Research note · synthetic intraday universe · reproducible

Statistical Arbitrage

A walk-forward pairs-trading engine that discovers cointegrated relationships, models their mean reversion, defends against data-snooping, and simulates execution down to the bid–ask spread. Every figure on this page is rendered from a live run of the engine.

Out-of-sample Sharpe	Max drawdown	Ticks / second	Pairs tested → kept
12.16	-0.05%	13.3M	66 → 4

Composite walk-forward equity curve, six out-of-sample windows stitched end to end.

01 Discovery   02 Dynamics   03 False discovery   04 Execution   05 Results

The system is split deliberately into a **research pipeline** that only ever sees the training window, and an **execution sandbox** that receives frozen parameters and replays the future tick by tick. That separation is the central defense against look-ahead bias — the failure mode that makes most backtests lie.

01
Cointegration screen
Engle–Granger · OLS hedge

→

02
OU spread model
AR(1) · half-life

→

03
Significance control
MacKinnon · BH / Bonferroni

→

04
Event-driven execution
frictions · Numba loop

`freeze →` parameters estimated in-sample are locked before the out-of-sample window begins. The simulator never recomputes a mean or a threshold from data it has not yet "lived through".

01 Discovering pairs

Two assets are *cointegrated* when neither price is stationary, yet some linear combination of them is. We estimate that combination with an ordinary-least-squares hedge regression of one leg on the other,

$$Y_t = \beta X_t + \alpha + \varepsilon_t,$$

and treat the residual as the tradeable **spread**:

$$\varepsilon_t \; = \; Y_t - \beta X_t - \alpha.$$

If ε_t is mean-reverting, deviations are temporary and exploitable. The hedge ratio β sets how many units of X neutralise one unit of Y. For the representative pair below, OLS recovers $\beta = 1.211$ with $R^2 = 1.000$.

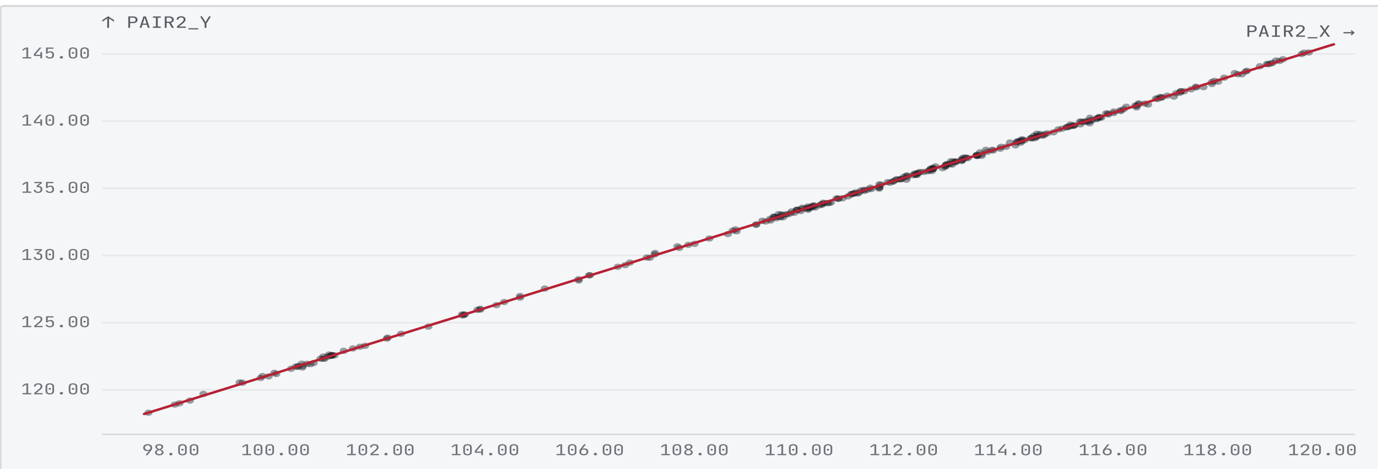

Fig. 1 In-sample hedge regression for PAIR2_Y on PAIR2_X. The crimson line is the fitted equilibrium $\hat{Y} = \beta X + \alpha$; vertical scatter about it is the spread.

