

Catch Alpha: Testing the Evaluator Before Testing the Trading Strategy

Joris Viaud
Independent Researcher
France
joris.viaud@edu.ece.fr

Abstract

A trading result is only as trustworthy as its evaluation. We first test our procedure on synthetic data with known truth. In this design, 0/200 zero-edge strategies pass, while detection rises with the planted edge; these finite tests do not cover every market. We then audit learning in a controlled market. Starting every policy flat, behavior cloning and implicit Q -learning recover 90% and 73% of a risk-adjusted analytic reference without costs. With quadratic costs, all ten fits for both methods lose money; median turnover is $6.03\times$ for cloning and $11.97\times$ for implicit Q -learning relative to the reference. Conservative Q -learning fails the no-cost prerequisite, so its costly result is not adjudicated. On twelve liquid cryptocurrency futures, neither the tested long-loser/short-winner reversal nor funding-carry implementations produce reliable net profit. The contribution is a reproducible workflow that measures power, counts attempts, and limits final-period access.

CCS Concepts

• **Computing methodologies** → **Machine learning**; *Reinforcement learning*; • **Applied computing** → *Economics*; Electronic commerce.

Keywords

trading evaluation, backtest overfitting, statistical power, reinforcement learning, cryptocurrency, negative results, reproducibility

ACM Reference Format:

Joris Viaud. 2026. Catch Alpha: Testing the Evaluator Before Testing the Trading Strategy. In *Proceedings of 7th ACM International Conference on AI in Finance (ICAIF '26)*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

A profitable backtest is not persuasive if it is merely the best of hundreds of variants. Repeated search inflates the apparent Sharpe ratio—average return relative to return variability, scaled to one

year [3, 4]; short samples can turn noise into spectacular annualized numbers [2]; and corrections applied only to the final strategy ignore choices made earlier. Deep reinforcement learning (RL) adds variation through random seeds and implementation choices [1, 8]. The evaluation procedure therefore needs evidence of its own reliability.

We learned this from a failed first program. Roughly 96 runs of methods trained from a fixed dataset or by imitation on three crypto assets and 60 days of 15-minute bars produced annualized Sharpe ratios between about -17 and $+17$. The sample was too short to support either success or failure: annualization had magnified noise. This paper replaces that workflow with a testable one.

We ask two questions. Can an evaluation procedure reject strategies with no edge while retaining power to detect a planted edge? When that calibrated procedure is applied to a pre-registered, free-data crypto panel, which claims are supported? The answer is a reusable measurement workflow and a narrowly scoped negative result, not a new trading strategy.

Our workflow has three stages.

- (1) **Calibrate the evaluator on known nulls and planted edges** (Section 2). Before testing a market hypothesis, we measure how often the required checks jointly accept 200 simulated strategies with zero expected return and how its detection rate changes as a known edge grows. They accept **0/200** in this finite Gaussian design and detect 85% at a planted annualized Sharpe of 2.0. Further heavy-tail, volatility-clustering, and serial-correlation tests also produce no acceptances. These experiments calibrate the tested generators; they do not bound every untested null process.
- (2) **Check whether the learning stack can use a known signal** (Section 3). A synthetic market controls the correlation between a signal and the next return. A closed-form Gârleanu–Pedersen (GP) trading rule [5] provides a risk-adjusted analytic reference, not the exact optimum of the reported profit score. A corrected evaluation starts every policy flat and tests ten fitted policies per method on the same 100 new paths. Behavior cloning (BC), which copies recorded actions, and implicit Q -learning (IQL), which learns action values from a fixed dataset [9], recover the signal without costs. Both then lose money in all ten independent fitted runs when quadratic trading costs are applied. Conservative Q -learning (CQL), which penalizes uncertain actions [10], fails the required no-cost check, so its costly result is descriptive.
- (3) **Apply the frozen procedure to real data** (Sections 4–5). We study twelve liquid Binance USD-margined perpetual

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICAIF '26, Milan, Italy

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

futures—contracts without an expiry date—at hourly frequency. Model selection uses only 2019–2024; the 2025–2026 final period is opened once at the end. Every signal screen and every gated strategy evaluation is counted. The tested long-loser/short-winner reversal and funding-carry implementations do not yield a reliable profit after costs. Two small on-chain correlations survive training and validation only; they are leads, not final-period trading claims.

The conclusion is deliberately limited to these assets, horizons, costs, and implementations. The training-and-validation sample fails the minimum-length check in Equation 1 at the stated annualized-Sharpe target even before search correction and the remaining acceptance checks. The calculation is not an end-to-end detection threshold. When the sample is too short, the procedure reports INSUFFICIENT DATA rather than treating a non-detection as proof that no smaller effect exists.

Related work. We operationalize safeguards against search-inflated backtests: the deflated Sharpe ratio (DSR) [4], the probability of backtest overfitting (PBO) [3], minimum backtest length [2], and the superior predictive ability (SPA) test [6]; multiple-testing concerns also motivate the penalties proposed by Harvey and Liu [7]. Our contribution is procedural: every attempted screen enters the trial count, insufficient evidence causes abstention, and final-period access is budgeted. Crypto return and funding premia are documented on broader universes and different horizons [11, 14]. We ask the narrower implementation question: can the tested constructions harvest those families after retail taker costs?

2 Evaluation protocol

Four checks are required for a positive decision: enough observations for the target effect; uncertainty intervals above a declared reference; a Sharpe ratio that remains credible after counting all attempts; and improvement over every baseline. A fifth, fold-based diagnostic estimates selection overfitting when enough comparable strategies and folds exist. Missing inputs can block a positive decision, but this diagnostic cannot by itself establish or overturn profitability. All thresholds are fixed before a run. Each evaluation writes an identifying record (dataset, split, metric family, experiment, commit) to an append-only trial ledger.

