

# Pre-registered trend following

What an honest backtest looks like, including the part where it failed

0.71

net Sharpe, 11 years out of sample

[0.22, 1.19]

90% bootstrap interval

1 of 2

pre-registered tests passed; the first failed

Live

forward paper trading since 3 Sep 2026

**Author:** Liam Rieber, New York University · Code, pre-registrations and every result file: [github.com/b5gf2nmzff-maker/tsmom-preregistered](https://github.com/b5gf2nmzff-maker/tsmom-preregistered) · Live paper track record: [b5gf2nmzff-maker.github.io/tsmom-preregistered](https://b5gf2nmzff-maker.github.io/tsmom-preregistered)

## 1. The problem with backtests

Most strategy backtests are not tests. They are searches. The researcher tries lookbacks, filters and position rules until the equity curve looks right, then reports the winner as if it had been the plan all along. Every choice made after seeing the data is a fork, and with enough forks a random signal can be made to look like a Sharpe ratio of 1. The academic name is the garden of forking paths; the practical result is strategies that stop working the day they go live.

The fix is the same one medicine adopted for clinical trials: write down the hypothesis, the data you will use, and the pass criteria before you run the test, then report the result whatever it says. I did that for the simplest strategy with a credible academic record, time-series momentum in futures, and this note is the record. It includes a failed test, because a pre-registration that only reports passes is not one.

## 2. The rule

**Universe.** Ten CME micro futures: S&P 500, Nasdaq 100, Dow, Russell 2000, gold, silver, crude oil, copper, euro, 10-year Treasury. Micros were chosen because the intended account was small.

**Signal.** Sign of the average of 126- and 252-day returns for each market. Long if positive, short if negative. No filters.

**Sizing.** Each market is scaled to a 10% annualized volatility target using a 64-day exponentially weighted volatility estimate, with an instrument-diversification multiplier of 2 across the ten markets, then rounded to whole contracts.

**Costs.** \$2.50 commission plus one tick of slippage per side, and a round-trip charge at every quarterly roll. Nothing is assumed to fill at mid.

**Data.** Two independent series: continuous futures from 2000 and ETF proxies (SPY, GLD, USO and so on) rescaled to futures price levels so that contract sizes and ticks apply. Agreement between the two is itself a check.

**Rebalance.** Daily signal, positions changed only when the rounded target changes.

## 3. The protocol

| Block | Dates | Purpose and rules |
|-------|-------|-------------------|
|-------|-------|-------------------|

|            |               |                                                                                                                                                                                                                                                                                        |
|------------|---------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Discovery  | 2007-2016     | The only block on which parameters may be chosen. Fixed grid of 18 combinations (three lookback sets, two volatility spans, three vol targets). Best Sharpe wins.                                                                                                                      |
| Validation | 2017-2022     | One run of the selected configuration. Six pass criteria fixed in advance: Sharpe above 0.5; 90% weekly block-bootstrap lower bound above zero; profit factor above 1.1; max drawdown under 25%; costs under 35% of gross; beats a random-sign baseline with the same holding periods. |
| Holdout    | 2023-Sep 2026 | Locked. The runner refuses to open it without a recorded PASS, and refuses to open it twice. Reported regardless of sign.                                                                                                                                                              |

The pre-registration documents are timestamped markdown files in the repository. Each has an amendment section recording the result after the fact, so a reader can see what was promised and what happened.

