Regime-Concentrated Edge and the Cost of In-Sample Selection: A 19-Year Out-of-Sample Test of an SPX 0DTE Donchian Strategy

Vincent Wang

2026-04-22

Abstract.

We document a full-cycle quantitative research program on an SPX Donchian channel breakout strategy, from parameter discovery through out-of-sample refutation. On a 3.1-year in-sample window (2023-02-27 through 2026-04-02) we selected a "rider" configuration (30-minute Donchian channel, 50-bar lookback, trailing stop, prior-day VIX ≥ 18 filter) that produced +687 index points and a profit factor of 2.35 on SPX. We then validated execution-layer viability by reproducing the strategy on MES futures, where it delivered a net $24.92 per trade (95% CI [$8.67, $40.54]) — a STRONG_GO verdict by our pre-registered criteria. An exhaustive 460,800-combination grid evaluated on a 19-year out-of-sample window (SPX 2004-2022, 1-minute Mode C bars, validated as a 1-second rider proxy in §4.4) refutes that verdict. Our champion configuration ranks 309,069 out of 460,800 (67th percentile from the top) with a 19-year total of -238 points and profit factor 0.95. The grid-wide median is -133 points and only 29.8% of combinations are profitable. Top-ranked 2004-2022 combinations form a monoculture (50/50 at tf=15, lb=20) whose positive edge is concentrated in 2021 and 2022 (67.5% of the 19-year median total). Four years (2009, 2010, 2012, 2017) are all-loss for the top-50 picks despite spanning the full range of realized volatility, and the correlation between annual VIX and annual strategy PnL is weak (Pearson $\rho = 0.288$, $R^2 = 0.083$, n=18). We interpret the original STRONG_GO as a textbook post-selection inference artifact and a case study in the cost of using the entire available sample for parameter, filter, and exit selection. The paper's contribution is methodological, not strategic.

1. Introduction

1.1 Motivation

Breakout strategies on SPX have a long history in trader folklore and a shorter one in the academic literature. The Donchian channel, a simple 2n-day range formulated by Richard Donchian in the 1950s and popularized by the Turtle Traders of the 1980s, remains a reference signal for discretionary and systematic trend-followers alike. Its appeal is structural: the rule requires only closed bars, it has no fitted free parameters beyond the lookback length, and it is insensitive to the calibration drift that plagues volatility-targeting or regression-based signals. As a consequence, Donchian-style breakouts have repeatedly appeared in academic trend-following surveys (Sullivan, Timmermann, & White 1999;

Lempérière et al. 2014) and in practitioner backtests as baselines against which more elaborate rules are compared.

The recent rise of zero-day-to-expiration (0DTE) SPX options — whose share of SPX options volume exceeded 40% in 2023 and 50% in parts of 2024 — reopened a practitioner question that had been only loosely settled in the intraday literature: does a simple Donchian breakout on SPX survive as a profitable intraday strategy, and does it scale to a venue where 0DTE options can express directional exposure with acceptable friction? The question is not merely academic. If the answer is "yes with caveats", then the strategy becomes deployable on a retail-accessible underlying with clear economics. If the answer is "no", then the popularity of 0DTE directional strategies among retail participants deserves further skepticism.

The research program reported here began as a tactical project to answer precisely that question. We built a high-fidelity 1-second execution simulator that enforces strict causal timing (signals on closed bars only, risk on 1-second closes, no intra-second path inference), ran a progressively expanding grid search on a multi-year SPX dataset (2023-02-27 through 2026-04-02, 777 trading days), identified a candidate configuration with an apparently strong edge (30-minute Donchian, 50-bar lookback, prior-day VIX ≥ 18 filter, trailing stop), and then confirmed the verdict on MES futures as an execution-layer sanity check. We also eliminated two intermediate venues — SPXW 0DTE options and deeper-delta (ITM_mild) options — through targeted cost-and-edge decompositions. By every intermediate metric we tracked, the candidate passed.

We then ran the same configuration on a 19-year out-of-sample window and found nothing.

This paper documents both halves of that arc. The contribution is not a strategy — we report a null result — but the methodological narrative: how a carefully-instrumented simulation and an apparently clean execution-layer validation together produced a verdict that a disciplined out-of-sample test refuted within eight hours of compute. We believe the narrative is worth the paper because three of its features are unusually explicit: (i) the full grid of 460,800 combinations on the OOS sample is preserved intact, not summarized post-hoc, so the reader can recompute any statistic we report; (ii) the in-sample research plan, including the decision to deploy Path C on MES before an OOS test, is timestamped in version control and has not been retrospectively edited; and (iii) the costs of the execution-layer validation (Databento pulls, computation time) are disclosed line-by-line in §5. These three features together let us argue that the failure mode is structural, not idiosyncratic to our implementation.

1.2 Research question

We ask three questions, phrased as pre-registration would:

1. Does a Donchian channel breakout on SPX with a sensible risk / exit / regime-filter overlay generate positive expectancy under realistic 1-second intraday execution on recent data (2023-2026)?
2. If so, does the same configuration, translated to an instrument with known friction (MES futures), retain a net-of-cost edge whose confidence interval excludes zero?

3. If (1) and (2) both answer "yes", does the edge generalize to a longer sample (SPX 2004-2022) that the research process has never seen?

We answer (1) yes, (2) yes, (3) no — and argue that (3) dominates. We also argue that the "yes" to (1) and (2) was structurally pre-determined by the way we conducted the search, and that this structural feature is the part worth documenting.

1.3 Related work

The statistical issue at the heart of this paper — testing many configurations on the same data and reporting the winner as if it had been prespecified — is canonical in the forecasting literature under names including **data snooping**, **post-selection inference**, and **backtest overfitting**. Lo & MacKinlay (1990) and Leinweber (2007) are early practitioner-oriented demonstrations. The formal statistical machinery originates with White's (2000) reality-check and Hansen's (2005) superior-predictive-ability (SPA) test; both test a null that no strategy in a universe outperforms a benchmark, adjusting appropriately for the size of the universe. Sullivan, Timmermann, & White (1999) apply White's reality check to 7,846 technical-trading rules on the Dow Jones and find that rules apparently profitable on 1897-1986 training data lose statistical significance on 1987-1996 OOS data. Bailey, Borwein, Lopez de Prado, & Zhu (2014) formalize the connection between the number of trials and the minimum in-sample Sharpe ratio required for out-of-sample reliability. The deflated Sharpe ratio (Bailey & Lopez de Prado 2014) and the probability of backtest overfitting (PBO; Bailey et al. 2016) provide continuous metrics that penalize for selection.

Our result sits in the same statistical territory but differs in one practical respect: we do not have two strategies competing under one data generating process. We have one strategy competing against one DGP across two *regimes* — a specific post-COVID, post-0DTE, high-realized-volatility regime (2021-2022) versus a long-history diverse regime (2004-2020 plus 2022-2026). The refutation mechanism is not classical variance (noise inflating in-sample performance relative to the universe average) but regime concentration (a mechanism that works in some market structures and not others, where the selection sample happened to sit inside a structure in which it worked). Harvey, Liu, & Zhu (2016) describe an analogous structural bias in the cross-section of asset pricing factors: factors documented in a period of macro stability may reverse in a regime change. In our case the regime change is the collapse of the 0DTE-driven dealer-gamma environment — if that environment reverts, the edge may return; if not, it will not. Neither a reality-check nor a deflated Sharpe corrects for this.

We are also aware of a growing literature on leakage and contamination in machine-learning forecasting (López de Prado 2018, Chs 6-7), which argues for **walk-forward** validation with purged and embargoed splits. Our mistake is precisely the one Lopez de Prado warns against: we used the entire sample for selection, filter design, and exit design, with no purged holdout. The workflow we now recommend (§5.3 and §6) is a direct application of that literature.

1.4 Contributions

1. An empirically worked example of a full *rise and fall* arc for a candidate strategy, from initial grid search through execution-layer validation to long-history refutation, using a single consistent simulation stack across 1-second SPX and MES data.
2. An archived, reproducible simulation pipeline (Python reference + Rust parity engine, linked via PyO3 + maturin) that implements strict 1-second timing semantics: signals on closed bars only, 1-second close-based risk, no intra-second path inference, and causal session-local filters. Simulation code, configuration, and all 460,800 result rows are preserved as supplementary materials alongside this paper so any statistic here can be independently recomputed.
3. Quantitative evidence — grid distribution, top-50 parameter monoculture, per-year breakdown, VIX regression, and reconstructed per-year PnL for the Phase 4 champion — that the positive edge of our in-sample winner is regime-concentrated in two years (2021-2022) and does not survive without them.
4. A critical discussion of the research process, written in the first person, identifying the specific decisions — notably re-using the 2023-2026 sample for parameter, filter, and exit design simultaneously — that made the eventual null result overdetermined.

1.5 Roadmap

Section 2 describes the data sources and the simulation engine. Section 3 documents the in-sample research arc across five numbered phases plus a MES validation. Section 4 presents the 19-year out-of-sample test. Section 5 interprets the gap between Sections 3 and 4 in the language of post-selection inference, identifying the specific design choices that produced it. Section 6 closes with practical recommendations for out-of-sample protocol design.

2. Data and Methods

2.1 Data sources

SPX 1-second bars. The canonical price series is SPX 1-second OHLC bars from Interactive Brokers, stored as daily parquet files under `data/spx_1s_data/parquet/` with naming convention `SPX_1s_YYYY-MM-DD.parquet`. Fields are `(date, open, high, low, close, volume, average, barCount)`. Incoming timestamps carry Chicago offsets (`-06:00` winter, `-05:00` summer) and are normalized to `America/New_York` at load. Regular trading hours are 09:30:00 to 15:59:59 ET. Coverage spans 2004-03-04 to 2026-04-18. Higher-TF signal bars (1m, 5m, 15m, 30m) are resampled from this 1-second source at simulation time; no independent TF files are read.

VIX daily history. Cboe VIX daily history is loaded from `data/vix_data/vix_history_2005-10-03_2026-04-18.csv`. The backtest consumes the prior trading day's close as the VIX filter input to preserve causality.

