Frictions All the Way Down: A Fully-Reproducible Audit of Time-Series Momentum from Frozen Specification to Integer Contracts*

Vincent W.†

June 2026 — working paper v0.1 (P1 draft)

Abstract

Can an individual investor systematically improve on holding an equity index, without abandoning the index position? We answer with a fully-reproducible audit of time-series momentum (TSMOM) that quantifies implementation friction layer by layer: strategy (daily ETF data, 2003-08–2026-05), instrument (CME futures), and contract granularity (integer contracts at a representative $500,000 account). All parameters were frozen from literature priors before estimation; look-ahead prevention is enforced by unit tests; every number in this paper is injected programmatically from the research repository's output files. Net of costs, the long/short ETF sleeve earns an excess Sharpe of 0.57 (Newey–West $t = 2.86$, bootstrap 95% CI [0.19, 0.96]) with -0.04 correlation to SPY and gains in 6/6 crisis windows (GFC: 16.1% vs. SPY -46.0%). Futures implementations track their ETF counterparts closely (strict-pair correlations 0.90–0.97), and a return-stacking overlay on a full SPY position improves risk-adjusted performance monotonically in the overlay multiple within the evaluated grid. Execution reality is harsher: integer-contract tracking error at the final configuration is 327 bp per year, decomposing into 110 bp of rounding-plus-commission noise and 355 bp of micro-versus-parent instrument basis. We disclose all failed acceptance criteria, a 100× contract-value mis-scaling that we discovered, corrected, and re-ran, and a cost-control incident in our own data-acquisition pipeline—arguing that in systematic investing, the audit chain itself is the product.

1 Introduction

Most retail attempts to "beat the index" fail at one of three gates: the strategy is overfit; the strategy is real but the chosen instrument leaks the edge through costs; or strategy and instrument are both adequate while the account is simply too small to hold the prescribed positions in integer contracts. The literature treats the first gate extensively, the second occasionally, and the third almost never. This paper walks all three gates in order, on one strategy, with one audit chain.

*Research and educational purposes only; not investment advice. All results are in-sample research on historical data; past performance does not predict future returns. Account sizes are *representative scales*, not disclosures of actual funds or positions. Build 60b39c4; every number in this paper is programmatically injected from the repository's output files.

†Methodology note: research execution used an AI-assisted workflow (frozen specification → agent implementation → human adjudication); see Section 10.

Our research question is deliberately narrow: *can an individual investor systematically improve on a buy-and-hold S&P 500 position without giving up that position?* The candidate answer is the oldest one in the managed-futures industry: a time-series momentum (TSMOM) overlay, sized by realized volatility, stacked on top of the index holding. The question is not whether TSMOM "works" in a backtest—that ground is well covered (Moskowitz et al., 2012; Hurst et al., 2017)—but how much of the paper edge survives the descent from specification to a deliverable order ticket.

1.1 What this paper contributes

A reproducibility chain, not a performance claim. Signal definitions, look-back horizons (3/6/12 months), the volatility window (60 trading days), the per-asset volatility target (10%), the leverage cap (1.50), cost assumptions (10 bp one-way; 50 bp annualized borrow on ETF shorts), and the single sample split (2015-01-01) were frozen in a written specification before any estimation, and the specification forbade re-optimization on this data. Look-ahead prevention is not a claim but a unit test: artificially doubling all future returns must leave current signals bit-for-bit unchanged. Every number in the text, including those in this introduction, is injected by the build system from the repository's output files; the LaTeX build fails on any undefined macro.

Friction quantified layer by layer. We measure the same strategy at three altitudes. At the *strategy* layer (daily ETF total-return data, 2003-08 to 2026-05, 274 months), the long/short sleeve nets an excess Sharpe of 0.57 with -0.04 monthly correlation to SPY. At the *instrument* layer, CME futures replications track their ETF counterparts with correlations of 0.90–0.97 on strictly comparable pairs—the difference between wrapper and strategy is small and measurable. At the *execution* layer, rounding prescribed weights to integer contracts at a representative $500,000 account produces an annualized tracking error of 327 bp against the frozen model, of which only 110 bp comes from rounding and commissions; the larger 355 bp component is instrument basis between micro contracts and their full-size parents. Tracking error, not alpha decay, is where the retail implementation story is decided.

Process results reported as results. Over the project we committed three classes of mistakes that conventional papers omit: a contract-value mis-scaling of exactly 100× on Treasury futures (discovered because it made bond legs look permanently untradable); two pre-registered predictions that the data subsequently falsified; and a cost-control failure in our own data-acquisition pipeline (cumulative spend briefly exceeded its authorized cap, which was later revised). Each is documented with its detection path, correction, and re-run deltas (Section 10). We argue these belong in the result set: a research pipeline's error-surfacing behavior is as much an empirical property as its Sharpe ratio.

1.2 Honest framing up front

All results are in-sample research. Parameter freezing from literature priors removes one form of overfitting but is not out-of-sample validation. The strategy's post-2015 softening is shown in full (the long/short sleeve's Sharpe falls from 0.73 to 0.40 across the frozen 2015-01-01 split), several pre-registered acceptance criteria FAILED and are reported as such, and nothing here constitutes investment advice.

2 Related Literature

Time-series momentum. Moskowitz et al. (2012) document persistent return predictability from an asset's own past 12-month excess return across 58 futures markets, with volatility-scaled positions; their 12-month sign signal and volatility targeting are the direct ancestors of our frozen specification. Hurst et al. (2017) extend the evidence to a century across asset classes, establishing the crisis-convexity property that motivates our stacking design. Babu et al. (2020) broaden the asset universe further. Our contribution is not new evidence on the anomaly but a complete, testable implementation audit of it at retail scale.

Implementation and capacity. Baltas and Kosowski (2020) examine how volatility estimators and trading rules affect deliverable TSMOM performance and address capacity at institutional scale. We work at the opposite end of the size spectrum, where the binding constraint is not market impact but integer-contract granularity—a constraint that, to our knowledge, has not been quantified as a tracking-error decomposition in the published literature.

Post-2015 trend softening. Practitioner and academic discussion of trend-following's weaker decade (e.g. Zakamulin and Giner, 2020) frames our split-sample disclosure. We take no stance on whether the softening is structural; we freeze the split date in the specification and report both halves in full.

Evidence layering. Following the centennial evidence cited above, we present our own results in three explicitly separated layers: a fifty-year monthly proof-of-concept (Section 5; coarse approximations, declared), a twenty-year full-specification daily implementation (Sections 6–8), and the century-scale literature (this section, citations only—no third-party data are re-plotted in this paper).

3 Data

3.1 ETF layer: three-tier architecture

The ETF backtest uses eight US-listed ETFs spanning equities (SPY, EFA, EEM), Treasuries (IEF, TLT), gold (GLD), broad commodities (DBC), and the dollar index (UUP). Daily total-return prices come from a three-tier architecture frozen in the specification: a primary source (adjusted closes), an official rates source (FRED `DGS3MO` for the risk-free leg, with a specification-fixed fallback chain), and an independent verification source (Databento consolidated US-equity daily bars) used to cross-validate unadjusted daily returns and repair bad prints. Assets enter the strategy only after listing plus a full signal warm-up; nothing is backfilled. The formal backtest starts when the third asset becomes available (2003-08) and runs through 2026-05 (274 months; the trailing partial month is dropped by rule).

Cross-validation results and all quality checks are reported in the repository's data-quality table: all eight tickers match the verification source with daily-return correlations of ≈ 1.0 on the overlapping window, zero bad prints required repair, and the two flagged extreme-return days are verified historical events (October 2008 rallies), not data errors. One honest limitation: the verification dataset's history begins in mid-2024, so cross-validation covers only the recent segment; earlier integrity rests on the primary source's internal checks (zero missing days within listing ranges; hard failure on non-positive prices).

3.2 Futures layer: measured availability, not assumed

The futures implementation uses eight CME contracts (ES, NQ, ZN, ZB, GC, CL, 6E, 6J) from the Globex MDP3 daily dataset. Every availability date in this paper was *measured* via the vendor's metadata API rather than assumed: the dataset begins 2010-06-06 for all eight roots; micro contracts arrive later and unevenly (MES/MNQ 2019; MCL 2021; micro gold and FX from 2010); the micro Japanese-yen contract was delisted in 2024-03 and its successor had a measured 22-month listing gap. These instrument life events are inputs to the execution analysis of Section 8, not footnotes: one of them terminates an entire substitution design.

Data acquisition operated under a hard budget discipline (estimate-first, cumulative cap, parquet caching). The discipline itself failed once in a cross-process accounting gap, documented as a process result in Section 10.

4 Methodology: A Frozen, Testable Specification

4.1 Signal and sizing (frozen)

For asset i at month-end t with adjusted price $P_{i,t}$, the signal averages sign momentum over look-backs $L \in \{3/6/12\}$ months:

$$S_{i,t} \;=\; \tfrac{1}{3}\textstyle\sum_{L} \operatorname{sign}\big(P_{i,t}/P_{i,t-L} - 1\big), \qquad S_{i,t} \in [-1, 1]. \tag{1}$$

Positions scale inversely with realized volatility ($\hat{\sigma}_{i,t}$: 60-day daily-return standard deviation, annualized) toward a per-asset target $\sigma^* = 10\%$, capped at $1.50\times$ leverage:

$$w_{i,t} \;=\; S_{i,t} \cdot \min\big(\sigma^*/\hat{\sigma}_{i,t},\, 1.50\big). \tag{2}$$

Two variants are always reported together: long/short (V-LS) and long/flat (V-LF, signals floored at zero with cash earning the risk-free rate). All parameters are literature priors (Moskowitz et al., 2012), frozen in the specification document before estimation; the robustness grid of Section 6 exists to display sensitivity, not to select.

