



Catch Alpha: Testing the Evaluator Before Testing the Trading Strategy

An evaluation rule calibrated on synthetic nulls and planted edges before it is applied to market data.

Joris Viaud · Independent Researcher, France · ORCID 0009-0000-5757-269X · joris.viaud@edu.ece.fr

ICAIF '26
Milan, 16–17 Nov 2026
Paper 121

Summary

Anyone who tries enough trading ideas will find one that looks great on past data, even if none of them works. We built a judge that counts every attempt and refuses to rule when the data are too short, and we tested the judge first: on simulated markets with no real edge it almost never accepted one, and it detected planted edges once they were large enough. It then judged four real strategies on twelve cryptocurrencies: the data show predictive patterns, but no tested portfolio established a reliable net profit after fees under the protocol.

Conclusions

1. Calibrate the decision rule on known nulls before touching market data.
2. Count every attempt; here the trial count is the safeguard that matters after selection.
3. Abstain when the sample is too short: ≥ 6.2 years of hourly data for a Sharpe of 1.
4. Predictability $\neq$ net profit: reversal and funding are real; the tested portfolios do not net a validated profit.

Motivation: search-inflated backtests

- **Search:** the best of 200 variants looks good even if none of them works.
- **Short samples:** annualizing scales noise as much as signal ($\times 16$ daily, $\times 94$ hourly).
- **Late corrections:** correcting only the final strategy ignores earlier choices.

Origin: ~ 96 runs on 60 days of 15-min bars; Sharpe from -17 to $+17$.

Evaluation protocol: four checks and one diagnostic

1 Enough data?

Minimum length, one-sided mean test at $\alpha = 0.05$, power 0.80 ($P = \text{periods/yr}$, $S_{\min} = \text{target annualized Sharpe}$):

$$n \geq 6.18 \frac{P}{S_{\min}^2}$$

Hourly, $S_{\min} = 1 \Rightarrow 54,160 \text{ obs} \approx 6.2 \text{ years}$. Our panel: 46,590 $\Rightarrow$ INSUFFICIENT DATA: no Sharpe-based verdict.

2 Excess over benchmark

Stationary-bootstrap 95% confidence intervals (CIs) on mean and Sharpe of (candidate – benchmark): both lower bounds > 0 . Blocks are resampled to preserve serial dependence; 400 replications (150 in the large calibration sweep).

3 Deflated Sharpe ratio (DSR)

Sharpe vs. the expected maximum of N zero-edge trials:

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

$$\text{pass if } \alpha \left(\frac{SR - SR_0}{\sigma_{SR}} \right) \geq 0.95$$

$V \approx$ the candidate's own Sharpe variance; $\gamma \approx 0.5772$; $\sigma_{SR} =$ standard error of the Sharpe.

N comes from an append-only trial ledger counting every screen: real-data ledger: 353 recorded evaluations (352 selection evaluations + the single final-period run); 344 signal screens + 9 portfolio evaluations.

4 SPA (superior predictive ability) test

vs. every declared baseline, worst $p \leq 0.10$. Baselines declared before the run: cash, equal-weight buy-and-hold, and one strategy-specific comparator (e.g. the trend mirror for reversal, 168h momentum for raw carry).

5 PBO (probability of backtest overfitting)

Bailey et al. Build the matrix of per-fold scores (strategies $\times$ 6 folds); split the folds into every train/test half: $\binom{6}{3} = 20$ combinations. In each, take the strategy best on the train folds and record its relative rank ω on the test folds; $\ln \frac{\omega}{1-\omega} < 0$ means the in-sample winner fell below the out-of-sample median. PBO = share of the 20 combinations with a negative logit. Gate: PBO ≤ 0.50 ; a missing matrix blocks acceptance. Here: only two reversal variants have complete fold scores (2×6 matrix). **PBO = 0.00**; the in-sample winner's median out-of-sample rank is 0.67. Not statistically informative: it shows the diagnostic executed, not that selection bias is absent.

Procedure

- **Append-only trial ledger:** every signal screen and every gated evaluation is recorded; the ledger is the only source of N .
- Six chronological folds with 168-hour gaps at every boundary; a claim must repeat across folds.
- Final-period access budgeted to 3 looks; **1 used**, on frozen portfolios only.