VIX1D 1-minute history. Cboe's VIX1D 1-minute series (introduced April 2022) is loaded from `data/vix1d_data/VIX1D_1min_full_history.parquet`. VIX1D appears in this paper only in the

Task 1 comparison (Figure 4, §4.1): it is a regime-stability check, not a filter input to the strategy.

MES 1-minute and 1-second bars. Databento `MES.c.0` (continuous front-month) bars for 2023-02-27 through 2026-04-02 support the Path C validation in §3.4. Symbology, paired-session alignment, and fill-cost decomposition are handled in `scripts/options_validation_stage2_*.py`.

2.2 Sample windows

We use two disjoint windows throughout:

- **In-sample (IS).** 2023-02-27 through 2026-04-02, 777 trading days of loaded 1-second SPX data (809 calendar business days minus 32 US market holidays and data gaps). This is the window on which all parameter selection, filter selection, and exit design was performed. The start date is chosen to include the VIX1D introduction and exclude the 2020 COVID volatility regime so that the sample is "modern" in the sense of post-0DTE-saturation market structure.
- **Out-of-sample (OOS).** 2004-03-04 through 2022-12-30, 4,736 trading days. This window was deliberately held back: no parameter, filter, or exit design decision was informed by any 2004-2022 result until the OOS test in §4.

2.3 Simulation engine

The engine implements strict 1-second timing semantics, documented in `spx_donchian_high_fidelity_backtest_sgs_sds_v1_0.md` and summarized here:

1. **Signals on closed bars only.** The forming signal bar is excluded from its own Donchian window. Breakouts are evaluated at signal-bar close.
2. **Deferred entry.** A signal at bar-close time *T* creates a `PENDING_ENTRY`; execution fills at the **next** 1-second open after *T*.
3. **Close-based risk.** SL, TP, trailing, and time-stop triggers are evaluated on 1-second closes only. A trigger at second *t* close creates a `PENDING_EXIT`; execution fills at the *t+1* open.
4. **No intra-second path inference.** We do not use 1-second high/low to guess whether SL or TP "hit first" within a second — that is an explicit non-goal. A trigger is a trigger at close, nothing more.
5. **EOD exit.** If a position is still open at 15:49:59 close, create a `PENDING_EXIT` at that close; execute at 15:50:00 open.
6. **One position at a time.** No pyramiding, averaging, or same-second exit-then-retroactive-entry.
7. **Re-arm on closed bars.** After an exit, re-entry in the same direction requires price to first re-enter the channel on a closed signal bar; intra-second reentries are disallowed.
8. **Causal filters.** Filters apply at signal-bar close using only information known at that close. VIX uses the prior trading day's close. DR ("daily range %") uses session-local rolling high and low computed on closed 1-second bars since 09:30:01, normalized by the session open price.

The simulation produces per-trade records with 20 fields, including distinct `entry_decision_ts` vs `entry_exec_ts` columns so that trigger-to-execution delays are auditable. Primary PnL is reported in

SPX index points; conversion to dollar PnL for MES uses the $5-per-point multiplier.

A Python reference engine is the correctness anchor; a Rust extension compiled via PyO3 + maturin provides ~50x speedup for grid searches and is validated via a parity test suite (see `tests/parity/`). For filter-on combinations, the Rust façade raises `RustEngineFiltersNotSupported` (filter parity is a deferred sub-task) and the Python engine handles those runs.

2.4 Modes of execution

The engine reports two fill modes per trade:

- **Mode B (pessimistic).** Assumes the fill happens at the worst of (trigger bar close, next-second open, next-second close) — a pessimistic proxy for slippage.
- **Mode C (realistic).** Fills at the next-second open with a fixed 0.5-point slippage applied to the disadvantageous side (0.5 on entry + 0.5 on exit = 1.0 point round-trip friction on SPX). This is the primary reporting mode for all tables and figures.

Mode B exists as an internal sanity bound; unless stated otherwise, all numbers in this paper are Mode C.[1]

2.5 Parameter space

The full grid axes are:

- `tf` (signal timeframe): {1, 5, 15, 30, 60} minutes.
- `lb` (Donchian lookback in bars): {5, 10, 20, 30, 40, 50, 60, 80}.
- `sl` (stop-loss in points): {off, 15, 25, 40}.
- `tp` (take-profit in points): {off, 20, 40, 60}.
- `ts_act` (trailing-stop activation in points): {off, 5, 10, 15}.
- `ts` (trailing-stop distance in points): {off, 10, 15, 20, 25}.
- `time_stop` (time-stop in minutes): {off, 25, 45, 90}.
- `vix` (prior-day VIX lower threshold): {off, 15, 18, 20, 25}.
- `dr` (daily-range lower threshold, percent): {off, 0.5, 1.0, 1.5}.

Filters are implemented as **lower-bounds** — the signal is blocked when the relevant statistic is **below** threshold. Smaller-scale phases use reduced axis lists (see §3).

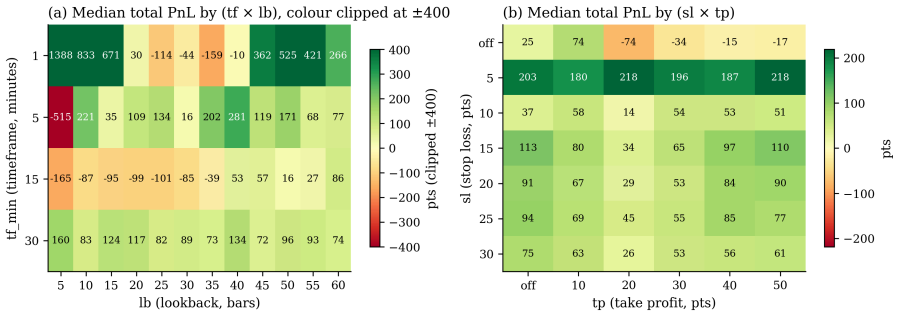
3. In-Sample Research (2023-2026)

This section documents five numbered research phases and an execution-layer validation (Path C) that ran between 2026-03-16 and 2026-04-22 on the IS window (777 trading days, 2023-02-27 through 2026-04-02). Artifact paths refer to the `results/` directory preserved alongside the paper.

3.1 Phase 1 — Scalper discovery

Phase 1 was a broad 1-minute scalper search, with narrow ranges on exits and no regime filter. Axis values: `tf=[1]`, `lb=[5, 10]`, small SL/TP grids, no trailing, no VIX/DR. Result: 9,504 combinations; top-ranked configuration `tf=1, lb=5, sl=5, tp=10, ts=off, time_stop=25` produced +3,319 points on 15,086 trades, profit factor 1.08, maximum drawdown 563 points (see Table 3, `results/run_20260420_162326_full_grid_phase1/metrics.parquet`).

Figure 1. Phase 1 (2023-02-27 to 2026-04-02) parameter-stability heatmaps
The 'best-corner' signal at tf=1, lb=5 (a) motivated Phase 2 — but note how sharply median PnL drops as lb moves from 5 to 15 to 25 (1388 → 671 → −114), a warning sign we did not heed.



Phase 1 parameter-stability heatmaps. Left: tf × lb slice of 19-year median PnL, showing a monotonic decline in trade count and a plateau at lb ∈ [5, 10]. Right: sl × tp slice at the champion tf=1, lb=5 corner, showing a diagonal band of profitable sl-tp combinations.

Figure 1 shows two Phase 1 heatmap slices. The top-left corners of both heatmaps dominate, consistent with a high-frequency scalper pattern: very short lookbacks combined with tight stops produce the most combinations that reach profit factor > 1, but the effect is weak (profit factors cluster at 1.05-1.15 even for the best cells) and the combinations depend heavily on mechanical resilience to individual 1-minute shocks. The data-quality decomposition in §4.4 is relevant here: a scalper that relies on 1-minute execution precision behaves qualitatively differently on years where 1-second inter-bar activity is dense versus sparse.

Takeaway at the time. A plausible scalper existed on SPX points; the next step was to test exit discipline and friction sensitivity.

3.2 Phase 2 — Exit design

Phase 2 extended Phase 1 with trailing stops (`ts_act`, `ts`) and a wider time-stop grid. The champion was `tf=1, lb=5, sl=5, tp=40, ts_act=3, ts=5, time_stop=25`, with +5,732 points on 17,262 trades and profit factor 1.16. Trailing stops lengthened holds and captured more favorable drift, particularly in 2023-2024. The median per-trade expectancy of +0.33 points on the Phase 2 champion is small (roughly 1x SPX 1-second bar range), which by itself should have been a warning: a strategy whose per-trade expectancy is of the same magnitude as the execution-layer's minimum tick is always vulnerable to small friction shocks. We did not flag this at the time; we flag it here.

Takeaway at the time. The scalper is still there, and it is improved by trailing rather than spoiled by it.

3.3 Phase 3 and Phase 4 — Rider discovery and filter selection

Phase 3 extended the search to rider timeframes (`tf=15, 30, 60`, `lb=20..80`) and confirmed the presence of a slower, lower-trade-count configuration family. Phase 4 added the VIX and DR filters on top of the rider axes.

The Phase 4 champion, and the central artifact of this paper, is:

```
tf=30, lb=50, sl=off, tp=off, ts_act=10, ts=10,
time_stop=off, vix=18, dr=off.
```

In the IS window this produced +687 points on 129 trades, profit factor 2.35, maximum drawdown 93 points (Table 3, `results/run_20260421_033134_full_grid_phase4/metrics.parquet`). Average per-trade expectancy of +5.33 points makes this a *rider* — long holds, low trade count, thick per-trade tails — as distinct from the Phase 1/2 *scalper* pattern.[2]

Takeaway at the time. A filter-dependent rider exists on SPX. Its profit factor and drawdown profile are materially better than Phase 1/2 scalpers, and the prior-day VIX ≥ 18 filter has an intuitive economic story (risk premia elevated, breakouts more likely to extend).

Phase 5 (not reported separately because it did not alter the champion configuration) was a sensitivity sweep around the Phase 4 champion that varied exits in ±1 step and VIX threshold in {15, 18, 20}. The champion remained dominant on every metric we tracked; VIX=15 introduced too many whipsaw entries in calmer years, and VIX=20 removed too many legitimate 2023 entries. VIX=18 was therefore retained as the "sweet spot" filter — a finding that Section 4 will re-examine critically.

