Order Blocks Contain No Information: A Preregistered Falsification of a Smart-Money-Concepts Trend-Continuation and Retest Strategy

Shiyu Guan[a], IRIS[a]

[a]*EROS Quantitative Laboratory, China*

Abstract

Smart Money Concepts (SMC) is among the most widely taught entry methods in retail trading education. Its central claim is mechanical: after a break of structure (BOS) and a displacement leg confirm a higher-timeframe trend, price returns to the "order block" that originated the displacement, reverses there, and continues with the trend, so a resting limit order with a tight stop earns a high reward-to-risk ratio. It has almost never been tested without look-ahead and with realistic costs. We mechanize it unambiguously and test it on 19 OKX USDT perpetual futures from January 2021 to July 2026 (5.5 years, 3.65 million 15-minute bars) under preregistered gates, stratified randomized controls, and a full-cost backtest. The findings come in three layers. *Literally true*: after price retests an order block it reaches +1 ATR before −1 ATR 55.5% of the time, and the 90% interval excludes 50%. *Wrongly attributed*: that rate is fully explained by the depth of pullback waited for—immediate entry gives exactly 50.3% and size-matched random zones give more (59.2%)—not by memory attached to the level, and the incremental effect over matched random zones is −0.134 ATR with an interval excluding zero. *Too small*: the reversal's gross expectation is 68% of one round-trip cost, all 48 cells of the full-cost backtest are negative (deflated Sharpe = −3.41), and equity falls from 1.00 to 0.28. Three supplementary experiments show the verdict insensitive to bootstrap block length (24/24 checks pass), replicated with the same sign on a second venue, and statistically equivalent to zero within a bound of one round-trip cost—a falsification backed by power, not an underpowered null. We also report a fill-bar convention error that reverses the conclusion: before the fix, order blocks showed a spurious gross Sharpe ratio of 3.41.

Preprint submitted to Working paper

Keywords: order blocks, smart money concepts, technical analysis, event study, preregistration, equivalence testing, transaction costs, cryptocurrency perpetual futures
JEL: G14, G17, C12, G40

---

1. Introduction

1.1. The question

Smart Money Concepts (SMC) is one of the most popular entry frameworks on YouTube, in crypto communities, and in retail trading education. The canonical recipe has four steps. Establish direction on a higher timeframe, say four hours. Wait for a break of structure (BOS) on an intermediate timeframe and require a *displacement* leg to confirm momentum. Mark the last opposite-direction candle before that leg began as the *order block*. Rest a limit order in that zone, wait for price to retest it, and pair the entry with a tight stop to obtain a high reward-to-risk ratio. The claim advanced by its proponents is explicitly causal: the order block is the footprint left by institutional accumulation, and price reverses there because unfilled orders remain.

The method travels widely; it has almost never been tested rigorously, without look-ahead, and with realistic costs. This gives it a different status from the rules that the technical-analysis literature usually examines. Our object is not an abstract moving-average or channel-breakout rule[2, 3] but a *specific, operational method that a large number of traders actually use*. We therefore ask a falsifiable question: **conditional on the same higher-timeframe trend state, the same volatility, and the same pullback depth, is the forward return after price retests an order block materially better than the forward return after price retests a matched control zone?** If the answer is no, the claim that order blocks carry information is false, however high their reversal rate may be.

1.2. Related work

Whether technical analysis contains information is an old debate, and it has always been conducted in the shadow of the efficient-markets benchmark[1]. Brock, Lakonishok, and LeBaron find moderate excess returns for moving-average and trading-range-break rules on the Dow [2]; Lo, Mamaysky, and Wang provide a kernel-regression framework for statistical

inference on chart patterns[3]. In crypto markets, Urquhart documents significant autocorrelation in Bitcoin returns and rejects weak-form efficiency, leaving room for technical rules[4].

That tradition has always been accompanied by a data-snooping problem. Sullivan, Timmermann, and White show, using White's reality check, that the Brock et al. rules cease to be significant once multiple testing is accounted for[5, 6]. Harvey and Liu document systematically how selection bias and overfitting inflate backtested Sharpe ratios[7, 8]. The deflated Sharpe ratio of Bailey and López de Prado discounts that inflation directly[9]; we use it in the backtest stage. McLean and Pontiff show that factor returns decay after publication, a substantial part of which is a data-mining artifact[10].