2.1 Abstain when the evidence is too weak

Missing or underpowered evidence produces no positive decision. Let P be the number of return periods per year and $S_{\min}$ the smallest annualized Sharpe ratio targeted by a simple one-sided mean test. Significance $\alpha = 0.05$ and power $1 - \beta = 0.80$ give the minimum-length approximation

$$n \geq \left(\frac{z_{1-\alpha} + z_{1-\beta}}{S_{\min}} \right)^2 P \approx 6.18 \frac{P}{S_{\min}^2}. \quad (1)$$

This calculation covers the mean test only; search correction and the other acceptance checks can require more data. At $S_{\min} = 1$, hourly data require 54,160 observations, or about 6.2 years. Our training-and-validation period contains 46,590 hourly portfolio returns, so the rule reports INSUFFICIENT DATA. The count is by timestamp, not by asset: twelve correlated assets do not become twelve

independent observations. For overlapping multi-hour returns, we further inflate the requirement by the moving-average dependence implied by the overlap horizon.

2.2 Uncertainty, search breadth, and benchmarks

For candidate return r_t and benchmark return b_t , a stationary bootstrap [13] resamples blocks to preserve serial dependence and generates 95% confidence intervals for both the mean and the Sharpe ratio of $r_t - b_t$. Both lower bounds must be positive. The main implementation uses 400 seeded bootstrap replications; the large calibration sweep uses 150 for runtime.

The deflated Sharpe ratio (DSR) then corrects the winning Sharpe for the number of strategies considered [4]. If N variants were tried, the null reference is the expected maximum across those trials:

$$SR_0 = \sqrt{V} \left[(1 - \gamma) \Phi^{-1} \left(1 - \frac{1}{N} \right) + \gamma \Phi^{-1} \left(1 - \frac{1}{Ne} \right) \right], \quad (2)$$

where $\gamma \approx 0.5772$ and V is the variance of Sharpe estimates across trials. The ledger records which evaluations were run, not a Sharpe ratio for each (most are signal screens), so the implementation approximates V by the candidate's own estimated Sharpe sampling variance. On the Gaussian calibration design below, the measured cross-trial variance is 4.6×10^{-4} against an approximation of 5.6×10^{-4} , and no zero-edge strategy passes under either. The candidate passes when the gap between its estimated Sharpe and SR_0 , divided by the estimated standard error of that Sharpe, has one-sided normal probability at least 0.95.

The crucial input is N : it comes from the ledger and counts every recorded evaluation on the panel, including upstream signal screens, not merely the final strategies. The frozen real-data ledger holds 353 records (352 selection evaluations and the single final-period run): 344 signal-horizon screens (128 price, 201 external, and 15 market-cap-adjusted on-chain pairs) and nine portfolio evaluations, of which three re-emit earlier verdicts and one is the final-period run. Counting re-emissions can only raise SR_0 . Later synthetic audits enlarge the ledger but not this count. Every real-data evaluation at the default target stopped at the minimum-length check, so no real-data verdict rests on a DSR value. As an illustration, for a per-period Sharpe of 0.06 over 3,000 observations, counting three trials instead of 347 moves DSR from 0.63 to 0.99 and reverses the decision.

A pairwise superior predictive ability (SPA) test [6] compares the candidate with each declared baseline using resampled performance differences (lower recentring, studentized by the marginal standard deviation). We require the largest baseline-specific p -value to be at most 0.10. Finally, the probability of backtest overfitting (PBO) [3] asks whether the strategy that wins on one group of chronological folds tends to disappoint on the remaining folds. Each of the $\binom{6}{3} = 20$ splits of the folds into two halves selects the best strategy on one half and records its relative rank ω on the other; PBO is the share of splits with $\ln(\omega/(1 - \omega)) \leq 0$ (winner below the out-of-sample median) and must not exceed 0.50. This

check is asymmetric: a missing fold-by-strategy matrix blocks acceptance, but a low PBO value cannot by itself establish profitability. Section 5 explains why the small matrix available here is only an implementation smoke test.

Access to the final period is controlled separately. At most three evaluations were allowed for the registered real-data panel, and this paper uses one: the closing test in Section 5. Final-period runs evaluate an already frozen claim and never enter model selection.

2.3 Test the evaluator before the market

Published corrections can still be implemented incorrectly or interact in an unexpected way. We therefore measure the four required checks jointly on data with known truth (Fig. 1). For the Gaussian calibration, returns are treated as daily observations ($P = 252$) and $S_{\min} = 1$; every 1,800-observation run exceeds the exact 1,559-observation minimum from Eq. 1. We apply the four checks to 200 independent draws. Each draws one new zero-mean candidate, benchmark, and two baselines, while DSR uses the fixed count $N = 200$. None passes: 0/200, with a Wilson 95% upper confidence bound of 1.9%. No winner is selected among competing candidates in this design, and PBO is not exercised.

The same zero count occurs with Student- t_3 innovations, a GARCH(1,1) volatility process with coefficients 0.08 and 0.90, and an AR(1) process with coefficient 0.2, each tested at 1,800, 5,000, and 50,000 observations. All three are scaled to 1% unconditional standard deviation; the Student- t draws are variance-normalized, while GARCH starts at that variance and uses intercept $0.02(0.01)^2$. Across all twelve settings, 0/2,400 pass; each setting’s Wilson upper bound is 1.9%, and no pooled bound is claimed because some settings share random streams. A market-derived null, created by shifting the real signal by at least 26 weeks, also yields 0/200; its 276 weekly observations exceed its declared minimum of 144 at $S_{\min} = 1.5$. These finite tests calibrate the tested zero-edge processes; they do not establish a false-positive bound for every possible market process.