4.2 Look-ahead prevention as a unit test

Signals computed at month-end t earn returns only in month $t{+}1$; the lag is implemented as one isolated `shift` step in code. The specification mandates the test, not just the property: *double all returns after a cutoff and assert that signals, volatilities, and positions up to the cutoff are bit-for-bit identical.* The repository's test suite enforces this for the ETF pipeline, the futures pipeline, and the continuous-contract construction (roll-day splice equality is asserted contract-by-contract on real data, not only on synthetic fixtures). The risk-free leg is lagged under the same convention.

4.3 Costs

ETF costs: 10 bp one-way on turnover $\sum_i |\Delta w_i|$, plus 50 bp annualized on short notional (V-LS only); sensitivity at three cost levels is reported in full. Futures costs are structural rather than ad valorem: one tick per side on rebalancing turnover, a full round-trip (two ticks) on each roll, with tick sizes and multipliers taken from exchange definition records (never hard-coded), plus fixed per-contract commissions in the integer-contract analysis.

4.4 Continuous futures construction

Per-contract daily series are stitched by *return splicing*: each day's continuous return is the held contract's own return, with rolls executed at the earlier of (a) the second consecutive day the next cycle contract's volume exceeds the held contract's, or (b) five trading days before expiry. Contract identity uses canonical keys (root and expiry year-month) because exchange security IDs are recycled and far-month symbols are renamed—both failure modes we hit and tested against. Roll-rule incidence by product is reported in Appendix B.

4.5 Statistical protocol

All Sharpe ratios are excess-return Sharpes. Significance uses Newey–West t-statistics (lag 6) and circular block-bootstrap 95% confidence intervals (block 6 months, 10000 draws, fixed seed). Two consecutive full rebuilds of every artifact in this paper are byte-identical; the build commit is stamped on the title page.

5 Results 0: A Fifty-Year Proof of Concept

Before the full-specification implementation, we present the project's original proof-of-concept, ported into the same audited repository (experiment G18). It is deliberately coarse: three monthly proxy assets—a Shiller-based S&P 500 total-return series, a 10-year Treasury return constructed from yields by duration approximation, and London gold—with a single 12-month sign signal, a 36-month volatility window, a 10% per-asset volatility target, no leverage (cap 1.0), and 10 bp turnover costs. The port reproduces the prototype exactly: on identical cached inputs, all nine gate metrics deviate by 0.00% (the pre-registered halt threshold was 10%).[1]

Over 603 months (1976–2026), the three-asset TSMOM composite returns 6.3% per year at 5.8% volatility (Sharpe 1.09 on the no-risk-free convention; roughly 0.8 after the declared adjustment), with a maximum drawdown of -15.9% and 0.21 monthly correlation to the equity proxy (buy-and-hold proxy Sharpe: 0.98 on the same convention). The pre/post-2000 split shows the familiar softening (1.22 $\rightarrow$ 0.96) while remaining positive in both halves. Figure 1 shows the half-century paths.

Evidence layering. This section is the first of three explicitly separated evidence layers: (1) this fifty-year proof of concept (coarse, approximations declared); (2) the twenty-year full-specification daily implementation of Sections 6–8 (frozen parameters, tested look-ahead prevention, measured costs); (3) the century-scale literature (Hurst et al., 2017; citations only, no third-party data re-plotted). Conclusions in this paper rest on layer (2); layers (1) and (3) establish that the phenomenon is not an artifact of the recent sample.

[1] **Declared approximations, in full:** (i) bond total returns use the modified-duration approximation $TR \approx y/12 - D\,\Delta y$, omitting convexity; (ii) the dividend series' recent gap is filled by carrying forward the last observed dividend yield; (iii) *no risk-free rate is deducted*—Sharpe ratios in this section are overstated by roughly 0.2–0.3 relative to the excess-return convention used everywhere else in this paper; (iv) monthly granularity; (v) indices are not directly investable. These approximations are why this section is labeled a proof of concept and kept strictly separate from the full-specification results.

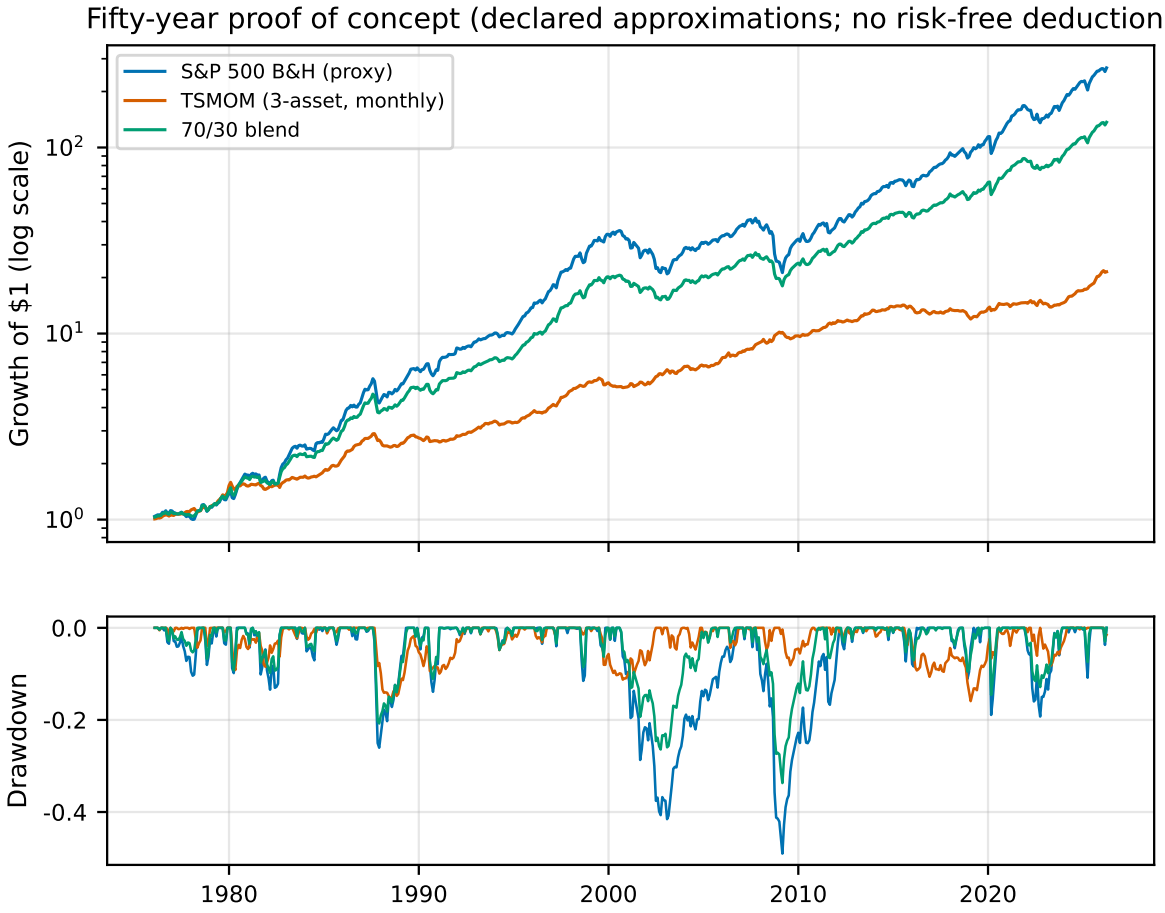

Figure 1: Fifty-year proof of concept (G18): growth of $1 (log scale) and drawdowns for the equity proxy, the three-asset TSMOM composite, and a 70/30 blend. Declared approximations apply; no risk-free deduction.

6 Results A: The ETF Implementation

6.1 Headline results, both variants, no selection

Table 1 reports both strategy variants beside the two benchmarks over the full sample and both halves of the frozen 2015-01-01 split. Nothing is selected: the specification requires that every variant, every period, and (below) every grid cell appear in print.