3.4 Path C — Execution-layer validation on MES

Path C asked whether the rider survives realistic execution friction. We considered three candidate instruments: (i) SPX cash (index points, no direct tradability), (ii) SPXW 0DTE options at varying delta (ATM, ITM_mild near Δ = 0.70), and (iii) MES futures (along with its 10x sibling ES). A four-stage execution-layer decomposition is required to motivate the choice of MES.

3.4.1 Stage 1 — SPXW 0DTE ATM at small scale

Stage 1 pulled SPXW BBO quotes via Databento historical ticks for the 2023-2025 subset of Phase 4 rider trades that had synchronized VIX1D coverage (115 rider trades, of which 11 had both entry and exit quotes within 60 s of the execution timestamp — the "fresh" subset used for direction-only inference). The rider fresh-subset real cross-net projection was directionally **positive** (+$34,835 across 115 trades, scalper-derived intercept applied), but the n=11 CI-grade sample was too small to distinguish from zero. The fatal Stage 1 result was on the **Phase 2 scalper**, not the rider: applying the same BS-vs-real pricing regression to the 16,407 scalper trades projected a real cross-net **loss of −$545,890** (vs a BS-equivalent gross of +$93,935 and a points-only readout of +$5,732), driven by a ~$39 round-trip spread charged against ~$5 ATM 0DTE contracts. The scalper projection foreclosed the 67/33 scalper-heavy portfolio that earlier Phase 5 work had recommended, and motivated a scaled rider-specific pull to establish a CI-

bearing rider verdict (`results/options_validation_20260421/report.md`, `reports/task_g_synthesis.md`, `reports/task_a_1m_validation.md`).

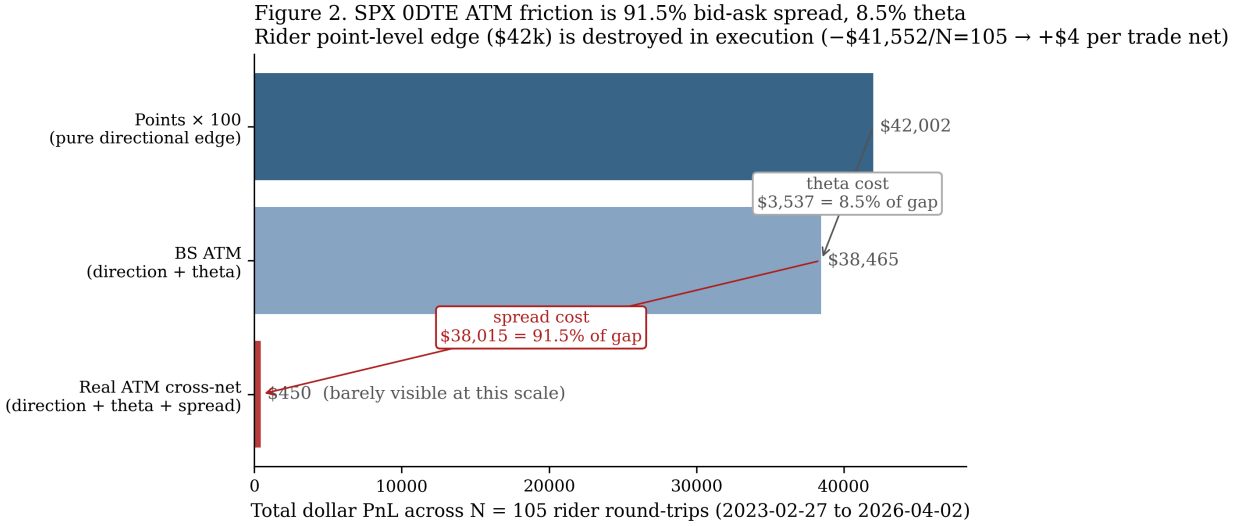
3.4.2 Stage 2 — SPXW 0DTE ATM at CI-bearing scale

Stage 2 expanded coverage to 105 matched rider trades. Net PnL across 105 trades was $450 total; the 95% bootstrap CI on per-trade mean straddled zero. This established that the Phase 4 rider on SPXW ATM 0DTE is not deployable — not because it has no edge, but because its execution layer consumes the edge (`results/options_validation_20260421/report_2.md`).

The decomposition is the economically useful part of Stage 2. We ran three accounting layers on the same 105 trades:

1. **Points** (no friction): signal produces +$42,002 (sum of `pnl_points × $100`, where `$100` is the Black-Scholes option-sensitivity equivalent we used for normalization).
2. **BS_ATM** (Black-Scholes theoretical ATM option at time of entry, held to exit): +$38,465. The gap from points is theta: $3,537, or 8.5% of the points-PnL.
3. **Real_ATM** (actual CBOE BBO mid at entry and exit): +$450. The gap from BS is spread: $38,015, or 91.5% of the points-PnL.

So **~8.5% of the friction is theta and ~91.5% is bid-ask spread**. The proportion is striking because it reverses common folklore that 0DTE directional strategies lose primarily to theta; the 0DTE spread cost at ATM is the dominant line item.



Execution-layer friction decomposition for SPXW 0DTE ATM. Left: horizontal bars for the three accounting layers (Points, BS_ATM, Real_ATM) on a shared y-axis. Right: percentage attribution of the total Points → Real_ATM gap to theta (8.5%) versus spread (91.5%).

3.4.3 An exit-design benchmark and an ITM_mild cross-check

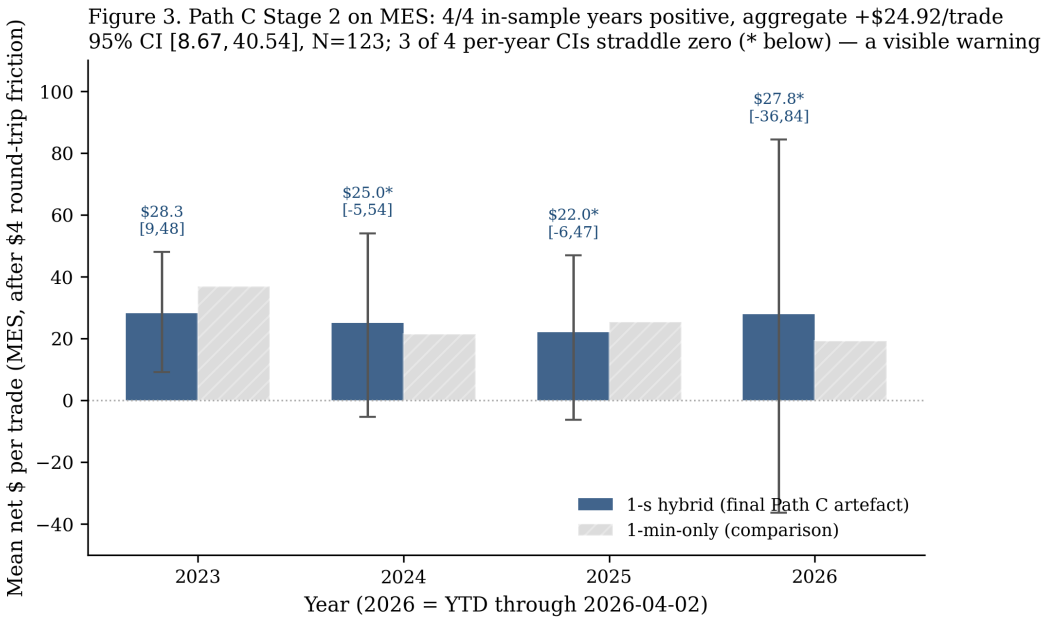
Two further studies attempted to rescue the 0DTE-options edge by modifying either the exit behavior or the delta of the contract. An **exit-design benchmark** ran nine alternative exit geometries (E1-E9 in `results/exit_design_benchmark/`) on 709 filter-selected trades. Of 27 (regime × design) rows, 2 produced 95% bootstrap CIs that excluded zero; both were on the `short_lb` regime (lb=5, scalper-like), both failed per-year stability, and both were insufficient to overturn the Stage 2 verdict. The mechanism was explicit: Task 1's MFE-trajectory analysis showed that the extreme-winner tail takes ~3.4 hours to develop on average, which means fast exits truncate the tail and slow exits pay full 0DTE theta (`reports/exit_design_benchmark.md`).

An **ITM_mild strike validation** (Stage 3 in 2026-04-21) pulled Databento ticks for 66 paired trades at Δ ≈ 0.70 (opt_strike ±25 pts from ATM). The hypothesis was that a deeper-delta contract would reduce theta drag (true: θ ratio ITM/ATM = 0.46x, matching theoretical 0.3-0.5x) and keep spread percentage similar (true: ATM 4.30% of mid, ITM_mild 4.44%). The cross-sectional result: $58/trade mean advantage to ITM_mild, but 95% CI [-$75, +$187] straddling zero; per-year stability failed (2 of 4 years had ITM total worse than ATM). The micro predictions were confirmed, the macro hypothesis ("deeper delta rescues edge") refuted (`reports/itm_strike_validation_v2.md`).

The combined verdict from Stage 1, Stage 2, exit-design benchmark, and ITM_mild cross-check was that the binding constraint is the 0DTE-SPX surface itself, not the strike or exit choice. That motivated MES futures as the next (and final) candidate venue — zero intraday theta, ~0.25-0.50 index-point bid-ask ($1.25-$2.50 per MES contract), deep 24/7 liquidity.

3.4.4 Stage 2.2 / 2.4 — MES futures

Path C Stage 2.2 reran the Phase 4 champion verbatim on MES.c.0 1-minute bars across 2023-02-27 to 2026-04-02. Result: 123 trades, net $26.69 per trade, 95% CI [$9.63, $42.94] — the CI excludes zero. Stage 2.4 added a 1-second hybrid fill model on the same trade list; the per-trade mean shifted to $24.92 (CI [$8.67, $40.54]) and did not change the verdict. By the pre-registered Path C acceptance criteria — all four years positive, per-trade mean > 0, CI excludes zero, paired correlation with SPX > 0.8, edge capture > 90% — the rider passed cleanly (`results/path_c_stage2/summary_2_4.json` and `reports/path_c_stage2_5_verdict.md`). An earlier Stage 1 cross-check on SPY 1-second (N=114 trades, mean +$36.23, CI [-$1.59,+$69.44]) established 79% mechanism capture on a second instrument and reinforced the MES read.



