Statistical Arbitrage in Cryptocurrencies (4H Bars)

Momentum vs. Reversal with Transaction Costs, Robustness, and Liquidity Filtering

Author: James T. Susanto

Date: 12*/31/2025*

Abstract / Overview

This project tests whether simple, interpretable "statistical arbitrage" signals in cryptocurrency markets can predict returns **after realistic trading frictions**. Statistical arbitrage ("stat arb") looks for repeatable patterns that produce a **tradable edge**—a statistical advantage that **survives transaction costs** and persists beyond a single backtest window.

I test two canonical cross-sectional patterns on liquid Binance spot USDT pairs using **4-hour bars**:

- **Momentum:** recent winners keep outperforming (trend continuation)
- **Reversal:** recent losers bounce back (mean reversion)

Using a cross-sectional long/short framework, I run **horizon sweeps** for both signals, apply a backtest with a **1-bar execution lag** (to avoid look-ahead bias), incorporate **turnover-based transaction costs** (baseline **20 bps**), and evaluate performance (CAGR, volatility, Sharpe, max drawdown). I then validate the signal mechanism using a bucket diagnostic, stress-test transaction costs, examine train/test stability (cutoff **2024-01-01**), and finally add a **liquidity filter** (trade only the top 50% by volume rank each bar).

Headline results:

- Momentum fails decisively (negative Sharpe across all tested horizons).
- Reversal is strong at short horizons; best performance is at **H = 3** (≈ 12 hours).
- The reversal effect is supported by a **monotonic bucket diagnostic** (signal strength predicts next-bar returns).
- Profits are **highly cost-sensitive** (20 bps works well; 40–60 bps largely destroys profitability).

- A liquidity filter improves realism but reduces performance; execution remains the key bottleneck.

Plain-English definitions (used throughout)

- **Edge:** a repeatable statistical advantage that survives realistic costs and can be traded profitably.
- **Regime:** a market environment where behavior changes (e.g., trending vs. mean-reverting periods).
- **Turnover:** how much the portfolio changes each rebalance; higher turnover usually implies higher costs.
- **Gross exposure:** sum of absolute weights (how "big" the book is regardless of long/short).
- **Dollar neutrality:** net exposure near zero (longs roughly offset shorts), reducing market-direction dependence.
- **In-sample vs. out-of-sample:** in-sample is used for selection; out-of-sample checks if results persist later.

Note: I include "edge" and "regime" because your results explicitly show (1) a strong statistical relationship (edge), and (2) performance differs across time (train vs. test), which is consistent with changing regimes—even if the strategy itself doesn't explicitly model regimes.

Motivation

Crypto markets are structurally different from equities: 24/7 trading, retail-driven flows, episodic liquidation cascades, and sharp reactions to news. These features can cause short-term price “overshoots” that later mean-revert. The goal is to measure whether this creates a **tradable edge** and whether it holds up under realistic costs and out-of-sample checks.

Research questions:

- Does cross-sectional momentum exist in liquid crypto assets on 4-hour horizons?
- Does cross-sectional reversal exist, and what horizon is strongest?
- Does the signal show a mechanism-level footprint (bucket diagnostic) rather than a backtest artifact?
- How sensitive are results to transaction costs (execution realism)?
- Does performance persist out-of-sample (train/test split)?
- Can a liquidity filter improve realism without destroying the edge?

Data

Source & frequency:

- Binance spot OHLCV data on **USDT pairs**, sampled at **4-hour bars**.

Universe & cleaning:

- Start from a larger set of liquid USDT pairs.
- Filter for strong coverage and align on common timestamps.
- Final universe in this run: **47 symbols**
- Shapes (as recorded):
    - `close_px` shape: **(6559, 47)**
    - `rets` shape: **(6558, 47)**

Return definition:

- $$r_{i,t} = \frac{P_{i,t}}{P_{i,t-1}} - 1$$

Why volume matters:
Volume (especially quote volume in USDT) is a practical proxy for liquidity. Higher volume generally implies tighter spreads and lower slippage, which is critical for turnover-heavy strategies.

Methodology

Signal definitions

Momentum (lookback L):

- $$\mathrm{mom}_{i,t}(L) = \frac{P_{i,t}}{P_{i,t-L}} - 1$$

Reversal (horizon H):

- $$\mathrm{rev}_{i,t}(H) = \sum_{k=1}^{H} r_{i,t-k}$$

  Interpretation: coins that went down recently score higher (more “oversold”) and are expected to bounce.

Portfolio construction (cross-sectional long/short)

At each rebalance:

- Rank assets by signal cross-sectionally
- Long top **20%** and short bottom **20%** (q = 0.2)
- Equal weight within each bucket

Scaling:
Raw long/short buckets naturally create ~200% gross exposure (longs sum to +1, shorts to −1). I scale weights by 0.5 so gross exposure is approximately **1.0** (100% gross), making costs/turnover easier to interpret.