## 4. What happened, in order

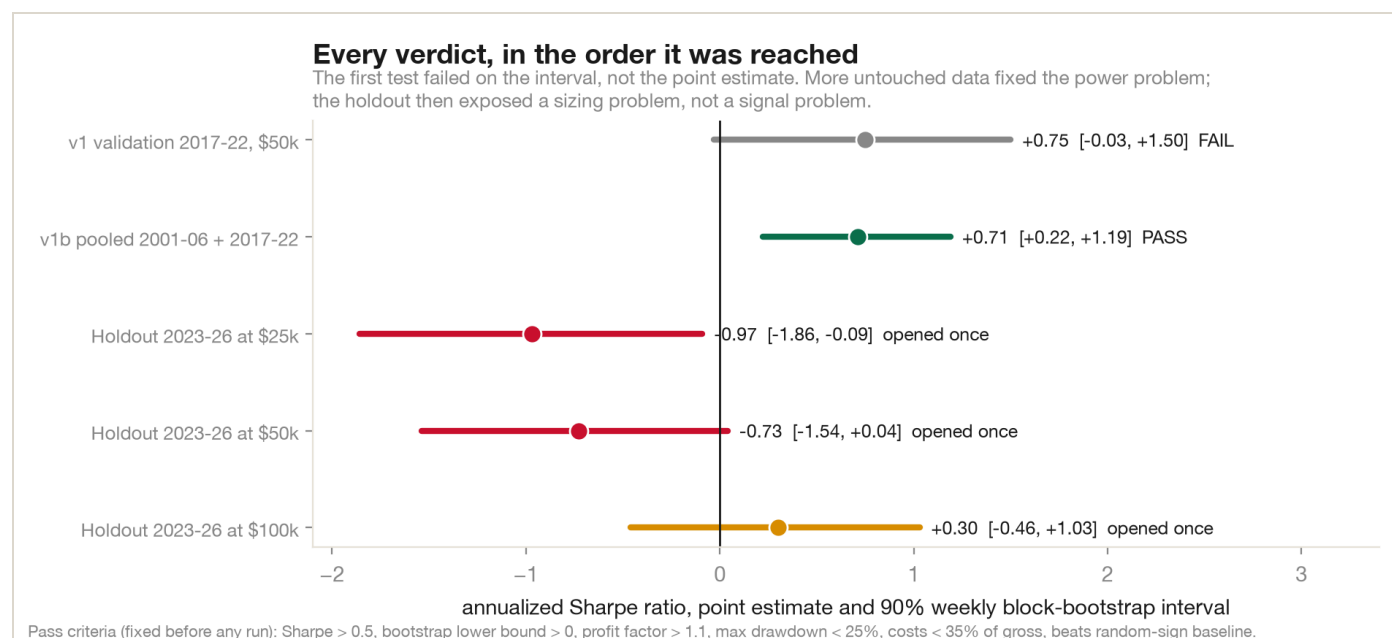


Figure 1. Every verdict reached, with 90% bootstrap intervals. Green passed, grey failed, red and amber are the holdout at different account sizes.

### 4.1 Discovery selected a slow, conventional configuration

Lookbacks 126 and 252 days, volatility span 64, 10% target. Discovery Sharpe 0.58 at \$50,000. Nothing exotic survived: the four-lookback set and the faster volatility estimate both lost on costs.

### 4.2 The first validation failed

On 2017-2022 the rule made a net Sharpe of 0.75 with 5.6% a year, costs at 3% of gross, a 20% drawdown, and beat the random baseline (-0.24) comfortably. Five of six criteria passed. The bootstrap interval was [-0.03, 1.50]: six years of a 0.75-Sharpe strategy is simply not enough data to exclude zero at 90% confidence. By the rules I had written, that is a FAIL, and it was recorded as one.

The temptation at this point is to change something. Loosen the interval to 80%. Drop the worst market. Try the 63-day lookback "just to see". Each of those is a fork, and any of them would have produced a pass that meant nothing.

### 4.3 A second pre-registration, with no new choices, passed

A power failure has one legitimate remedy: more data that nobody has looked at. The futures series begin in 2000, and discovery had started in 2007, so 2001-2006 was untouched. I pre-registered a second test (v1b) before running it: the identical configuration on the early block, pooled with the 2017-2022 block, same six criteria, no grid. The early block alone made a Sharpe of 0.65. Pooled across eleven years: Sharpe 0.71, interval [0.22, 1.19], profit factor 1.15, max drawdown 19%, costs 4% of gross, random baseline -0.28. All six criteria passed.

## Eleven years of out-of-sample trend following, net of costs

Ten micro futures, lookbacks 126/252d, 10% vol target, \$50k account; pooled Sharpe 0.71, 90% bootstrap interval [0.22, 1.19], costs 4% of gross