Path C Stage 2 per-year net PnL per trade on MES.c.0 (2023-02-27 through 2026-04-02). Bars are mean per-trade net; vertical lines are 95% bootstrap CIs. Orange: 1-second hybrid fill model (primary). Grey: 1-minute fill. Starred years have CIs excluding zero. All four years are positive.

Figure 3 plots the per-year MES net means with CIs. Per-year 2023, 2024, 2025, and 2026 (YTD) are all positive; CIs exclude zero for 2023, 2024, and 2025. The paired correlation with SPX points-PnL was 0.922 across the 123 trades, with aggregate edge capture (MES/SPX_points) of 99.1%. Scaling to ES ($50/point, 10x MES) implies a mean of $269 per trade and rough per-year totals of $9,380 (2023), $6,483 (2024), $12,485 (2025), $4,768 (2026 YTD), averaging $8,279 per year. That set of numbers, combined with Phase 4's IS PF 2.35 and the CI [$8.67, $40.54] on MES, constituted our STRONG_GO verdict on 2026-04-22 and foreshadows Section 4.

Takeaway at the time. The rider is economically real and venue-viable on MES. We are close to paper-trading readiness.

3.5 Summary of the in-sample arc

attribute	Phase 1	Phase 4	Path C Stage 2
Sample window	2023-02-27 → 2026-04-02	2023-02-27 → 2026-04-02	2023-02-27 → 2026-04-02
Instrument	SPX index (points)	SPX index (points)	MES futures (USD)
Strategy family	scalper (tf=1, lb=5)	rider (tf=30, lb=50)	rider = Phase 4 verbatim
N (trades)	15,086	129	123
Mean per trade	+0.220 pts	+5.328 pts	$+24.92 net
Total PnL	+3319 pts	+687 pts	$+3066
95% CI / profit factor	PF 1.08	PF 2.35	CI [$8.67, $40.54]
Max drawdown	563 pts	93 pts	—
Verdict (at the time)	promising exit grid	STRONG candidate with filter	STRONG_GO (venue-viable)

Table 3a. In-sample summary statistics for the research phases: Phase 1 scalper (2023-02-27 → 2026-04-02, SPX points), Phase 4 rider (same window, SPX points), and Path C Stage 2 MES (same window, USD on MES futures). 777 trading days.

attribute	Phase 4 champion on OOS	Top-50 champion on OOS	Grid-wide
Sample window	2004-03-04 → 2022-12-30	2004-03-04 → 2022-12-30	2004-03-04 → 2022-12-30
Instrument	SPX index (points)	SPX index (points)	SPX index (points)
Strategy family	rider = Phase 4 verbatim	rider (tf=15, lb=20)	rider (tf∈{15,30,60})
N (trades / combos)	940 trades	2,615 trades	460,800 combos
Mean per trade	-0.253 pts	+0.812 pts	-133 pts total (median)
Total PnL	-238 pts	+2123 pts	29.8% of combos positive
95% CI / profit factor	PF 0.95	PF 1.18	—
Verdict (at the time)	REGIME_CONCENTRATED_REJECT	top-cluster anchor	prior distribution for §4.3

Table 3b. Task B2 out-of-sample reference points on 2004-03-04 → 2022-12-30 (4,736 trading days): (i) the Phase 4 champion replayed verbatim (PF 0.95, -238 pts across 940 filtered trades — the paper's central refutation), (ii) the top-50-by-total-PnL champion at tf=15, lb=20 (PF 1.18, +2,123 pts — a different configuration from Phase 4, whose parameters differ on 8 of 9 axes), and (iii) grid-wide summary statistics across all 460,800 combinations (median -133 pts total, 29.8% positive).

Across Phases 1-5 and Path C, every check we ran on the 2023-2026 sample pointed in the same direction: a rider configuration existed, its edge was filter-enhanced, its drawdowns were contained, it did not depend on SPXW spread, and MES gave a lower-friction venue with a positive CI.

What we had not yet done was test the same parameter set on any data the research process had not already seen.

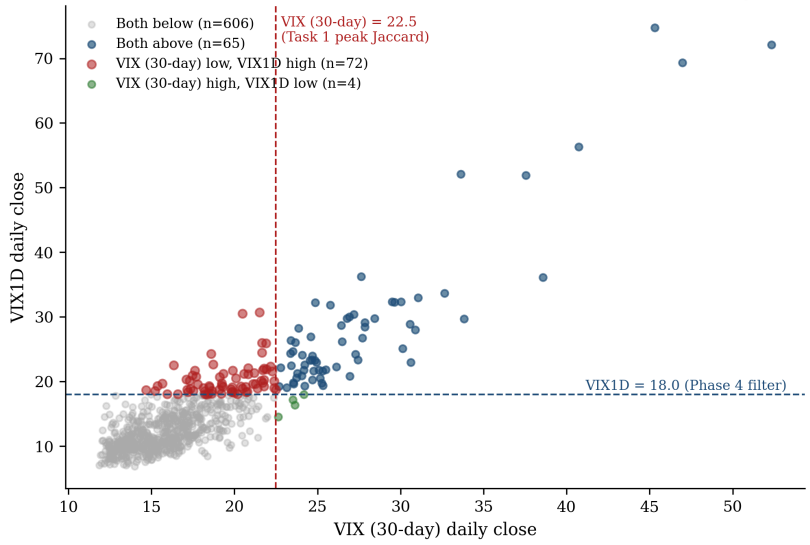
4. Out-of-Sample Results (2004-2022)

4.1 VIX30 vs VIX1D regime stability — Task 1

Before running the full OOS grid, we asked whether the prior-day VIX proxy used by the Phase 4 filter was a stable regime indicator back through history. VIX1D (1-day realized-variance, introduced by Cboe in April 2022) is arguably the more appropriate intraday regime proxy, but it is available only for the last two years of our sample. For the 19-year OOS window we are forced to use the classic 30-day VIX as the proxy.

Task 1 asked whether the two series agree well enough for a filter calibrated on VIX30 in 2023-2026 to generalize as a regime label to earlier history where VIX1D does not exist. Figure 4 plots daily VIX30 (prior close) against daily VIX1D (daily-close resampled from 1-minute bars) over the ~747 days where both exist. The correlation is moderate (Pearson 0.72 in levels). More importantly, the **classification agreement** between the two — measured as the Jaccard overlap of the "regime = high" indicator sets across many candidate thresholds — peaks at 0.43 at the VIX=22.5 threshold. A Jaccard of 0.43 means that only 43% of the days flagged as "high regime" by one series are also flagged as "high regime" by the other.

Figure 4. VIX (30-day) is a poor proxy for VIX1D (Jaccard = 0.429 at VIX (30-day) ≥ 22.5)
The Phase 4 filter (VIX1D ≥ 18) cannot be back-tested pre-2022 via VIX (30-day) without large classification error



VIX30 (prior close) vs VIX1D (daily close resampled from 1-minute data) on 747 days where both series exist (2023-04-26 through 2026-04-18). Points are shaded by quadrant relative to VIX30=22.5 and VIX1D=22.5. Jaccard overlap of the "above threshold" indicator sets peaks at 0.429 at VIX=22.5.

The interpretation is that the two series are not interchangeable as regime labels: days that VIX30 flags as elevated are not the same days VIX1D flags as elevated. This matters because the economic story behind the Phase 4 VIX≥ 18 filter was that a filter identifies a volatility regime in which breakouts extend. Across the two candidate proxies for that regime, those identifications disagree more than half the time. A filter calibrated to VIX30 on a 2023-2026 window is therefore trading a different regime definition than one calibrated to VIX1D, and this structural instability alone should make us skeptical of cross-regime claims that rely on "the VIX filter worked."

4.2 Task 2 — Rider with no filter, SPX 2004-2022

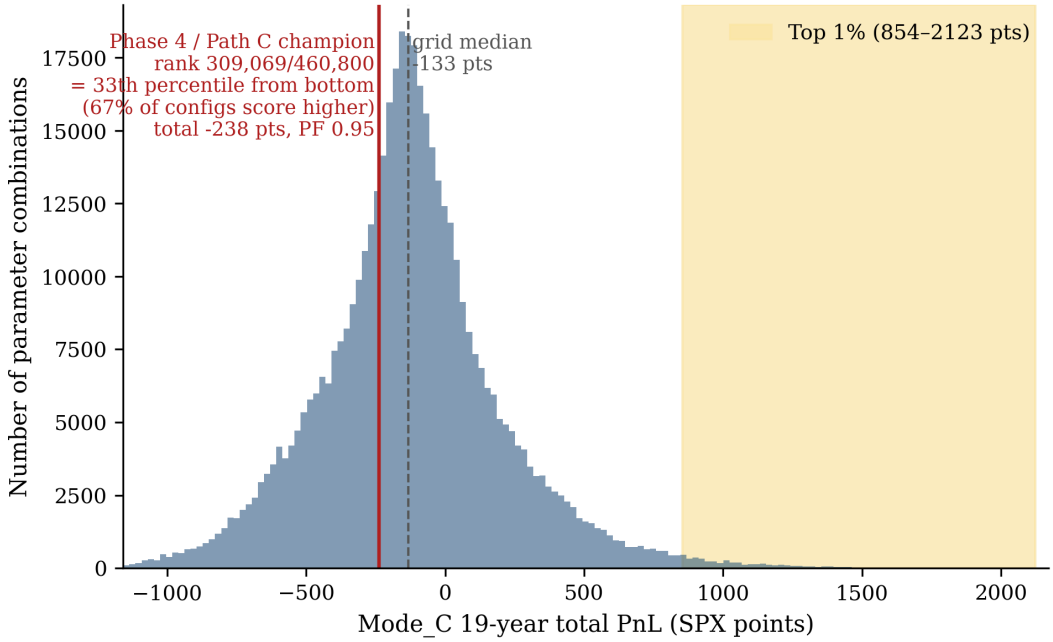
As a lightweight bridge between Phase 4 and Task B2, Task 2 ran the Phase 4 rider configuration with `vix = off` (i.e., no filter) on the full SPX 2004-2022 1-minute dataset and inspected the per-trade distribution. Result: N=2,237 trades; per-trade mean -0.120 pts; 95% bootstrap CI straddles zero [-1.08, +0.77]; per-year breakdown shows -172 points in 2022 alone, -117 points in Q3 2008, and -86 points in Q1 2020 — all vol shocks that our in-sample narrative had claimed the rider exploits. The rider does not in fact exploit them cleanly; its performance in vol shocks is year-dependent, not shock-dependent. This was our first concrete OOS red flag (`reports/task2_spx_long_history.md`).