Backtest mechanics

- **Daily rebalancing:** every 6 bars (6 × 4H = 24H)
- **1-bar execution lag:** weights decided at time ttt are applied from t+1t+1t+1
- **Turnover-based costs:**
  - $$\text{Turnover}_t = \sum_i \left| w_{i,t} - w_{i,t-1} \right|$$
  - $$\text{Cost}_t = \text{Turnover}_t \times \frac{\text{cost\_bps}}{10{,}000}$$
- **Net return** (net_ret = gross_ret − cost):

$$r_t^{\text{net}} = r_t^{\text{gross}} - \text{Cost}_t$$

- **Equity curve**: cumulative product of (1 + net_ret)

Annualization note:
For 4H bars, a common assumption is **2190 bars/year** (365 × 6). This affects annualized Sharpe/Vol/CAGR levels but not the direction of results.

Results

A) Momentum horizon sweep (L ∈ {6,12,24,42,84,168})

Momentum strongly fails across all tested horizons. Sharpe is negative everywhere and equity decays toward zero.

Momentum sweep From Figure [1] (exact):

L	CAGR	Vol	Sharpe	MaxDD	AvgTurnover	FinalEquity
168	-0.687622	0.305265	-3.657896	-0.971480	0.063666	0.030498
84	-0.743572	0.308986	-4.248745	-0.984809	0.087296	0.016872
42	-0.847462	0.306236	-5.984106	-0.996768	0.119805	0.003552
24	-0.904874	0.316831	-7.261905	-0.999172	0.151554	0.000862
12	-0.943366	0.323901	-8.695676	-0.999837	0.198491	0.000182
6	-0.969928	0.331632	-10.391627	-0.999973	0.272069	0.000027

Momentum Sweep Figure [1]:

-

	CAGR	Vol	Sharpe	MaxDD	AvgTurnover	L	FinalEquity
5	-0.687622	0.305265	-3.657896	-0.971480	0.063666	168	0.030498
4	-0.743572	0.308986	-4.248745	-0.984809	0.087296	84	0.016872
3	-0.847462	0.306236	-5.984106	-0.996768	0.119805	42	0.003552
2	-0.904874	0.316831	-7.261905	-0.999172	0.151554	24	0.000862
1	-0.943366	0.323901	-8.695676	-0.999837	0.198491	12	0.000182
0	-0.969928	0.331632	-10.391627	-0.999973	0.272069	6	0.000027

B) Reversal horizon sweep (H ∈ {1,2,3,6}) — unfiltered

Reversal performs strongly, with best results at **H = 3** (≈12 hours).

Reversal sweep From Figure [2] (exact):

H	CAGR	Vol	Sharpe	MaxDD	AvgTurnover	FinalEquity
3	2.062342	0.326027	3.595908	-0.261067	0.255222	28.703794
6	1.669618	0.338336	3.071809	-0.226235	0.271765	19.017470
2	1.378375	0.331296	2.780925	-0.255719	0.256863	13.448364
1	0.728423	0.291809	2.020843	-0.226980	0.258690	5.162284

Primary conclusion: Best unfiltered reversal is **H = 3** with Sharpe **3.60** and Final Equity **28.70×** (net of 20 bps costs).

Reversal Sweep Figure [2]:

	CAGR	Vol	Sharpe	MaxDD	AvgTurnover	H	FinalEquity
2	2.062342	0.326027	3.595908	-0.261067	0.255222	3	28.703794
3	1.669618	0.338336	3.071809	-0.226235	0.271765	6	19.017470
1	1.378375	0.331296	2.780925	-0.255719	0.256863	2	13.448364
0	0.728423	0.291809	2.020843	-0.226980	0.258690	1	5.162284

C) Mechanism check: bucket diagnostic (H = 3)