Same Sharpe (0.06/period, 3,000 obs): DSR 0.99 if you count 3 trials, 0.63 if you count 347.

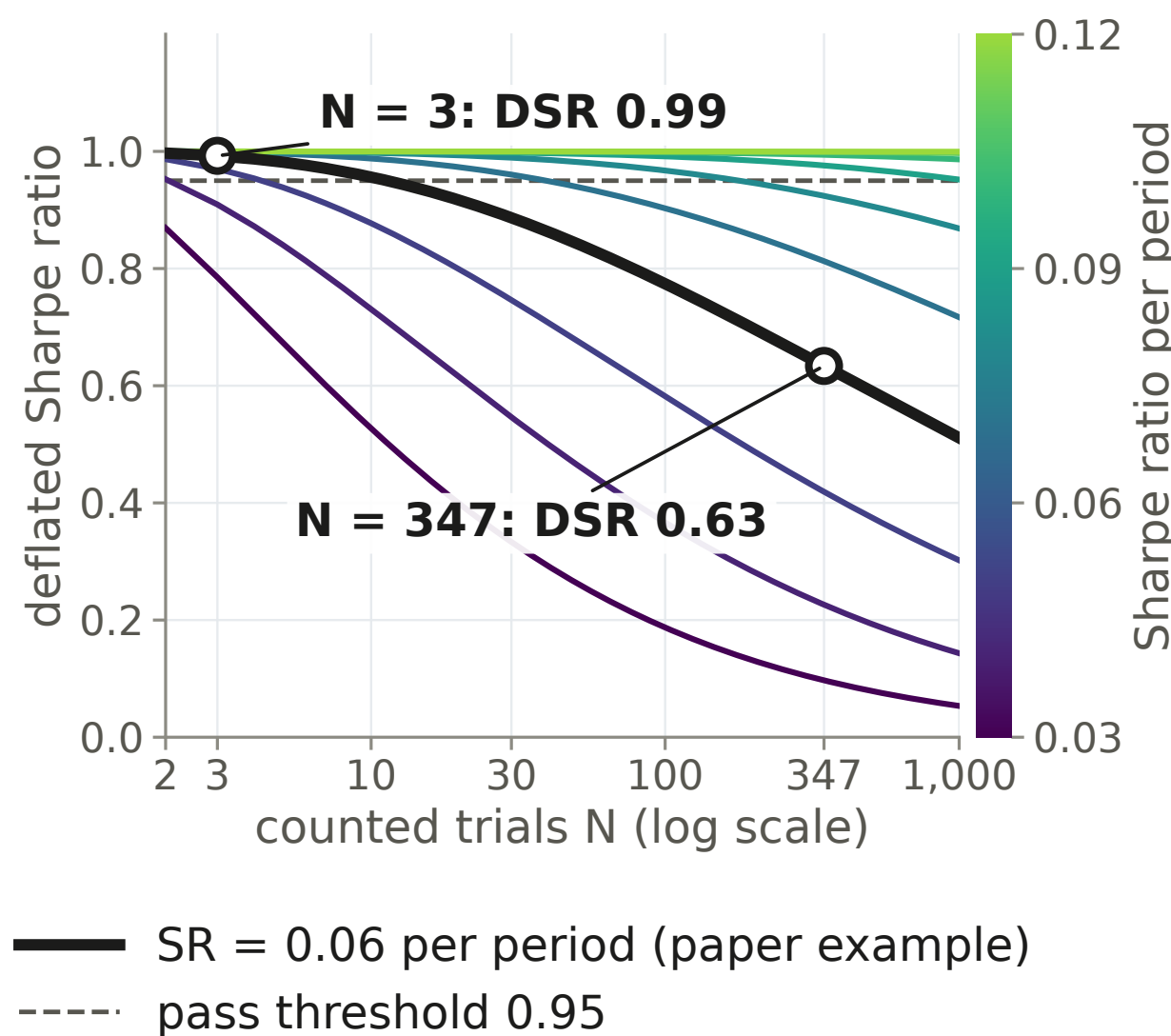


Fig. 1 – Same Sharpe, different trial count: DSR 0.99 at $N = 3$, 0.63 at $N = 347$ (347 = ledger count at the time of the paper's example; frozen count 353; 3,000 obs; computed from the paper's DSR formula).