Three features matter for the stacking design that follows. First, the long/short sleeve is *uncorrelated* with equities (-0.04 monthly over the full sample), while long/flat retains a moderate equity tilt (0.26)—the diversification raw material differs by variant. Second, both variants clear conventional significance on the full sample: V-LS earns Sharpe 0.57 (Newey–West $t = 2.86$; bootstrap 95% CI $[0.19, 0.96]$) and V-LF earns 0.85 ($t = 4.45$; $[0.49, 1.24]$).[2] Third, the decade

[2]These use the standard Moskowitz–Ooi–Pedersen convention: the position implied by the month-end-t signal is taken at that close and earns the close-to-close return of month $t+1$. Execution timing is sensitive to the month-boundary entry gap—under a more conservative one-trading-day-delayed fill (first-trading-day close to month-end close) the full-sample excess Sharpes fall to 0.32 (V-LS) and 0.58 (V-LF). This is a disclosed convention, not look-

Table 1: ETF implementation: full sample and frozen split. All Sharpe ratios are excess-return Sharpes; costs included.

series	period	n_months	ann_return	ann_vol	sharpe_excess	max_drawdown	corr_spy
SPY	full	274	11.3	14.6	0.69	-50.8	1.00
SPY	pre	137	8.7	14.0	0.57	-50.8	1.00
SPY	post	137	14.0	15.1	0.81	-23.9	1.00
60/40 SPY-IEF	full	274	8.5	8.9	0.76	-29.5	0.96
60/40 SPY-IEF	pre	137	7.9	8.1	0.80	-29.5	0.95
60/40 SPY-IEF	post	137	9.0	9.7	0.73	-20.5	0.96
V-LS	full	274	4.2	4.3	0.57	-6.6	-0.04
V-LS	pre	137	4.7	4.5	0.73	-6.6	0.02
V-LS	post	137	3.7	4.2	0.40	-5.2	-0.10
V-LF	full	274	4.7	3.5	0.85	-4.7	0.26
V-LF	pre	137	5.1	3.9	0.96	-4.7	0.22
V-LF	post	137	4.4	3.2	0.73	-3.3	0.31

Table 2: Crisis-window cumulative returns (windows frozen in the specification; month-end convention).

window	start	end	SPY	V-LS	V-LF
GFC (Oct 2007 - Mar 2009)	2007-10	2009-03	-46.0	16.1	10.4
Euro debt (Aug - Oct 2011)	2011-08	2011-10	-2.5	0.6	2.1
Aug 2015 (single month)	2015-08	2015-08	-6.1	-0.5	-0.9
Q4 2018	2018-10	2018-12	-13.5	-1.2	-1.5
COVID (Jan - Mar 2020)	2020-01	2020-03	-19.4	1.9	-1.0
2022 hiking bear (full year)	2022-01	2022-12	-18.2	8.4	3.4

split is not flattering and is shown anyway: V-LS softens from 0.73 to 0.40 and V-LF from 0.96 to 0.73 after 2015-01-01—consistent with the post-2015 trend-following discussion, and central to the forward-expectation framing of Section 9.

6.2 Crisis windows: the economic case

The strategy's economic value concentrates where the index holder needs it. Across the six pre-specified crisis windows (Table 2, Figure 3), V-LS beats SPY in 6/6 and V-LF in 6/6: in the GFC window the long/short sleeve returned 16.1% against SPY's -46.0%. This is the textbook crisis-convexity profile—earned here after costs, under tested lag discipline, in a frozen specification.

6.3 Robustness: the full grid, no cherry-picking

The frozen protocol varies single look-backs, volatility windows, and leverage caps in a full cross of 108 cells, all shown in Figure 4 and listed in Appendix C. 100% of cells have positive net Sharpe; the worst cell still earns 0.36. There is no cliff: the headline result does not depend on the frozen parameter choices being lucky.

ahead: the position uses only information through close t (`tests/test_no_lookahead.py`).

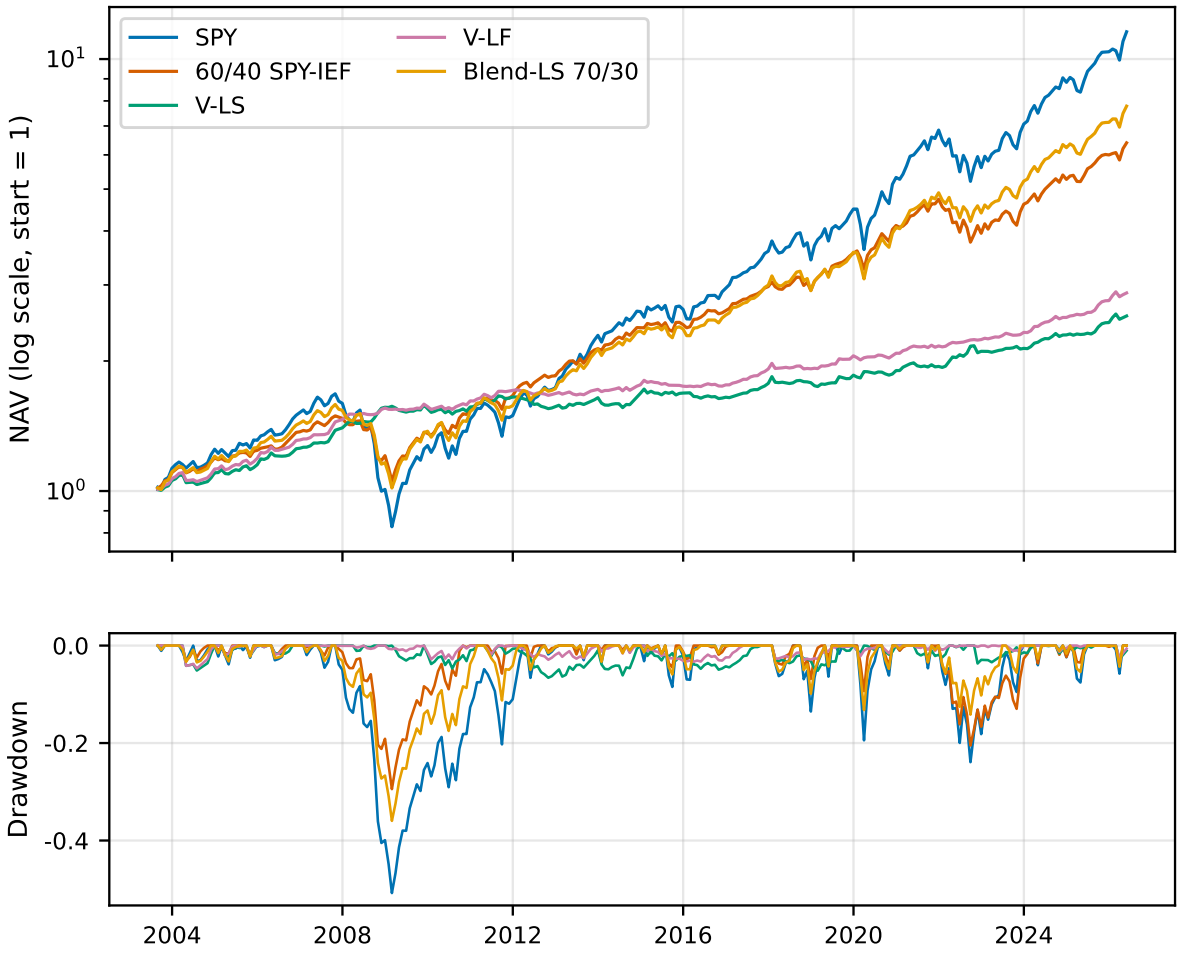

Figure 2: Log NAV (start = 1; log scale stated) and drawdowns: benchmarks, both variants, and the 70/30 blend.

6.4 Blends, costs, and rolling behavior

Mixing the sleeve into an SPY portfolio at fixed weights improves risk-adjusted performance along the de-risking path: the 70/30 SPY/V-LS blend earns Sharpe 0.76 versus SPY's 0.69, with maximum drawdown -36.0% versus -50.8%. Cost sensitivity is linear and survivable—at a doubled 20 bp one-way assumption the long/short sleeve still nets 0.48. Rolling 36-month Sharpe and correlation (Figure 5) show the sleeve's lean years in full rather than smoothing them away.

7 Results B: Futures and the Implementation Gap

7.1 Same strategy, different wrapper

The futures implementation reuses the frozen signal, sizing, and lag code verbatim; only the instrument layer changes (self-built continuous contracts, tick-based costs, full-collateral accounting). Table 3 reports the wrapper effect directly: on the four strictly comparable pairs

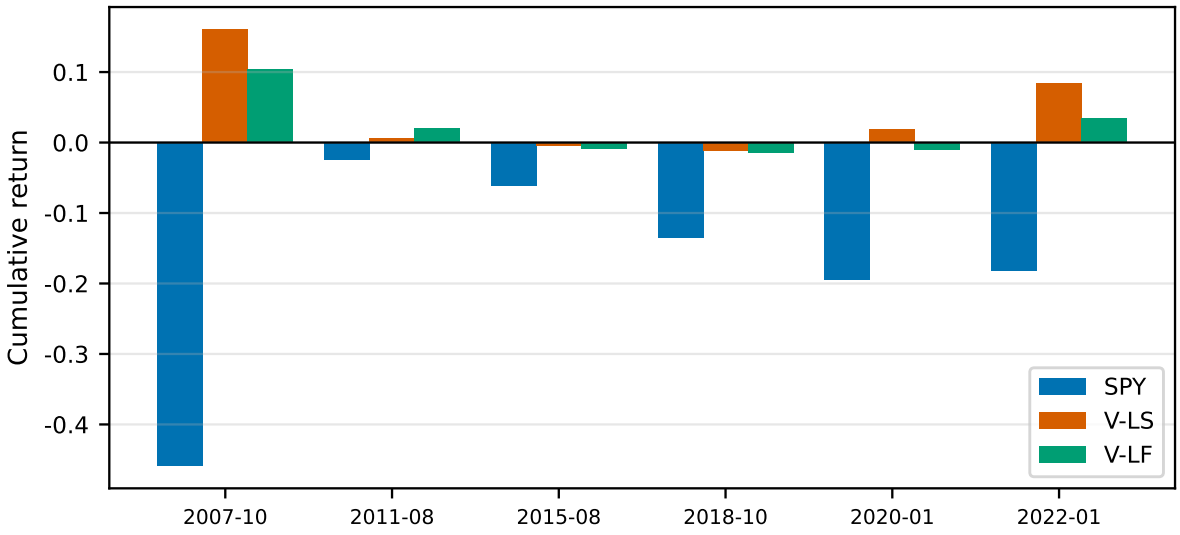

Figure 3: Crisis windows: SPY versus both TSMOM variants.