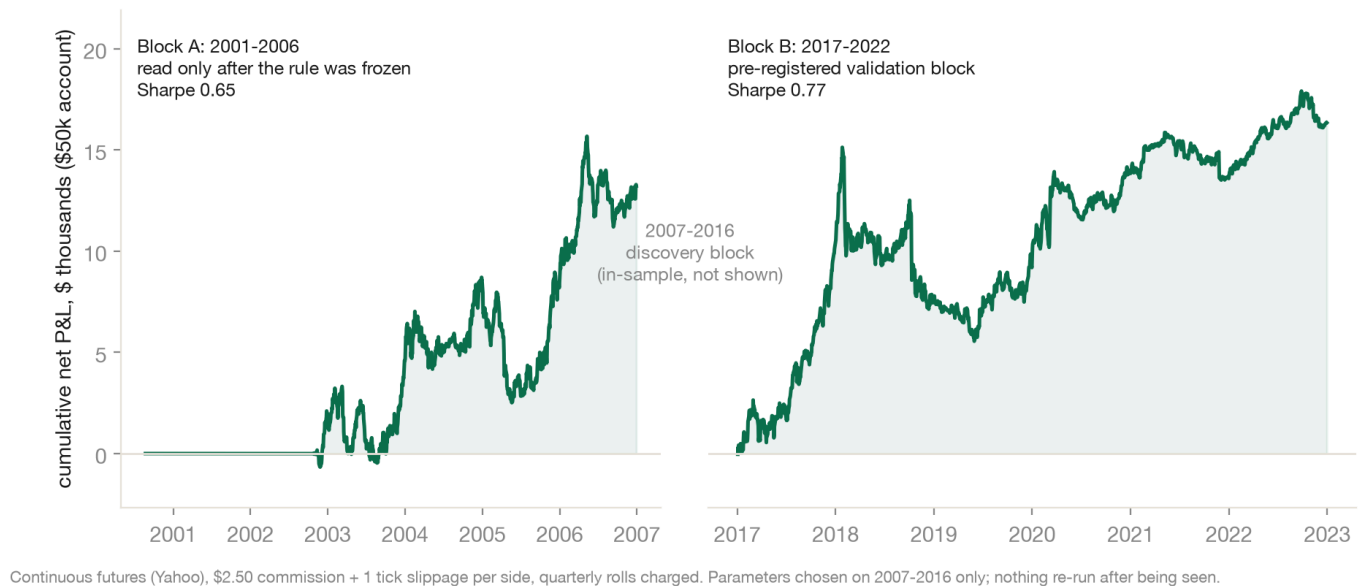


Figure 2. The two out-of-sample blocks that make up the eleven-year verdict. The 2007-2016 discovery block between them is deliberately not shown.

## 5. The holdout: a sizing problem, not a signal problem

A PASS licensed one opening of the 2023-2026 holdout. I opened it for each account size under consideration and for both data series. At \$25,000 the rule lost money (Sharpe -0.97). At \$50,000 it lost money on both series (-0.73). At \$100,000 it made a modest profit (+0.30, 3 of 4 years positive). At \$250,000, run afterwards for the chart, +0.64.