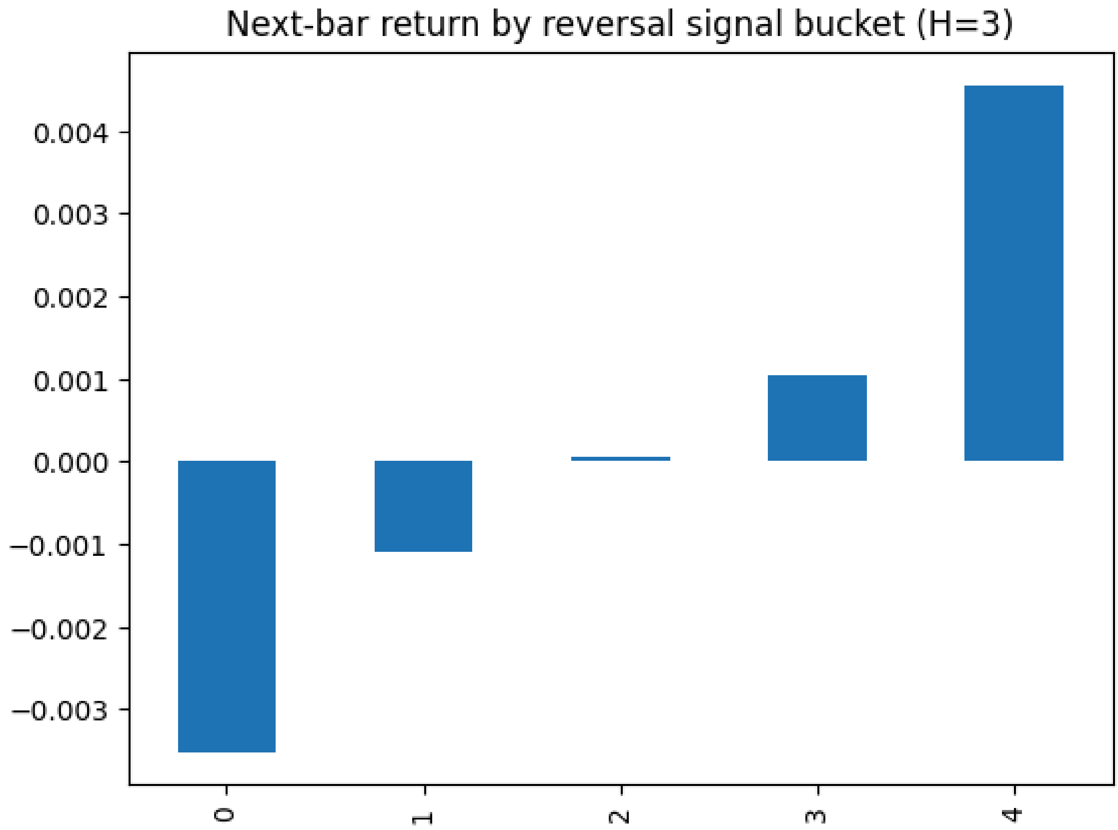
To test whether reversal is a real predictive effect (not just a lucky equity curve), I run a bucket diagnostic. At each bar, I rank coins by the reversal signal and split them into 5 equal buckets. Bucket 0 is the least oversold group and Bucket 4 is the most oversold group. I then compute the average next-bar return for each bucket. If the signal is real, Bucket 4 should have the highest next-bar return and Bucket 0 the lowest.

Bucket means Results From Figure [3] (exact):

- Bucket 0: −0.004244
- Bucket 1: −0.001411
- Bucket 2: −0.000062
- Bucket 3: 0.001161
- Bucket 4: 0.004871

Interpretation: Strong monotonic spread: most oversold assets (bucket 4) bounce the most; least oversold (bucket 0) underperform. This supports a genuine mean-reversion edge.

Bucket Diagnostic Result Figure [3]:



-

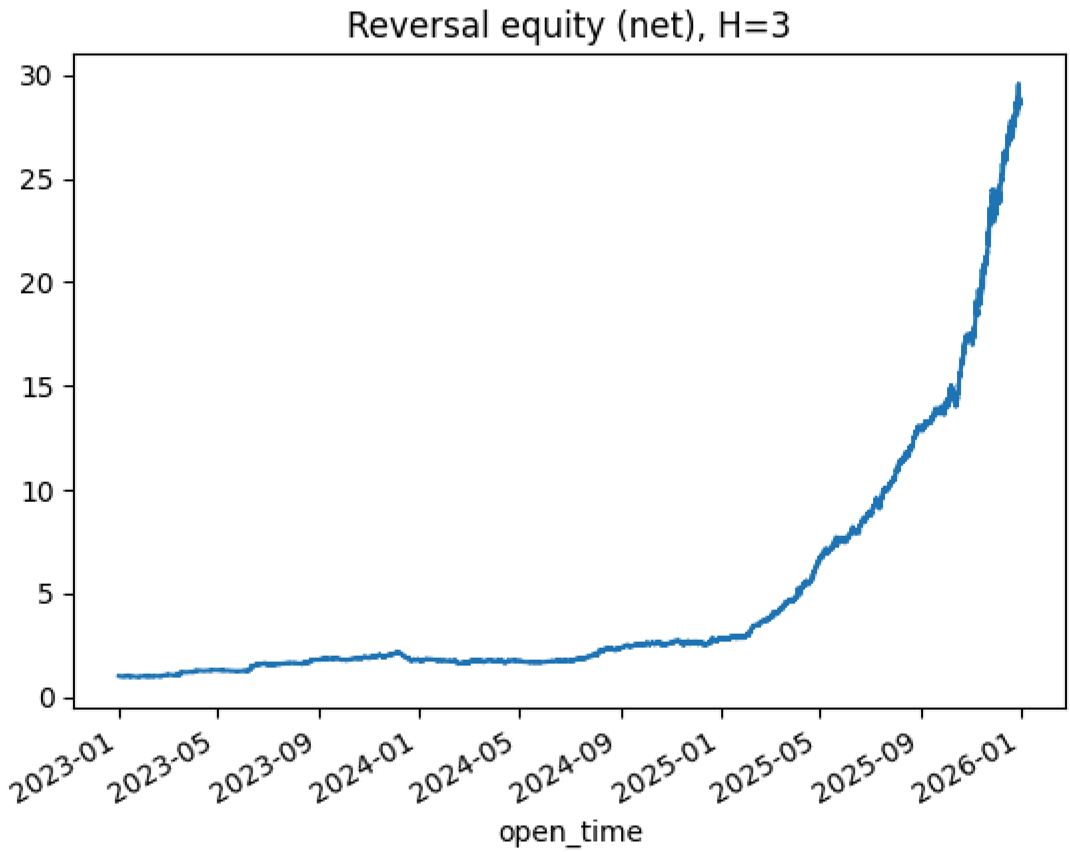
Robustness checks

D) Equity / turnover / cost plots (H = 3, unfiltered)

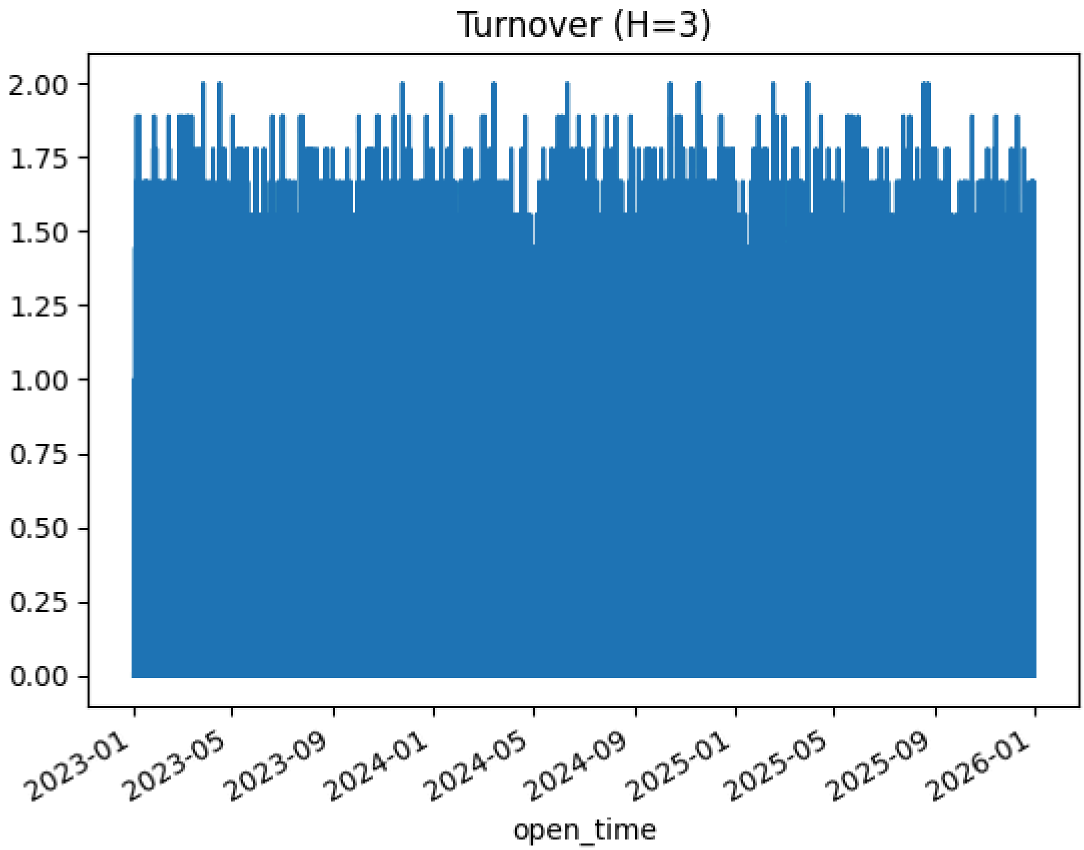
These plots show whether performance is consistent and whether costs/turnover are persistent.

