

When AUC Survives but Portfolios Do Not: Leakage Accounting Errors and Selection Instability in a Retail-Scale Machine-Learning Equity Study

AVTORJI: Parker Fawcett

POVZETEK: Raziskava ocenjuje, kako prediktivna razlika (AUC) in portfeljna performanca v zvezi s strojnimi učenji v finančnih trgovinah. Narejeno je reproducibilno iskustvo z umetno inteligenco, ki ocenjuje verjetnost, da se akcijska varnost konča v najboljši kategoriji v 20-trading-dnevnom obdobju. Rezultati izkažejo, da se AUC lahko ustrezno poveča, dokler se ne spremeni poročevalna metrika, kar povzroča neustaljeno izbiro portfeljev. Prispevki so, da poročevalne neskončnosti lahko vplivajo na finančne rezultate, če se ne odpravi, in da implementacijska preglednost je ključna za malo kvantitativno raziskovanje.

METODA: Raziskava uporablja drevesni razvrstevalec, ki ocenjuje verjetnost, da se akcijska varnost konča v najboljši kategoriji. Poročevalna neskončnost je izkazana, ko se ne odpravi poročevalni koda. Kontrola ponovitvenega obnovega protokola spremeni le pravilo poročevanja, zato da se izvede preglednost. Testni AUC se spremeni minimalno, vendar se spremeni poročevanje portfeljev.

KLJUČNE UGOTOVITVE:

- AUC lahko poveča, dokler se ne spremeni poročevalna metrika, kar povzroča neustaljeno izbiro portfeljev.
- Poročevalne neskončnosti lahko vplivajo na finančne rezultate, če se ne odpravi.
- Implementacijska preglednost je ključna za malo kvantitativno raziskovanje.

POMEN ZA TRGOVANJE: Raziskava podpira, da je implementacijska preglednost ključna za uspešno strojno učenje v finančnih trgovinah. Če se poročevalne neskončnosti ne odpravi, lahko povzročijo neustaljeno izbiro portfeljev, kar vpliva na finančne rezultate. Torej je pomembno, da se implementacijska preglednost uveljavlja pri raziskovanju in razvoju strojnega učenja v finančnih trgovinah.

When AUC Survives but Portfolios Do Not

Leakage, accounting errors, and selection instability in a retail-scale machine-learning equity study

Parker Fawcett

Submission manuscript, September 2026 Prospective extension registered and in progress

Abstract

Machine-learning equity studies often report predictive discrimination and portfolio performance as if improvement in one supports the other. This paper audits that assumption in a reproducible U.S. equity experiment. A tree classifier ranks stocks by their probability of finishing in the top cross-sectional quintile of 20-trading-day return relative to a market benchmark. The audit found validation-label leakage, incorrect aggregation of security-level log returns, incomplete turnover measurement, and risk controls that changed concentration rather than exposure. A controlled refit changes only the purge rule while holding the price snapshot, universe, folds, labels, test observations, model code, and random seed fixed. Test AUC changes from 0.553 to 0.551, but the probability-rank correlation is 0.762 and mean Jaccard overlap of selected portfolios is only 0.389. Portfolio conclusions change despite nearly constant headline accuracy. A frozen 12-minus-1-month momentum rule is more stable and dominates the tested machine-learning portfolio and a fixed ensemble after the ledger is corrected. Its advantage over matched-window SPY is directional rather than statistically distinguishable from zero in a moving-block bootstrap. A dynamic, point-in-time cloud implementation remains positive. A fixed-universe cloud run produces a monthly return path correlated 0.933 with the local implementation across 183 complete months. The evidence supports a methodological conclusion rather than a new-alpha claim: portfolio selection near a model cutoff can be unstable even when aggregate AUC appears robust, and implementation audits can matter more than model choice in small quantitative research programs.

Keywords: machine learning, equity selection, predictive multiplicity, backtest overfitting, look-ahead bias, momentum, reproducibility, portfolio accounting

JEL Classification: C52, C55, G11, G17

1. Introduction

Financial machine learning has two evaluation problems. A statistical model must generalize to future observations, and the decisions made from its output must survive portfolio construction, trading costs, and market frictions. A model can perform acceptably on a global classification statistic while being economically unstable at the narrow cutoff where securities enter a portfolio. Conversely, a correct signal can be misrepresented by an incorrect ledger.

This study documents both failures in one end-to-end equity research project. The project began as an independent reconstruction of a publicly documented retail stock-prediction system. It eventually tested price-based tree models, insider features, long-short portfolios, risk overlays, ensembles, fundamental value, multi-asset trend, and cross-sectional momentum. Negative results and superseded results were retained. An external review then identified errors that required the main strategies to be replayed without retuning.

The most useful result comes from a controlled purge ablation. The original walk-forward routine approximated a 20-trading-day label horizon with calendar days and failed to purge validation

labels that crossed the test boundary. Validation outcomes therefore entered calibration or early stopping even when training observations were separated from the test period. Repairing that rule on an otherwise identical data and model pipeline changes aggregate AUC by only 0.002. It nevertheless replaces most names in the investable portfolio each month and changes the reported returns of previously selected overlays.

This distinction matters because AUC averages pairwise ranking quality across all positive and negative observations. A portfolio uses only a small tail of that ranking and often imposes a probability threshold as well. Two models can therefore have similar global discrimination while disagreeing around the investment boundary. In this experiment, mean selected-name overlap is about 39%, despite similar AUC.

The study makes three contributions. First, it gives a concrete, reproducible decision-level case study in which a valid correction leaves headline predictive accuracy nearly unchanged while materially changing portfolio membership and economic results. This is not a claim that predictive multiplicity itself is new. The narrower contribution is to identify the effect of one validation-code repair while holding the data snapshot, folds, test observations, model, and seed fixed, then follow that effect from AUC to cutoff selections and portfolio results. Second, it separates model error from accounting error through controlled replays. Third, it compares the machine-learning pipeline with a frozen, textbook momentum rule and with a dynamic point-in-time cloud implementation. The comparison shows why a known simple signal can be a more demanding baseline than a weakly informative machine-learning model.