References. Bailey & López de Prado 2014, deflated Sharpe ratio, *J. Portf. Manag.*; Bailey, Borwein, López de Prado & Zhu 2017, probability of backtest overfitting, *J. Comput. Finance*; Hansen 2005, superior predictive ability, *JBES*; Politis & Romano 1994, stationary bootstrap, *JASA*; Gârleanu & Pedersen 2013, dynamic trading with predictable returns and transaction costs, *J. Finance*.

Calibration on synthetic data

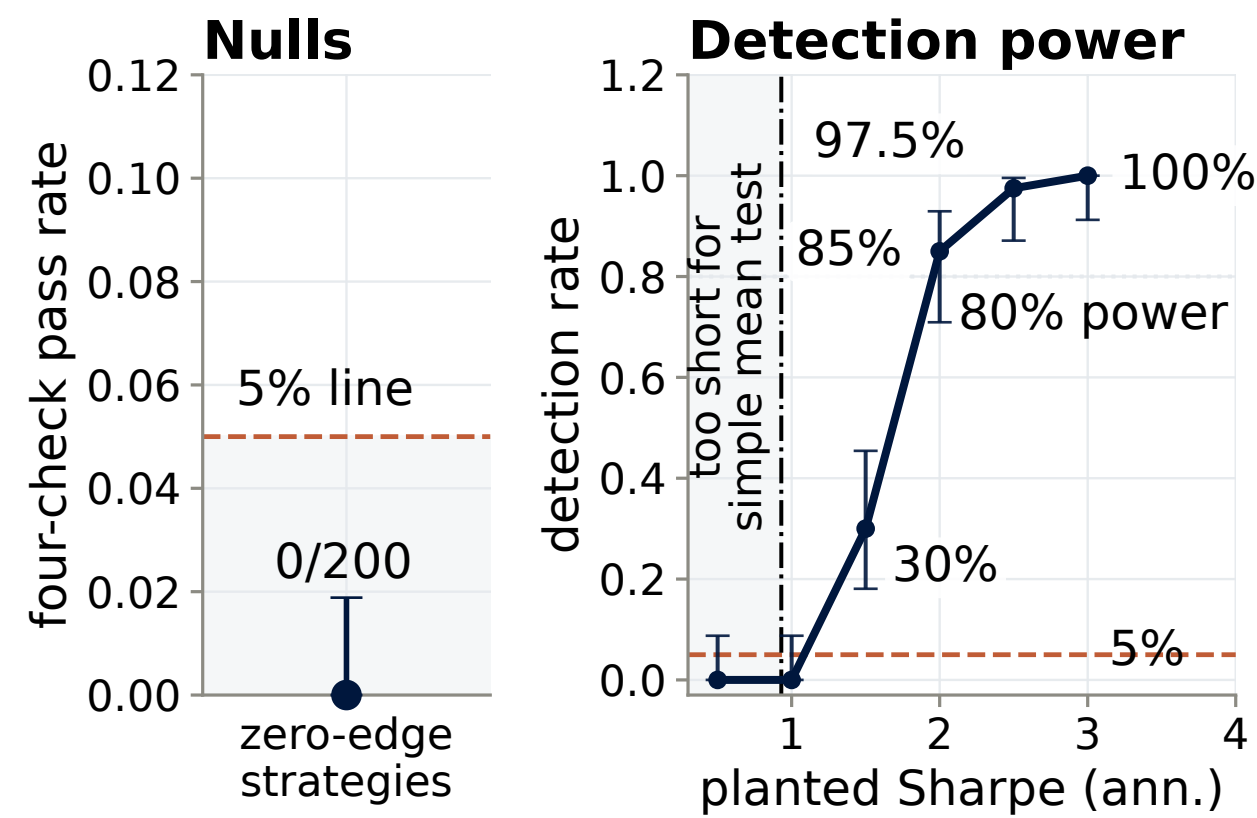


Fig. 2 – 0/200 zero-edge strategies accepted; detection rises with the planted edge. DSR counts $N = 200$ attempts; 40 draws per planted edge (paper Fig. 1).