(ES/SPY, ZN/IEF, ZB/TLT, GC/GLD), per-asset strategy returns correlate at 0.90–0.97 over the overlap, with Sharpe differences within a few hundredths. At the sleeve level the correlation is 0.81—lower only because the two universes differ by construction (NQ/CL/6E/6J have no one-to-one ETF counterparts). The conclusion is the one that matters for execution design: *the wrapper costs little; the strategy ports.*

7.2 Return stacking: keep the index, add the sleeve

Because futures consume margin rather than capital, the sleeve can sit *on top of* a full SPY position. Stacking k units of sleeve excess return on SPY behaves as designed (Figure 7): correlation to SPY falls monotonically in k (from 1.00 at k=0 to 0.79 at k=2.5 for the long/short sleeve) while risk-adjusted performance rises (Sharpe 0.90 $\rightarrow$ 1.06 at k=1). One pre-registered acceptance criterion FAILED here and is discussed in Section 9: within the futures-era sample no k narrows maximum drawdown by the required ten percentage points, because the sample lacks a GFC-scale equity event and the 2020 crash was too fast for monthly trend signals.

7.3 Long-history splice and decision geometry

Splicing the ETF sleeve (pre-2011) with the futures sleeve (post-2011) into a synthetic long history—explicitly *not* a tradable series—restores the GFC to the stacking sample: at k=1.5 the stacked portfolio's GFC-window drawdown is -41.2% versus SPY's -50.8% (Figure 8). The two-dimensional grid of Figure 9 (base allocation $\times$ overlay multiple) then locates the configuration adopted for execution analysis: an 80% SPY / 20% cash base with a 1.5$\times$ long/short overlay earns 14.7% at 0.96 Sharpe with -31.8% maximum drawdown over the 2003–2026 synthetic sample, against 11.3% and -50.8% for pure SPY. De-risking the base while adding overlay is the only direction in the grid that simultaneously raises return, lowers volatility, and narrows drawdown.

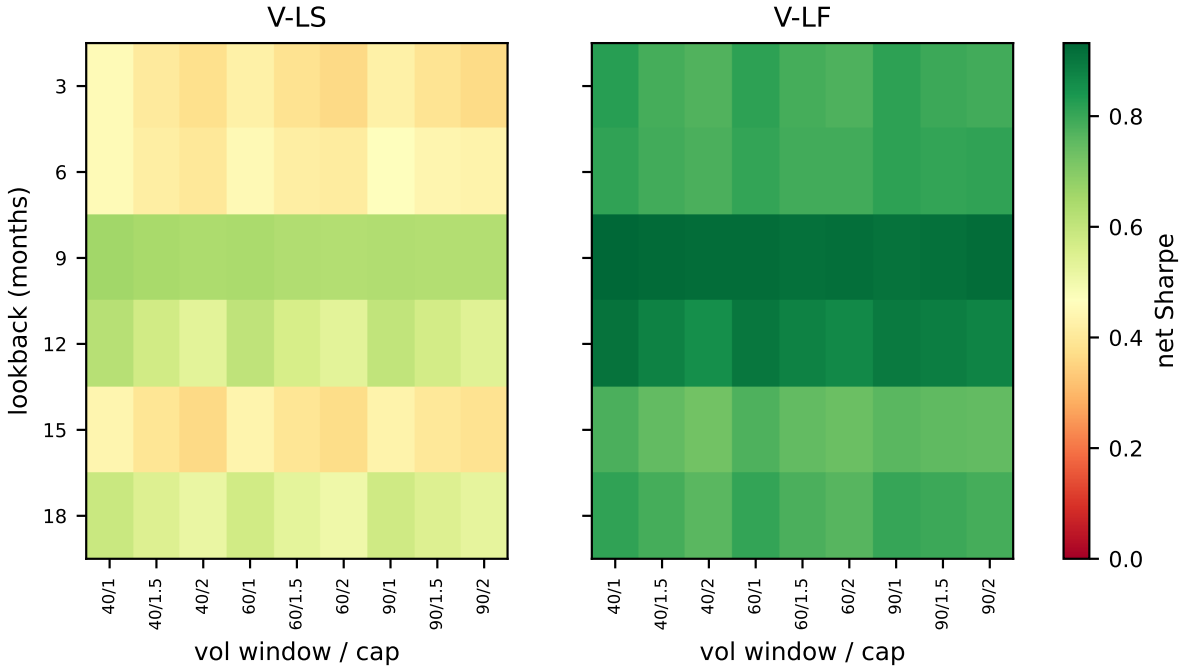

Figure 4: Net Sharpe across the full robustness grid (every cell displayed).

8 Results C: Execution Reality

Two tracking-error frames, stated once. Throughout this section we distinguish TE_{exec}—tracking error against a *same-instrument fractional* replay (isolates rounding and commissions)—from TE_{model}—tracking error against the *frozen parent-contract model* (adds micro-versus-parent instrument basis; this is what monthly shadow runs compare against). Every TE number below carries its frame.

8.1 Integer contracts and minimum viable scale

Translating prescribed weights into integer contracts is where paper strategies meet account statements. At a \$100K account the discretized corner configuration tracks its fractional model with $TE_{exec} = 413$ bp/yr; \$200K improves to 228 bp (same frame) (Figure 10). The curve's structural floor is set not by account size but by legs that lack micro contracts—a fact we initially mis-measured: a 100× contract-value scaling defect on Treasury futures (Section 10) made bond legs appear permanently untradable. After correction, bond legs execute from roughly \$500K.

8.2 Instruments die: substitution under pre-registered gates

Execution design must survive instrument life events. Our liquidity health check declared the 30-year Micro Yield future dead (last data 26 business days stale at measurement; 21-day median volume 1 lots), the micro JPY future was delisted mid-project, and its successor had a measured 22-month listing gap (Figure 11). Substitution candidates were evaluated under pre-registered correlation gates—MJY mapped to the yen leg at 0.951 monthly (gate 0.95), the 10-year micro-yield duration mapping reached 0.983 (gate 0.90), while the 30-year mapping's daily correlation

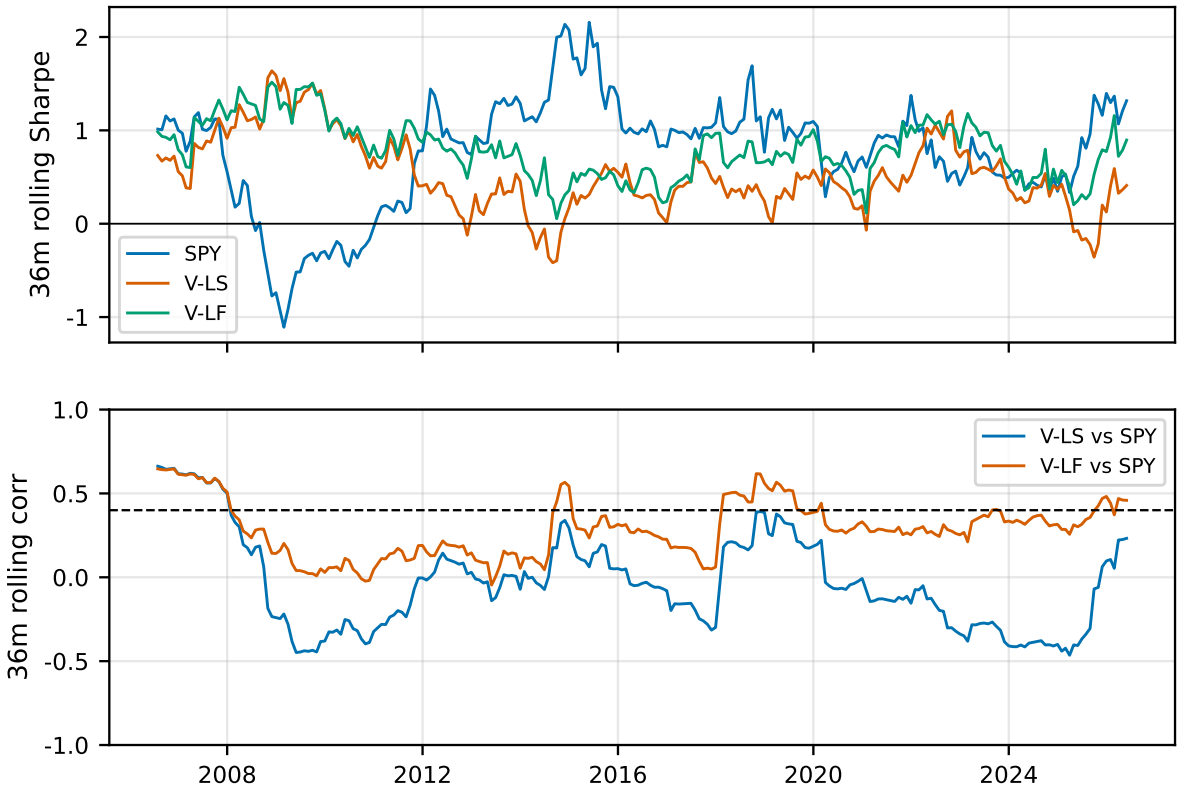

Figure 5: Rolling 36-month Sharpe and correlation to SPY.

of 0.824 flagged it as the weakest link before liquidity killed it outright. Evaluation never implied adoption: the final mapping uses native full-size Treasury futures.