We do not hand-pick pairs. The screen regresses every combination in the universe, keeps only those with a credible fit and a sane mean-reversion speed, then submits the residuals to a stationarity test (§03). Of 66 candidate pairs in the first training window, a handful survive the pre-filters:

Pair	ADF p	R^2	Half-life	Score	Kept
● PAIR2_Y/PAIR2_X	2.0e-30	1.000	7.5	0.178	✓
● PAIR0_Y/PAIR0_X	2.0e-30	0.999	7.6	0.175	✓
● PAIR3_Y/PAIR3_X	9.3e-30	1.000	8.9	0.150	✓
● PAIR1_Y/PAIR1_X	1.8e-28	1.000	11.0	0.121	✓
○ NOISE3/PAIR2_Y	0.0318	0.503	296.4	0.004	–

Pair	ADF p	R^2	Half-life	Score	Kept
○ NOISE3/PAIR2_X	0.0325	0.504	297.6	0.004	—
○ PAIR1_Y/PAIR0_Y	0.152	0.442	482.0	0.003	—
○ PAIR1_X/PAIR0_Y	0.157	0.441	492.6	0.003	—
○ PAIR1_Y/PAIR0_X	0.168	0.442	499.3	0.003	—
○ PAIR1_X/PAIR0_X	0.173	0.440	510.3	0.003	—

Tbl. 1 Surviving candidates, ranked by ADF p-value. ● genuine cointegrated pair · ○ spurious / noise combination. Half-life is in event-ticks.

02 Modeling the spread

A stationary spread is modeled as an **Ornstein–Uhlenbeck** process — the continuous-time description of a quantity pulled back toward a long-run mean:

$$dX_t = \theta\,(\mu - X_t)\,dt + \sigma\,dW_t.$$

Here θ is the speed of mean reversion, μ the equilibrium level, and σ the volatility of shocks. Rather than a black-box MLE, we fit the exact discrete analogue — an AR(1) transition — and read the parameters off in closed form:

$$X_{t+1} = \phi X_t + c + \eta_t, \qquad \theta = -\tfrac{1}{\Delta t}\ln\phi, \qquad \mu = \frac{c}{1-\phi}.$$

The single most useful number that falls out is the **half-life** of mean reversion — the expected time to close half of a deviation — which sets the natural holding period and screens out spreads that drift too slowly to trade:

$$t_{1/2} = \frac{\ln 2}{\theta}.$$

For the representative pair that works out to `7.5 ticks` — a fast, tradeable spread.

θ	μ	σ	φ
0.0928	0.000	0.034	0.9114
reversion speed	equilibrium	shock vol	AR(1) coef

Trading thresholds are expressed in standard deviations of the equilibrium distribution, $z_t = (\varepsilon_t - \mu) \,/\, \sigma_{\text{eq}}$ with $\sigma_{\text{eq}} = \sigma/\sqrt{2\theta}$. Crucially, μ and σ_{eq} are frozen from the training window — the z-score at time t is never normalised with information from t's own future. The engine enters when $|z|$ exceeds `3.50`, takes profit back inside `1.00`, and stops out at `5.0`.