- 200 zero-edge Gaussian strategies, 1,800 obs, $N = 200$: **0/200 accepted** (Wilson upper bound 1.9%). Calibrated on the tested generators only.
- Gaussian, Student- t_3 , GARCH(1,1), AR(1) at 1,800 / 5,000 / 50,000 obs: **0/2,400** (1.9% upper bound per setting). Market-derived null (real signal shifted ≥ 26 weeks): 0/200.
- Planted annualized Sharpe 0.5 $\rightarrow$ 3.0, 40 draws each: detection **0 / 0 / 30 / 85 / 97.5 / 100%**.

Selection-aware calibration and ablation

3,300 campaigns; 200 equicorrelated candidates; winner = best development Sharpe. Rates in %; detection at planted $S = 1.5$.

Rule	Null false accept. Detection			
	$\rho=0$	0.3	0.7	$\rho=0$
On selection data				
Uncorrected t-test	100.0	83.6	48.2	86.5
Four checks, $N=200$	0.2	0.0	0.0	35.0
Four checks, $N=1$ (clamped to 2)	36.4	20.6	8.0	74.5
DSR alone, $N=200$	0.2	0.0	0.0	35.0
On an untouched holdout				
Four checks, $N=200$	0.0	0.0	0.0	24.5
t-test alone	5.0	5.6	7.6	86.0

The trial count provides the control: DSR alone $\approx$ four checks; drop N and **8–36%** of null winners pass. A clean single-use holdout is near nominal (5.0–7.6%) with more power: the correction protects selection on reused data; on a truly independent, single-use holdout a plain test is more powerful here.