Rejecting everything would be useless, so a second control uses the same 1,800-observation design and plants annualized Sharpe ratios from 0.5 to 3.0, with 40 draws per effect size. Detection rates are 0/0/30/85/97.5/100%. The evaluator is conservative after 200 attempts, but its detection rate rises with a sufficiently large planted edge.

Selection among correlated candidates. The design above evaluates one candidate per draw. In a follow-up registered after review, each campaign draws 200 equicorrelated zero-edge candidates (pairwise correlation 0, 0.3, or 0.7), selects the largest development-window Sharpe, and evaluates that winner on the selection data and on an independent holdout: 500 null campaigns per correlation, and 200 campaigns per planted edge in which one candidate carries the edge. Table 1 compares the four checks with simpler rules on the same draws. An uncorrected test on the selection data accepts 48–100% of zero-edge winners. The four checks with $N = 200$ accept 1 of 1,500 (worst-cell upper bound 1.1%); the same winners pass in 8–36% of campaigns when N is set to 1 (clamped to 2 by the implementation), and the DSR check alone reproduces the four-check outcome, so the trial count carries the false-acceptance control here. A one-sided test on an untouched

Table 1: Acceptance rates (%) after selecting the best of 200 correlated candidates. Null columns: zero-edge winners accepted, 500 campaigns each. Detection columns: planted candidate selected and accepted, 200 campaigns, correlation 0. “Sel.” = selection data; “hold.” = independent holdout.

Rule	Null, by correlation			Detection	
	0	0.3	0.7	$S=1.5$	$S=2.0$
Uncorrected t -test, sel.	100.0	83.6	48.2	86.5	97.5
Four checks, $N=200$, sel.	0.2	0.0	0.0	35.0	81.5
Four checks, $N=200$, hold.	0.0	0.0	0.0	24.5	81.5
Four checks, $N=1$, sel.	36.4	20.6	8.0	74.5	95.0
DSR alone, $N=200$, sel.	0.2	0.0	0.0	35.0	82.0
t -test alone, hold.	5.0	5.6	7.6	86.0	97.5

holdout is near nominal (5.0–7.6%) with higher power: at planted Sharpe 1.5 it detects 86%, against 24.5% for the four checks on the same holdout and 35% on the selection data. That is the price of the correction when a clean, single-use holdout exists. All five registered predictions held. Selection is over a fixed family; adaptive search is not simulated.

2.4 Chronological evaluation without leakage

Real-data selection uses six chronological walk-forward folds. A 168-hour gap at every fold boundary prevents overlapping targets and lagged features from crossing between training and validation. Every validation-window comparison enters the ledger. A claim must repeat across folds rather than rely on one fortunate interval.

3 Can the learner use a known signal?

The statistical calibration tests the decision rule. We next test the learning pipeline in a market where the signal is controlled. Each of five simulated assets has an independent persistent predictor, $s_{i,t+1} = 0.9s_{i,t} + \eta_{i,t+1}$, with $\eta_{i,t+1} \sim \mathcal{N}(0, 1 - 0.9^2)$. Returns follow $r_{i,t+1} = 0.0005s_{i,t} + \epsilon_{i,t+1}$, where $\epsilon_{i,t+1} \sim \mathcal{N}(0, 0.01^2(1 - 0.05^2))$. Innovations are independent across assets and time, and the observed feature equals the latent signal. Total return volatility is therefore 0.01 and the expected correlation between $s_{i,t}$ and $r_{i,t+1}$ —the information coefficient (IC)—is 0.05.

A policy chooses portfolio weights w_t . On a 1,000-period evaluation path we score raw profit after quadratic trading costs,

$$S(\pi) = \sum_t \left[w_t^\top r_{t+1} - \frac{\lambda}{2} \|w_t - w_{t-1}\|_2^2 \right], \quad (3)$$

where r_{t+1} is the next asset-return vector and λ controls how strongly changes in the portfolio are penalized. Turnover is the total absolute change in portfolio weights. The case without costs uses $\lambda = 0$; the case with costs uses $\lambda = 1$.

The comparison policy is the discrete-time GP rule of Gârleanu and Pedersen [5]. It moves gradually toward a desired position to balance expected return, risk, and trading costs. We normalize a learner’s score as