Equity [4], Turnover [5], Cost [6] Figures Below:

1. **Equity curve [4]:**

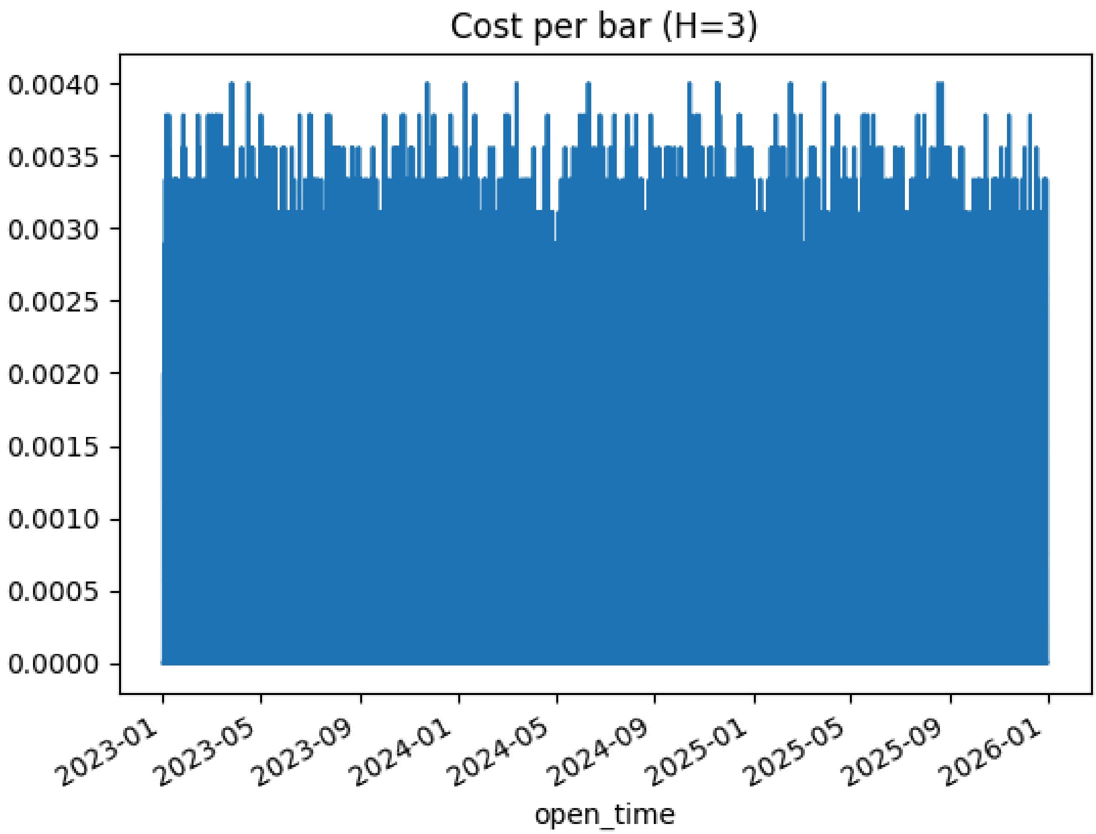


2. **Turnover [5]:**



Turnover (H=3)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
2023-01
2023-05
2023-09
2024-01
2024-05
2024-09
2025-01
2025-05
2025-09
2026-01
open_time

3. **Cost per bar [6]:**



What to notice:

- Equity compounds steadily to ~28.7× net equity under baseline costs.
- Turnover is persistently high → costs matter.
- Cost per bar is relatively stable over time.

E) Cost stress test (H = 3; 20/40/60 bps)

This test checks cost sensitivity—whether the edge survives worse execution assumptions.

Cost stress test from Figure [7] (exact):

cost_bps	CAGR	Vol	Sharpe	MaxDD	AvgTurnover	FinalEquity	AvgCostPerBar
20	2.062342	0.326027	3.595908	-0.261067	0.255222	28.703794	0.000510
40	-0.003789	0.342076	0.159290	-0.727987	0.255222	0.988677	0.001021
60	-0.677064	0.365471	-2.909630	-0.974159	0.255222	0.033696	0.001531

Interpretation:

- At 20 bps, the strategy is extremely strong.
- At 40 bps, it becomes roughly flat with much worse drawdowns.
- At 60 bps, the strategy is strongly negative.

Cost Stress Test Figure [7]:

	H	cost_bps	CAGR	Vol	Sharpe	MaxDD	AvgTurnover	FinalEquity	AvgCostPerBar
0	3	20	2.062342	0.326027	3.595908	-0.261067	0.255222	28.703794	0.000510
1	3	40	-0.003789	0.342076	0.159290	-0.727987	0.255222	0.988677	0.001021
2	3	60	-0.677064	0.365471	-2.909630	-0.974159	0.255222	0.033696	0.001531

F) Train/test split (cutoff = 2024-01-01; best_H = 3)

This checks whether results persist beyond the selection period.

Train/test From Figure [8] (exact):

Split	Sharpe	CAGR	Vol	MaxDD	AvgTurnover
Train (< 2024-01-01)	2.010	0.765	0.305	-0.190	0.257
Test (>= 2024-01-01)	4.323	3.033	0.336	-0.134	0.254

Interpretation:
Performance persists out-of-sample and improves post-2024, consistent with a stronger reversal regime in the later period.