Methodologically we borrow two tools that upgrade a "null result" into a falsification with power behind it. The first is *equivalence testing*. Lakens argues for two one-sided tests (TOST) against a preset smallest effect size of interest (SESOI), so that "not significant" is not misread as "no effect"[11, 12]. We set the SESOI to one round-trip transaction cost: an effect smaller than the cost of getting in and out is economically indistinguishable from no effect even if it is real. The second is the *stationary bootstrap*[13], used as an unequal-length counterpart to calendar-month block resampling in order to test whether the verdict survives the subjective choice of block length.

Relative to this literature the paper differs in three ways. The object of study is a concrete retail method rather than an abstract rule. The design places the *event study before the P&L backtest*, so that the verdict is reached at the level of effect size. And falsification is the explicit goal: every design choice—preregistration, frozen gates, stratified randomized controls, equivalence testing—serves one question, namely whether we can say convincingly *where* the method is wrong if it is wrong.

1.3. Contribution

The paper contributes three things. *Infrastructure*: a multi-timeframe structure engine that passes 31 automated tests—including bit-for-bit truncated-history replay and a deliberately leaking negative control—and turns SMC vocabulary into a computable, look-ahead-free research object reusable for any "multi-timeframe structure plus retest" hypothesis (Section 2). *Protocol*: a complete falsification pipeline combining preregistration, frozen gates, stratified randomized controls, monthly-block bootstrap, deflated Sharpe, and event study before P&L; the negative answer already arrives in the event study (P2), and the backtest (P3) merely

confirms it (Section 3). *Findings and a case study*: a three-layer verdict on a popular method—literally true, wrongly attributed, too small—plus a fill-bar convention error that reverses the conclusion and carries a warning for technical-strategy backtests generally (Sections 4–6). Section 7 discusses mechanism, methodological lessons, and limitations.

2. Data and the mechanization of the strategy

2.1. Data

The universe is 19 core OKX USDT perpetual contracts, all of which exist throughout the sample, so no survivorship screen is applied; ten later-listed contracts are held in a separate subsample and do not enter the main results. The window runs from 2021-01-29 to 2026-07-25: 66 calendar months and roughly 3.65 million 15-minute bars. Structure is organized on three timeframes. Four hours is the higher timeframe (HTF) and fixes trend direction; one hour is the intermediate timeframe (MTF) and carries BOS events, displacement, and order blocks; 15 minutes is the lower timeframe (LTF) and serves as the substrate for retest detection and path reconstruction. The one-hour and four-hour series are aggregated from 15-minute bars. One-minute Bitcoin data are used only to calibrate the conservative assumption applied when entry and stop are touched inside the same bar; they never enter a signal. Funding is taken from Binance perpetual funding as a proxy for OKX—the two venues' rates are highly correlated—and is charged as a carry cost. Table 1 summarizes the data and the structural inventory.

Table 1: Data and structural inventory.

Universe	OKX USDT perpetuals, 19 core symbols
Sample window	2021-01-29 to 2026-07-25 (66 months)
Base bars	15-minute, $\approx$ 3.65 million
Structure timeframes	HTF = 4h, MTF = 1h, LTF = 15m
BOS events	41,620
Trend-aligned qualified order blocks	9,460
First-retest fills (main analysis sample)	6,662
Share of time with a definite HTF trend	31.1% up, 30.0% down
Order blocks per BOS	1.000

Structural counts come from the P1 acceptance run. The remaining 38.9% of the time the HTF trend is a range, in which no position is opened. That order blocks per BOS equals exactly 1.000 means that "an order block exists" is an automatic consequence of a BOS rather than a screening condition (Section 2.4).

2.2. Making SMC precise

SMC is stated verbally and loosely; testing it requires writing it down without ambiguity. Table 2 gives computable definitions of the core concepts, with the full set of definitions and parameter values in Appendix A. Three points matter. First, every structural object—pivots and BOS events alike—carries a *confirmation delay*: a pivot becomes real only after the L bars to its right have closed, and the event stream is processed in order of confirmation time rather than bar label. Second, an order block is defined operationally as the last opposite-direction candle before the displacement leg begins, with zone boundaries taken from that candle's high and low (the wick convention, used in the main results) or its open and close (the body convention, reported in the appendix), and with width normalized by ATR and capped. Third, order blocks have an explicit life cycle: they terminate on first touch, expiry, trend reversal, or a close through the zone, and second and third touches are recorded separately and never pooled with the first.

Table 2: Computable definitions of the core SMC concepts (summary).