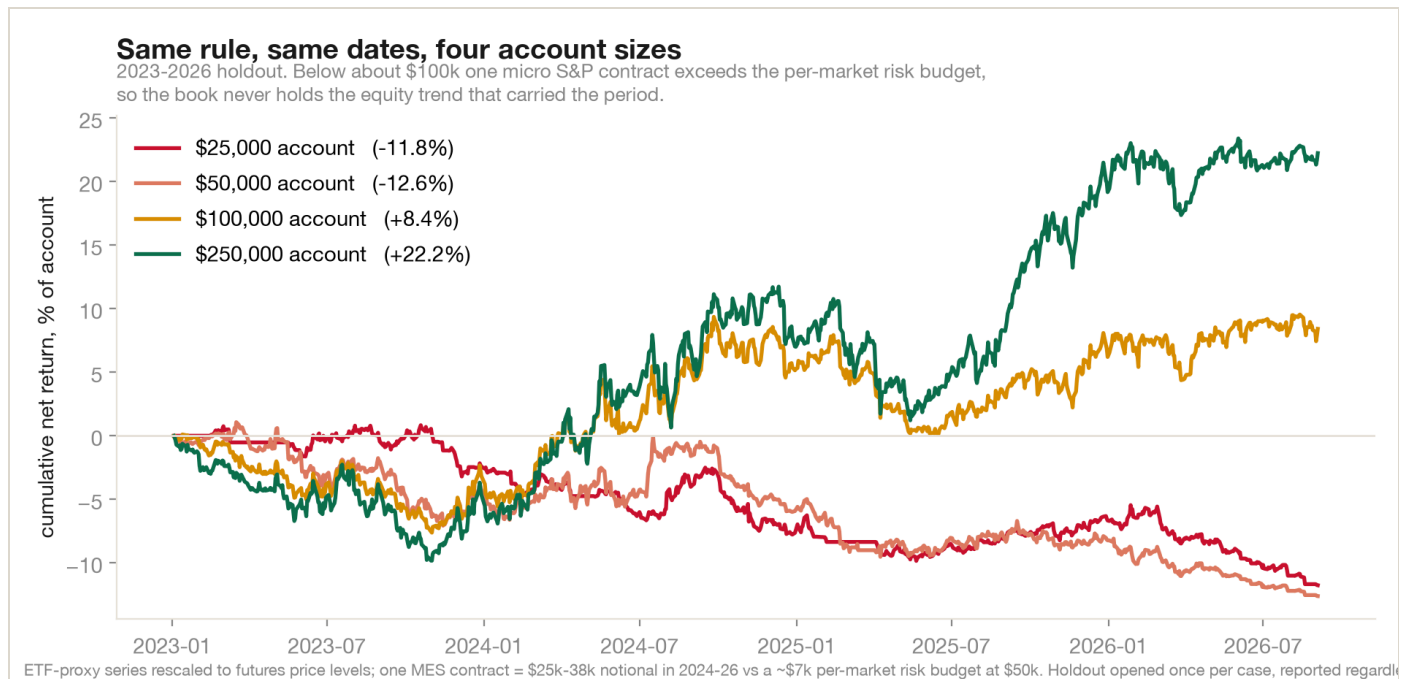


Figure 3. Identical rule and dates; only the account size changes.

The per-market breakdown explains it. In 2024-2026 one micro S&P contract represents \$25,000 to \$38,000 of notional. At a \$50,000 account with a 10% volatility target, the per-market risk budget is about \$7,000 of notional. Rounding sends the target position to zero, so the book never held the S&P 500 or Nasdaq, the two strongest trends of the period, and was left concentrated in Treasuries, the euro and crude, three markets that chopped. The same rule held SPY for 583 days during 2017-2022, when the contract was half the size.