$$Q(\pi) = \frac{S(\pi)}{S(\text{GP})}. \quad (4)$$

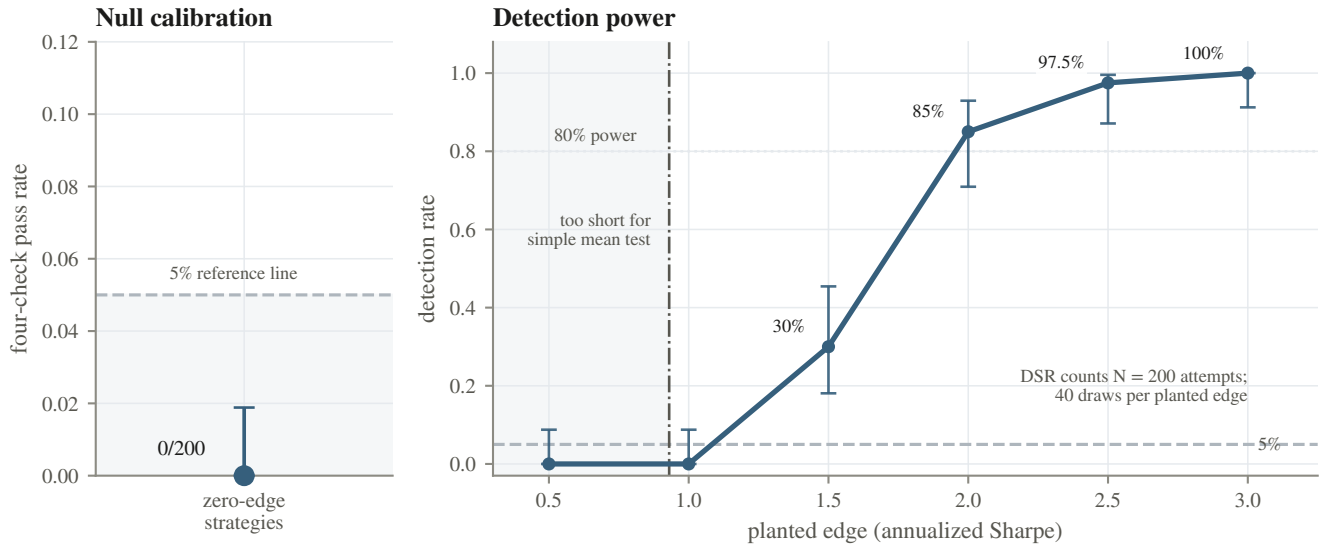


Figure 1: Finite calibration on known truth. Left: none of 200 Gaussian zero-edge strategies passes (95% upper bound 1.9%). **Right:** detection rises from 0% at planted Sharpe 0.5–1.0 to 100% at 3.0, with 40 draws per planted effect; error bars are 95% Wilson intervals. Additional tested null generators also produce zero acceptances. All rates and bounds are conditional on the tested generators and sample sizes; they do not imply a universal false-positive rate.

GP is a risk-adjusted analytic reference, not the exact optimum of the raw profit score in Eq. 3; Q therefore measures performance relative to GP, not exact regret.

A corrected comparison with identical starts. An audit found that the original experiment did not give every policy the same initial state: learned policies started the scored window flat, whereas GP carried a position from earlier data. The value 10 was also used both as GP’s risk-aversion coefficient and, incorrectly, as the learner’s future-reward discount for IQL and CQL. We withhold those original ratios. The pre-registered correction starts every policy from $w_{-1} = 0$, separates risk aversion 10 from learner discount 0.99, and evaluates each fitted policy on the same 100 new paths. Nine pre-specified configurations use ten independent random training seeds each, giving $9 \times 10 = 90$ fitted policies and 9,000 policy-path evaluations. No evaluation path overlaps a training path.

Controlled configuration. Each training seed generates a 2,000-period history and fits on its first 1,000 periods. The offline actions come from GP plus Gaussian noise of standard deviation 0.05. Policies receive a causal 16-period window, use hidden width 64, batch size 32, and train for eight epochs; every weight is clipped to $[-2, 2]$. Seven adjudicating configurations cover BC, IQL, and CQL without costs; the same three methods with costs; and a registered BC rescue that includes the previous position in its input. The costly conclusions below concern only BC and IQL at $\lambda = 1$ without that position input; the no-cost fits are prerequisites, and CQL’s costly arm is descriptive. Two diagnostics repeat costly IQL and CQL with the old discount error; they isolate that error and cannot support a positive capability conclusion. Primary IQL and CQL use discount 0.99.

What passes and what does not. Without costs, the median scores relative to GP are 0.900 for BC, 0.731 for IQL, and 0.335 for CQL. The pre-specified no-cost rule requires at least 0.50 and positive mean profit for at least eight of ten fitted runs. BC and IQL pass; CQL does not. CQL’s costly outcome is therefore descriptive rather than an adjudicated failure.

With $\lambda = 1$, BC and IQL have median mean profits -0.995 and -4.036 . All ten fitted runs are negative for both methods; their median turnovers are $6.03\times$ and $11.97\times$ that of GP, respectively. They satisfy the pre-registered collapse rule. The reference mean is $+0.0461$, with a one-sided 95% lower bound of $+0.0395$, so the sign comparison is not an unstable division by a reference score close to zero. Adding the previous position to BC does not rescue the result: median mean profit is -0.960 , all ten fitted runs are negative, and median turnover remains $5.92\times$ GP.

This is a negative result for the two tested offline configurations under the specified signal strength, cost model, and horizon, and does not generalize to offline RL as a whole. BC reproduces a noisy GP-derived behavior policy rather than discovering the signal independently. CQL fails the required no-cost check, so we draw no conclusion about its costly case. The experiment does not cover larger budgets, online policy training, or architectures that directly constrain turnover. Excess turnover accompanies the BC and IQL losses, but the experiment does not by itself prove that turnover is the only mechanism.