Concept	Computable definition
Pivot	Local extremum, strictly greater on the left and not less on the right (ties resolved to the first bar), passed through an ATR amplitude filter to yield an alternating high–low zigzag; confirmed only after L bars have closed to its right.
Trend state (HTF)	Up if the last two highs and lows form HH-HL, down if LH-LL, range otherwise; no new position is opened in a range.
BOS	A close beyond the most recent confirmed swing extreme that has not already been consumed by an earlier BOS.
Displacement	Net travel on the BOS leg of at least $d_{\min}$ ATR with directional efficiency of at least $\text{eff}_{\min}$; otherwise the event is recorded as a "no-displacement BOS" and retained as a control.
Order block	The last opposite-direction candle before the displacement leg begins; the zone spans its high–low (wick convention) or open–close (body convention), with width at most $w_{\max}$ ATR.
Life cycle	Activated at the BOS close; terminated by first touch, expiry, trend reversal, or a close through the zone; first touches are tabulated separately from second and third touches.

Full definitions, parameter values, and the reasoning behind them are in Appendix A and in the frozen design document. All parameters were frozen before any data were run and take the midpoint, not an endpoint, of the candidate range from the upstream design.

2.3. The causality constraint and 31 automated tests

The most common source of look-ahead in a multi-timeframe strategy is an `asof` merge performed directly on coarse-timeframe labels, which makes a four-hour bar visible before it has finished forming. We eliminate it with a *single* causality constraint: for a bar on any timeframe with label t and length D, every quantity derived from it—OHLC, ATR, pivots, BOS events—becomes available at avail $= t + D$, and a decision taken at lower-timeframe instant u may use only quantities with avail $\leq u$. Every higher-timeframe feature table therefore carries an `avail_at` column; alignment to the lower-timeframe grid is only ever `merge_asof(timestamp, avail_at)`, and the merged result is asserted to satisfy `avail_at` $\leq$ `timestamp`.

Around that constraint the engine passes 31 automated tests in four groups (listed in Appendix B): structural self-consistency; causality, including bit-for-bit truncated-history replay and a *deliberately leaking negative control* that guarantees the causality tests are not vacuously true; fill and measurement semantics; and the requirement that controls and the main sample be anchored to the same events. This test suite is the foundation of whatever credibility the study has.

2.4. A structural fact: the existence of an order block is not a filter

The P1 acceptance run delivers one fact that has to be stated up front: order blocks/BOS events = 1.000. In other words, *every BOS necessarily traces back to an opposite-direction candle.* "Finding the order block"—which sounds like the identification of a scarce structure—is in fact a tautological consequence of a BOS and screens nothing at all. The screening work is done by displacement quality and trend alignment. This determines that the decomposition of the filter chain must be done *stepwise* (Section 4.2): without stepwise attribution, the contribution of the trend filter would be misattributed to the order block.

2.5. A fill-bar convention that reverses the conclusion

Before reporting any result we record an implementation error caught during the work that, left uncorrected, *reverses the conclusion entirely*; it carries a warning for technical-strategy backtests generally. The first version of the first-crossing resolver *allowed the fill bar's own high to reach the target.* At 15-minute resolution this is undecidable: a long fill occurs at the moment the bar trades down to the near edge of the zone, and the same bar's high may perfectly well have printed before its low—that is, before the fill happened.

The bias *benefits only order blocks*, because the order-block stop radius R is just 0.91 ATR (against 3.0 ATR for the pullback rules), so a distance of $1R$ is small enough to be covered inside a single 15-minute bar. The consequence: before the fix, `order_block|1R|h48` reached its target inside the fill bar in 24.5% of cases (against 0.15% for the pullback rules), with a gross Sharpe ratio of **3.41** and a 55.5% win rate—the only standout cell in the entire grid, and in direct conflict with the negative verdict of the event study. After the fix the gross mean turns negative and the Sharpe ratio is -0.86; the conflict disappears. We record the rule of thumb: *whenever only the tightest-stop cell looks exceptional, check the fill-bar convention first.*

3. Research design: effect size first, P&L second

3.1. The core question and its derivatives

The design does not presume that order blocks predict anything. It organizes a set of derived questions around the core question Q0, whether order blocks carry independent information. Q1: does the higher-timeframe structure filter have value on its own? Q2: does the intermediate-timeframe BOS add anything on top of the trend filter? Q3: do displacement magnitude and efficiency add anything? Q4: does the order-block zone add anything relative to an arbitrary zone of the same width—the core of Q0? Q5: is the first retest better than the second and third? Q6: does waiting for a lower-timeframe change of character (CHOCH) raise or lower net expectation relative to a resting limit order? Q7: is a structural exit better than a fixed R? Q8: does any of these advantages survive costs? Each derived question is unpacked one filter layer at a time, and each added layer must answer two questions: how much forward expectation it adds, and how much sample it costs.