Train/Test Split Figure [8]:

```
best_H = 3
```

	Sharpe	CAGR	Vol	MaxDD	AvgTurnover
Train (< 2024-01-01)	2.010	0.765	0.305	-0.190	0.257
Test (>= 2024-01-01)	4.323	3.033	0.336	-0.134	0.254

Execution improvement: Liquidity filter (volume rank ≥ 0.5)

G) Liquidity filter definition

To improve realism, I restrict the tradable universe each bar to the **top 50% by volume rank** and set weights to zero for non-tradable assets, then re-normalize to maintain gross exposure ≈ 1.

This reduces reliance on “thin” assets where real execution may be poor.

H) Liquidity-filtered reversal sweep (exact)

Liquidity-filtered reversal H sweep from Figure [9]:

H	CAGR	Vol	Sharpe	MaxDD	AvgTurnover	FinalEquity	AvgCostPerBar	AvgGrossRet
3	1.294391	0.445846	2.086009	-0.366244	0.279228	12.073625	0.000558	0.000983
2	0.765649	0.448040	1.492498	-0.360680	0.279723	5.503012	0.000559	0.000865
1	0.755279	0.457007	1.457460	-0.388450	0.283529	5.406636	0.000567	0.000871
6	0.361651	0.473715	0.886483	-0.462364	0.286497	2.524275	0.000573	0.000765

Liquidity Sweep Table Figure [9]:

	CAGR	Vol	Sharpe	MaxDD	AvgTurnover	H	FinalEquity	AvgCostPerBar	AvgGrossRet
2	1.294391	0.445846	2.086009	-0.366244	0.279228	3	12.073625	0.000558	0.000983
1	0.765649	0.448040	1.492498	-0.360680	0.279723	2	5.503012	0.000559	0.000865
0	0.755279	0.457007	1.457460	-0.388450	0.283529	1	5.406636	0.000567	0.000871
3	0.361651	0.473715	0.886483	-0.462364	0.286497	6	2.524275	0.000573	0.000765

Interpretation:

- Under the liquidity filter, reversal remains profitable, with best horizon still **H = 3** in this run.
- Performance decreases versus unfiltered reversal (expected—some edge may come from less liquid coins).
- Turnover remains high; execution still matters.

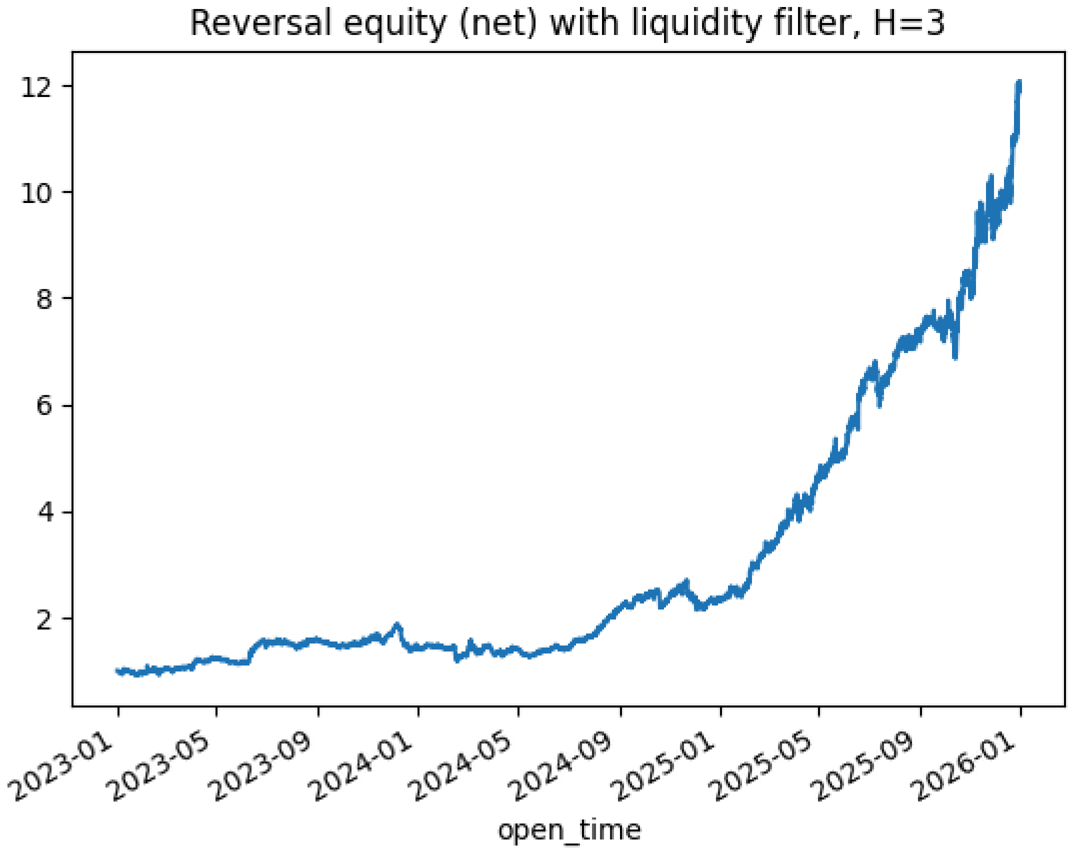
Robustness checks

D) Equity / turnover / cost plots (H = 3, unfiltered)