4 Real-data study

Data and split. The panel contains twelve liquid Binance USD-margined perpetual futures at hourly frequency: BTC, ETH, BNB,

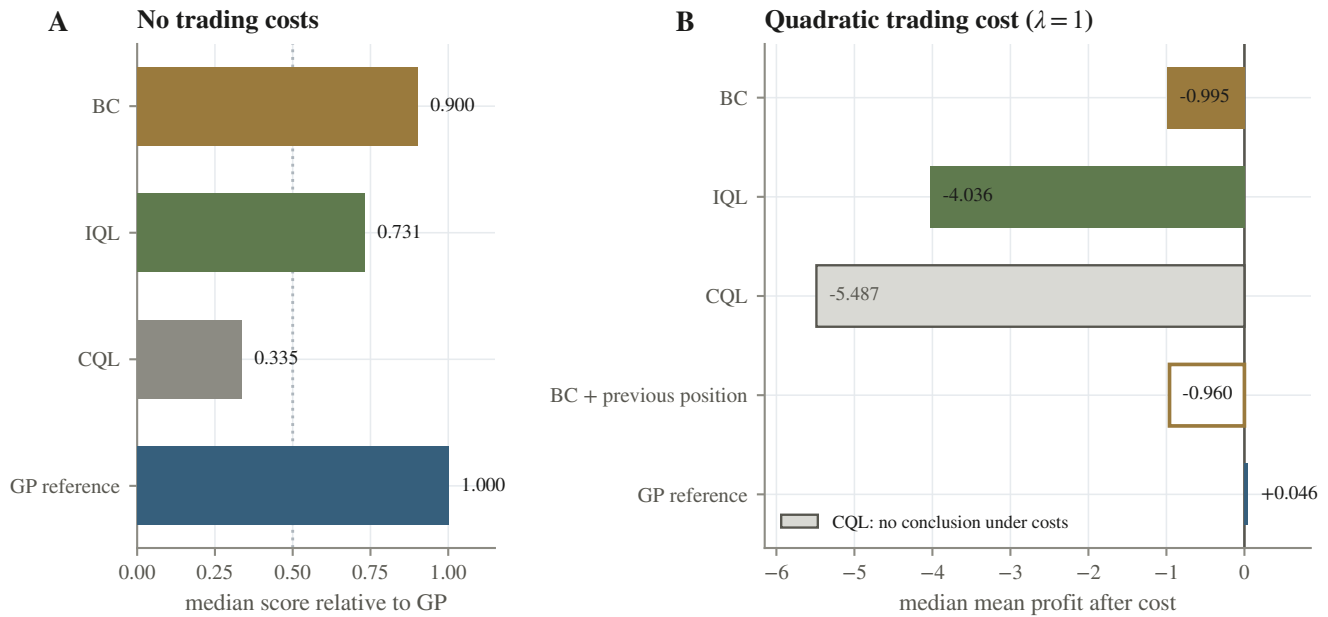


Figure 2: Corrected learning test at $IC = 0.05$; every policy starts flat. Left: median score relative to GP without trading costs; the dotted line is the pre-specified 0.50 minimum. BC and IQL pass, while CQL does not. Right: median mean profit after quadratic costs. BC and IQL are negative for all ten fitted runs; their median turnovers are $6.03\times$ and $11.97\times$ that of GP, respectively. CQL is shown in gray because its failed no-cost check prevents a conclusion under costs. Adding the previous position to BC does not rescue it.

XRP, ADA, LTC, LINK, DOGE, SOL, AVAX, DOT, and ETC. Funding history—periodic cash transfers between long and short holders that help keep a perpetual near its spot price—runs from each contract’s listing through 11 June 2026. The 2019–2023 period is used for training, 2024 for validation, and 2025 onward is a locked holdout period opened once at the end. Six chronological folds, separated by 168-hour gaps, are used inside training and validation. Funding payments were cross-checked numerically against Binance Vision archives on sampled contracts and months.

Four portfolio rules were frozen before evaluation: raw funding carry; a beta-hedged carry that offsets broad market moves using exposure estimated only on training data; 24-hour long-loser/short-winner reversal; and delta-neutral cash-and-carry. Every strategy is compared with cash and an equal-weight buy-and-hold portfolio. Reversal is also compared with the opposite trend ranking; raw carry with 168-hour momentum; and both hedged carry and cash-and-carry with raw carry. All trades use a frozen five-basis-point (0.05%) immediate-execution cost.

The 46,590 training-and-validation timestamps are fewer than the 54,160 required by Eq. 1 for a simple annualized-Sharpe target of 1.0. This is only a minimum-length approximation before the other acceptance checks. Consequently, strategy-level results may be INSUFFICIENT DATA; a weaker edge remains unresolved rather than disproved.

Price signals point to short-horizon reversal. We screen 32 price signals at 1, 8, 24, and 72-hour forecast horizons, for 128 pre-registered pairs. For each pair we compute the cross-sectional

Spearman rank correlation between the signal and the future return. A positive value means that assets ranked highly by the signal subsequently tend to outperform; a negative value means the ranking reverses. A pair is called stable when its 95% confidence interval, adjusted for serial dependence by the Newey–West method [12] with lag equal to the forecast horizon, excludes zero in at least four of six folds.

Sixty-nine pairs meet that rule. All qualifying short-horizon momentum signals have a negative sign: 24-hour momentum predicts the next hour with rank correlation -0.032 in all six folds (median fold $t = -7.79$). The positive relations are slower; inverse volatility reaches $+0.046$ at an eight-hour horizon in all six folds. On this universe and horizon, the evidence supports mean reversion, not “buy recent winners.” This does not contradict longer-horizon crypto momentum on broader universes [11].

A stable correlation need not make a profitable portfolio. We next trade the most extreme quarter of the ranking (three of twelve assets per side): long recent losers and short recent winners. This long-loser/short-winner portfolio loses -1.45 (24-hour rebalance) and -3.02 (8-hour rebalance) in gross price terms over the window, and no fold has a positive lower confidence bound. The mirror-image trend portfolio gains the same amount gross, so the extreme assets trend at the tested 8- and 24-hour rebalancing horizons even though the full-ranking one-hour correlation is negative; we do not identify where in the ranking the reversal sits. The pre-registered prediction of a profitable long-loser/short-winner portfolio is therefore refuted even before transaction costs.