3.2. Controls A–E

Falsification hinges on the controls. All five control families are *anchored to the same events* as the main sample; family D, the stratified random pseudo-zone, is the direct test of Q0 (Table 3). D is matched by strata of (symbol, calendar month, HTF trend state, expanding-window ATR quantile bucket), constructs pseudo-zones of identical geometry to the real order blocks, and draws 20 samples per real event so that the control is a distribution rather than a point.

Table 3: Definition of controls A–E and the question each answers.

Family	Construction	Question answered
A	Market entry immediately on trend confirmation, no waiting for a retest	Does waiting for the retest raise expectation, or only the apparent reward-to-risk ratio of the trades that do fill?
B	Triggered by a pullback of $\{0.5, 1.0, 1.5\}$ ATR from the recent extreme	Are order blocks better than an ordinary volatility-pullback entry?
C	Retest of EMA20 or EMA50	Are order blocks better than the most naive trend-pullback tool?
D	**Random pseudo order block**: a random bar within the same stratum, made into a zone of the same ATR width and geometry, run through an identical touch and forward-measurement pipeline	Does the zone itself carry independent information? (Direct test of Q0)
E	Displacement origin zone without requiring a BOS	Separates the contributions of displacement and of the structural break

All families are matched on (symbol, calendar month, HTF trend state, expanding-window ATR quantile bucket). The ATR quantile used for stratification is an expanding-window quantile, not a full-sample quantile backfilled onto history, which would inject look-ahead into the control itself. The headline estimand Δ_{simple} is measured against the *strongest* of the seven A/B/C controls, the setting least favorable to order blocks.

3.3. Preregistration and acceptance gates

The four gates of the event study (P2) were frozen before any result existed. G2-0, power: sample size at least 500. G2-1, informativeness: the 5th percentile of Δ_D above zero. G2-2, superiority over simple rules: the point estimate of Δ_{simple} above zero with a 5th percentile above -0.05. G2-3, tradability: the gross order-block effect at least three times the round-trip cost, that is 0.42%. Four possible outcomes, from "worth continuing" to "order blocks falsified", were mapped to ex post actions in advance. In the backtest stage (P3) the 48-cell grid is pinned at `import` time by `assert len == 48`, and the ordering—design period fully resolved before validate or test is read—is enforced in code. The P3 preregistration document, however, was written after the results were seen; it is weaker than P2 and has been explicitly downgraded and disclosed (Section 3.6).

3.4. Statistical conventions

All uncertainty is estimated by *calendar-month block bootstrap*: 2,000 resamples, whole months drawn jointly across all symbols, a shared resampling matrix so that between-group comparisons are paired, and the random seed frozen at `0xA57E`. We report 90% intervals and no significance stars, and event counts accompany every subsample. The backtest stage discounts the 48-cell search with the deflated Sharpe ratio[9]. The *power boundary* of the study should be stated in advance: returns to trend strategies are dominated by the right tail and are highly correlated across symbols, so the effective number of independent observations is closer to the number of independent market regimes, on the order of 100–200. What the study can answer reliably is therefore whether an effect is large and whether its sign is stable, not whether a Sharpe ratio is 1.8 or 2.1.

3.5. Why the event study precedes the P&L backtest

We deliberately put the event study first, for three reasons. (a) The event study uses *every touch* rather than only filled trades, which raises power considerably. (b) It decouples signal from execution, so that an effect can be attributed to a single link in the chain; otherwise trend, displacement, order block, and exit remain four entangled unknowns. (c) Exit models can be computed ex post from the same path table, which removes half of the backtesting work. Our own experience shows that the ordering is essential: had the backtest come first and the signal search second, the project might well have stopped at the spurious Sharpe ratio of 3.41 in Section 2.5. The overall workflow is shown in Figure 1.

3.6. Multiple testing and honest disclosure

Research discipline is enforced by code rather than by self-restraint, and is logged line by line in a trial ledger. The event study's `n_configs` is recorded as 2 because two zone conventions, wick and body, were both inspected; the P3 grid is recorded as 48. Two post-hoc analyses—the body convention and CHOCH confirmation—are labeled as such and never tabulated alongside preregistered results. The P3 preregistration was written after the fact and is downgraded accordingly. These disclosures, together with the limitations in Section 7, form the framework within which the evidence in this paper should be graded.