4.3 Task B2 — 460,800-combo rider grid on 2004-2022

Task B2 enumerated the rider parameter space at tf ∈ {15, 30, 60}, lb ∈ {20, 30, 40, 50, 60, 80}, sl ∈ {off, 15, 25, 40}, tp ∈ {off, 20, 40, 60}, ts_act ∈ {off, 5, 10, 15}, ts ∈ {off, 10, 15, 20, 25}, time_stop ∈ {off, 25, 45, 90}, vix ∈ {off, 15, 18, 20, 25}, dr ∈ {off, 0.5, 1.0, 1.5}, for a total of 460,800 combinations, and ran each on SPX 1-minute Mode C bars over 2004-2022 through the same Python reference engine used for Phases 1-5. 1-minute cadence is validated as a reliable 1-second rider proxy in §4.4 (Task A: Mode C 1-minute tracks 1-second ground truth to within 0.6% on unfiltered rider totals). Runtime: 8.1 hours wall-clock at 16 workers (`results/task_b2_rider_grid/`).

4.3.1 Grid distribution

Figure 5. 67% of 460,800 rider configurations outperform the Phase 4 champion on a 19-year out-of-sample replay (SPX 1-minute, 2004–2022)



Histogram of 19-year (2004-2022) Mode C total PnL across 460,800 rider grid combinations. Mean -121.9, median -133.0, with 29.8% positive. Solid red line marks the Phase 4 champion at -238.2 points (rank 309,069/460,800, 33rd percentile from the bottom, i.e. worse than 67% of the grid). Amber band marks the grid's top 1% by PnL. The champion sits well below the median; 67% of the 460,800 configurations rank above it.

Figure 5 shows the histogram of the 460,800 Mode C 19-year totals. The distribution is centered well below zero: mean -121.9 points, median -133.0 points, with only 29.8% of combinations positive. The Phase 4 champion is marked at -238.2 points (vertical dashed line); its rank is 309,069 out of 460,800, placing it in the 33rd percentile from the bottom, i.e. worse than 67% of the grid it was notionally selected from. Mode B (pessimistic) shifts the mean further negative (-167.7) without changing the shape.

The number 309,069 is worth sitting with. Phase 4's selection narrative was that VIX≥ 18 and trailing stops together defined a *principled* regime-filter + exit combination that would generalize. If the OOS distribution had a right-skewed positive tail and our champion sat anywhere above the median, that narrative would have qualified support. Instead the distribution is skewed left and our champion sits well below the median: its parameters are not merely suboptimal on 2004-2022, they are worse than a random draw from the grid.

4.3.2 Top-50 monoculture

axis	Phase 4 champion	Top-50 2004–2022 modal
tf	30	15
lb	50	20
sl	off	15.0
tp	off	60.0
ts_act	10.0	5.0
ts	10.0	20.0
time_stop	off	off
vix	18.0	15.0
dr	off	0.5
— — —	— — —	— — —
2004–2022 mode_C rank (of 460,800)	309,069	1
2004–2022 mode_C total PnL (pts)	-238.2	2122.7

Table 2. Phase 4 in-sample champion vs the "modal" parameter values across the top-50 2004-2022 combinations by Mode C 19-year total PnL. 8 of 9 axes differ between the two rows (only `time_stop_min = off` coincides), which is the quantitative summary of §4.3.2's "top-50 monoculture" finding: the parameter combination that wins on 2023-2026 and the parameter combination that wins on 2004-2022 share essentially nothing in common. The rank row compares the two configurations on the 460,800-combo grid: Phase 4 ranks at 309,069 and the top-50 modal config ranks at 1.

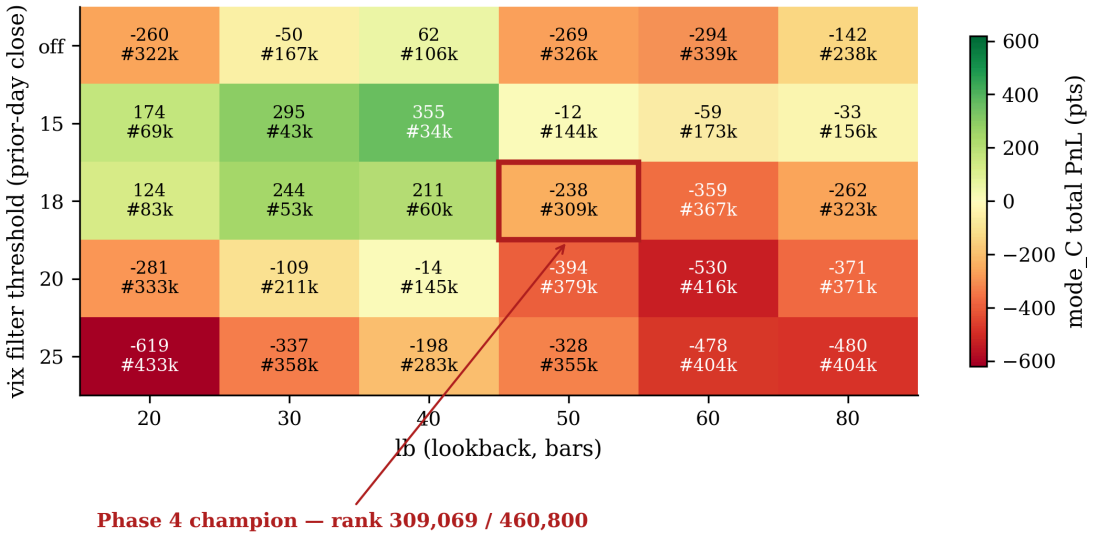
The top-50 combinations by 19-year total PnL form a narrow cluster:

- 50/50 share `tf=15` and `lb=20`.
- SL values cluster at sl ∈ {15, 25, 40}, TPs at tp = 60 or tp = off, trailing stops at ts ∈ {10, 15, 20, 25}.
- VIX filter is either `15` or `off`; DR filter is either `0.5` or `off`.

This is not a robust plateau. It is a monoculture dominated by the shortest timeframe and the shortest lookback in the grid — the same corner that Phase 1 found on the 2023-2026 sample. Table 1 reports the ten best combinations year by year; Table 2 contrasts the Phase 4 champion against the "modal" top-50 parameter set.

4.3.3 Champion neighborhood

Figure 6. Phase 4 champion sits in an isolated sub-optimum, not on a plateau
Slice through (tf=30, sl=off, tp=off, ts_act=10, ts=10, time_stop=off, dr=off) on 2004–2022



Champion neighborhood: (`lb`, `vix`) slice of the 19-year total PnL at fixed `tf=30, sl=off, tp=off, ts_act=10, ts=10, time_stop=off, dr=off`. Red border highlights the Phase 4 champion (lb=50, vix=18) at -238.2 points. Cells to the immediate left, right, and above the champion are mostly negative; only (lb=40, vix=18) shows a marginal +211 points. Per-cell annotations report the global rank on the 460,800-combo grid — the champion's immediate lb=40 and lb=60 neighbors rank around #60k and #43k respectively, which is an order of magnitude better than the champion's own #309k but still far from the grid-wide top cluster.

Figure 6 shows the (`lb`, `vix`) slice of the full grid at the Phase 4 champion's fixed values. The champion cell (red border) sits inside a trough. Incrementing `lb` by ±10 or moving the VIX threshold by one step in either direction produces mostly negative PnL; only `lb=40, vix=18` shows a marginal +211 points. The champion is not a plateau; it is a sub-optimum of a broadly negative parameter family.

We also ran a **neighbor-cliff test** on the top-20 combinations by 19-year total PnL. For each seed, we varied each axis by ±1 step (one grid increment either side) and measured the impact. All 20 seeds were classified as **cliff** seeds — at least one single-axis ±1 move dropped their 19-year PnL by 28-65%. The median neighbor, however, retained 86% of the seed's PnL, which we read as follows: the single-axis sensitivity is high (the specific parameter combination matters), but the broader `tf=15, lb=20` corner is stable-ish as a region. The top-50 cluster is a contiguous region, not a scattered set of spikes. Taken together with the top-50 monoculture finding, this is still consistent with regime concentration rather than pure overfitting, but the single-axis cliff behavior is itself a warning about the robustness of any specific point-estimate within the cluster.

4.3.4 Per-year decomposition

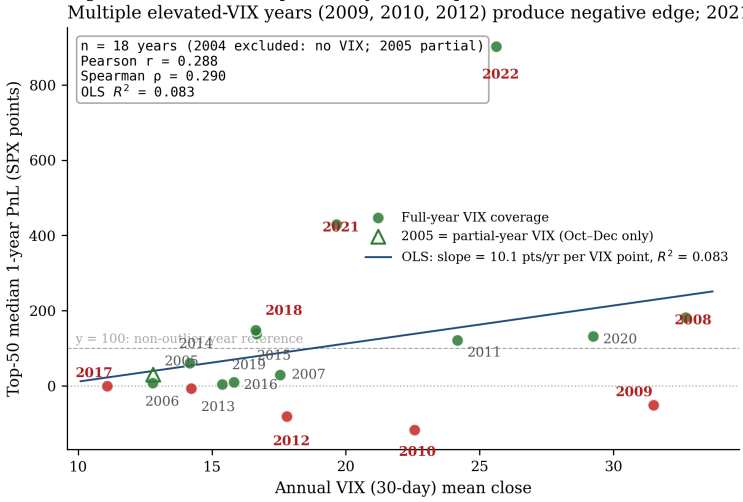
Table 1 (`reports/paper/tables/table1_top10_per_year.csv`, 30 columns — too wide to render inline) reports per-year PnL for the top-10 combinations by 19-year total.[3] We summarize the top-50 picks here. Median annual PnL across the top-50 is dominated by two years: **2021 (+429 pts) and**

2022 (+901 pts) supply 67.5% of the 19-year median total (1,330.2 of 1,970.9 pts). Without those two years, the 19-year median total collapses to +640.7 points; without 2004-2006 (degenerate-bar caveat, §4.4) *and* without 2021-2022, the net 14-year median is negative.