The paper does not establish a deployable trading strategy. The local equity universes use current constituents, the fixed-universe cloud comparison retains that same bias, cloud and local fills and costs are not identical, and the provenance-locked prospective series has not produced its first observation. The paper's submitted claim is the completed controlled audit. The registered prospective series is a separate extension and does not support the results reported here.

2. Related literature

The project sits at the intersection of empirical asset pricing, financial machine learning, momentum, and research-reliability studies.

Gu, Kelly, and Xiu compare machine-learning methods for expected-return prediction and find that trees and neural networks can extract nonlinear interactions from large characteristic panels. They also identify momentum, liquidity, and volatility among the dominant signals. Their institutional-scale setting—roughly 30,000 stocks and hundreds of predictors—is much broader than the present experiment and provides a useful boundary on what a small public- data system should be expected to reproduce [1].

Cross-sectional momentum is a well-established benchmark. Jegadeesh and Titman document continuation among past winners over intermediate holding horizons [2]. The present paper does not claim momentum as a new anomaly. It uses a 12-minus-1-month winner portfolio as a transparent control against which the more complicated model must add value.

The risk of false discovery rises when many strategies are inspected. Harvey, Liu, and Zhu argue that conventional significance thresholds are inadequate in a literature containing hundreds of tested factors [3]. Bailey and López de Prado develop the Deflated Sharpe Ratio to adjust an observed Sharpe ratio for selection bias, non-normal returns, and the number of trials [4]. This project keeps

a lifetime proposal budget and an archive of failures and reports a Deflated Sharpe Ratio using both the formal and broader trial counts.

The broad stability problem has substantial precedent outside and inside finance. Marx, Calmon, and Ustun define *predictive multiplicity*: models with nearly equal accuracy can give conflicting predictions for the same cases [9]. D'Amour et al. describe the related problem of underspecification, in which pipelines with similar held-out performance can behave differently in deployment [10]. Masum et al. provide a recent leakage-controlled, fold-isolated equity-prediction benchmark, but do not study transaction costs, turnover, or a controlled before-and-after portfolio cutoff [11]. Grądzki is one close financial precedent: repeated deep-reinforcement-learning fits can have unstable Sharpe ratios and portfolio weights, motivating allocation-level evaluation and deflated performance statistics [12]. That experiment varies random seeds and stochastic optimization. The present experiment varies one causal validation repair in a supervised cross-sectional classifier while holding the seed and evaluation sample fixed. Li et al.'s August 2026 PC-Audit preprint is also close in purpose: it treats validation-based model selection as a decision-reliability problem and audits transfer, costs, execution, and multiplicity in index ETFs [13]. It compares candidate-model selection across time; it does not measure a before-and-after leakage repair or the stability of stocks selected at a cross-sectional probability cutoff.

Positioning against the closest prior work

Study	Source of disagreement	Domain
Marx et al. (2020) [9]	Near-optimal classifiers	General classification
D'Amour et al. (2022) [10]	Underspecified pipelines	Multiple applied domains
Masum et al. (2026) [11]	Leakage-controlled model benchmark	Short-horizon equities
Grądzki (2026) [12]	Random seeds and stochastic training	Financial deep reinforcement le
Li et al. (2026), preprint [13]	Validation-selected candidate models	Index-ETF forecasting
This study	One label-purge repair, all named controls fixed	Cross-sectional equities

The present contribution differs from a factor-discovery paper. It asks how model-validation details propagate through a selection threshold into realized portfolio decisions. That focus also motivates the public audit trail: incorrect results remain visible with explicit supersession labels instead of disappearing from the record. Its novelty claim is therefore the controlled propagation path, not the general existence of predictive multiplicity, momentum, or unstable financial models.

3. Research questions and hypotheses

The historical analysis is organized around four questions.

RQ1. Does correcting label purging materially change aggregate predictive discrimination?

H1. The corrected and legacy pipelines have similar out-of-fold AUC.

RQ2. Can similar aggregate discrimination conceal unstable investable rankings?

H2. Selected-name overlap is materially lower than the correlation of the full probability rankings.

RQ3. Does the tested machine-learning rank improve a simple momentum portfolio?

H3. A 50/50 machine-learning and momentum ensemble does not improve both CAGR and Sharpe relative to momentum alone on the same periods.

RQ4. Does the direction of the local momentum result survive removal of the current-constituent universe restriction?

H4. A frozen 252/21 momentum rule remains positive in a dynamic, point-in-time U.S. equity universe.

These hypotheses describe the recorded experiments; they are not newly preregistered claims. The separate **prospective-v2** protocol is the project's first immutable forward test.

4. Data

4.1 Local equity panels

Local experiments use adjusted daily OHLCV data downloaded from Yahoo and cached by ticker. The small-cap sample contains 100 names sampled from the September 2026 S&P SmallCap 600 constituent list, with **LEG** subsequently excluded because the endpoint returned only six stale observations. The mid-cap replication contains 100 names sampled from the September 2026 S&P Mid-Cap 400 list. IWM supplies the market-relative label and benchmark for both the small-cap model and the mid-cap momentum results reported in Table 3. An earlier, separate mid-cap machine-learning generalization test (documented in **FINDINGS.md**, not reported in this paper) used MDY as its market-relative label; it is not the source of any table here. SPY is used for matched-window buy-and-hold comparisons.

These lists are known today and therefore contain survivorship and membership look-ahead bias when projected backward. Local return magnitudes are provisional. The fixed lists remain useful for controlled comparisons in which the same observations and securities appear on both sides.