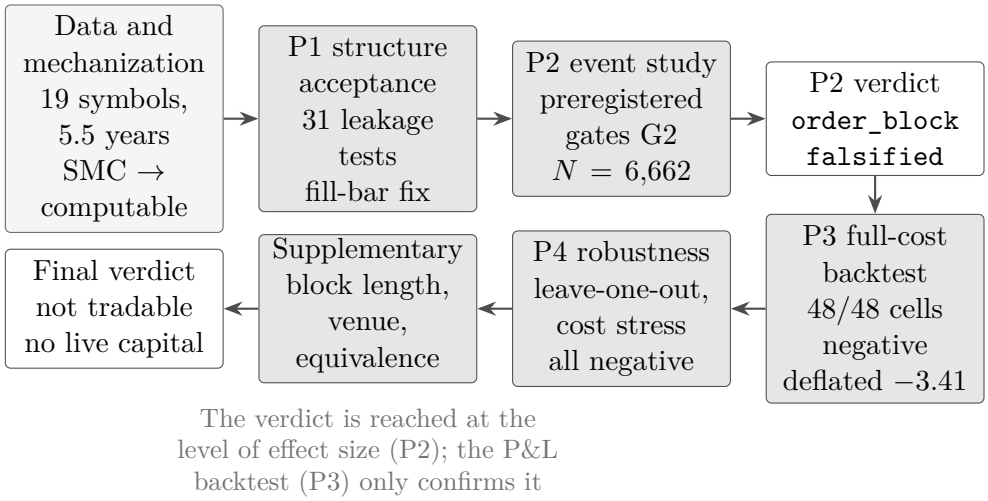

Figure 1: Research workflow. The event study (P2) delivers the verdict at the level of effect size and the P&L backtest (P3) merely confirms it. Every stage carries its own ex ante constraint: preregistered gates, a frozen grid, stratified randomized controls, and a multiple-testing discount.

4. Event-study results (P2)

4.1. The verdict

The test statistic is the *trend-direction return over the 12 hours after the fill*, normalized by the one-hour ATR at the moment of the fill. The incremental effect of order blocks relative to matched random zones is $\Delta_D = -0.1343$ ATR (90% CI $[-0.2079, -0.0542]$, $N = 6{,}662$); relative to the strongest simple control anchored at the same BOS instant it is $\Delta_{\text{simple}} = -0.0186$ (CI $[-0.0941, +0.0240]$, which spans zero). The order block's own gross effect is $\mu_{\text{OB}} = 0.0199$ ATR $= 0.0246\%$, only 18% of a single round-trip cost (14 bp ≈ 0.113 ATR). One of the four preregistered gates is passed (Table 4), and the verdict is `order_block_falsified`. That the gate passed is G2-0 matters: *this is a rejection backed by power, not a complaint about sample size.*

Table 4: P2 gate decisions.

Gate	Preregistered criterion	Realized	Verdict
G2-0 power	$N \geq 500$	$N = 6{,}662$	Pass
G2-1 informativeness	5th pct. of $\Delta_D > 0$	$\Delta_D = -0.1343$, CI $[-0.2079, -0.0542]$	Fail
G2-2 better than simple rules	Point estimate > 0 and 5th pct. > -0.05	$\Delta_{\text{simple}} = -0.0186$, CI $[-0.0941, +0.0240]$	Fail
G2-3 tradability	$\mu_{\text{OB}} \geq 3\times$ round-trip cost $= 0.42\%$	$\mu_{\text{OB}} = 0.0246\%$	Fail

The strongest simple control is family B at 1.5 ATR ($\mu = +0.0385$). Cost basis: 5 bp taker in, 5 bp taker out, 2 bp slippage per side, for 14 bp round trip, converted at a median ATR/price of 0.01238 to 0.113 ATR. Robustness statistics for μ_{OB}: winsorized mean -0.0086, median -0.0799, median across symbols -0.0187, and 9 of 19 symbols positive.