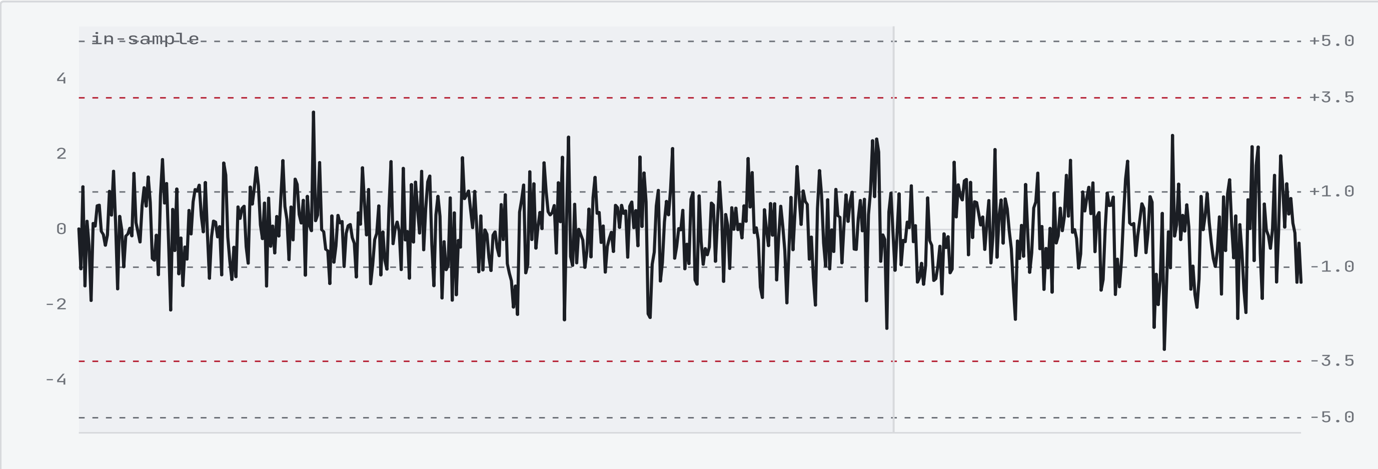

Fig. 2 The normalised spread (z-score) for the representative pair. Shaded region is the in-sample training window; everything to its right is out-of-sample. Dashed bands mark the entry, exit, and stop thresholds — all set before the out-of-sample data was seen.

---

03 Guarding against false discovery

This is where most pairs-trading research quietly fails. Two distinct statistical traps manufacture alpha that does not exist; the engine defends against both.

The right reference distribution

Stationarity is judged with an Augmented Dickey–Fuller regression on the spread,

$$\Delta\varepsilon_t = \gamma\,\varepsilon_{t-1} + \textstyle\sum_i \delta_i\,\Delta\varepsilon_{t-i} + c + u_t,$$

testing $\gamma = 0$ (a unit root) against $\gamma < 0$ (mean reversion). The subtle error — common even in published work — is to read the p-value off a Student-t table. Under a unit root the statistic does *not* follow Student-t; it follows the non-standard Dickey–Fuller distribution, whose 5% critical value with a constant is about `-2.86`, not `-1.65`. Using Student-t makes the test roughly ten times too eager to declare cointegration. We score the statistic against the **MacKinnon** response surface instead.

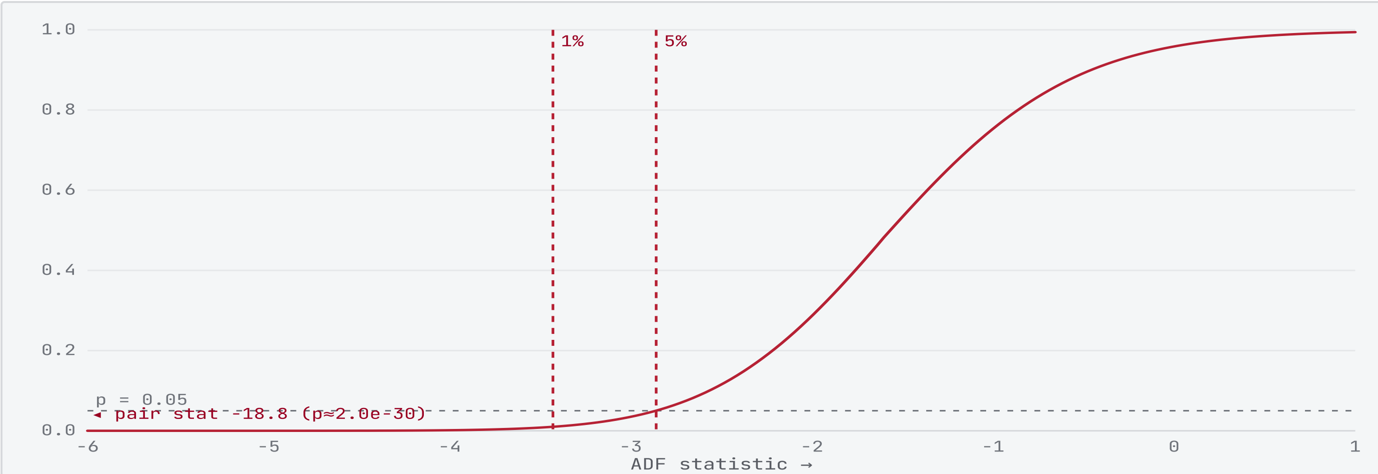

Fig. 3 MacKinnon p-value as a function of the ADF statistic. The 1% and 5% critical values are marked. The representative pair's statistic sits far to the left (-18.76), giving $p \approx$ 2.0e-30 — overwhelming evidence of mean reversion.