The main momentum panel covers 188 monthly decisions from February 28, 2011, through September 10, 2026. It is split into two equal 94-month halves, with the boundary between November 30 and December 31, 2018. The terms *tune* and *holdout* are retained from the experiment log, although the paper treats the second half cautiously because the wider project inspected it during later development.

4.2 Dynamic-universe cloud check

The QuantConnect implementation selects, at each universe update, the 200 most liquid U.S. equities that have fundamental data, price above \$5, and daily dollar volume above \$500,000. QuantConnect's Security Master tracks splits, dividends, delistings, mergers, and ticker changes, while dynamic universe selection reduces the selection bias inherent in a current-constituent list [5]. The cloud experiment uses QuantConnect data and its Interactive Brokers brokerage model.

The cloud universe, dates, order timing, and fee model differ from the local test. It is a directional external check rather than a matched replication.

4.3 Fixed-universe cloud comparison

A second QuantConnect experiment fixes the security list to the same 99 September 2026 small-cap constituents used by the local run after excluding **LEG**, and covers January 2011 through June 17, 2026. This comparison retains survivorship bias by design. Its purpose is narrower: test whether independent QuantConnect data and execution produce a similar magnitude when universe and broad calendar period are held constant.

The implementations are still not observation-for-observation identical. QuantConnect converts market orders submitted with daily data to market-on- close or market-on-open orders, applies its Interactive Brokers fee model, adjusts symbols to their available mapping and factor-file start dates, and ends with a partial June holding interval. The local implementation charges a fixed 25 basis points and measures 21-session returns from its month-end decision dates. These differences are recorded rather than silently treated as matched.

5. Methods

5.1 Prediction target and features

For security (i) on decision date (t), the target is an indicator for whether its 20-trading-day forward log return in excess of the market lies in the top cross-sectional quintile:

[$y_{i,t} = \mathbb{1}(\mathbf{r}_{i,t \rightarrow t+20} - \mathbf{r}_{m,t \rightarrow t+20}) \text{ is in the top 20\% at } t$.]

The eleven price and volume features are trailing 1-, 5-, and 20-day log returns; 20-day simple momentum; 20-day return volatility; daily range and its 20-day mean; volume relative to its 20-day mean; 14-day RSI; and price gaps relative to 10- and 50-day moving averages. The feature set intentionally stays small and public.

The classifier is a gradient-boosted decision-tree model. The implementation uses LightGBM when installed and otherwise uses scikit-learn's histogram gradient boosting with isotonic calibration. A saved result must therefore record the backend and software versions; the newest cache and prospective-run formats do so.

5.2 Walk-forward evaluation

Each fold uses approximately three years of training data, six months of validation, and six months of test data. Test windows advance sequentially. Because the target spans 20 trading sessions, a row's label end is computed from the actual twentieth subsequent observation for that ticker.

The corrected split removes a training row if its label ends inside or near the validation period and removes a validation row if its label ends inside or near the test period. A 20-calendar-day embargo is added beyond the measured label end. The test set is never used for calibration or early stopping.

The legacy routine made two different choices: it approximated the label end with calendar time and purged training observations only. The controlled ablation reconstructs that legacy routine without changing any other component.

5.3 Controlled purge ablation

Both ablation arms are refit on one frozen current price snapshot. They share:

- all 26 fold boundaries;
- identical training candidates before purging;
- identical test rows, labels, closing prices, and security universe;
- the same feature code, classifier implementation, backend, and seed; and
- the same portfolio and accounting code.

Only the split-purging rule differs. This design identifies the effect of the purge repair within this pipeline. It does not identify a general causal effect for other models or datasets.

We report pooled out-of-fold AUC, Pearson and Spearman probability association, monthly rank association, and selected-name Jaccard overlap. The stability analysis uses the last cached observation in each calendar month, matching the portfolio rebalance rule. For selected sets (A_t) and (B_t), overlap is

$$[J_t = \frac{|A_t \cap B_t|}{|A_t \cup B_t|}.]$$

The selection rule takes up to 20 names whose predicted probability is at least 0.25. This is the economically relevant tail used by the replayed overlays. A month in which both arms select no names has no portfolio union and is excluded from the mean Jaccard calculation.

5.4 Portfolio accounting

All corrected local strategies use simple security returns and compound portfolio equity as ($E_t = E_{t-1}(1+R_t)$). An equal-weight portfolio return is the arithmetic mean of constituent simple returns. This replaces the earlier practice of averaging log returns and exponentiating, which does not represent an equal-weight rebalanced portfolio.

Turnover is one half of the (L_1) distance between old and new security weights:

$$[T_t = \frac{1}{2} \sum_i |w_{i,t} - w_{i,t-1}|.]$$

The local momentum study charges 25 basis points per unit of turnover. Risk scaling changes total invested weight and holds the remainder in cash. Stops, where used by an overlay, exit at the first observed close beyond the threshold and use that actual close rather than the threshold price. Maximum drawdown is computed from an initial equity value of 1.0 so that an immediate first-period loss is counted.

5.5 Momentum and ensemble baselines

The frozen cross-sectional momentum rule ranks securities monthly by return from 252 sessions before the decision to 21 sessions before it. It holds the top decile, subject to at least five names, for 21 trading sessions with equal weights. The rule is not fitted.

The ensemble averages the monthly returns of a frozen machine-learning book and the momentum book at 50/50 weight. It introduces no optimized mixing parameter. The relevant test is whether it improves both return and risk-adjusted return relative to momentum alone on identical dates.

5.6 Performance measures

The paper reports compound annual growth rate, monthly-return Sharpe ratio annualized by ($\sqrt{12}$), maximum drawdown, turnover, and exposure. The historical tables also report moving-block-bootstrap confidence intervals, CAPM and Fama-French-five-plus-momentum factor alpha, and a Deflated Sharpe Ratio for the searched self-improvement loop.