Two clarifications belong before any interpretation. First, the order block's point estimate of $+0.0199$ looks positive, but the interval is wide enough to span zero, the winsorized mean and the median are both negative, and only 9 of 19 symbols are positive—*even the weakest possible claim, that order blocks have positive expectation in themselves, is unsupported*. Second, Δ_D has an identification defect: family D is not matched on distance to the nearest swing extreme. An order block sits at the origin of the displacement leg, a location price reaches by turning back after having already travelled, whereas a random pseudo-zone may fall in the middle of a trend leg. Δ_D therefore mixes in a positional effect, namely that the origin of a displacement leg is a worse place to be than the middle of a trend. We accordingly treat Δ_{simple} as the **headline estimand**: families A, B, and C are anchored at the same BOS instant, making the comparison strictly within-event, with the only difference being which price level one waits for. Δ_D is demoted to a reference benchmark. That Δ_D remains significantly negative still says something: size-matched random zones do not produce worse 12-hour trend-direction returns than order blocks, and all 19 symbols are positive—the positive mean coming mainly from the drift of the crypto market itself over 2021–2026 rather than from any entry method.

4.2. An ablation ladder: which layer carries the effect

Stacking SMC's three filter layers one at a time, with each layer a subset of the one before, gives the 12-hour trend-direction returns to "enter immedi-

ately once this layer passes" shown in Table 5 and Figure 2. **Not one of the six rungs has an interval that excludes zero**: the point estimate rises from +0.003 to +0.028 and falls back to +0.006, and the whole excursion lies within noise. Reading this ladder as evidence that "displacement magnitude helps" or that "order blocks hurt" would over-interpret it in either direction.

Table 5: Ablation ladder L1–L6: 12-hour trend-direction return to immediate entry once each layer passes.

Layer	N	μ (ATR)	90% CI	Increment
L1 HTF trend	557,831	0.0029	$[-0.0414, 0.0465]$	—
L2 + BOS	12,235	0.0141	$[-0.0639, 0.0920]$	+0.0112
L3 + displacement magnitude	10,732	0.0278	$[-0.0554, 0.1133]$	+0.0137
L4 + displacement efficiency	10,731	0.0274	$[-0.0558, 0.1131]$	−0.0004
L5 + order block exists	9,458	0.0058	$[-0.0842, 0.0973]$	−0.0216
L6 + wait for the first retest	6,662	0.0199	$[-0.0761, 0.1102]$	+0.0141

L1–L5 enter at the BOS close as soon as the layer passes; L6 waits in addition for a retest fill. Two incidental findings: the `min_efficiency` filter eliminates a single observation (10,732 → 10,731) and is therefore a dead parameter; and the sample loss from L4 to L5 comes from zone-width and other quality gates, not from the order-block concept itself, whose existence is tautological (Section 2.4).

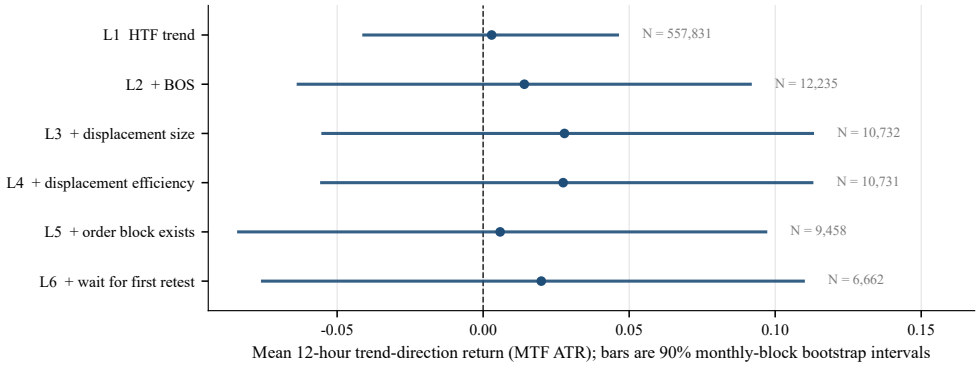

Figure 2: Point estimates and 90% monthly-block bootstrap intervals for the ablation ladder L1–L6. No rung has an interval excluding zero: nowhere along the chain "trade with the higher-timeframe trend, wait for a BOS, require displacement, locate the order block, wait for the retest" is any link's effect size distinguishable from zero.

4.3. The reversal-frequency test: is the literal claim correct?

Everything above measures *returns*, whereas SMC's literal claim is narrower—*price is more likely to reverse at an order block*—which is a claim about *frequency* and could hold while the strategy still loses money. It deserves a separate test. The statistic is simple: walk forward along the 15-minute path from the fill price and record whether +1 ATR or −1 ATR arrives first. Four properties recommend it. The no-information benchmark is exactly 50/50, so the number can be read on its own. The statistic is symmetric and does not depend on a stop definition. Both barriers sit well outside the zone, so this is not a restatement of "did the zone hold". And the ATR unit is scale-free, so all eight rules can be placed on one axis. Results are in Table 6 and Figure 3.