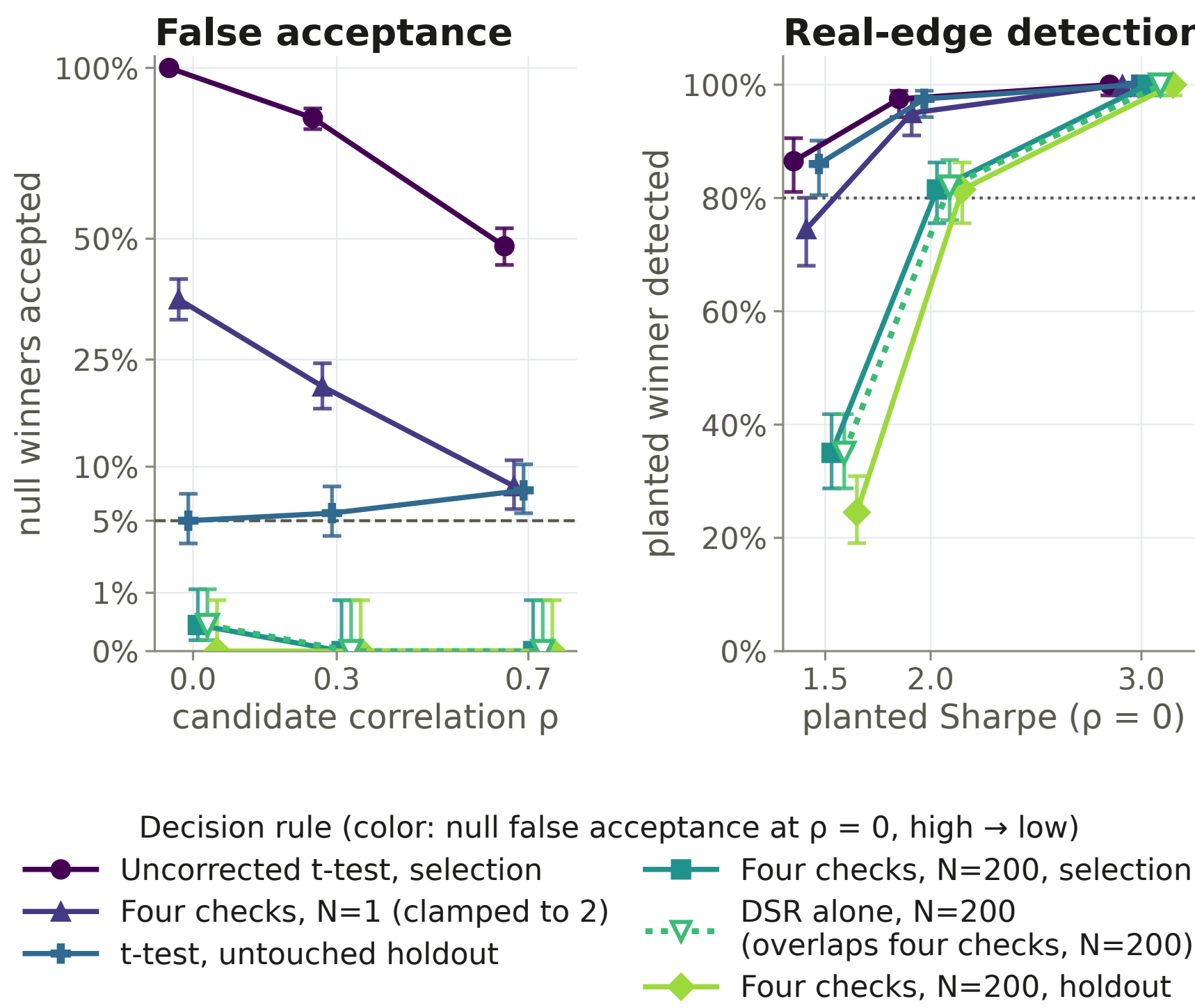


Fig. 3 – After selecting the best of 200 correlated candidates; 3,300 campaigns, Wilson 95% CIs, square-root vertical scale in the null panel (computed from the paper's committed data).

Scope of the evidence

- **Supported:** on the tested generators the four checks reject zero-edge winners (1 of 1,500 accepted after selecting the best of 200 correlated candidates); the trial count carries that control.
- **Not shown:** a false-positive bound for untested market processes, or that any adaptive search is covered. The simulated family is fixed.
- **Not shown:** a profitable strategy; the learning result is configuration-specific and the real-data result is conditional on this panel, horizon and 5 bps cost.