Cross-sectional rank correlation with future returns (stable folds annotated)

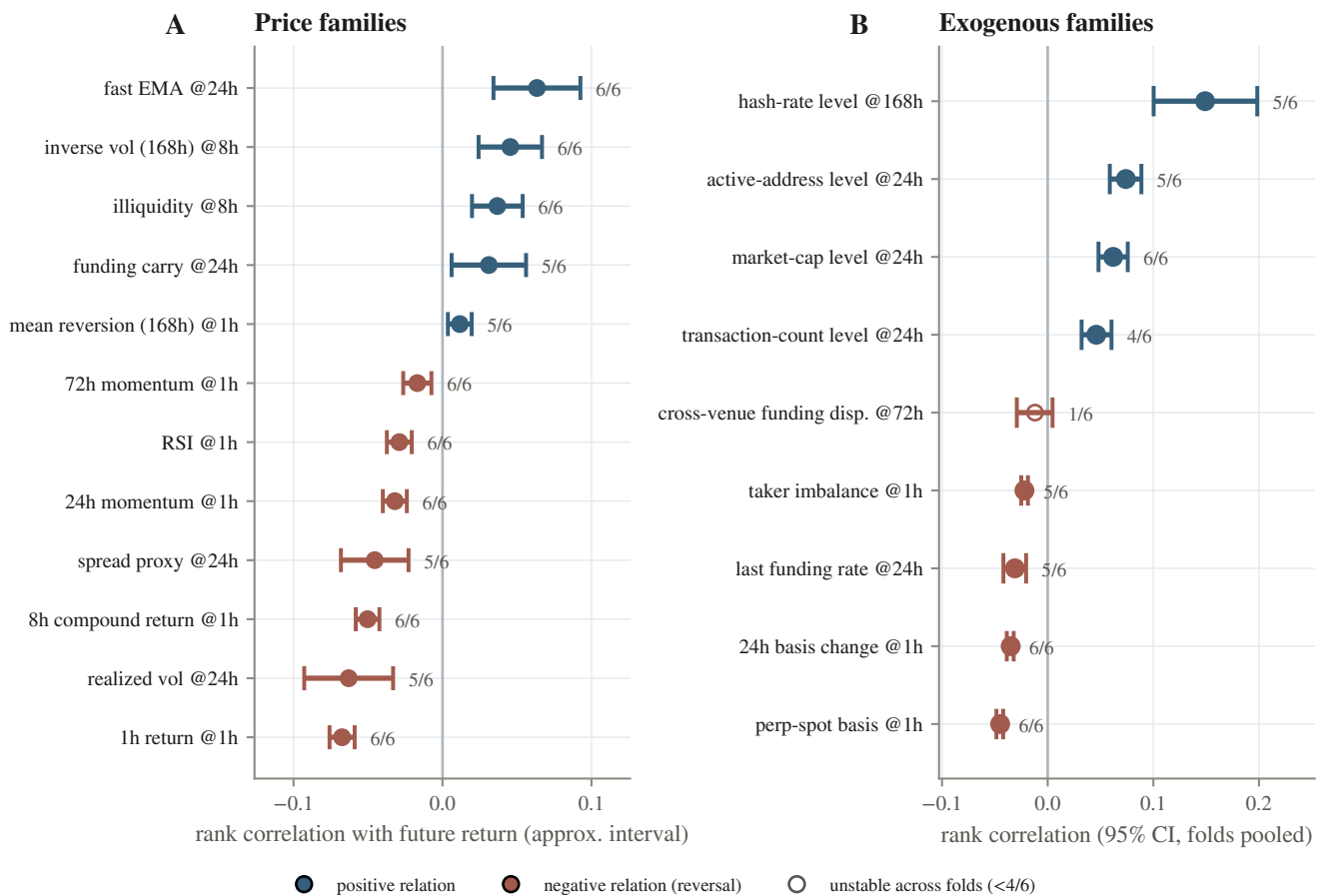


Figure 3: Cross-sectional signal screen on training and validation data. Points show average rank correlation with future returns. Bars are approximate intervals: for the price panel, mean fold correlation $\pm 1.96 |\text{mean}| / |\text{median fold } t|$, a heuristic rather than a 95% confidence interval; for the external panel, normal intervals that combine fold-level Newey–West standard errors treating folds as independent. Labels give the number of folds whose Newey–West interval excludes zero, out of six. The on-chain panel precedes market-cap residualization. *Left*: short-horizon momentum is negative (reversal). *Right*: the largest stable non-price relations are slow on-chain size measures. Open points are unstable in more than two folds.

Funding income is positive but implementation costs dominate. Funding income is positive in every fold, but three pre-registered trading constructions fail to harvest it reliably.

Raw carry holds assets with attractive funding and remains exposed to price moves. No fold has a net-return confidence interval above zero. *Delta-neutral cash-and-carry* shorts the perpetual and buys spot every eight hours. It receives +0.491 in funding, but changes in the price gap between the perpetual and spot contribute -0.023 and two-leg taker fees contribute -0.732, leaving -0.263. Figure 4 varies the fee with trades unchanged: net return falls linearly from +0.468 at zero fee to -0.995 at ten basis points and crosses zero near 3.2 basis points per leg, below the frozen five basis points and below a two-basis-point maker fill (+0.176, one

fold). Even at zero fee only five of six folds have a positive lower confidence bound: the fee sets the sign of the total, not fold-stable evidence. A *beta hedge fitted on training data* also fails: no fold has a positive lower confidence bound, and funding supplies only 0.443 of the portfolio's gross components versus the pre-registered 0.6 requirement.