6. Results

6.1 AUC stability conceals portfolio instability

Table 1 presents the central controlled result.

Table 1. Controlled purge ablation

Measure	Legacy purge	Corrected purge	Cross-pipeline agreement
Pooled out-of-fold AUC	0.553	0.551	—
Probability Pearson correlation	—	—	0.762
Probability rank correlation	—	—	0.762
Mean selected-name Jaccard	—	—	0.389

The headline AUC declines by only 0.002 after the repair. The portfolios are far less stable: at the threshold and top-20 cutoff, only about 39% of the union of selected names appears in both arms on an average decision date. The result supports H1 and H2. Aggregate classification quality is insufficient to establish decision stability.

Figure 1 separates time variation from proximity to the selection boundary. Across 155 monthly decisions, 154 have a nonempty union of selected names. Mean Jaccard is 0.389, median Jaccard is 0.419, and only two nonempty months select exactly the same portfolio. Membership changes for 41.4% of security-months within one probability point of the arm-specific boundary and 38.4% between one and two points away. The rate falls to 12.6% at two to five points, 0.9% at five to ten points, and 0.3% beyond ten points. This monotone concentration at the cutoff supplies direct mechanism evidence for H2.

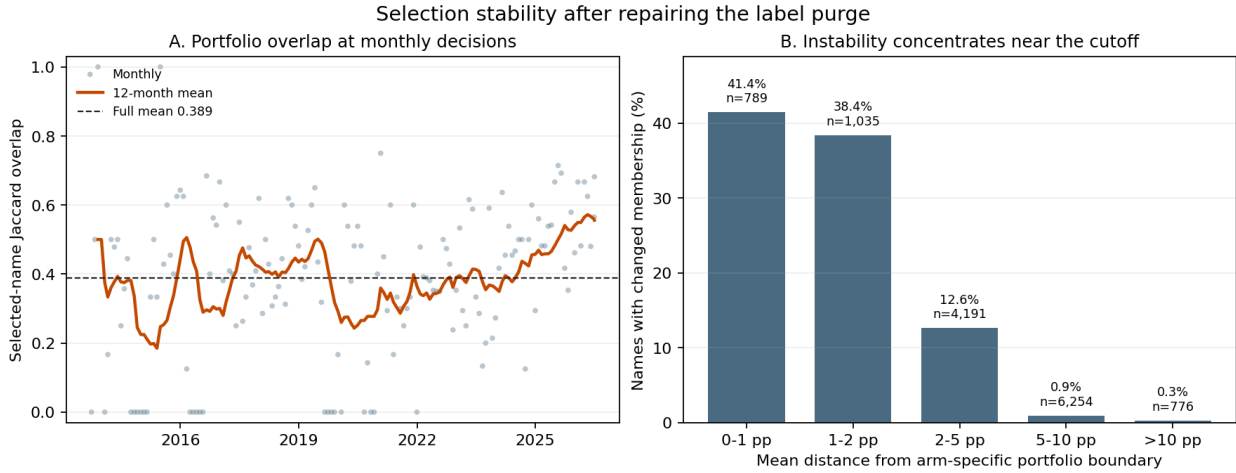


Figure 1. Selection stability after repairing the label purge. Panel A shows monthly selected-name overlap and its trailing 12-month mean. Panel B groups all common security-months by their mean probability distance from each arm's effective probability/top-count boundary. Counts are security-months.

Previously promoted portfolio overlays also change materially, as shown in Table 2. Neither ablation arm clears the original tuning gate with confidence (one cell lands almost exactly on it), so these results are diagnostics and must not be interpreted as retroactive strategy promotions.

Table 2. Previously selected overlays under the controlled purge ablation

Overlay	Legacy tune			Legacy holdout			Corrected tune		
P5: lower probability gate	+0.8%	0.14	-43.2%	+11.9%	0.57	-35.6%	+4.8%	0.35	-33.0%
P8: inverse-volatility weights	+4.2%	0.32	-40.4%	+6.7%	0.40	-39.4%	+5.6%	0.40	-31.1%
P9: top 30	+4.3%	0.32	-40.2%	+6.6%	0.40	-39.7%	+5.5%	0.40	-30.4%
P12: top 40	+4.3%	0.32	-40.2%	+6.4%	0.39	-39.7%	+5.7%	0.41	-29.9%

Cells report CAGR / annualized monthly Sharpe / maximum drawdown. “Tune” and “holdout” describe the project's recorded split. Corrected-tune P12 (0.41) lands almost exactly on the original tune Sharpe requirement rather than clearly under it; the other five tune cells stay clearly below it. (Tune cells revised Sep 14, 2026 after fixing a NaN-propagation bug in `leg_simple`, documented in `FINDINGS.md`; holdout cells were re-verified and are unchanged.)

The direction of the holdout difference is unexpectedly favorable to the corrected purge. That observation does not mean leakage was conservative in general. Purging changes fitted rankings, and a single realized market path can reward either set. The identified finding is instability, not the sign of the change.

6.2 Momentum is stronger than the machine-learning portfolio

Correcting the equal-weight return ledger materially raises the measured momentum result without changing its rule, dates, or securities. The old formula is reproduced on the same observations, which attributes the change to accounting rather than market-data drift.

Table 3. Frozen 12-minus-1-month momentum after ledger correction

Universe	Tune CAGR	Tune Sharpe	Tune max DD	Holdout CAGR	Holdout Sharpe	Holdout max DD
Small cap	17.8%	0.84	-26.5%	21.7%	0.78	-37.1%
Mid cap	9.8%	0.54	-27.8%	24.7%	0.81	-31.0%

Matched-window SPY buy-and-hold returns 11.9% in the first half and 16.9% in the second. Small-cap momentum exceeds SPY by 5.9 and 4.8 percentage points, respectively. Mid-cap momentum trails by 2.1 points in the first half and exceeds SPY by 7.8 points in the second. These comparisons use identical decision dates and annualization conventions, but the constituent lists still contain survivorship bias.