Learning test: offline RL on a planted signal

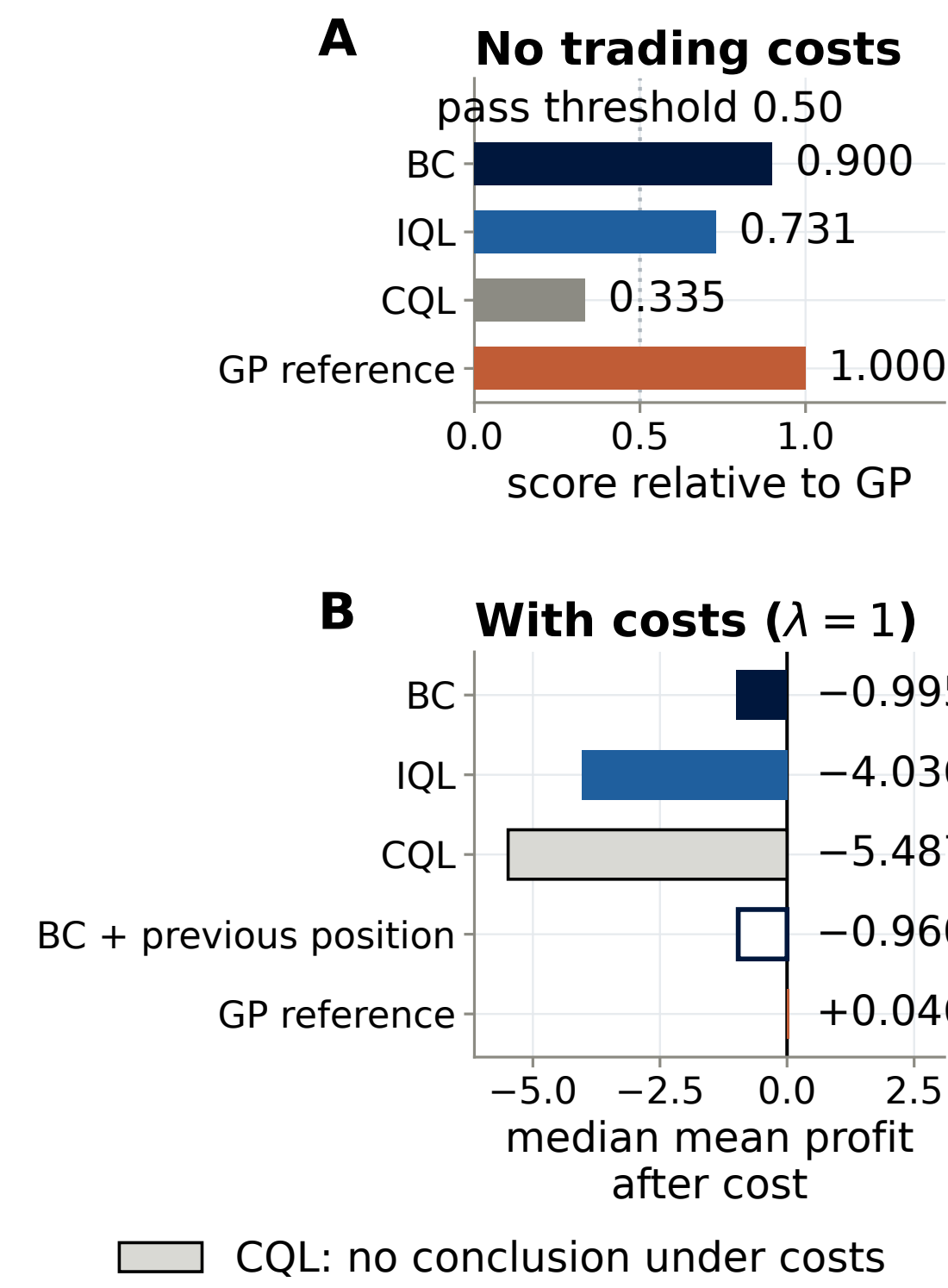


Fig. 4 – BC and IQL learn the signal but lose under quadratic costs in this configuration (paper Fig. 2).

Market (per asset): $s_{t+1} = \phi s_t + \eta_{t+1}$, $r_{t+1} = b s_t + \epsilon_{t+1}$ with $\phi = 0.9$, $\eta \sim N(0, 1 - \phi^2)$, $b = 0.0005$, $\sigma_\epsilon = 0.01\sqrt{1 - 0.05^2}$, so IC = 0.05.

$$\text{Score: } S(\pi) = \sum_t [w_t^\top r_{t+1} - \frac{\lambda}{2} \|w_t - w_{t-1}\|^2], \quad Q(\pi) = \frac{S(\pi)}{S(\text{GP})}$$

Q is the relative score (threshold 0.50).

GP rule (partial adjustment): $w_t = w_{t-1} + A(\text{aim}_t - w_{t-1})$, $\text{aim}_t = m(\phi) \frac{b s_t}{g}$

with $g = \gamma_r \sigma_\epsilon^2$, risk aversion $\gamma_r = 10$. $\lambda = 0$: $A = 1$, $m = 1$ (pure Markowitz, aim = $0.50 s_t$). $\lambda = 1$: $A = 0.031$, $m = 0.243$, aim = $0.12 s_t$; **GP moves 3% of the way per period**; BC/IQL turnover is $6.03\times / 11.97\times$ that of GP.

GP optimises a risk-adjusted objective; S is raw profit net of costs, so Q is distance to a reference, not exact regret.