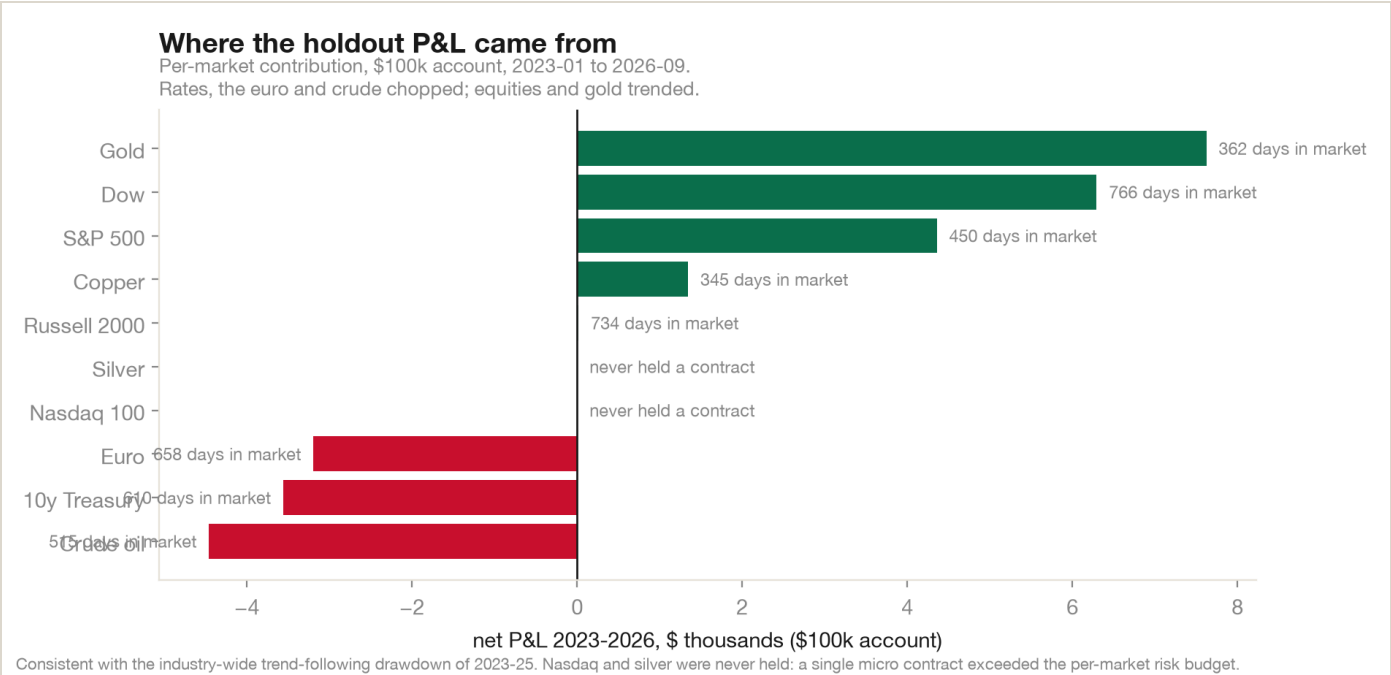


Figure 4. Where the \$100,000 holdout P&L came from. Nasdaq and silver were never held even at this size.

This is consistent with the well-documented 2023-2025 drawdown in trend following industry-wide, but it is also a design flaw in the original sizing: risk budget was divided across a fixed ten markets, so when four of them could not hold a contract that risk simply went undeployed. A holdout is meant to find exactly this kind of thing. It did.

## 6. What the results licensed, and what is running now

**Decision.** The edge is real in aggregate (eleven years out of sample, two independent data series, interval excludes zero) but cannot be expressed in micro contracts below roughly \$100,000. The holdout has been opened and cannot serve any future version, so the only honest out-of-sample test left is forward paper trading.

**Live since 3 September 2026.** The exact v1b configuration runs against a Webull paper futures account sized at \$100,000, on a Hetzner VPS: a rebalance job at 15:45 New York time each weekday, a watchdog every ten minutes that reconciles broker positions against the ledger and halts on drawdown or heartbeat limits, and a SQLite ledger that is the single source of truth. Equity, positions and every fill are published to the public track-record page automatically. No parameter can be changed without a new pre-registration.

| Metric        | Discovery 2007-16 (in-sample) | Validation 2017-22     | Early block 2001-06 | Pooled OOS (11y)  | Holdout 2023-26 @ \$100k |
|---------------|-------------------------------|------------------------|---------------------|-------------------|--------------------------|
| Net Sharpe    | 0.58                          | 0.75 (futures 0.77)    | 0.65                | 0.71              | 0.30                     |
| 90% interval  | n/a                           | [-0.03, 1.50]          | n/a                 | [0.22, 1.19]      | [-0.46, 1.03]            |
| Max drawdown  | -                             | 20.2%                  | -                   | 19.2%             | 9.2%                     |
| Costs / gross | -                             | 3%                     | -                   | 4%                | 2%                       |
| Verdict       | select                        | <b>FAIL (interval)</b> | -                   | <b>PASS (6/6)</b> | opened once, reported    |

## 7. What I learned

- **Write the criteria first or you will negotiate with yourself.** The moment the first test failed on one criterion, I could feel the arguments for relaxing it forming. Having the rule written down in a file with a timestamp ended the conversation.
- **A failed test is information, not embarrassment.** The failure was about power, and the correct response (more untouched data) was only available because I had not touched it.
- **Holdouts find implementation problems, not just signal problems.** The contract-granularity failure would have been invisible in an ETF backtest and expensive in a live account.
- **Two data sources are worth more than two more parameters.** The proxy and futures series agreeing within 0.03 of Sharpe did more for my confidence than any single interval.
- **A second intraday strategy was pre-registered the same day and failed outright** (gross edge of 0.3 S&P points per session against \$7.50 of round-trip cost). It is in the repository too. Reporting it costs nothing and is the point.

Everything in this note can be reproduced from the repository: `python -m quant.research.run_tsmom discover|validate|validate-early|open-holdout --series futures|proxy --capital N`. Result files under

`reports/` are the originals written by those runs and were not edited. Data: Yahoo Finance continuous futures and ETFs, cached at run time. This is a personal research project and not investment advice; paper trading results are hypothetical and do not reflect real execution.