,195

Worst loss rank out of 1,195 two-month windows since 1927.

PROFITABILITY RETURN

+11.1%

Profitability (RMW) in July 2026, its third-best month out of 758 since 1963.

INTERPRETATION

A factor can give back most of a year's gain in two months. A forward window shorter than one regime would never show this. Evaluators need long out-of-sample data to see how strategies behave when the environment changes.

IMPLICATION

A forward test should span more than one market environment. On the regulatory side, the SEC charged nine advisers \$850,000 in September 2023 over hypothetical performance ^[11], and SEC staff keeps updating the Marketing Rule FAQ (latest 15 January 2026) ^[12].

Forward data reduces forecast error and exposes backtest decay

We examined published stock predictors to measure performance on data outside the original sample. Returns in the post-publication period earned 55.3% less than the backtest. The decay begins immediately.

- 01 When evaluated on unchanged rules, the forward window earned 29.4% less than the original backtest.
- 02 The strongest fifth kept 37.3%.
- 03 The best of 100 zero-skill ten-year backtests shows 0.80 on average ^[06].
- 04 A Sharpe ratio of 0.5 needs 11.0 years of monthly data to be told apart from zero at 95% confidence ^[07]. Short records mislead.
- 05 Judged together with a forward window, forecast error fell 14.5% on average ^[01]. The benefit is not uniform across cohorts. All of it occurred among the 81 predictors published after 2006.

FORWARD VALIDATION SUMMARY

MEASURE	WHAT IT SHOWS	VALUE
Forward window relative return	Return shortfall compared to the original sample	-29.4%
Post-publication relative return	Return shortfall after the research appears in print	-55.3%
Strongest fifth retention	Share of performance maintained by the best backtests	37.3%
Random backtest Sharpe	Expected ratio from multiple attempts with zero skill	0.80
Years to confirm Sharpe	Minimum track record needed at standard confidence levels	11.0 y
Change in forecast error	Improvement in predicting post-publication returns using forward data	-14.5%

Before allocating capital to a systematic strategy, demand the number of variants tried and a frozen rule. Evaluate these alongside years of forward data judged together with the backtest.

Selected sources

-
- [01] **Open Source Asset Pricing data, October 2025 release**
<https://www.openassetpricing.com/data/>
 We pulled the monthly long-short returns of 212 predictors. The dataset provides original sample years, publication dates and liquidity-screened versions.
-
- [02] **Chen and Zimmermann, Open Source Cross-Sectional Asset Pricing**
<https://www.federalreserve.gov/econres/feds/files/2021-037pap.pdf>
 This paper outlines the methodology behind the replication.
-
- [03] **Kenneth French Data Library**
https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
 This repository supplied the returns for value, size, momentum, profitability and investment factors through August 2026.
-
- [04] **McLean and Pontiff, Does Academic Research Destroy Stock Return Predictability? (Journal of Finance 2016)**
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2156623
 They documented the original 26% out-of-sample decay and 58% post-publication decay.
-
- [05] **Harvey, Liu and Zhu, ...and the Cross-Section of Expected Returns (Review of Financial Studies 2016)**
<https://www.nber.org/papers/w20592>
 The authors argue a newly discovered factor requires a t-statistic above 3.0.
-
- [06] **Bailey and Lopez de Prado, The Deflated Sharpe Ratio (Journal of Portfolio Management 2014)**
<https://www.davidhbailey.com/dhbpapers/deflated-sharpe.pdf>
 Their math yields the expected maximum Sharpe ratio from N trials with zero true skill.
-
- [07] **Bailey and Lopez de Prado, The Sharpe Ratio Efficient Frontier (Journal of Risk 2012)**
<https://www.davidhbailey.com/dhbpapers/sharpe-frontier.pdf>
 We used their formula for calculating the minimum track record length.
-
- [08] **Suhonen, Lennkh and Perez, Quantifying Backtest Overfitting in Alternative Beta Strategies (Journal of Portfolio Management 2017)**
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2757113
 They found a 73% median decay in live Sharpe ratios across 215 bank-promoted strategies.
-
- [09] **Wiecki, Campbell, Lent and Stauth, All That Glitters Is Not Gold (Journal of Investing 2016)**
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2745220
 Testing 888 trading algorithms revealed a backtest Sharpe ratio predicts out-of-sample results with an R-squared below 0.025.
-
- [10] **Jacobs and Mueller, Anomalies across the globe: Once public, no longer existent? (Journal of Financial Economics 2020)**
<https://www.sciencedirect.com/science/article/pii/S0304405X19301618>
 Looking at 241 anomalies in 39 markets showed the US is the only market with a reliable post-publication decline.
-
- [11] **SEC press release 2023-173 (September 2023)**
<https://www.sec.gov/newsroom/press-releases/2023-173>
 The regulator charged nine advisers a combined \$850,000 for advertising hypothetical performance without required policies.
-
- [12] **SEC Division of Investment Management, Marketing Compliance FAQ**
<https://www.sec.gov/rules-regulations/staff-guidance/division-investment-management-frequently-asked-questions/marketing-compliance-frequently-asked-questions>
 Staff published their latest interpretations on 15 January 2026.
-

This document is research, not investment advice. Returns appear gross of costs and taxes. Past decay does not place a limit on future decay.

Independent research, introductions clearly defined.

BlackRidge is an independent research bureau introducing private investors to quantitative traders through multiple strategy providers. We publish quantitative research. Investors access strategies through PAMM accounts at the broker. We are not a fund or broker and never hold client money.

01

Your funds stay at the broker. The account is opened in your own name. BlackRidge never receives, holds, or has withdrawal rights over client capital.

02

No fee on your profits. A one-time access payment of 6.7% of the agreed trading level, paid to the selected strategy provider. No management fee, no performance fee, no profit share, in any year.

03

You fund the risk, not the exposure. Allocations are notionally funded: you agree a trading level and fund the margin and drawdown allowance behind it. Trading losses can exceed the deposit without applicable negative balance protection; the separate 6.7% access payment is nonrefundable.

CONTINUE READING

blckridge.com/research; published research. [Notional funding](#) and [PAMM accounts](#) explained in full, and [answers](#) to the questions this raises.

We use AI models to gather and aggregate source material and to help prepare each report. Read it with its cited sources, sample, methods and limitations, and send corrections through [research methodology and corrections](#).

The client pays the selected strategy provider a one-time access payment of 6.7% of the agreed trading level. BlackRidge receives an introduction fee from that provider and does not receive broker compensation. [About BlackRidge](#).

Minimum funded capital \$25,000. Strategy authorship, live records, the broker, the selected strategy provider and the full commercial terms are presented on an introductory call and confirmed in writing before any payment or deposit.

This report is published for information only. It is not investment advice and does not take account of your circumstances. Trading leveraged instruments including CFDs carries substantial risk and is not suitable for all investors; you may lose the capital you fund and, depending on your broker's terms, may owe more than your deposit. Findings in this report may combine third-party evidence, historical calculations and illustrative scenarios; they are not verified live trading results of any strategy provider. Past performance does not indicate future results.