Real-data study: twelve Binance perpetuals

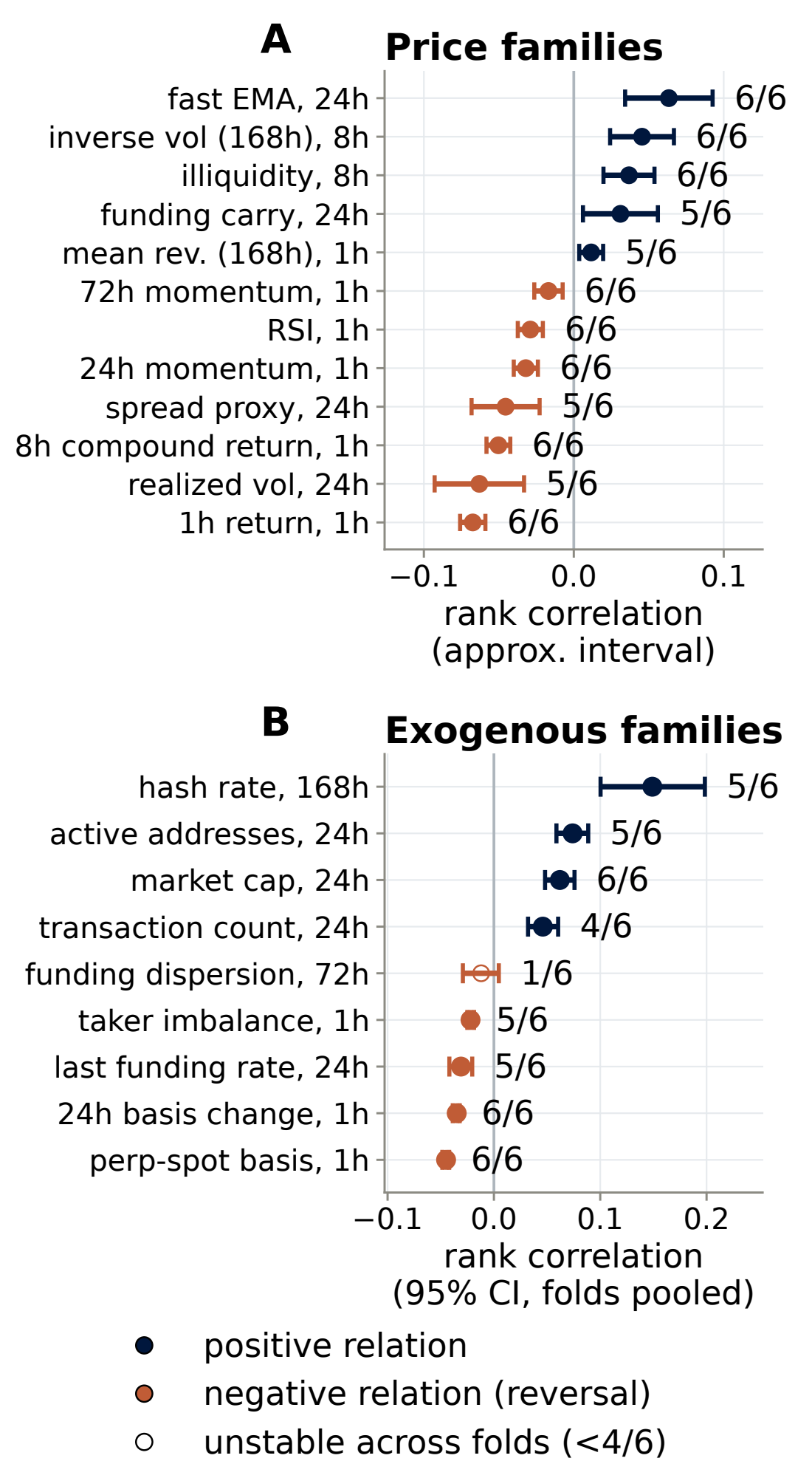


Fig. 5 – Rank correlation with future returns (paper Fig. 3). Price-panel bars are mean fold IC $\pm 1.96 \cdot |\text{mean IC}| / |\text{median fold t}|$, an approximation rather than a 95% CI; external-panel bars combine fold-level Newey–West standard errors treating folds as independent; $n/6 =$ folds whose Newey–West interval excludes zero; on-chain panel shown before market-cap removal.