Table 6: Reversal frequency against cost, by descending reversal rate.

Entry rule	N	Reversal	90% CI	Breach	Gross	Net
D random pseudo-zone	118,649	**59.24%**	[58.55, 59.93]	27.96%	0.1375	**+0.0251**
C EMA50	7,682	56.71%	[55.38, 58.07]	—	0.0936	−0.0188
Order block	6,666	**55.54%**	**[54.11, 56.97]**	**30.30%**	**0.0761**	**−0.0364**
B 1.5 ATR pullback	9,061	53.25%	[52.05, 54.47]	—	0.0530	−0.0595
C EMA20	8,812	52.61%	[51.60, 53.65]	—	0.0321	−0.0803
B 1.0 ATR pullback	9,299	50.75%	[49.55, 51.88]	—	0.0110	−0.1015
A immediate entry	9,460	50.28%	[49.09, 51.50]	—	0.0036	−0.1088
B 0.5 ATR pullback	9,442	49.99%	[48.84, 51.15]	—	−0.0015	−0.1139

The reversal rate is the share of *adjudicated* observations—those that neither breach both barriers within one bar nor time out—that reach +1 ATR first. "Breach" is the share of zones through which price passes completely. Gross and net expectations use the symmetric ±1 ATR bracket on the full sample (a win counts +1, a loss −1, a same-bar double touch as a loss, a timeout as 0). Cost is one round trip of 14 bp = 0.1124 ATR and excludes funding. The reversal-rate gap between order blocks and random pseudo-zones is −3.69 percentage points, 90% CI [−5.08, −2.35] under paired resampling. The only rule with positive net expectation is the untradable random control.

Four observations have to be read together. *First, the order block's reversal rate really is significantly above 50%* (55.54%, with an interval excluding 50)—the one result in the entire project that supports the method literally. *Second, this is not the order block's own contribution*: the whole table is ordered monotonically in the depth of the pullback waited for, immediate entry lands at exactly 50.28% (which is also a self-check on the measurement), and the order block's rate is precisely what its pullback depth predicts. The

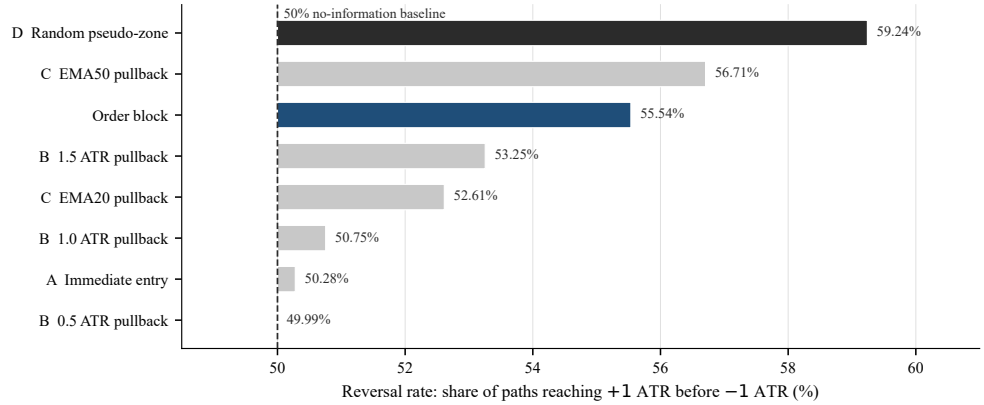

Figure 3: Reversal rates ordered monotonically in the depth of the pullback waited for. Immediate entry lands at exactly 50.28%, which doubles as a check that the measurement is unbiased; the order block sits precisely at the level its pullback depth predicts and contributes nothing beyond it; size-matched random pseudo-zones reverse more often. The dashed line is the 50% no-information benchmark.

mechanism is the *origin of measurement*: the deeper one waits, the closer +1 ATR is to where price has just been, and the easier it is to reach. *Third, size-matched random zones reverse more often* (59.24%, a gap of −3.69 percentage points with an interval excluding zero), and the share of order blocks that price breaches *completely* (30.30%) is if anything slightly higher than for random zones (27.96%)—order blocks are no more "solid" than a randomly chosen zone of the same size. *Fourth, and decisively, there is the magnitude*: the gross expectation of the reversal is 0.0761 ATR against a round-trip cost of 0.1124 ATR.