6.3 Adding the machine-learning leg dilutes momentum

Table 4 compares the model, momentum, and their equal-weight ensemble on the same window.

Table 4. Machine learning, momentum, and the fixed 50/50 ensemble

Strategy	Tune CAGR	Tune Sharpe	Tune max DD	Holdout CAGR	Holdout Sharpe	Holdout max DD
Machine learning	-0.1%	0.13	-61.5%	8.9%	0.38	-37.1%
Momentum	17.3%	0.86	-25.3%	20.3%	0.71	-31.0%
50/50 ensemble	9.8%	0.60	-42.3%	16.5%	0.63	-36.5%

The ensemble reduces CAGR and Sharpe relative to momentum alone in both halves and increases maximum drawdown in the first. H3 is supported for this fixed comparison. The model's weak discrimination does not translate into incremental portfolio value.

6.4 Dynamic-universe QuantConnect check

The corrected QuantConnect strategy applies the same 252/21 momentum formation rule to a dynamic top-200 liquid U.S. equity universe. The cloud result reports 3,455.56% total return, 25.97% annualized return, 53.70% daily-path maximum drawdown, and a 0.695 daily-path Sharpe ratio. It completes without the missing-price order errors seen in an earlier implementation. H4 is supported: the direction remains positive when the current-constituent list is removed.

This result is intentionally not placed beside the local CAGRs as a direct replication. Its dynamic universe contains larger and more liquid companies, the backtest spans January 2011 through the cloud data available in 2026, and QuantConnect applies a different execution and fee model. The magnitude could reflect those differences.

6.4.1 Fixed-universe comparison The fixed 99-name QuantConnect run reports 2,557.48% total return, 23.617% compound annual return, 48.10% daily-path maximum drawdown, and a 0.690 daily-path Sharpe ratio. End equity is \$2,657,483.15 from \$100,000, with \$18,335.09 in reported fees across 1,733 orders. The local reference on the comparable survivor list and window is 21.65% CAGR, 37.15% maximum drawdown, and 0.85 Sharpe.

Table 5. Fixed-survivor implementation comparison by sampling frequency

Implementation and sampling	CAGR	Sharpe	Maximum drawdown
Local Yahoo, 183 common 21-session legs	20.86%	0.83	-37.15%
QuantConnect, 183 common calendar months	23.75%	0.93	-39.30%
QuantConnect, official daily path	23.62%	0.69	-48.10%

The official full-span 1.97-percentage-point CAGR difference is small relative to the return magnitude, which is encouraging evidence that the local result is not solely a Yahoo pricing artifact. The apparent 10.95-point drawdown gap in the original comparison was mostly a frequency mismatch. At month-end, the gap is 2.15 points and QuantConnect's Sharpe is higher by 0.11. Across 183 complete months from March 2011 through May 2026, the two return paths correlate 0.933 (R-squared 0.871), with 2.15 percentage points mean absolute monthly difference.

The free-tier log cap preserves 142 comparable pick records through December 2022. After mapping historical tickers to their current companies, mean pick-set Jaccard overlap is 0.866 and 61 months match exactly. Exact-pick months still have 1.83 points mean absolute return difference, consistent with the remaining price-history, fill-timing, and fee differences. This establishes close path-level replication but says nothing new about survivorship because both implementations use the same current-constituent list.

6.5 Factor exposure and resampling uncertainty

Section 9 items 1-2 are now complete. We regress each momentum series' monthly excess return (over the one-month T-bill) on the Fama-French five factors [7] plus a momentum factor [8], using Newey-West standard errors with three lags to account for autocorrelation from the 21-trading-day

holding period [6]. Each decision-date return is assigned to the calendar month in which it is realized.

Table 6. CAPM and Fama-French-5-plus-momentum regressions, full sample (185 months, March 2011-July 2026)

Universe	CAPM alpha (annualized)	(t)	FF5+Mom alpha (annualized)	(t)	(R ²)	Momentum beta
Small cap	+5.8%	1.09	+9.0%	2.14	0.65	0.43 ((t)=4.9)
Mid cap	+1.5%	0.29	+0.9%	0.18	0.62	0.70 ((t)=5.7)

Small-cap momentum retains a marginally significant alpha after controlling for market, size, value, profitability, investment, and the academic momentum factor itself, at conventional (uncorrected) significance. Mid-cap alpha is statistically indistinguishable from zero once those factors are priced in: most of its apparent edge loads on known factors, principally momentum and size, rather than on something beyond them. Neither figure carries a multiple-testing correction. Table 8 adjusts the self-improvement loop's selected Sharpe ratio; it does not correct these regression coefficients.

We next construct a moving-block bootstrap (block length six months, 5,000 resamples, circular resampling to preserve series length) over the holdout half only, for mean monthly return, CAGR, annualized Sharpe ratio, and mean monthly excess return over SPY measured on the same decision dates.

Table 7. Moving-block bootstrap 95% confidence intervals, holdout half (94 months)

Universe	Mean monthly return	CAGR	Sharpe	Mean monthly excess vs.
Small cap	+2.06% [+0.58%, +3.68%]	+21.7% [+1.9%, +47.6%]	0.78 [0.22, 1.46]	+0.59% [-0.66%, +1.85%]
Mid cap	+2.30% [+0.70%, +4.27%]	+24.7% [+3.7%, +54.5%]	0.81 [0.27, 1.35]	+0.79% [-0.53%, +2.36%]

The first three columns use all 94 holdout months. The SPY-excess column uses the 92 months with a complete matched 21-session SPY return.