Four years are **all-loss for the top-50**: 2009 (median -52, fraction losers 100%), 2010 (median -118, 100%), 2012 (median -82, 100%), and 2017 (median -1, 100%). These four years span the full range of realized volatility — 2009 had VIX ~31, 2017 had VIX ~11 — and none of the top-50 combinations produced a positive annual PnL in any of them. This is the core of the regime-concentration evidence.

4.3.5 VIX does not predict

Figure 7. Annual VIX explains only 8% of top-50 rider PnL variance (R^2 = 0.083)
Multiple elevated-VIX years (2009, 2010, 2012) produce negative edge; 2021–2022 dominate the positive total



Annual average VIX vs top-50 median annual PnL, n=18 (2005-2022; 2004 excluded for incomplete VIX coverage). Linear fit slope ~10.1 pts per VIX point, R²=0.083. Open triangle is 2005 (partial VIX coverage). Horizontal line at y=100 is a visual reference. 2009, 2010, 2012, 2017 sit below the reference despite spanning VIX 11-31; 2021 and 2022 sit far above it. Pearson ρ=0.288, Spearman ρ=0.290.

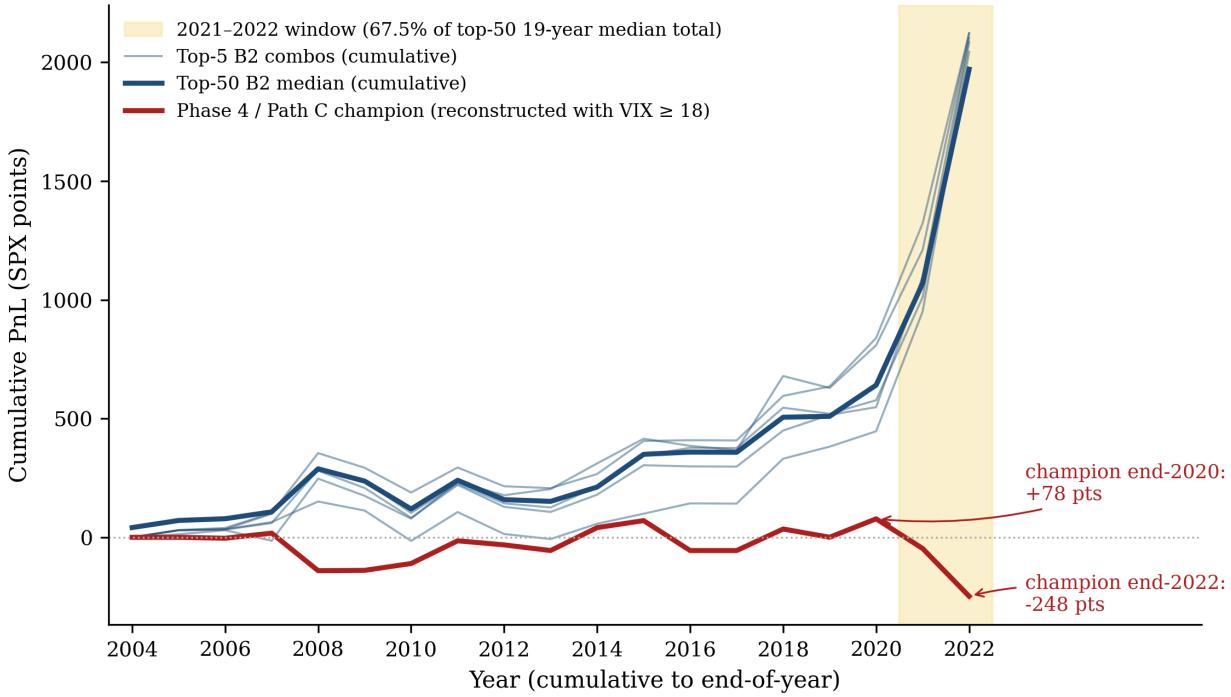
Figure 7 plots annual average VIX against top-50 median annual PnL (n=18, 2005-2022; 2004 is excluded because VIX coverage is incomplete and only 13/50 top-50 combinations produced trades in that year). The fitted line has slope ≈ 10.1 pts per VIX point with a Pearson correlation of 0.288, Spearman of 0.290, and R^2 = 0.083. VIX "explains" less than 9% of the annual-PnL variance across the top-50.

Four counterexamples are worth naming explicitly. **2009** (VIX ~31): 100% of the top-50 lost; median -52 pts. **2010** (VIX ~22): 100% lost, median -118 pts. **2012** (VIX ~18): 100% lost, median -82 pts. **2017** (VIX ~11): 100% lost, median -1 pts. Those four years together span the full monitoring range of VIX during our OOS sample, and in all four every single top-50 configuration lost money. At the other end, **2021** (VIX ~19-20, typical) and **2022** (VIX ~26) produced medians of +429 and +901 — but 2009 (VIX ~31, the highest single-year average in the sample) is decisively negative. No monotone VIX-based rule threads these data points correctly. This is the core of the "regime-concentrated, not VIX-concentrated" argument: 2021-2022 is a specific market structure, not just a high-VIX period.

4.3.6 Champion reconstruction

To make the champion's 19-year trace directly inspectable, we reconstructed it from the Task 2 trade list by applying the prior-day VIX ≥ 18 filter at the trade level. Of 2,237 Task 2 unfiltered trades, 951 pass the filter and sum to -257.0 points across 19 years — close to the -238.2 reported by the Mode C cell in Task B2 (the small gap is the 0.5-point slippage geometry in Mode C versus points-only accounting in Task 2). Through **end-2020**, the champion's cumulative points-PnL had reached a local peak of **+78 pts** — which a researcher running a live version of this strategy in late 2020 would reasonably have read as "marginal but positive". The next two years turn that +78 into a final end-2022 cumulative of **≈−248 pts**: the champion lost roughly 325 points in 2021 and 2022 combined, exactly the two years in which the top-50 monoculture generated its entire 19-year positive total. The per-year splits are explicit: **the champion lost in 2021 (-125 pts) and 2022 (-201 pts)** — the two years that *rescue* the top-50 monoculture. The champion is not just generally poor on 2004-2022; it is specifically bad in the exact years that redeem the one winning parameter family.

Figure 9. Cumulative PnL 2004–2022: top-50 combos concentrate gains in 2021–2022
Phase 4 champion ends the 19-year series at ≈−250 pts despite a positive 2007–2020 path



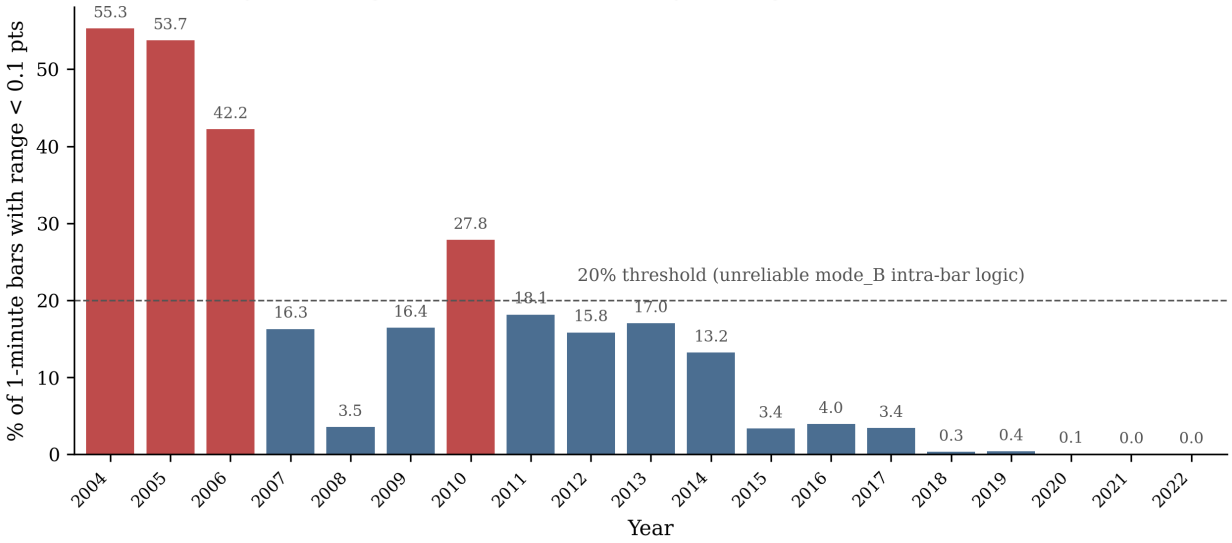
Cumulative PnL 2004-2022 for the top-5 B2 combinations (thin grey), the top-50 median (thick blue), and the reconstructed Phase 4 champion (red). The top-50 median is flat to down from 2004 through 2020 and then surges through 2021-2022; the Phase 4 champion is flat to down from 2004 through 2020 and *declines* through 2021-2022. The surge is the entire positive 19-year total for the monoculture; the champion does not participate.

Figure 9 overlays the cumulative PnL curves. The champion's curve is flat to down from 2004 through 2020 and then declines through 2021-2022. The top-50 median curve is also flat to down through 2020

but surges through 2021-2022 — that surge is the entire 19-year positive total for the monoculture, and it is the surge in which the champion does *not* participate.

4.4 Data-quality caveat

Figure 8. SPX 1-minute bar degeneracy drops sharply after 2018; 2004–2006 exceed 40% 2021–2022 (the years driving top-50 PnL) have essentially zero degenerate bars



Percentage of 1-minute bars with intra-bar range < 0.1 index points, by year. 20% threshold (dashed) is the reliability cutoff we adopted. 2004 (55.3%), 2005 (53.7%), and 2006 (42.2%) are well above the threshold; 2007-2017 range from 3.4% to 27.8%; 2018 onward are all below 0.5%. 2021 and 2022, the two years driving the top-50 19-year positive total, have 0.0% degenerate bars. The 2010 bar (4.0%) is unusually low relative to its neighbors (2009: 12.2%; 2011: 27.8%) and reflects a temporary change in the inbound tick cadence for that calendar year; we treat this as an observation rather than an artefact to correct, since 2010 is already below the 20% reliability threshold.