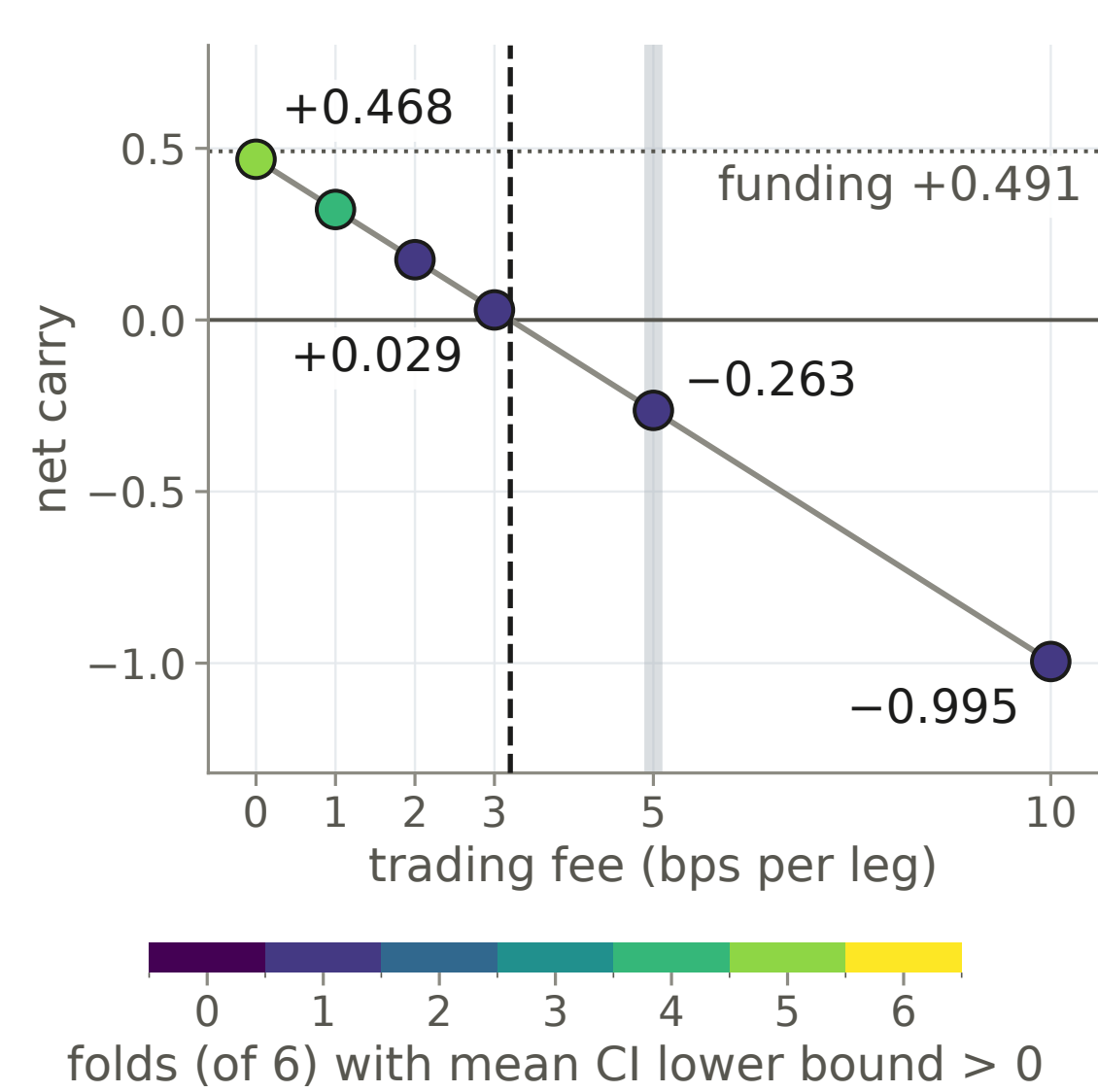


Fig. 6 – Cash-and-carry net return vs fee per leg; funding +0.491 is lost above about 3.2 bps/leg (computed from the paper's committed data).

5 assets, IC = 0.05 (information coefficient: corr(signal, next return)); every policy starts flat; 10 seeds $\times$ 100 paths.

- No costs, median Q (score $\div$ GP score): BC 0.900, IQL 0.731, CQL 0.335 (threshold 0.50).
- $\lambda = 1$ (trading-cost weight): BC -0.995 , IQL -4.036 , **negative in 10/10 seeds**; turnover $6.03\times / 11.97\times$ that of GP; GP itself $+0.046$.
- BC with the previous position as input: -0.960 .
- Learning the signal does not make it tradable after costs.

BC = behaviour cloning, IQL = implicit Q-learning, CQL = conservative Q-learning; GP = Gârleanu–Pedersen analytic reference.

- 128 price signal–horizon pairs; 69 fold-stable.
- 24h momentum $\rightarrow$ next hour: rank corr. **-0.032 in 6/6 folds** (median fold $t = -7.79$): short-horizon reversal.
- Tail portfolio (long losers, short winners): **-1.45 (24h) / -3.02 (8h)** before costs; the mirror trend book gains the same gross, so the tails trend at the tested horizons.
- 216 external pairs mostly reduce to size; after market-cap removal two small on-chain leads remain (hash rate 24h: 0.049; active addresses 72h: 0.061). These are reported as leads, not strategies.
- Final period (Jan 2025–Jun 2026, 12,648 h), one look: no interval above zero; the sample is too short.

- Funding positive in 6/6 folds; cash-and-carry receives $+0.491$, pays -0.023 basis and -0.732 fees: **nets -0.263** .
- Break-even $\approx 3.2 \text{ bps/leg}$ (frozen cost 5 bps).