Mean return, CAGR, and Sharpe intervals stay entirely positive in both universes; the result that momentum makes money in the holdout period is resampling-robust. The excess-over-SPY interval crosses zero in both universes. A positive point estimate against a matched benchmark is not the same claim as a statistically distinguishable one, and Section 6.2's comparison should be read with that distinction: the point estimate and the window-matching methodology are correct, but the outperformance itself is directional, not confidently established, at this sample size.

6.6 Deflated Sharpe Ratio for the self-improvement loop

Section 9 item 3 is complete. The Deflated Sharpe Ratio [4] targets the self-improvement loop specifically, not momentum: momentum is a frozen, unfitted rule (Section 5.5), so its Sharpe is not the maximum of a search the way a promoted loop proposal's is. We rerun all 12 formal loop proposals on the tune half of the legacy (pre-registered) fold cache under the fully corrected code and take the empirical maximum monthly Sharpe as ($\widehat{\text{SR}}$).

Under the corrected code the tune-half maximum is P3 (a signal blend never promoted historically despite clearing the original Sharpe and drawdown thresholds on record), not the historically promoted P5. We do not resolve why; either an undocumented criterion filtered P3 out at the time,

or it was a selection oversight. DSR is computed on the actual current maximum (P3) regardless, since that is what a Sharpe-maximizing search surfaces today. P3's 82-month tune return series has skewness +0.417 and kurtosis 3.94 (Pearson; normal is 3).

Table 8. Deflated Sharpe Ratio under three trial-count assumptions

Trial count	($\sigma(\text{SR})$) source	($E[\max \text{SR} \mid N]$) (ann
(N=12), formal loop as run	all 12 tune Sharpes	
(N=12), formal loop	11 tune Sharpes, P1 artifact excluded	
(N=38), broad (all reported configurations)	same 11, held fixed	

The naive probabilistic Sharpe ratio, which ignores the selection process entirely, reports 98% confidence that the true Sharpe of the best trial is positive. That figure is not the relevant one. Correcting for the 12 trials actually conducted lowers this to 0.57-0.59: statistically indistinguishable from a coin flip on whether the loop's best result is genuine skill rather than the expected maximum of 12 noisy draws. Extending the trial count to every distinct configuration this project has reported ((N=38), including the four later exploratory rules) lowers it further to 0.36 -- below even chance, meaning the observed best Sharpe is now smaller than the expected maximum of 38 pure-noise trials. The (N=38) row holds ($\sigma(\text{SR})$) at the formal estimate rather than re-measuring it from the informal set, so it is a sensitivity bound rather than an independently estimated figure; the direction is unambiguous regardless of that simplification.

This result is consistent with every other machine-learning finding in this paper: nothing about the self-improvement loop's output survives being asked whether it is distinguishable from what N trials looks like under a null of no skill.

After this retrospective DSR analysis was completed, four additional rules (P13-P16) were explored on the already inspected corrected-cache tune period. The best, P15, requires both the machine-learning and momentum ranks to be in the cross-sectional top 40%. It reports 15.32% CAGR, 0.903 Sharpe, and -22.2% maximum drawdown versus 11.33%, 0.762, and -21.6% for a same-run momentum control. These four return series are excluded from Table 8's historical 12-proposal reconstruction because they were designed later. They are counted in the broader (N=38) trial-count sensitivity row. They remain ineligible for promotion: their design follows inspection of the period and they have no untouched holdout. P15 is a hypothesis for future data, not evidence supporting the paper's historical conclusions.

P15 was subsequently subjected to one external validation on the frozen mid-cap `mc_price` cache. The validation rule and gate were committed before execution: P15 had to beat a same-run momentum control on both CAGR and Sharpe, with no more than five percentage points of additional maximum drawdown. Across 177 months, P15 returned 13.53% CAGR, 0.698 Sharpe, and -35.9% drawdown versus 15.40%, 0.692, and -30.7% for momentum. It failed on CAGR and exceeded the drawdown tolerance by 0.2 point. The small-cap exploratory improvement therefore does not generalize to this independent current-constituent universe. The mid-cap cache predates the manifest system; its ordered fold contents are locked in the result registry by a SHA-256 digest, but its original generation environment cannot be reconstructed as completely as newer caches.

7. Discussion

The main result is a disconnect between global predictive stability and local decision stability. AUC summarizes ranking performance across the full sample. The portfolio acts on a small set near two discontinuities: a probability gate and a top-count cutoff. Small probability changes can move many securities across those boundaries even when the full-sample statistic barely moves.

This mechanism explains why “AUC survived the correction” is not sufficient reassurance. The corrected classifier may retain weak information in aggregate, yet a particular portfolio derived from it can remain fragile. Reporting the overlap of actual selections, turnover, and return sensitivity alongside AUC would make this fragility visible in other financial classification studies. This result should be read as a finance-specific, controlled instance of predictive multiplicity and under-specification [9,10]. Grądzki's seed-driven allocation instability provides the nearest portfolio-level comparison [12], while Li et al. provide the nearest validation-decision audit [13]. The additional evidence is the causal structure of the ablation: a single validation repair, identical test observations and seed, and a measured chain from nearly invariant AUC to low selected-name overlap and changed economic results.

The accounting corrections carry a separate lesson. Averaging security-level log returns and exponentiating creates the geometric mean of constituent gross returns, not the return of an equal-weight portfolio. With dispersed small-cap outcomes, the difference becomes economically large. Similarly, measuring turnover from additions and deletions misses changes in position size, and reducing the number of names without scaling weights can increase concentration under the label of risk reduction. These are implementation errors rather than subtle statistical disagreements, but they changed the project's conclusions more than a new model did.

Momentum's relative success is a baseline result, not a novel anomaly claim. It shows that complexity failed to earn its place in this experiment. The machine-learning leg did not improve the simple rule, and its promotions were sensitive to the probability pipeline. The proper inference is limited to the tested public-data setup.

8. Limitations

Five limitations constrain the current evidence.