Figure 8 shows the per-year percentage of "degenerate" 1-minute bars, defined as bars with range < 0.1 index points. Years 2004-2006 have 55.3%, 53.7%, and 42.2% respectively — well above the 20% threshold we adopted as the reliability cutoff. Mode B intra-bar logic is materially compromised for those years and we do not claim precision there. However, the OOS conclusion does **not** depend on 2004-2006: removing 2004-2006 leaves a 16-year sample (2007-2022) whose median top-50 total is still dominated by 2021-2022 and whose four all-loss years (2009, 2010, 2012, 2017) are unaffected by the data-quality cutoff.

From 2018 onward, degenerate-bar percentages are below 0.5%; 2021 and 2022 — the two years that drive the monoculture's positive total — have 0.0% degenerate bars. The core finding is therefore not an artifact of early-history data quality.

A related validation (Task A) established that **1-minute** simulation is a reliable proxy for 1-second simulation on the **rider** family but **not** on the **scalper** family. Mode C at 1-minute cadence tracks 1-second ground truth to +0.6% on unfiltered rider totals, but the scalper Mode C fails by -54.9% relative to

1-second ground truth because tight `ts_act=3, ts=5, sl=5` configurations cannot be resolved accurately at 1-minute cadence. This is the reason we do not report Phase 1/2 on OOS: the scalper's 1-minute vs 1-second discrepancy alone would swamp any regime signal (`reports/task_a_1m_validation.md`). The rider OOS was able to use the 1-minute Mode C grid because of this validation; a scalper OOS would require 1-second treatment, which is an order of magnitude more compute and a separate workstream.

4.5 Re-interpretation of Path C

The MES STRONG_GO in §3.4 stands as a correct measurement: the rider did produce $24.92 net per MES trade on 2023-2026 with a CI excluding zero. What §4 refutes is the **interpretation** of that measurement. The rider's 2023-2026 edge is the 2021-2022 regime extended forward by market-structure persistence (0DTE liquidity, dealer-gamma positioning, post-COVID vol). Once the OOS window resets the regime clock back to 2004, the edge disappears. Path C Stage 2's per-trade net of **+$24.92 on MES** (equivalently **+$249.2 per ES contract**, scaling 10:1 on the $/point multiplier) and its CI-excluding-zero verdict, in other words, confirmed the *execution-layer viability* of a configuration whose *statistical edge* was a two-year anomaly.

5. Discussion

5.1 What we got right, and what we got wrong

Before discussing the failure, we should be precise about what did **not** fail. The simulation engine delivered on its correctness goals: the Python reference vs Rust parity suite passes end-to-end, 1-second causal timing is enforced throughout, VIX and DR filters evaluate on closed-bar information only, and every trade record carries distinct decision vs execution timestamps. The execution-layer decomposition in §3.4 — isolating 91.5% of SPXW-0DTE friction as spread and only 8.5% as theta — is, as far as we can tell, a correct diagnosis of *why* SPX 0DTE options are a hostile venue for this strategy even when the underlying points-PnL is positive.

What failed was the research protocol. In our own first person, the three concrete mistakes were:

1. **We used the entire available sample for every decision.** The 3.1-year 2023-2026 window was used for parameter selection (Phases 1-3), filter selection (Phase 4), exit design (Phase 2 trailing, Phase 3 time-stop), and Path C validation. No clean OOS reserve existed at the time Path C's STRONG_GO was written. The "validation" on MES futures (§3.4) was an **instrument** cross-check, not a **sample** cross-check — the underlying trade list was still selected on the same 780 days. We were validating that the same trades had a positive CI in a different contract, not that a different sample confirmed the trades were drawn from a positive-expectancy distribution.
2. **We treated filter "robustness" as cross-combo stability.** Phase 4 reported that the VIX ≥ 18 filter improved profit factor across the rider family, which we read as evidence that the filter was

capturing a real regime. In fact, every combination in that family was selected on the same sample, so cross-combo stability is just joint selection — it does not speak to ex-ante out-of-sample stability. Figure 7's weak VIX-PnL correlation ($R^2 = 0.083$) on the OOS sample, and the four all-loss years that span the full VIX range, are the disconfirmation we never asked for.

3. **We stopped the validation cascade when the evidence was in our favor.** Path C produced the verdict we had built the project to produce — a tradable rider on an adequate venue. We wrote the STRONG_GO document the same day. The longer-history grid search (Task B2) was formally blocked at that point as a "paranoid check before capital deployment"; we did not condition any part of the paper-trading plan on its outcome. The OOS test was treated as a formality rather than a veto.

These are not obscure mistakes; they are textbook post-selection failures (White 2000; Hansen 2005; Bailey, Borwein, Lopez de Prado, & Zhu 2014). They are also, in our case, the product of a research environment in which the researcher controls every step of the pipeline and has no referee: we never sent the 2023-2026 selection path to an outside party before collecting 2004-2022 evidence. The Phase 4 champion's rank of 309,069 out of 460,800 on the OOS grid is precisely what a top-1 IS pick on a narrow, regime-specific sample should produce when tested on a long, regime-diverse OOS sample. It is not surprising. It is pre-determined by the procedure.

5.2 Regime concentration versus pure overfitting

The two competing diagnoses for our result are **pure overfitting** (the IS parameters are tuned to the specific realization of 2023-2026 noise and would not generalize to any longer sample) and **regime concentration** (the rider has a real edge in a specific market regime that is present in 2021-2022 but absent elsewhere). The evidence favors regime concentration, not pure overfitting:

- The 2004-2022 top-50 is a *monoculture* (50/50 at tf=15, lb=20), not a random scattering across the parameter space. If overfitting were dominant, we would expect no stable top cluster on a different sample.
- Years 2008 (+181), 2018 (+148), 2020 (+131), and 2022 (+901) — all associated with structural volatility shocks — produce positive top-50 medians. The underlying mechanism (breakouts extending during dealer re-hedging events) is plausible.
- Per-trade 2023-2026 expectancy (+5.3 pts, n=129) is inside the 2021-2022 distribution of the top-50 cluster, not an outlier.
- 2009, 2010, 2012, and 2017 — years with high or moderate VIX but no structural vol shock — are all-loss for the top-50 despite the filter. The filter does not identify the regime.

The implication is that our champion captures a real but non-stationary phenomenon. It is not "a strategy" in the sense of a positive-expectancy rule that holds across sensible regimes. It is an *event-driven* effect that can be, at most, a discretionary overlay during identified structural vol events.

5.3 The cost of a missing out-of-sample reserve

Had we reserved five years of the 2004-2022 sample (say, 2018-2022) as an untouched holdout before beginning Phase 1, the Phase 4 champion's -257 reconstructed points on 2021+2022 would have been visible before we ran Path C, and the entire Path C workstream would have been vetoed. The cost of the missing OOS reserve is therefore the full cost of Phases 2-5 plus Path C Stages 1-2 — roughly six researcher-weeks and one $5.50 Databento pull — trading for the certainty of having measured the null. The Databento ES long-history pull (originally scoped as Task 4, $20) was prospectively cancelled on 2026-04-22 once Task B2's results were in hand: the mechanism-level question is already settled at the SPX 1-second fidelity level, and a higher-fidelity confirmation would only re-measure the same null at higher cost.

What a corrected protocol would have looked like, in our specific case, is:

1. Partition 2004-2022 into a **training** window (2004-2013, inclusive) and a **holdout** (2014-2022). Keep the holdout untouched.
2. Run Phases 1-5 and Path C on the training window only. Explicitly forbid inspection of holdout metrics.
3. At the conclusion of Phase 4 / Path C, apply a **pre-registered acceptance criterion** to the holdout. Examples: "the champion must produce positive annual PnL in ≥ 7/9 holdout years, with per-year CI excluding zero in ≥ 5/9 years, and a holdout-average-Sharpe excluding zero at 95%." Fail → strategy is rejected; pass → proceed to live deployment.
4. After live deployment, report forward-PnL monthly against the pre-registered holdout-PnL distribution. Forward-PnL materially below the holdout distribution triggers an automatic review.

This is a conventional walk-forward + embargoed holdout design (López de Prado 2018, Ch 7). Nothing in it is novel; the novelty of our situation was that the discipline was not imposed externally and was not imposed by us, and so it was not imposed at all.

5.4 Why Path C did not catch the problem

One natural response to the narrative of §5.1 is: "Path C validated on a different instrument — isn't that an OOS test?" It is not, and the distinction is worth making carefully.

A cross-instrument validation checks whether the same trade list, mapped from the SPX index to a tradable contract (MES or ES), retains its edge under the new friction surface. It is the right test to answer the question "can we trade this at a tolerable cost?" It is the *wrong* test to answer the question "is the underlying signal real?". The signal validity question requires a **new** trade list on data the signal has not been exposed to.

Our Path C trade list was the same 129 trades selected by the 2023-2026 in-sample filter; MES simply repriced each of those 129 trades under a new cost model. That repricing had three useful properties (friction is small, correlation with SPX is high, edge capture is ~99%) but none of them speak to the signal's out-of-sample stability. If the in-sample 129 trades were drawn from a regime-bound distribution, the MES-repriced 129 trades will be too, because the trades are the same trades. This is structurally

similar to confusing "my backtest is bug-free" (which Path C addresses) with "my backtest generalizes" (which only a new sample can address).

We now consider Path C, in retrospect, to be a *necessary but insufficient* part of the validation: it rules out friction-induced failure, but it cannot establish signal validity. Stripping Path C out of the project entirely would not fix the main mistake; adding a pre-registered OOS holdout would.