8.3 Legislated thresholds, human capital decisions

A pre-committed rule (bond-leg zero-rounding $< 40\%$) selected the native full-size universe at \$1M (ZN 0.0%, ZB 17.5%) but not at \$500K (ZB 43.9%); the account scale itself was then *decoupled* from the threshold and decided by the human—a process lesson we record deliberately (Section 10). At the decided representative scale of \$500,000, per-leg sizing kept all six micro-capable legs micro (target-to-contract ratios at most 1.22 against a $2\times$ full-size rule), the yen-leg rule produced a statistical tie (TE difference of one basis point, both candidates inside the 5%-of-volume impact constraint; MJY mean impact 2.0% versus zero for the full-size contract)—resolved by *human adjudication on liquidity depth* in favor of the full-size contract, with both the rule outcome and the override recorded (Section 10). A duration-equivalent fold of the ZB leg into ZN was *rejected by its own pre-registered rule*: folding raised tracking error to 377 bp against 327 bp for accepting partial tracking loss, despite an acceptable combined zero-rounding share of 8.8% (hedge ratio $h = 1.72$).

Table 3: Implementation gap: ETF versus futures, overlap period.

level	variant	pair	strict	n_months	corr	sharpe_etf	sharpe_fut
asset	LS	ES/SPY	True	179	0.95	0.50	0.44
asset	LS	ZN/IEF	True	179	0.90	0.07	0.14
asset	LS	ZB/TLT	True	179	0.94	0.12	0.12
asset	LS	GC/GLD	True	179	0.94	0.24	0.20
asset	LS	CL/DBC	False	179	0.71	0.30	0.23
asset	LS	6E/UUP	False	179	0.73	-0.17	0.01
asset	LS	6J/UUP	False	179	0.38	-0.17	0.44
sleeve	LS	TSMOM-ETF/TSMOM-FUT	False	179	0.81	0.39	0.63
asset	LF	ES/SPY	True	179	0.97	0.68	0.62
asset	LF	ZN/IEF	True	179	0.95	0.16	0.21
asset	LF	ZB/TLT	True	179	0.94	0.21	0.24
asset	LF	GC/GLD	True	179	0.94	0.40	0.37
asset	LF	CL/DBC	False	179	0.70	0.19	0.12
asset	LF	6E/UUP	False	179	-0.12	0.12	-0.42
asset	LF	6J/UUP	False	179	-0.03	0.12	-0.28
sleeve	LF	TSMOM-ETF/TSMOM-FUT	False	179	0.80	0.72	0.67

Table 4: G16: native versus all-micro universes at two account scales. TE column is TE_{exec} (each universe versus its own fractional replay).

universe	nav	te_ann	cagr	sharpe_excess	max_drawdown	total_commission_usd
A_native	500000.000000	108	11.6	0.62	-13.4	1258.500000
A_native	1000000.000000	46	11.6	0.61	-14.6	2484.750000
B_micro	500000.000000	47	12.4	0.68	-14.5	2120.500000
B_micro	1000000.000000	24	12.3	0.67	-14.2	4224.500000

8.4 Tracking-error decomposition and the acceptance channel

The final configuration's TE_{model} is 327 bp/yr, decomposed in Figure 12: only 110 bp is TE_{exec} (rounding and commissions); the dominant 355 bp is *instrument basis*—micro contracts versus their full-size parents in roll timing and venue microstructure (components co-vary; they are not additive in quadrature). The shadow-run acceptance channel is anchored at $[261, 653]$ bp, i.e. $[0.8, 2.0]\times$ baseline: paper trading that lands inside this band is consistent with the model plus known basis, not evidence of decay.

9 Limitations and Honest Disclosures

The publication specification enumerates disclosure clauses that must each appear in this paper. They are collected here as subsections rather than scattered as hedges.

9.1 All results are in-sample

Every statistic in this paper is computed on historical data with full knowledge of the sample. Freezing parameters from literature priors before estimation removes one route of overfitting—the parameters could not adapt to this data—but it is not out-of-sample validation, and no

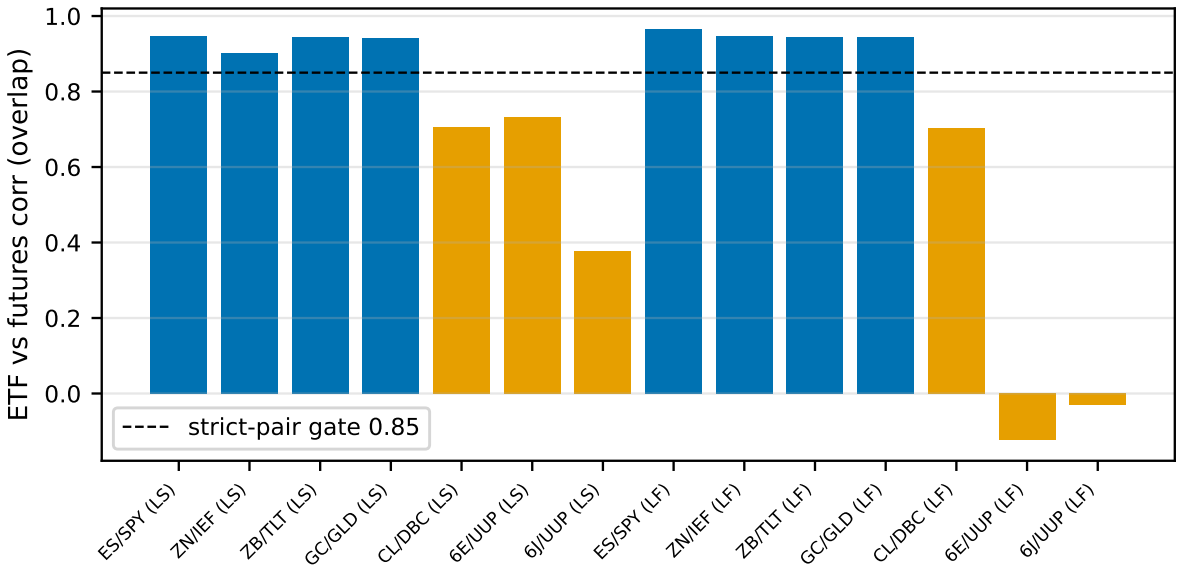

Figure 6: Per-pair strategy-return correlations; strict pairs highlighted against the 0.85 acceptance gate.

Table 5: G17 leg sizing under the 2× full-size rule.

leg	median_target_usd	full_cv_median_usd	ratio	selection
ES	68781.000000	240638.000000	0.286	micro
NQ	48979.000000	340075.000000	0.144	micro
GC	54799.000000	205720.000000	0.266	micro
CL	15526.000000	75880.000000	0.205	micro
6E	113967.000000	136225.000000	0.837	micro
6J	104982.000000	86119.000000	1.219	full (human adjudication)

claim of out-of-sample performance is made.

9.2 A pre-registered acceptance criterion FAILED

The v1.5 acceptance protocol required some stacking multiple to narrow maximum drawdown by at least ten percentage points versus pure SPY within the futures-era sample. None did: the best configuration narrowed drawdown by roughly six points. The structural reading, confirmed on the spliced long history, is that a monthly-rebalanced trend overlay on a *full* index position buffers deep equity crises by roughly 6–10 points at $k \leq 1.5$; larger relief requires de-risking the base, not just stacking more overlay.

9.3 The spliced series is not tradable

The 2003–2026 long history splices two different implementations (ETF sleeve, then futures sleeve; overlap-period sleeve correlation ≈ 0.81). It is a research view for crisis context, not a return stream anyone could have held.

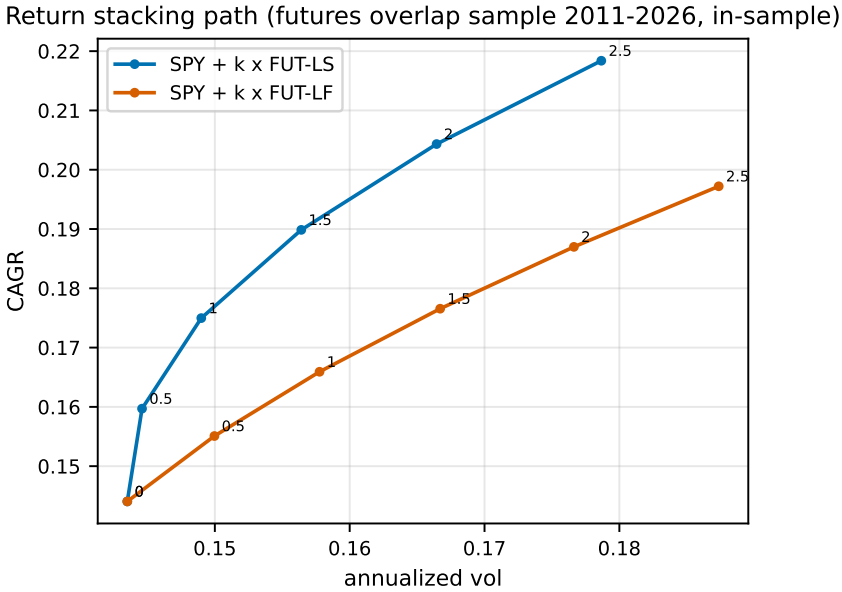

Figure 7: Stacking path in (vol, CAGR) space; k annotated. Futures overlap sample (2011–2026), in-sample moments.

9.4 Post-2015 softening, shown in full

V-LS: 0.73 → 0.40; V-LF: 0.96 → 0.73 across the frozen 2015-01-01 split. The fifty-year layer shows the same direction (1.22 → 0.96 around 2000, no-risk-free convention). We do not adjudicate between crowding, regime, and chance; we report the split as frozen.

9.5 Forward expectations belong near 0.8, not 0.96

The full-sample corner Sharpe (0.96 on the synthetic long history) is an in-sample, GFC-inclusive number. The honest forward anchor combines the post-split eras and execution drag: a corner-configuration Sharpe in the vicinity of 0.8, with tracking error of order 327 bp/yr around the model.