1. **Current-constituent local universes.** The local small- and mid-cap lists are sampled from September 2026 membership. Delisted and removed securities are absent from earlier years.
2. **Residual external implementation differences.** The fixed-universe QuantConnect result controls the names and broad period, and monthly returns correlate 0.933, but fill timing, fees, mapping start dates, and the final partial holding interval still differ. The dynamic run validates direction without current-constituent selection; the fixed run compares magnitude without removing survivorship bias.
3. **Research-path dependence.** Many ideas were evaluated during the wider project. The 20-proposal lifetime budget has spent 16 trials: 12 in the historical promotion loop and four later exploratory trials that cannot be promoted. Informal design choices also consume researcher degrees of freedom.
4. **Incomplete robustness analysis.** Bootstrap intervals, factor regressions, and a Deflated Sharpe Ratio are complete. A probability-of-backtest- overfitting analysis remains open.
5. **No clean prospective outcomes yet.** The immutable `prospective-v2` series begins at the next completed month end. Historical paper rows were generated under a superseded

dating and grading implementation and are excluded from forward evidence. No prospective result is used to support the controlled audit claim in this submission.

These limitations prevent claims of investable alpha or live deployability. They do not invalidate the controlled finding that similar AUC can coexist with low portfolio overlap.

9. Completed robustness work and registered extension

The following robustness analyses were completed before submission without changing the frozen momentum or machine-learning decision rules.

1. **Done (Section 6.5).** Moving-block-bootstrap confidence intervals for mean monthly return, CAGR, Sharpe, and momentum-minus-SPY return.
2. **Done (Section 6.5).** CAPM and Fama-French five-factor-plus-momentum regressions on monthly portfolio returns, with Newey-West standard errors.
3. **Done (Section 6.6).** Deflated Sharpe Ratio using the documented research-trial count and return skewness and kurtosis, with a sensitivity range extending to informal trials.
4. **Done (Section 6.4.1).** `qc_momentum_matched.py` fixes the cloud universe to the local 99-ticker list and broad 2011-2026 window. `qc_reconcile.py` extracts and aligns the result: 183 complete monthly returns correlate 0.933, and 142 logged monthly pick sets have mean Jaccard overlap of 0.866.
5. **Done (Section 6.1).** `selection_stability.py` reports monthly portfolio overlap and membership disagreement by probability distance from the cutoff.
6. **Done (Appendix C).** `paper_provenance.py` locks every numbered table and figure to its displayed block, source artifacts, generating code, parameters, reproduction commands, and software environment. The prospective protocol in Section 10 is a registered extension rather than a condition of the present paper. A later version may report it after at least 12 monthly cohorts, while preserving this initial empty-series manuscript and all run manifests. The extension will test forward performance; it cannot alter the completed controlled ablation reported here.

The four remaining improvement-loop trials are banked for a genuinely new data window or independent universe. No exploratory result from already inspected history is eligible for promotion.

10. Prospective protocol

The frozen prospective system contains two sleeves: `wide-seed0-v1` and `mom12-1-v1`. A valid prediction run must use completed month-end data and a clean pipeline. It writes an immutable manifest containing UTC creation time, git state, input hashes, model backend and versions, feature list, training and validation dates, parameters, and the complete signal set. Each ledger row stores the manifest hash.

Positions enter at the next session's open and are observed for 20 trading sessions. A position exits at the first actual close at or below the 15% stop, or at the final close if no stop occurs. Round-trip costs are charged. Grading recomputes and verifies the manifest hash and signal fields and is idempotent. The clean records live separately from the superseded historical paper log.

At submission, the clean series contains no mature observations. The historical audit is evaluated on its controlled evidence. Future prospective outcomes will be reported as an extension and will not be used to rewrite the submitted historical record.

11. Conclusion

This audit finds that a repair to financial-model validation can leave aggregate AUC almost unchanged while replacing most investable selections and moving portfolio outcomes. In the controlled ablation, AUC changes from 0.553 to 0.551, while mean selected-name Jaccard overlap is 0.389. The result exposes a blind spot in workflows that stop at predictive discrimination.

Correct portfolio accounting also changes the magnitude of the evidence. Once simple-return aggregation, weight-based turnover, exposure scaling, gap-aware stops, and initial-equity drawdown are enforced, a frozen 12-minus-1-month momentum baseline dominates the tested machine-learning portfolio and its ensemble. A dynamic-universe cloud implementation preserves the positive direction, and a fixed-universe cloud run produces a similar CAGR with worse risk statistics. Comparable month-end sampling narrows the risk difference, and the monthly paths correlate 0.933. These return results remain provisional and do not carry the paper's main methodological claim. Momentum's advantage over matched-window SPY buy-and-hold, however, is directional rather than statistically distinguishable from zero under a moving-block bootstrap, and the self-improvement loop's best observed Sharpe ratio is not distinguishable from the expected maximum of a 12-trial search under a null of no skill -- and falls below it once the search is widened to every configuration this project tested. The machine-learning side of this study has produced no result that survives being asked whether it is better than what the number of trials conducted would produce by chance.

The publishable finding is therefore methodological and deliberately narrow. Predictive multiplicity, financial allocation instability, and validation- decision audits are already known [9,10,12,13]. This study adds a controlled case showing that one leakage repair can leave AUC nearly invariant while replacing most cutoff-selected stocks and changing economic conclusions. In small financial machine-learning studies, researchers should report decision- level stability and preserve corrected failures alongside headline accuracy.