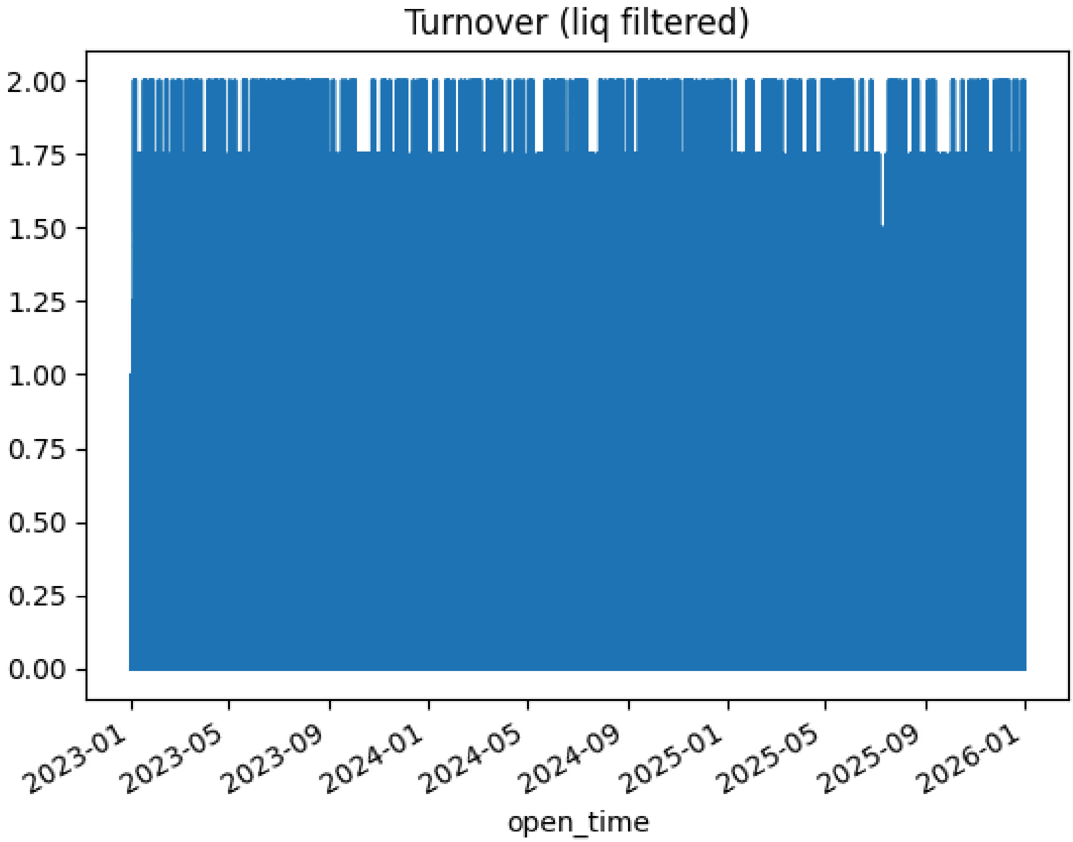
These plots show whether performance is consistent and whether costs/turnover are persistent.

Equity [10], Turnover [11], Cost [12] Figures Below:

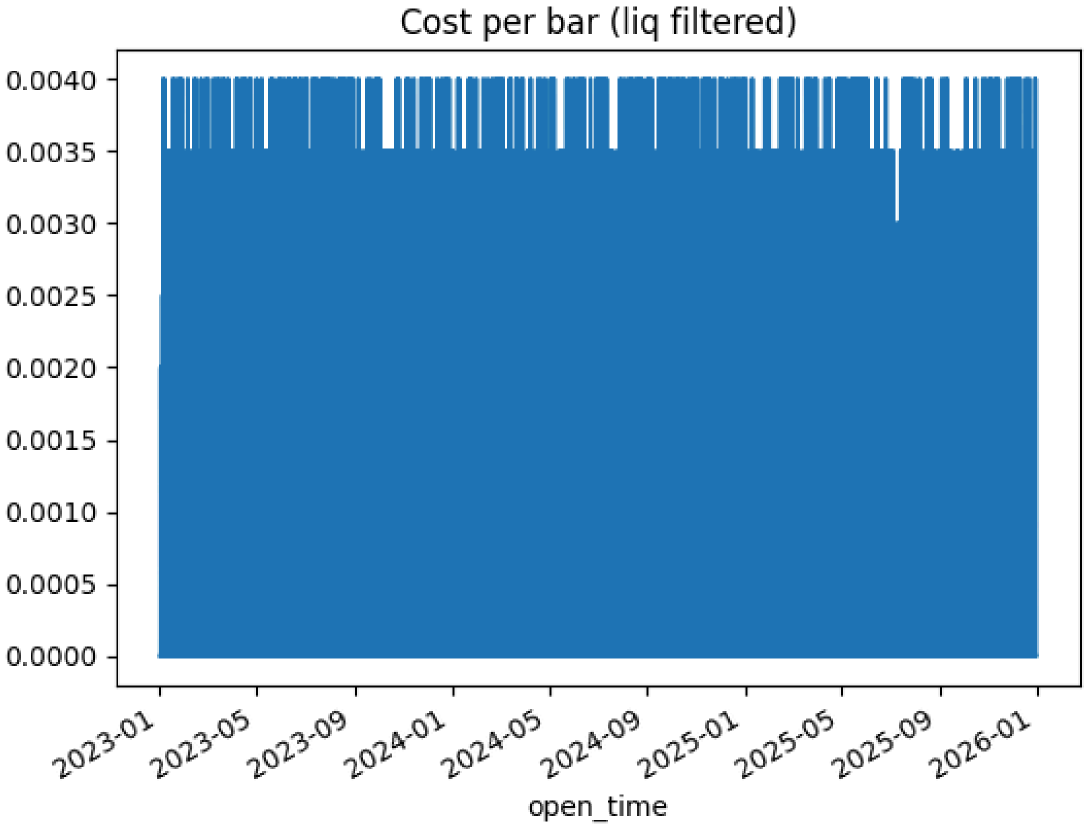
- **Equity curve with liquidity filter Figure [10]:**



- **Turnover with liquidity filter Figure [11]:**



- **Cost per bar with liquidity Figure [12]:**



I) Liquidity-filtered cost stress test (optional, if you include it)

Cost stress test with liquidity filter Figure [13]:

	H	cost_bps	CAGR	Vol	Sharpe	MaxDD	AvgTurnover	FinalEquity	AvgCostPerBar
0	3	20	0.257182	0.245919	1.053680	-0.235882	0.143409	1.986778	0.000287
1	3	40	-0.330215	0.251971	-1.464508	-0.735612	0.143409	0.300529	0.000574
2	3	60	-0.643564	0.261508	-3.813059	-0.957225	0.143409	0.045305	0.000860

Discussion: what the results mean

Why momentum likely failed here

- The cross-section appears dominated by **mean reversion** rather than trend continuation at these horizons.
- Momentum can be fragile under costs if rankings reshuffle frequently.
- Crypto behavior changes across regimes; momentum may work in other windows/features, but **in this sample it does not survive realistic frictions**.

Why reversal worked (and why H = 3 stands out)

- The bucket diagnostic shows a clean oversold→bounce pattern (direct signal-level evidence).
- H = 3 may balance “enough overshoot to matter” while still capturing the short snap-back window.

Why costs are the main constraint

- Short-horizon stat arb often earns small returns per bar.
- With high turnover, doubling/tripling costs quickly removes the edge.
- The strategy is real in this dataset, but **execution quality is the bottleneck**.

Limitations

- **Simplified cost model:** linear bps × turnover; real slippage is nonlinear and size-dependent.
- **No limit order modeling:** limit orders can reduce costs but introduce fill risk.
- **Shorting realism:** spot shorting is limited; perps add funding and liquidation dynamics.

- **Universe selection bias:** if the universe is defined from "currently liquid" names, it may inflate results.
- **Regime dependence:** performance can change as market structure changes.

Next steps

- Add a slippage model using volume/spread proxies (not only turnover).
- Model market vs limit-order execution trade-offs.
- Add liquidity/capacity constraints (top-N by quote volume, position limits).
- Add risk targeting (vol scaling / portfolio volatility targeting).
- Explore conditional reversal (e.g., stronger after dislocation/high-vol periods).

Conclusion

This project evaluated cross-sectional momentum and reversal on liquid crypto assets using 4-hour bars with realistic backtest mechanics (daily rebalancing, 1-bar lag, turnover-based costs). Momentum failed decisively across horizons. In contrast, short-horizon reversal showed a strong and explainable edge, with best performance at **H = 3 (~12 hours)**. The bucket diagnostic confirms the signal's predictive structure: more oversold assets have higher next-bar returns.

Robustness checks strengthen confidence in the signal but highlight that profitability is **highly cost-sensitive**, making execution the primary bottleneck. A liquidity filter improves realism but reduces performance, which is expected when restricting to more tradable names. Overall, the key takeaway is that in this dataset, **short-horizon mean reversion dominates** and can be traded (provided costs and execution are handled carefully).