The conclusion is about implementation, not the existence of the funding premium: funding is positive in all six folds, but the tested portfolios do not isolate and retain it.

External signals mostly reduce to size. We also screen 216 pairs from basis, positioning, on-chain activity, taker flow, options volatility, term structure, and cross-venue dispersion. The largest stable non-price group consists of slow on-chain size measures.

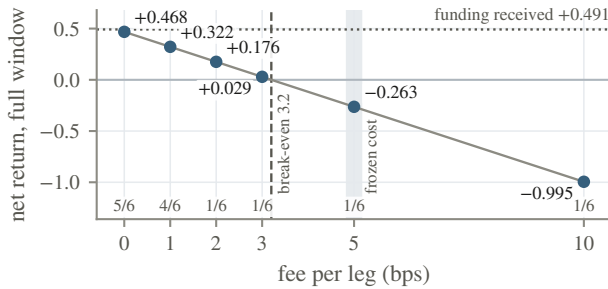


Figure 4: Cash-and-carry net return over the training-and-validation window versus fee per leg (dotted: funding received; band: the frozen five-basis-point assumption). Labels: folds of six with a positive net-return lower bound.

To ask whether this merely restates market capitalization, we remove the linear market-cap component using parameters fixed before the run. Most of the relation disappears. Two training-and-validation pairs remain: hash rate at a 24-hour horizon (rank correlation 0.049, stable in four folds, average return difference +0.0017) and active addresses at 72 hours (0.061, four folds, group-return difference +0.0002). They are reported as leads for later work, not as strategies and not as holdout results.

Together, the real-data study finds no validated, net-of-cost trade among the tested long-loser/short-winner reversal and funding-carry constructions. The two residual on-chain relations remain explicitly provisional.

5 What the evidence supports

We use three outcomes (Table 2). A *supported* prediction passed; a *refuted* prediction was contradicted by the data; an *unresolved* prediction lacks enough evidence. This prevents “we did not detect an effect” from becoming “no effect exists.”

The overfitting calculation is only an implementation check. The only strategies with complete fold scores are two reversal-tail variants over six folds. Splitting those folds into equal train/test groups gives twenty comparisons and a PBO value of 0.00, which is not statistically informative. It shows that the calculation executed; it is not positive evidence that selection bias is absent. The matrix is too small, and the winning variant’s mean per-fold annualized Sharpe is -1.98 . Under the protocol, a missing matrix can block acceptance, but this small matrix cannot certify profitability.

A conventional benchmark illustrates why abstention matters. A pre-registered weekly momentum portfolio has 276 weekly observations, short of the 323 required by its simple power calculation. Its naive mean is $+0.56\%$ per week, and its unadjusted bootstrap interval clears zero by only 6×10^{-5} . A conventional analysis might report a positive momentum result. Our procedure stops at the length check, reports INSUFFICIENT DATA, and records the surprise without tuning another variant; with 353 recorded real-data evaluations behind it, so thin a margin is not evidence.

One controlled look at the final period. After all choices were frozen, we rebuilt the raw-carry, beta-hedged-carry, 24-hour-reversal, and cash-and-carry portfolios on the previously untouched interval from 1 January 2025 to 11 June 2026, containing 12,648 hourly returns. No portfolio’s bootstrap interval lies above zero. Cash-and-carry is entirely below zero (-2.88×10^{-5} per hour), as expected from two-leg fees; the equal-weight average of the four portfolios includes zero. The sample fails the simple mean-test length check before search correction and the remaining checks. The final-period result does not contradict the scoped negative result, but it cannot rule out smaller effects. Exactly one of three permitted final-period evaluations was used.

The claim is therefore narrow and auditable: on this panel, horizon, and cost model, the evaluated reversal and funding-carry portfolio constructions do not produce a validated net-of-cost trade. The two on-chain residuals remain training-and-validation leads.

6 Limitations

Power is not a single end-to-end number. Equation 1 provides a minimum-length approximation for a simple one-sided mean test under fixed assumptions. The implied annualized Sharpe values—1.08 for training and validation and 2.1 for the shorter final period—are descriptive consequences of sample length. They do not define decision thresholds, detection guarantees, or power for the full evaluation procedure. Abstention also does not resolve whether smaller economic effects exist.

The null calibration is finite. Zero of 200 Gaussian null strategies pass, and the tested heavy-tail, volatility-clustered, serially correlated, and market-shifted designs also give zero acceptances. These results validate the implementation on those settings; they do not bound false positives for every possible market process. The selection experiment draws a fixed family of 200 candidates once; adaptive search, where later variants depend on earlier results, is not simulated, and a clean single-use holdout had higher power than the corrected rule.

Market coverage is narrow. Twelve liquid, highly correlated perpetual futures represent the most efficient corner of crypto. Early folds are nearly Bitcoin-only. Conclusions may not transfer to smaller assets, lower frequencies, maker execution, or other market regimes.

GP is a reference, not exact regret. The GP policy optimizes a risk-adjusted objective, whereas our reported score is raw profit net of trading costs. A value relative to GP therefore measures distance from that reference, not regret to the exact raw-profit optimum.