9.6 Discretization error is first-order at retail scale

At \$100K, execution-frame tracking error of 413 bp/yr amounts to roughly 63% of the overlay leg's volatility budget (6.6% at k=1.5)—execution noise at small scale is the same order as the risk being deployed. Below \$500K the bond legs are effectively untradable in full-size contracts.

9.7 The yield-futures detour was motivated by our own bug

The micro yield-futures substitution study was undertaken to solve a bond-leg execution problem that the 100× contract-value defect had exaggerated; the eventual liquidity death of the 30-year micro made the detour moot. We report it because the evaluation gates worked as designed, and because the episode illustrates how a measurement error can redirect research effort.

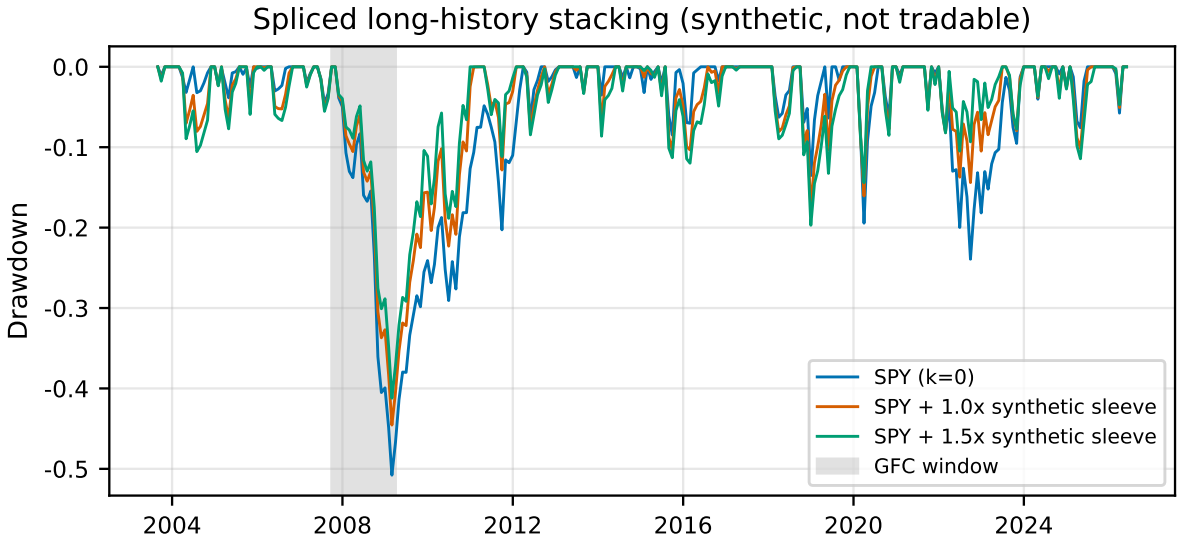

Figure 8: Drawdowns through the GFC on the spliced synthetic history.

9.8 Not investment advice

This is a research and educational document. Account sizes are representative scales; no brokerage details, actual funds, positions, or order timestamps are disclosed. Past performance does not predict future returns.

10 Erratum and Process Notes

Three process results are reported with the same care as performance results, because a pipeline's error behavior is an empirical property.

10.1 The 100× contract-value erratum

Detection. A scale comparison showed Treasury legs zero-rounded in *all* months at \$500K–\$1M—arithmetically implausible given known contract values. **Diagnosis.** The exchange definition field `unit_of_measure_qty` encodes *face value* for Treasury futures (quoted in percent-of-par), not point value; contract values were inflated 100× (apparent ZN value ≈\$10.9M versus true ≈\$110K). **Correction.** Contract values and tick costs are now derived as `price`×`qty`×s, with $s = 0.01$ when the unit-of-measure is a face-value currency—driven entirely by definition records, no hard-coding. **Impact restatement.** Table 6 fixes the pre-correction values (historical, non-regenerable by the fixed code) beside live post-correction values; the \$200K tracking error moved from 353 to 228 bp and the "plateau pinned by bond legs" interpretation was withdrawn in favor of "bond legs execute from ∼\$500K."

10.2 Two falsified predictions

(i) We predicted folding the 30-year leg into the 10-year leg (duration-equivalent execution) would reduce tracking error; it *raised* it by 50 bp—the constant-hedge-ratio basis cost exceeds

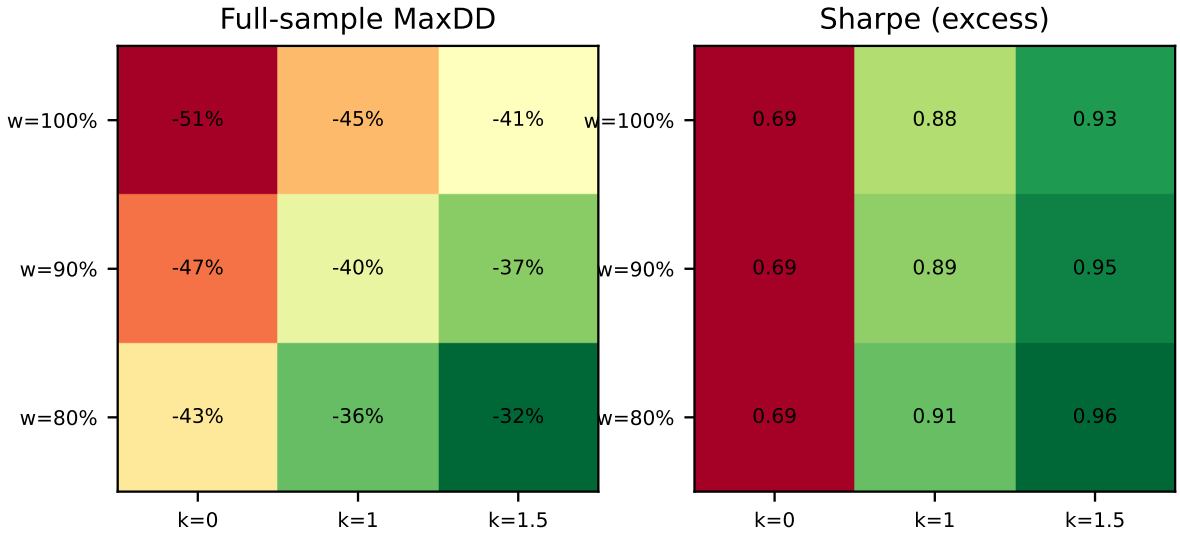

Figure 9: Decision geometry: base allocation × overlay multiple.

the rounding loss it removes. (ii) We predicted micro-yield substitution would be the bond-leg solution; corrected contract values plus an instrument death falsified the premise. Both predictions were written down before the data arrived, which is why their failure is reportable rather than invisible.

10.3 A cost-control incident in our own pipeline

Two compounding defects—a cache key containing the live dataset end-date, and per-process budget baselines that could not see each other—caused cumulative data spend to briefly exceed the then-authorized cap, which was revised upward once the breach was disclosed. The fix is structural: a persistent append-only spend ledger that every guard reads at initialization, incremental tail top-ups, and narrowed refresh semantics. The single-process estimate-first discipline had worked throughout; the failure surface was cross-process accounting. The dollar amount is immaterial; the reportable result is the failure mode—locally correct guards in a multi-process agent pipeline need not compose into a correct global constraint. We include it because cost discipline is part of reproducibility, and because this is exactly the kind of result selective reporting would omit.

10.4 Workflow: arithmetic can be legislated; capital cannot

The project ran as a frozen-specification, agent-implemented, human-adjudicated workflow. Pre-registered thresholds worked well for execution questions (roll rules, substitution gates, sizing rules)—they prevented in-flight rationalization. They failed once by *overreaching*: a threshold written for execution feasibility briefly entangled the account size itself, and the human decision-maker had to decouple capital allocation from the rule. A second, smaller instance: the yen-leg instrument rule ended in a one-basis-point tie—statistical noise—and was settled by human adjudication on liquidity depth, with the rule outcome and the override both recorded in the decision grid. The division of labor we ended with—machines hold the frozen arithmetic, humans

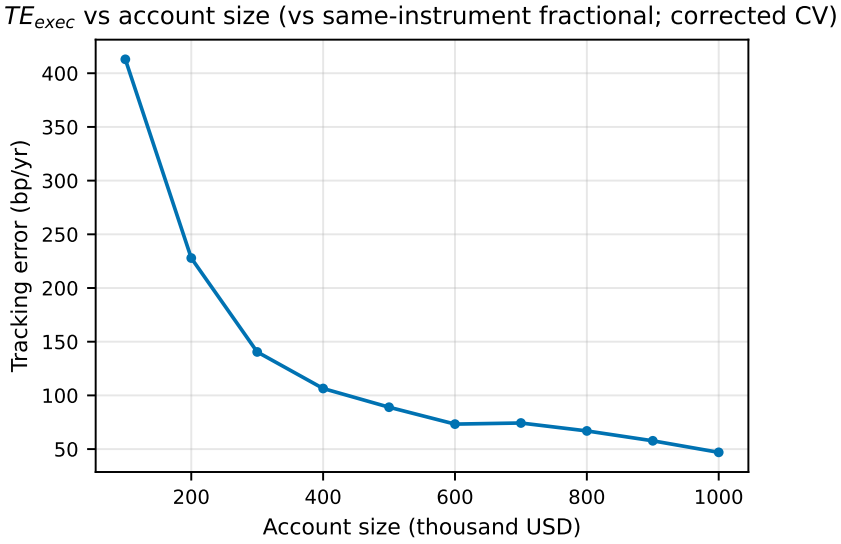

Figure 10: Tracking error versus account size, corrected contract values.

hold the capital and the spec-change pen—is, in our view, the transferable methodological result.