The multiple-testing problem

Search enough pairs and some will look significant by chance alone. Testing 66 candidates at $\alpha = 0.05$ would admit several false positives. We control this with the Benjamini–Hochberg false-discovery-rate procedure (keep $p_{(k)} \leq \frac{k}{m}\alpha$) and offer the stricter Bonferroni bound α/m as an alternative.

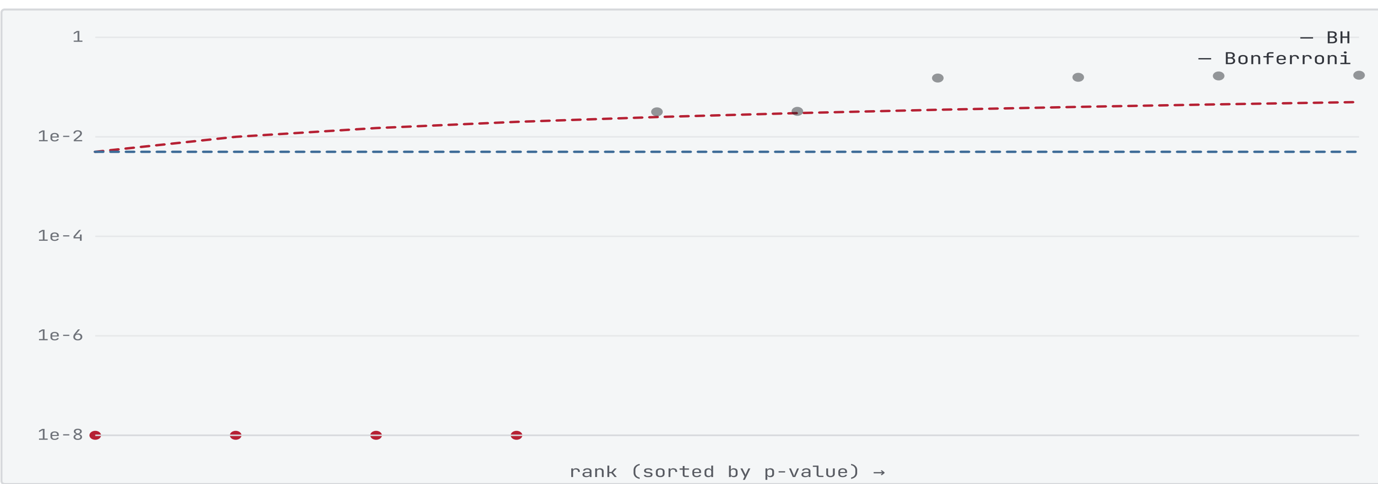

Fig. 4 Sorted candidate p-values against the Benjamini–Hochberg and Bonferroni acceptance frontiers. Without correction, 6 pairs clear $\alpha = 0.05$; BH retains 4. The gap is the false positives the correction removes.

Regime stability

Cointegration that held in-sample can break under a regime shift. Each surviving pair carries diagnostics — a Chow structural-break test, a CUSUM statistic, and rolling parameter-stability checks — and the simulator liquidates gracefully if a spread blows past its stop or a rolling ADF test starts failing. For the representative pair:

θ CV	CUSUM	Chow F	β s.e.
0.091	2.637	36.66	1.7e-4
reversion-speed drift	cumulative-sum stat	max break F	hedge-ratio uncertainty

04 Execution realism

Phantom alpha dies on contact with real costs. The out-of-sample simulator is a path-dependent event loop — compiled to machine code with Numba — that fills at the adverse side of the quoted spread and charges every friction a real desk pays:

- `bid/ask` fills cross the spread: buys at the ask, sells at the bid.
- `commission` per share, tiered like a real broker.
- `slippage` + `market impact` scaling with participation in volume.
- `borrow` fees accrued daily on the short leg.
- `latency`, stop-outs, borrow recalls, and drawdown liquidation.