The learning claim is configuration-specific. The corrected test uses ten fitted policies per configuration and 100 independent evaluation paths per fit, always from the same flat starting position. BC and IQL pass the no-cost prerequisite and fail the pre-specified test with costs. CQL fails the prerequisite, so its costly result is descriptive. The test does not imply that RL in general cannot learn cost-aware trading. Larger budgets, online algorithms, or architectures that directly constrain turnover remain outside this paper [1, 8].

Data and cost assumptions matter. The study uses public proxies because liquidation history, broad exchange flows, option skew,

Table 2: Evidence summary. Predictions and stopping rules were fixed before each experiment.

Test	Question	Result	Interpretation
Evaluator calibration	Does the rule reject zero-edge strategies and detect planted edges?	Gaussian design: 0/200 nulls pass; detection is 85% at Sharpe 2.0 and 100% at 3.0.	Calibrated on the tested generators only.
Controlled learning	Can learners use a known signal and survive costs?	No cost: BC/IQL have median scores 0.900/0.731 relative to GP. With costs, 10/10 fits lose; median turnover is $6.03\times/11.97\times$ GP. CQL fails the prerequisite.	BC/IQL fail with costs; CQL is not adjudicated.
Price ranking	Does short-horizon momentum predict returns and its tail portfolio profit?	Rank correlation is -0.032 in 6/6 folds; the tail portfolio loses -1.45 to -3.02 before costs.	Reversal correlation is supported; the pre-registered tail portfolio is refuted.
Funding carry	Can three implementations retain funding after risk and fees?	Funding is positive in 6/6 folds. Cash-and-carry receives $+0.491$, pays -0.732 in costs, and nets -0.263 ; the others lack positive fold evidence.	The premium is observed; tested implementations are unsupported.
On-chain residuals	Do size-related signals survive market-cap removal?	Two pairs remain stable in 4/6 folds, with small group-return differences, only in training and validation.	Provisional leads, not trading claims.
Final period	Do frozen portfolios overturn the earlier negative result?	No interval clears zero; the short sample fails the minimum-length check.	No contradiction; smaller effects remain unresolved.

survivorship-free universes, and verified account-specific fees are not freely available. Five basis points per leg is conservative for a retail taker and determines the carry result (Fig. 4). The fee sweep rescales one linear fee on unchanged trades; nonlinear market impact is not tested.

The overfitting matrix is thin. The PBO matrix contains only two strategies and six folds. Its value of 0.00 is not statistically informative: it shows that the diagnostic executed, not that selection bias is absent. The append-only trial count and deflated Sharpe ratio carry the main search-breadth correction.

7 Conclusion

We test the evaluator before trusting the backtest. In finite synthetic tests, none of 200 Gaussian zero-edge strategies passes, and detection rises with a planted edge. Other tested null generators also give zero acceptances; untested market processes remain outside this calibration.

In the controlled market, BC and IQL learn the frictionless signal, but all 10/10 fits lose under quadratic costs; their median turnovers are $6.03\times$ and $11.97\times$ the GP reference. CQL fails the no-cost prerequisite, so its costly case is not adjudicated. These failures are configuration-specific, not a general statement about RL.

On twelve liquid crypto perpetuals, no tested reversal or carry implementation produces a validated net-of-cost trade. Funding is positive in every fold, but the portfolios do not retain it after price risk and retail taker fees. Two on-chain relations remain provisional. The single planned final-period evaluation does not contradict the result, but is too short for stronger claims about small effects.

The reusable contribution is the workflow: minimum sample size, a count of every attempted variant, declared baselines, acceptance checks calibrated on known nulls and planted edges, and rationed final-period access.

References

- [1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. 2021. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34. Curran Associates, Inc., Virtual, 29304–29320.
- [2] David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. 2014. Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance. *Notices of the American Mathematical Society* 61, 5 (2014), 458–471.
- [3] David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. 2017. The Probability of Backtest Overfitting. *Journal of Computational Finance* 20, 4 (2017), 39–69.
- [4] David H. Bailey and Marcos López de Prado. 2014. The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality. *The Journal of Portfolio Management* 40, 5 (2014), 94–107.
- [5] Nicolae Gârleanu and Lasse Heje Pedersen. 2013. Dynamic Trading with Predictable Returns and Transaction Costs. *The Journal of Finance* 68, 6 (2013), 2309–2340.
- [6] Peter Reinhard Hansen. 2005. A Test for Superior Predictive Ability. *Journal of Business & Economic Statistics* 23, 4 (2005), 365–380.
- [7] Campbell R. Harvey and Yan Liu. 2015. Backtesting. *The Journal of Portfolio Management* 42, 1 (2015), 13–28.
- [8] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep Reinforcement Learning That Matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Press, New Orleans, Louisiana, USA, 3207–3214.
- [9] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2022. Offline Reinforcement Learning with Implicit Q-Learning. In *International Conference on Learning Representations (ICLR)*.
- [10] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 1179–1191.
- [11] Yukun Liu, Aleh Tsyvinski, and Xi Wu. 2022. Common Risk Factors in Cryptocurrency. *The Journal of Finance* 77, 2 (2022), 1133–1177.
- [12] Whitney K. Newey and Kenneth D. West. 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55, 3 (1987), 703–708. doi:10.2307/1913610
- [13] Dimitris N. Politis and Joseph P. Romano. 1994. The Stationary Bootstrap. *J. Amer. Statist. Assoc.* 89, 428 (1994), 1303–1313.
- [14] Maik Schmeling, Andreas Schrimpf, and Karamfil Todorov. 2023. *Crypto Carry*. BIS Working Papers 1087. Bank for International Settlements.