11 Conclusion

A volatility-targeted time-series momentum overlay, frozen to literature parameters and audited end-to-end, improves an index holder's risk-adjusted position in-sample at every layer we measured: as an ETF sleeve (Sharpe 0.57, correlation -0.04 to SPY, winner in 6/6 crisis windows), as a futures replication (strict-pair correlations 0.90–0.97), and as an integer-contract execution at a representative $500,000 account (tracking error 327 bp/yr, dominated by instrument basis rather than rounding). The descent from paper to ticket cost real performance at every step, and measuring those steps— rather than asserting the top-line number—is what this paper adds.

The repository behind this paper rebuilds every number, table, and figure from raw outputs with one command, fails on any uninjected number, and reproduces byte-identically across consecutive builds. The mistakes we made along the way are in Section 10 with their detection paths and corrections. We commend the format: in systematic investing, the audit chain is the product.

References

Abhilash Babu, Ari Levine, Yao Hua Ooi, Lasse Heje Pedersen, and Erik Stamelos. Trends everywhere. *Journal of Investment Management*, 18(1):52–68, 2020.

Nick Baltas and Robert Kosowski. Demystifying time-series momentum strategies: Volatility estimators, trading rules and pairwise correlations. In *Market Momentum: Theory and Practice*, chapter 3. Wiley, 2020. doi: 10.1002/9781119599364.ch3. Earlier working paper: SSRN 2140091.

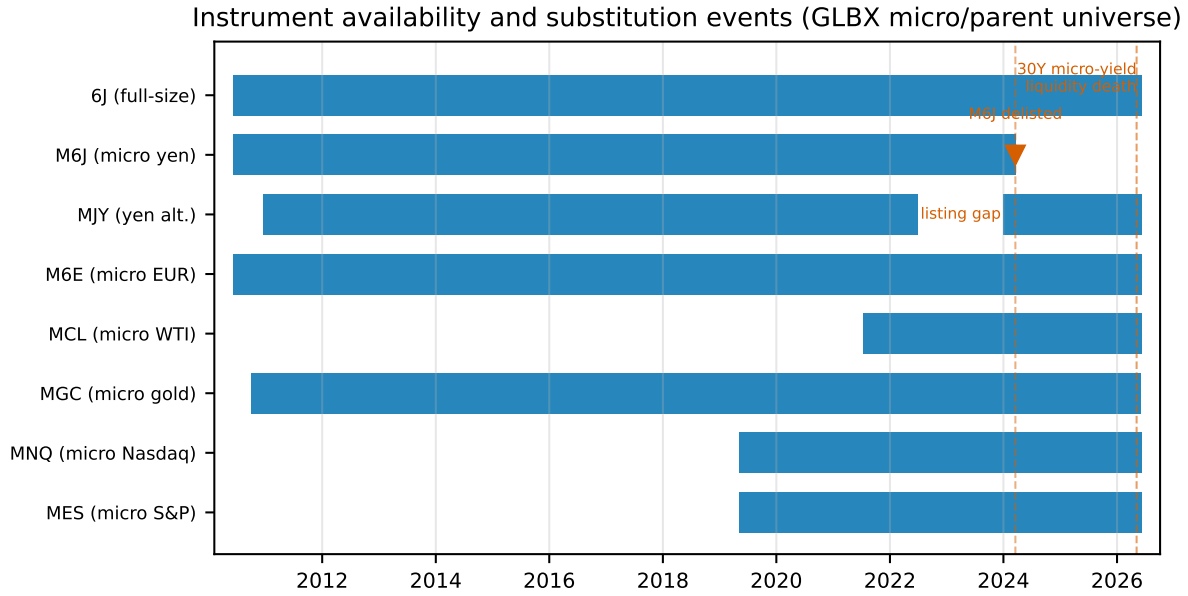

Figure 11: Instrument availability and substitution events across the futures universe: micro-contract listing dates (Databento definition probes), the mid-project M6J micro-yen delisting, the MJY listing gap, and the 30-year micro-yield liquidity death. All dates are injected from the data pipeline (discretization probe, liquidity-health snapshot, and the provenance ledger).

Brian Hurst, Yao Hua Ooi, and Lasse Heje Pedersen. A century of evidence on trend-following investing. *Journal of Portfolio Management*, 44(1):15–29, 2017. doi: 10.3905/jpm.2017.44.1.015.

Tobias J. Moskowitz, Yao Hua Ooi, and Lasse Heje Pedersen. Time series momentum. *Journal of Financial Economics*, 104(2):228–250, 2012. doi: 10.1016/j.jfineco.2011.11.003.

Valeriy Zakamulin and Javier Giner. Trend following with momentum versus moving averages: A tale of differences. *Quantitative Finance*, 20(6):985–1007, 2020. doi: 10.1080/14697688.2020.1716057.

A Assumptions Register

Every detail the strategy specification left undefined was resolved by the *simplest reasonable* implementation and recorded at decision time. That register—47 logged decisions spanning categories A–G (the v1.5 futures module is G1–G19)—is reproducibility-supporting material: each decision that bears on a result is already stated where it arises in the main text, and the register is the complete per-item log behind them. It is kept in the working language of the research log (Chinese) and reproduced verbatim as the Chinese edition's Appendix A; it is published in the same repository as this paper (`ASSUMPTIONS.md`) and is regenerated into that appendix by a single build command, consistent with the reproducibility claim of this work.

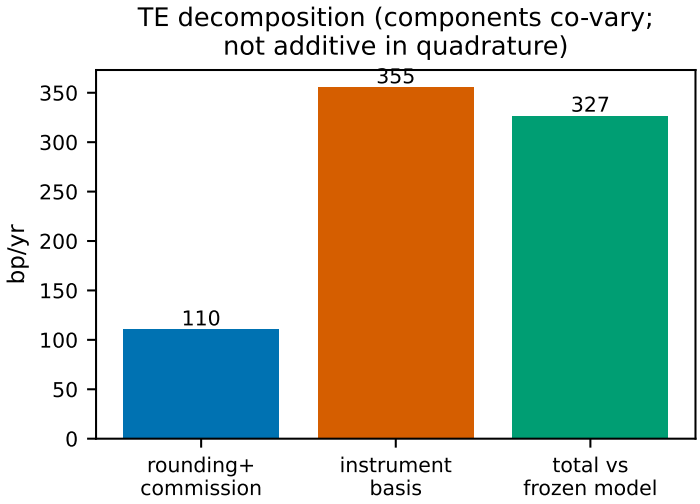

Figure 12: TE decomposition at the final configuration.

Table 6: Erratum record: pre-correction values are curated historical facts with commit references; post-correction values are read live from current outputs by the build system.

metric	scope	value_before	unit	commit_fix
te_full_universe	G12 at $100K	424	bp	11a2ae7
te_full_universe	G12 at $200K	353	bp	11a2ae7
te_reduced_universe	G12 reduced universe at $200K	250	bp	11a2ae7
te_sweep_plateau	G12 sweep plateau ($500K-$1M)	298-310	bp	11a2ae7
bond_zero_share_zn	G16 native universe at $500K	100.0	%	11a2ae7
bond_zero_share_zb	G16 native universe at $500K	100.0	%	11a2ae7
bond_zero_share_zn	G16 native universe at $1M	100.0	%	11a2ae7
bond_zero_share_zb	G16 native universe at $1M	100.0	%	11a2ae7
te_a_native	G16 native at $500K (TE)	251	bp	11a2ae7
te_a_native	G16 native at $1M (TE)	247	bp	11a2ae7
zn_contract_value_apparent	apparent value during G12-G16	1.09e7	USD	11a2ae7

B Roll Statistics

C Full Robustness Grid

variant	lookback_m	vol_window_d	leverage_cap	start	n_months	sharpe_net	ann_return	max_drawdown
LS	3	40	1.00	2002-11-30	283	0.45	3.9	-13.1
LS	3	40	1.50	2002-11-30	283	0.40	3.8	-14.8
LS	3	40	2.00	2002-11-30	283	0.38	3.7	-15.6
LS	3	60	1.00	2002-11-30	283	0.42	3.8	-13.4
LS	3	60	1.50	2002-11-30	283	0.38	3.7	-14.7
LS	3	60	2.00	2002-11-30	283	0.36	3.6	-15.3
LS	3	90	1.00	2003-01-31	281	0.42	3.8	-12.9
LS	3	90	1.50	2003-01-31	281	0.39	3.7	-13.7
LS	3	90	2.00	2003-01-31	281	0.37	3.7	-14.0