To see how much these matter, the same pair is run twice over the full out-of-sample series — once with the full friction stack, once frictionless. Costs erode the gross result to a fraction of the "phantom" curve:

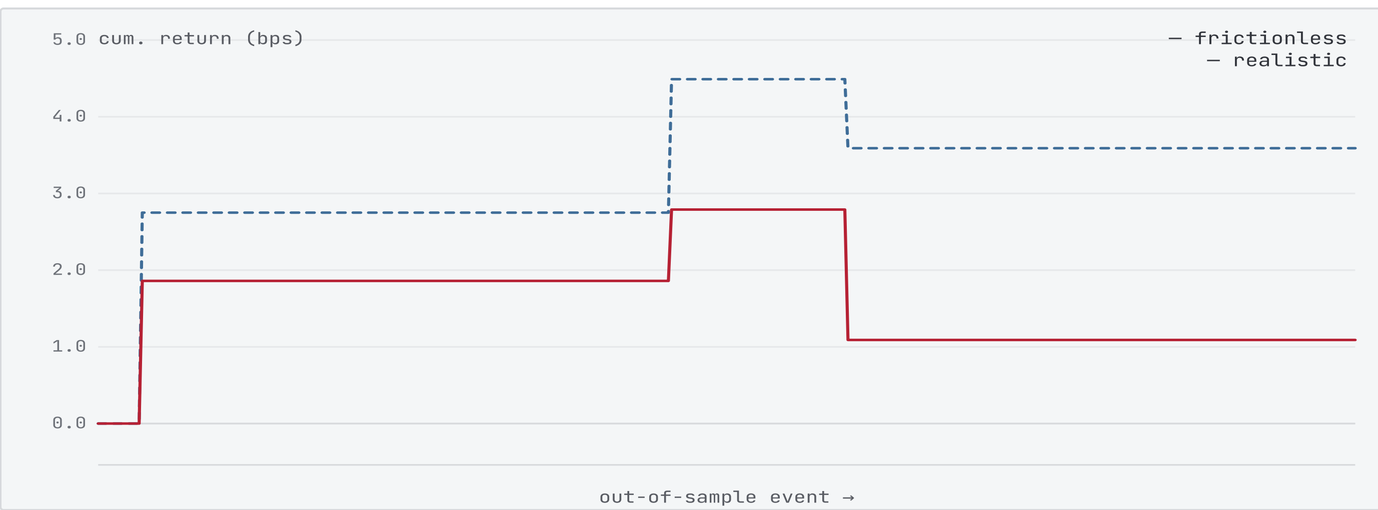

Fig. 5 Frictionless vs. realistic out-of-sample equity for the representative pair (normalised to 1.0 at entry). Realistic net return is `0.01%` against a frictionless `0.04%` — the difference is alpha that never survives the trade.

13.3M

events per second through the compiled loop. A full 10,000,000-tick replay projects to 0.75 seconds — comfortably inside the four-second target, leaving headroom for parameter sweeps.

05 Walk-forward results

Validation is fully out-of-sample. Parameters are fit on a rolling training window, frozen, and tested on the subsequent window; the windows then step forward and the process repeats over 6 slices. The equity curve below is the only one on this page that has never seen its own future.

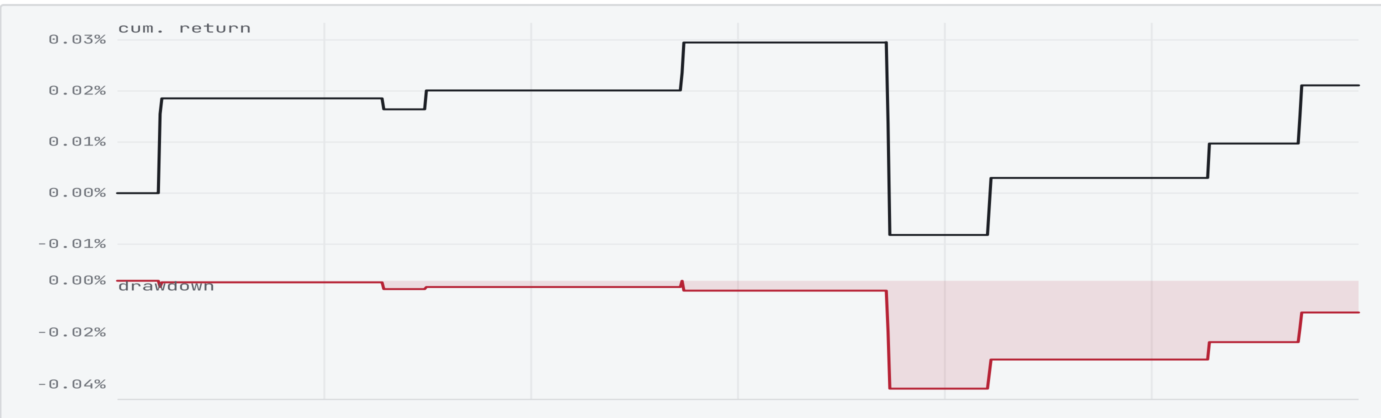

Fig. 6 Composite out-of-sample equity (top) and underwater drawdown (bottom). Vertical rules separate the six walk-forward windows.