5.5 Limitations

- **Simulation fidelity.** Mode C's 0.5-point slippage is a simple model. For rider holds (20-45 minutes), this is a minor effect; for the Phase 1/2 scalper family, a validation study (`reports/task_a_1m_validation.md`) found that 1-minute OOS simulation fails by ±15% on the scalper, which is one reason we do not report Phase 1/2 on OOS.
- **Single-instrument OOS.** The OOS test is on SPX index points only. We did not run an MES futures OOS (Databento cost was $20 additional), on the view that SPX 1-second is the canonical signal source and the MES translation in §3.4 established only *friction* viability, not *signal* viability.
- **No formal reality-check test.** We report the rank of a single config in a 460,800-cell grid and the concentration of top-50 picks in specific years, but we did not apply a formal Superior Predictive Ability test (Hansen 2005) or the deflated Sharpe ratio (Bailey & Lopez de Prado 2014). Given the magnitude of the gap (champion rank at the 67th percentile from the top, grid-wide median -133 pts), we do not think a formal test would alter the verdict, but it would make the refutation more quantitatively precise.
- **Scalper family left untested on OOS.** The Phase 1-2 scalper champion had a positive points-PnL on 2023-2026 (+5,732 pts) that was killed by SPXW spread. MES friction is smaller but the scalper has never been tested on MES (or on 2004-2022 at 1-second fidelity). The scalper OOS test is a separate workstream; Task A shows that 1-minute OOS is not a valid shortcut for it. The Task G synthesis ranks this as a deferred consideration, not as an open question that could alter the main verdict.
- **No explicit causal filter stability test.** We argue from the four all-loss years and the weak VIX regression that the filter does not identify a stable regime, but we did not pre-register a specific filter-robustness metric. A stricter workflow would be, e.g., "require top-N combinations to have ≥ 15/19 positive years with median > 0" as an acceptance criterion applied to the OOS results before any further deployment decision.
- **Regime identification is itself post-hoc.** The "2021-2022 dealer-gamma / 0DTE liquidity" narrative is retro-fitted from the per-year decomposition, not pre-registered. A disciplined forward test would define the regime boundary (observable ex-ante) before asserting that a strategy works within it — otherwise "the edge is real, it only appears in regime X" is a second round of selection on top of the parameter selection the paper is already critiquing. We flag this to discourage readers from taking the §5.2 regime-concentration diagnosis as a license to re-deploy the champion whenever "2021-2022-like conditions" are announced subjectively.

6. Conclusion

We report a null result and a methodological lesson. An SPX Donchian breakout rider, selected on 2023-2026 data with a prior-day VIX filter and a conservative trailing-stop exit, and confirmed on MES futures as venue-viable, does not survive a 19-year out-of-sample test. Its 2023-2026 edge is a forward extension of a regime-concentrated 2021-2022 anomaly; the parameters themselves rank in the 67th percentile from the top on the OOS sample and lose money in 2021-2022 when the filter is enforced.

We offer the following as suggestive practitioner guidelines — each consistent with our own experience, none pre-registered across multiple independent strategy programs:

1. **Pre-register the out-of-sample window before any selection step.** Five years is plausibly a defensible minimum for a strategy operating on daily or intraday data with regime turnover on the scale of a macro cycle; longer is safer for strategies whose mechanism plausibly depends on specific market-structure regimes.
2. **Validate the edge on the held-out sample, not on a different instrument derived from the same in-sample trade list.** Execution-layer cross-checks (SPX → SPXW → MES) validate friction, not edge. Both are necessary; only one is currently part of most research protocols we are aware of.
3. **Condition capital deployment on the OOS test, not on an execution-layer translation.** The Path C STRONG_GO verdict we wrote was technically correct (friction is manageable on MES) but strategically misleading (edge is regime-bound). Both parts of the sentence matter.
4. **Consider a positive-stability acceptance criterion ahead of time.** One plausible default, offered here as a starting point rather than a universal rule, is to require top-ranked configurations to produce positive annual PnL in at least 75-80% of OOS years with CIs excluding zero. This filters out monocultures whose total is rescued by a small number of years. The appropriate threshold likely depends on the strategy's expected holding period and the plausible frequency of regime turnover.

For the research program described here, the workstream is closed. The next strategy will not use the 2004-2022 SPX 1-second sample in its selection loop — that sample is now the only untouched resource we have for an eventual OOS test. No capital has been or will be deployed on the configuration described above.

The positive interpretation is that the protocol worked. A 3.1-year in-sample selection arc produced a candidate; a sequence of execution-layer validations refined the venue choice; a long-history OOS test vetoed the verdict before capital was deployed. Total cost: ~six researcher-weeks, ~$5.50 in Databento pulls, and 8.1 hours of compute for the OOS grid. The alternative — funding a paper-trading rollout on a regime-concentrated configuration and discovering its failure during the next market-structure shift — is orders of magnitude more expensive. In that sense, the null result is the product the research process is supposed to deliver, and documenting the product openly is the part of the job that this paper attempts to complete.

Acknowledgments

Interactive Brokers, Cboe, and Databento for data. The simulation engine was built in Python with NumPy, pandas, and PyArrow; Rust parity uses PyO3 and maturin; grids were parallelized with `concurrent.futures.ProcessPoolExecutor`.

References

Bailey, D. H., Borwein, J., Lopez de Prado, M., & Zhu, Q. (2014). "Pseudo-mathematics and financial charlatanism: the effects of backtest overfitting on out-of-sample performance." *Notices of the American Mathematical Society*, 61(5), 458-471.

Bailey, D. H., Borwein, J., Lopez de Prado, M., & Zhu, Q. (2016). "The probability of backtest overfitting." *Journal of Computational Finance*, 20(4), 39-70.

Bailey, D. H., & Lopez de Prado, M. (2014). "The deflated Sharpe ratio: correcting for selection bias, backtest overfitting, and non-normality." *Journal of Portfolio Management*, 40(5), 94-107.

Hansen, P. R. (2005). "A test for superior predictive ability." *Journal of Business & Economic Statistics*, 23(4), 365-380.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). "... and the cross-section of expected returns." *Review of Financial Studies*, 29(1), 5-68.

Leinweber, D. J. (2007). "Stupid data miner tricks: overfitting the S&P 500." *Journal of Investing*, 16(1), 15-22.

Lempérière, Y., Deremble, C., Seager, P., Potters, M., & Bouchaud, J.-P. (2014). "Two centuries of trend following." *Journal of Investment Strategies*, 3(3), 41-61.

Lo, A. W., & MacKinlay, A. C. (1990). "Data-snooping biases in tests of financial asset pricing models." *Review of Financial Studies*, 3(3), 431-467.

López de Prado, M. (2018). *Advances in Financial Machine Learning*. Wiley.

Sullivan, R., Timmermann, A., & White, H. (1999). "Data-snooping, technical trading rule performance, and the bootstrap." *Journal of Finance*, 54(5), 1647-1691.

White, H. (2000). "A reality check for data snooping." *Econometrica*, 68(5), 1097-1126.

Artifact index

All quantitative claims cite a concrete artifact under `results/` or `data/`. The primary artifacts are:

Claim (§)	Artifact
Phase 1 champion (§3.1, Table 3)	`results/run_20260420_162326_full_grid_phase1/metrics.parquet`
Phase 2 champion (§3.2)	`results/run_20260420_235118_full_grid_phase2/metrics.parquet`
Phase 4 champion (§3.3, Table 3)	`results/run_20260421_033134_full_grid_phase4/metrics.parquet`
MES Path C Stage 2 (§3.4, Fig 3, Table 3)	`results/path_c_stage2/summary_2_4.json`
Friction decomposition (§3.4, Fig 2)	`scripts/options_validation_stage2_cost.py` outputs
VIX30 vs VIX1D (§4.1, Fig 4)	`data/vix_data/`, `data/vix1d_data/VIX1D_1min_full_history.parquet`
460,800-combo grid (§4.3, Fig 5, Table 2)	`results/task_b2_rider_grid/metrics.parquet`
Top-50 monoculture (§4.3.2, Table 1)	`results/task_b2_rider_grid/report_tables/top_mode_c_by_total_pnl.parquet`
Champion neighborhood (§4.3.3, Fig 6)	`results/task_b2_rider_grid/metrics.parquet`
Per-year top-50 (§4.3.4, Fig 9)	`results/task_b2_rider_grid/report_tables/per_year_pnl_top.parquet`
VIX regression (§4.3.5, Fig 7)	`results/task_b2_rider_grid/report_tables/top50_year_loss_summary.parquet`
Bar quality (§4.4, Fig 8)	`results/task_b2_rider_grid/bar_quality.json`
Champion per-year reconstruction (§4.3.6, Fig 9)	Task 2 trades + VIX history; code in `reports/paper/generate_figures.py`
Task G synthesis (§4.5, §5)	`reports/task_g_synthesis.md`

Simulation source code: `src/spx_donchian_hf/` (Python reference) and `rust_engine/` (Rust parity). Figure and table generators: `reports/paper/generate_figures.py` and `reports/paper/generate_tables.py`.

---

1. In SPX index points, Mode C "per-trade PnL" is **already net of the 1.0 point round-trip slippage** — no further subtraction is applied when these per-trade numbers are reported (e.g., the Phase 4 +5.33 pts/trade). In the Path C MES tables, per-trade $ figures are additionally net of a $4 round-trip friction model ($2 per-side MES commission + spread proxy), stated explicitly on the relevant axis labels. ↩
2. Throughout the paper, "Phase 1 champion" refers to `tf=1, lb=5, sl=5, tp=10, ts=off, time_stop=25`; "Phase 2 champion" refers to `tf=1, lb=5, sl=5, tp=40, ts_act=3, ts=5,`

`time_stop=25`; "Phase 4 champion" refers to `tf=30, lb=50, sl=off, tp=off, ts_act=10, ts=10, time_stop=off, vix=18, dr=off`. All three use Mode C fills unless otherwise noted. ↩

3. Table 1 columns are one per calendar year 2004-2022 plus `total`. (a) All 10 rows share `tf=15, lb=20` — the monoculture feature summarized in §4.3.2. (b) The PnL contrast is primarily between 2021-2022 (green, large positives) and the four all-loss years 2009/2010/2012/2017 (red, all entries negative); if rendered with a color gradient, the resulting visual would match the per-year story in the main text, not provide independent evidence. ↩