References

- [1] Gu, S., Kelly, B., and Xiu, D. (2020). “Empirical Asset Pricing via Machine Learning.” *The Review of Financial Studies*, 33(5), 2223–2273. <https://doi.org/10.1093/rfs/hhaa009>
- [2] Jegadeesh, N., and Titman, S. (1993). “Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency.” *The Journal of Finance*, 48(1), 65–91. <https://doi.org/10.1111/j.1540-6261.1993.tb04702.x>
- [3] Harvey, C. R., Liu, Y., and Zhu, H. (2016). “... and the Cross-Section of Expected Returns.” *The Review of Financial Studies*, 29(1), 5–68. <https://doi.org/10.1093/rfs/hhv059>
- [4] Bailey, D. H., and López de Prado, M. (2014). “The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality.” *The Journal of Portfolio Management*, 40(5), 94–107. <https://doi.org/10.3905/jpm.2014.40.5.094>
- [5] QuantConnect (2026). “Auxiliary Data: US Equity Security Master” and “Research Guide: Survivorship Bias.” Accessed September 2026. <https://www.quantconnect.com/docs/v2/cloud-platform/datasets/quantconnect/auxiliary-data>
- [6] Newey, W. K., and West, K. D. (1987). “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica*, 55(3), 703–708. <https://doi.org/10.2307/1913610>

- [7] Fama, E. F., and French, K. R. (2015). “A Five-Factor Asset Pricing Model.” *Journal of Financial Economics*, 116(1), 1–22. <https://doi.org/10.1016/j.jfineco.2014.10.010>
- [8] Carhart, M. M. (1997). “On Persistence in Mutual Fund Performance.” *The Journal of Finance*, 52(1), 57–82. <https://doi.org/10.1111/j.1540-6261.1997.tb03808.x>
- [9] Marx, C., Calmon, F. P., and Ustun, B. (2020). “Predictive Multiplicity in Classification.” *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119, 6765–6774. <https://proceedings.mlr.press/v119/marx20a.html>
- [10] D’Amour, A., Heller, K., Moldovan, D., et al. (2022). “Underspecification Presents Challenges for Credibility in Modern Machine Learning.” *Journal of Machine Learning Research*, 23(226), 1–61. <https://jmlr.org/papers/v23/20-1335.html>
- [11] Masum, M., et al. (2026). “Audit-Ready Machine Learning for Short-Horizon Equity Prediction: A Dual-Target Benchmark With Fold-Isolated Preprocessing.” *Engineering Reports*, 8(6), e70893. <https://doi.org/10.1002/eng2.70893>
- [12] Grądzki, P. (2026). “Unstable Gains: Multiplicity-Aware Evaluation of Financial Deep Reinforcement Learning.” *The Journal of Finance and Data Science*, 12, 100205. <https://doi.org/10.1016/j.jfds.2026.100205>
- [13] Li, S., Zhang, W., Wang, Y., and Lei, Q. (2026). “PC-Audit: A Decision-Reliability Framework for Auditing Validation-Based Machine-Learning Model Selection in Weak-Signal Financial Time Series.” Preprints.org, 202608.2127, version 1. Preprint; not peer reviewed. <https://www.preprints.org/manuscript/202608.2127>

Declarations

The analysis code, manuscript source, derived evidence, and reproduction instructions are available in the companion repository: <https://github.com/Parker-Fawcett/Stock>. Large third-party price and fold caches are identified by SHA-256 hashes in the provenance manifest and can be regenerated with the documented commands. The author used AI-assisted tools for parts of code development and editorial revision, then reviewed the code, outputs, and manuscript.

Appendix A. Reproduction map

Result	Repository entry point	Evidence artifact
Controlled legacy/fixed purge refit	<code>purge_ablation.py</code>	<code>data/cache/purge_legacy_ablation</code>
Probability and selection agreement	<code>cache_compare.py</code>	Fold-level probability CSVs in the two
Monthly selection stability	<code>selection_stability.py</code>	<code>data/selection_stability/</code> and fig
Corrected momentum results	<code>mom_run.py</code>	Cached OHLCV files plus console sum
Corrected ML/momentum ensemble	<code>ens_test.py</code>	<code>data/cache/sc_full_v2/</code> plus consol
Dynamic-universe momentum check	<code>qc_momentum.py</code>	QuantConnect result recorded in FIND
Immutable prospective series	<code>paper.py</code>	<code>data/paper/runs/</code> , <code>data/paper/log_</code>
Factor regressions and bootstrap CIs	<code>factor_analysis.py</code>	Ken French factor cache (<code>predictor/</code>)
Deflated Sharpe Ratio	<code>deflated_sharpe.py</code>	<code>data/cache/sc_full/</code> (legacy, pre-reg
Fixed-universe QuantConnect comparison	<code>qc_momentum_matched.py</code>	Cloud backtest Determined Black Co
QuantConnect path and pick reconciliation	<code>qc_reconcile.py</code>	<code>data/qc_reconciliation/summary.j</code>
Paper artifact and environment verification	<code>paper_provenance.py</code>	<code>data/paper_provenance/manifest.j</code>

Appendix B. Evidence-status vocabulary

- **Controlled:** one named mechanism changes while the relevant observations, model, and evaluation remain fixed.
- **Holdout:** evaluated once after a rule passed a previously stated gate; later inspection can contaminate that status but cannot restore it.
- **Directional replication:** the sign survives in a materially different universe or execution system; magnitudes are not treated as estimates of one another.
- **Prospective:** the signal, implementation, and complete manifest are frozen before the outcome begins.
- **Superseded:** a result produced by a known-invalid implementation and kept only as part of the audit trail.

Appendix C. Artifact and environment lock

`data/paper_provenance/manifest.json` records every numbered table and figure, the SHA-256 hash of its displayed Markdown block, the hash of every source file and generating code file, its parameters, and exact reproduction commands. It also records the Python, operating-system, package, and `requirements.txt` versions used to create the lock. Large local price and fold caches are not stored in Git; their individual and aggregate hashes identify the exact external artifacts required for reproduction.

Run `python3 paper_provenance.py verify` from the repository root. The command fails if a displayed table or figure, source input, code dependency, or locked environment differs. Rebuild the manifest only after intentionally rerunning the affected analysis and reviewing the changed result.