Window	Train	Test	Tested	Kept	Allocated	Gross
01	0-8000	8000-10000	66	4	PAIR2 · PAIR0 · PAIR3	$1,000,000
02	2000-10000	10000-12000	66	4	PAIR0 · PAIR2 · PAIR3	$1,000,186
03	4000-12000	12000-14000	66	4	PAIR0 · PAIR2 · PAIR3	$1,000,201
04	6000-14000	14000-16000	66	4	PAIR3 · PAIR0 · PAIR2	$1,000,295
05	8000-16000	16000-18000	66	7	PAIR3 · PAIR0 · PAIR2	$999,918
06	10000-18000	18000-20000	66	6	PAIR0 · PAIR3 · PAIR2	$1,000,030

Tbl. 2 Per-window summary: pairs tested, accepted after correction, and the names allocated capital that window.

Returns are deliberately modest in absolute terms — the strategy trades a clean synthetic universe conservatively — so the honest read is risk-adjusted. The tail is what a risk manager checks first:

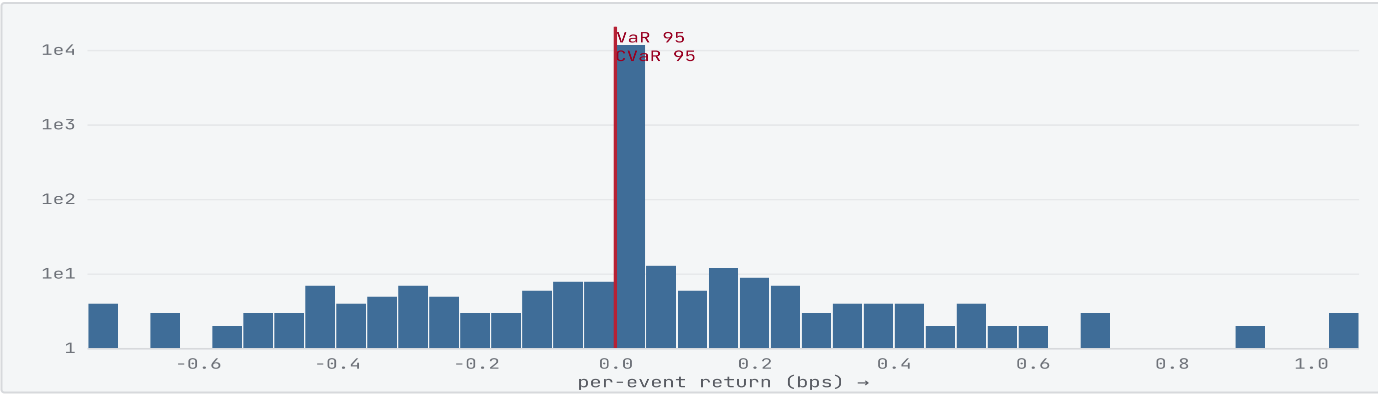

Fig. 7 Distribution of per-event returns. The crimson markers are the 95% Value-at-Risk and the Conditional VaR (expected shortfall beyond it).

Total return	0.021%	Annualized return	10.93%
Sharpe ratio	12.16	Sortino ratio	2.11
Max drawdown	-0.046%	Calmar ratio	235.4
Annualized volatility	0.85%	Trades	8
Win rate	87.50%	Avg trade P&L	$67
Turnover	$3,334,526	VaR 95% (per event)	0.000%
CVaR 95% (per event)	0.000%	Tail ratio	0.00

Tbl. 3 Full post-friction performance report.

Colophon

Pure Python on NumPy with one Numba-compiled execution kernel — no pandas, no statsmodels — so the mathematics stays explicit and auditable. The MacKinnon p-value, OU fit, multiple-testing correction, and friction model are hand-implemented and unit-tested. Every number here was produced by:

```
PYTHONPATH=src python3 site/export_data.py
```

Synthetic universe · 20,000 events · 12 symbols (4 planted + 4 noise) · deterministic for a fixed seed.