Continued on next page

variant	lookback_m	vol_window_d	leverage_cap	start	n_months	sharpe_net	ann_return	max_drawdown
LS	6	40	1.00	2003-02-28	280	0.45	3.8	-10.1
LS	6	40	1.50	2003-02-28	280	0.42	3.8	-10.2
LS	6	40	2.00	2003-02-28	280	0.40	3.7	-10.8
LS	6	60	1.00	2003-02-28	280	0.45	3.8	-9.7
LS	6	60	1.50	2003-02-28	280	0.42	3.7	-9.9
LS	6	60	2.00	2003-02-28	280	0.41	3.7	-10.3
LS	6	90	1.00	2003-02-28	280	0.46	3.8	-9.4
LS	6	90	1.50	2003-02-28	280	0.44	3.8	-9.3
LS	6	90	2.00	2003-02-28	280	0.43	3.9	-9.6
LS	9	40	1.00	2003-05-31	277	0.66	4.7	-6.9
LS	9	40	1.50	2003-05-31	277	0.65	4.9	-7.9
LS	9	40	2.00	2003-05-31	277	0.64	4.9	-8.3
LS	9	60	1.00	2003-05-31	277	0.64	4.6	-6.9
LS	9	60	1.50	2003-05-31	277	0.63	4.8	-7.8
LS	9	60	2.00	2003-05-31	277	0.63	4.8	-8.2
LS	9	90	1.00	2003-05-31	277	0.63	4.6	-6.6
LS	9	90	1.50	2003-05-31	277	0.63	4.7	-7.3
LS	9	90	2.00	2003-05-31	277	0.63	4.8	-7.6
LS	12	40	1.00	2003-08-31	274	0.62	4.6	-8.8
LS	12	40	1.50	2003-08-31	274	0.57	4.6	-9.6
LS	12	40	2.00	2003-08-31	274	0.53	4.4	-10.1
LS	12	60	1.00	2003-08-31	274	0.60	4.5	-8.0
LS	12	60	1.50	2003-08-31	274	0.56	4.4	-9.2
LS	12	60	2.00	2003-08-31	274	0.53	4.3	-9.6
LS	12	90	1.00	2003-08-31	274	0.60	4.4	-8.2
LS	12	90	1.50	2003-08-31	274	0.57	4.5	-9.7
LS	12	90	2.00	2003-08-31	274	0.54	4.4	-10.0
LS	15	40	1.00	2003-11-30	271	0.44	3.7	-12.1
LS	15	40	1.50	2003-11-30	271	0.39	3.6	-12.4
LS	15	40	2.00	2003-11-30	271	0.36	3.5	-12.5
LS	15	60	1.00	2003-11-30	271	0.43	3.6	-11.5
LS	15	60	1.50	2003-11-30	271	0.39	3.5	-11.8
LS	15	60	2.00	2003-11-30	271	0.37	3.5	-11.8
LS	15	90	1.00	2003-11-30	271	0.43	3.6	-10.9
LS	15	90	1.50	2003-11-30	271	0.40	3.6	-11.1
LS	15	90	2.00	2003-11-30	271	0.38	3.5	-11.1
LS	18	40	1.00	2004-02-29	268	0.59	4.3	-9.9
LS	18	40	1.50	2004-02-29	268	0.55	4.3	-11.6
LS	18	40	2.00	2004-02-29	268	0.51	4.2	-12.3
LS	18	60	1.00	2004-02-29	268	0.57	4.2	-10.1
LS	18	60	1.50	2004-02-29	268	0.53	4.2	-11.3
LS	18	60	2.00	2004-02-29	268	0.50	4.1	-11.9
LS	18	90	1.00	2004-02-29	268	0.58	4.2	-9.3
LS	18	90	1.50	2004-02-29	268	0.55	4.3	-11.3
LS	18	90	2.00	2004-02-29	268	0.53	4.2	-11.9
LF	3	40	1.00	2002-11-30	283	0.82	4.9	-6.3
LF	3	40	1.50	2002-11-30	283	0.79	4.9	-6.6
LF	3	40	2.00	2002-11-30	283	0.77	4.9	-6.7
LF	3	60	1.00	2002-11-30	283	0.81	4.8	-6.3
LF	3	60	1.50	2002-11-30	283	0.78	4.8	-6.5
LF	3	60	2.00	2002-11-30	283	0.77	4.9	-6.6
LF	3	90	1.00	2003-01-31	281	0.82	4.8	-6.1
LF	3	90	1.50	2003-01-31	281	0.79	4.8	-6.3
LF	3	90	2.00	2003-01-31	281	0.79	4.8	-6.3
LF	6	40	1.00	2003-02-28	280	0.81	4.8	-3.9
LF	6	40	1.50	2003-02-28	280	0.79	4.9	-4.3
LF	6	40	2.00	2003-02-28	280	0.78	4.9	-4.6
LF	6	60	1.00	2003-02-28	280	0.81	4.8	-3.9
LF	6	60	1.50	2003-02-28	280	0.79	4.9	-4.4
LF	6	60	2.00	2003-02-28	280	0.79	4.9	-4.4
LF	6	90	1.00	2003-02-28	280	0.81	4.8	-3.9
LF	6	90	1.50	2003-02-28	280	0.81	4.9	-4.3
LF	6	90	2.00	2003-02-28	280	0.81	5.0	-4.3

Continued on next page

variant	lookback_m	vol_window_d	leverage_cap	start	n_months	sharpe_net	ann_return	max_drawdown
LF	9	40	1.00	2003-05-31	277	0.93	5.3	-3.8
LF	9	40	1.50	2003-05-31	277	0.92	5.5	-4.4
LF	9	40	2.00	2003-05-31	277	0.92	5.5	-4.6
LF	9	60	1.00	2003-05-31	277	0.92	5.2	-3.9
LF	9	60	1.50	2003-05-31	277	0.91	5.4	-4.4
LF	9	60	2.00	2003-05-31	277	0.92	5.5	-4.5
LF	9	90	1.00	2003-05-31	277	0.91	5.1	-3.9
LF	9	90	1.50	2003-05-31	277	0.91	5.3	-4.3
LF	9	90	2.00	2003-05-31	277	0.92	5.4	-4.3
LF	12	40	1.00	2003-08-31	274	0.91	5.2	-5.0
LF	12	40	1.50	2003-08-31	274	0.88	5.3	-5.9
LF	12	40	2.00	2003-08-31	274	0.85	5.3	-6.0
LF	12	60	1.00	2003-08-31	274	0.90	5.1	-4.9
LF	12	60	1.50	2003-08-31	274	0.88	5.2	-5.6
LF	12	60	2.00	2003-08-31	274	0.86	5.2	-5.7
LF	12	90	1.00	2003-08-31	274	0.89	5.1	-5.0
LF	12	90	1.50	2003-08-31	274	0.89	5.2	-5.8
LF	12	90	2.00	2003-08-31	274	0.88	5.2	-5.8
LF	15	40	1.00	2003-11-30	271	0.78	4.7	-5.6
LF	15	40	1.50	2003-11-30	271	0.75	4.7	-6.5
LF	15	40	2.00	2003-11-30	271	0.73	4.7	-6.6
LF	15	60	1.00	2003-11-30	271	0.77	4.6	-5.4
LF	15	60	1.50	2003-11-30	271	0.75	4.7	-6.3
LF	15	60	2.00	2003-11-30	271	0.74	4.7	-6.4
LF	15	90	1.00	2003-11-30	271	0.76	4.6	-5.6
LF	15	90	1.50	2003-11-30	271	0.75	4.7	-6.4
LF	15	90	2.00	2003-11-30	271	0.75	4.7	-6.5
LF	18	40	1.00	2004-02-29	268	0.81	4.8	-6.5
LF	18	40	1.50	2004-02-29	268	0.78	4.9	-7.8
LF	18	40	2.00	2004-02-29	268	0.76	4.8	-7.9
LF	18	60	1.00	2004-02-29	268	0.80	4.7	-6.4
LF	18	60	1.50	2004-02-29	268	0.78	4.8	-7.6
LF	18	60	2.00	2004-02-29	268	0.76	4.8	-7.7
LF	18	90	1.00	2004-02-29	268	0.80	4.7	-6.6
LF	18	90	1.50	2004-02-29	268	0.79	4.8	-7.6
LF	18	90	2.00	2004-02-29	268	0.79	4.9	-7.7

D Shadow Ticket Sample

A *historical* sample month (June 2024 holding period, signals as of May 2024) generated under the final instrument mapping, shown to illustrate the paper-record format only. Consistent with the disclosure policy, this is *not* a current or intended position; the operational monthly tickets are not published.

leg	instrument	contract	close_used	multiplier	cv_usd	n_contracts	status
ES	MES	MES-2024-06	5299.500000	5.000000	26498.000000	3	OK
NQ	MNQ	MNQ-2024-06	18571.750000	2.000000	37144.000000	2	OK
GC	MGC	MGC-2024-06	2325.500000	10.000000	23255.000000	3	OK
CL	MCL	MCL-2024-07	77.160000	100.000000	7716.000000	7	OK
6E	M6E	M6E-2024-06	1.085600	12500.000000	13570.000000	-3	OK
6J	6J	6J-2024-06	0.006371	12500000.000000	79631.000000	-1	OK
ZN	ZN	ZN-2024-09	108.890625	100000.000000	108891.000000	-1	OK
ZB	ZB	ZB-2024-09	116.281250	100000.000000	116281.000000	-1	OK

Table 7: Continuous-contract construction stats.

root	n_contracts	n_rolls	rolls_volume	rolls_expiry	series_start	series_end	n_days	n_gap_issues
ES	68	65	43	22	2010-06-07	2026-06-07	4964	0
NQ	71	65	39	26	2010-06-07	2026-06-07	4964	0
ZN	67	66	65	1	2010-06-07	2026-06-07	4968	0
ZB	67	66	65	1	2010-06-07	2026-06-07	4968	0
GC	112	98	97	1	2010-06-07	2026-06-07	4965	0
CL	245	193	192	1	2010-06-07	2026-06-07	4967	0
6E	71	65	0	65	2010-06-07	2026-06-05	4963	0
6J	70	65	0	65	2010-06-07	2026-06-05	4963	0