4.4. Why a real reversal still loses money

The last column of Table 6 holds the most easily grasped number in the project. The order block's reversal is real, with a gross expectation of +0.0761 ATR, but the cost of getting in and out once is 0.1124 ATR: *a genuine edge worth 68% of its own cost.* The only rule with positive net expectation is the random control at +0.0251 ATR, and it is not tradable; all seven tradable rules are negative. This is the same fact as gate G2-3 in Section 4.1, where the gross effect is 18% of the round trip, computed under a different exit model; the two agree in sign.

4.5. Opportunity cost and return concentration

Waiting for the retest is not free. Of the 9,460 trend-aligned qualified zones, **29.5% are never retested before they terminate**, and the moves thereby forgone are not moves that failed to start: within 12 hours 87.4% had already travelled at least 1 ATR with the trend, with a median of 2.92 ATR, and within 24 hours 93.7% had, with a median of 4.05 ATR. This is why L2, immediate entry after the BOS, and L6, waiting for the retest, are indistinguishable in mean: *the better entry price bought by waiting is offset exactly by the smoothest trends that waiting misses.* In a competitive market, a better price and a lower fill probability are two sides of the same trade.

Return concentration is the single most important number in the paper. On the full sample $\mu = +0.0199$; dropping the best 1% of observations, 67 events, gives $\mu = -0.1176$, so **those 67 events contribute 685% of the total**. The mean that "looks positive" in the headline specification rests entirely on 1% of the sample, and the remaining 6,595 events sum to something significantly negative. Any claim that the method works must first explain why one should expect those 67 events to recur. Beyond that, even among the observations that eventually turned a profit, half had first moved more than 0.75 ATR against the position—the "precision entry" story does not survive contact with the data.

4.6. Two leads that point the right way and still fall short

Two results point in a direction favorable to the method. We report them and explain why they remain insufficient. *The body zone convention*: defining the order block by open–close rather than high–low improves every metric. μ_{OB} rises from 0.0199 to 0.0474, Δ_{simple} flips from -0.019 to $+0.009$ (the first time an order-block mean slightly exceeds the strongest simple control), while Δ_D remains -0.103 with an interval excluding zero. It still passes no gate—the gross effect is 42% of a round trip—and it is the second convention examined on the same data, which the ledger records honestly as `n_configs=2`. Seeing a quarter of a standard deviation of drift across two attempts is entirely expected. *Waiting for a lower-timeframe CHOCH*: the fill rate falls to 83.2% and the entry price is 0.43 ATR worse, in exchange for a 12-hour gross mean rising from 0.0181 to 0.0462 ATR (gross over cost from 0.18 to 0.42) and a net loss per opportunity narrowing from -0.092 to -0.054 ATR. The direction is right, but the interval on the improvement spans zero ($+0.028$, CI $[-0.042, +0.093]$) and the analysis is post-hoc (flagged POST-HOC in the ledger). Both leads stall at a gross-to-cost ratio of 0.42, still

more than a factor of two short of breakeven, and both improve by the same mechanism—trade less, and enter only after the move has begun—which is the same pattern as the grid optimum sitting on the parameter boundary in Section 5.

5. Backtest with realistic costs (P3–P4)

5.1. Forty-eight cells, all negative, and the multiple-testing discount

What P2 falsifies is *the order block as a specific price level*, not the more general practice of entering on a pullback with the trend. P3 therefore removes the order-block layer, retains the "HTF trend + BOS + pullback entry" frame, and backtests all 48 cells with realistic costs (14 bp round trip plus 3 bp per day of funding), realistic concurrency constraints, and out-of-sample splits. The result is *uniformly negative*: of 48 configurations, none has a positive `mean_R` in the design period and none has a positive Sharpe ratio. The configuration selected on design-period Sharpe is `atr_pullback_1|time|h96`, whose design Sharpe is −0.056 and whose **deflated Sharpe ratio is** −3.41: searching a 48-cell grid in which no cell has any edge at all would still deliver, on average, an inflated Sharpe ratio of +3.35 simply by selecting the best cell, and the best cell actually found is −0.056, far worse than the no-edge null (Table 7).

Table 7: Top eight of the 48 cells in the P3 design period (2021–2023), by Sharpe ratio.

Configuration	Trades	mean_R	Win rate	Sharpe	Max DD	Total R
atr_pullback_1|time|h96	3,373	−0.0111	38.7%	**−0.056**	−67.3%	−37.6
atr_pullback_0.5|time|h96	3,428	−0.0153	38.8%	−0.114	−67.1%	−52.5
atr_pullback_0.5|time|h48	4,112	−0.0209	42.5%	−0.282	−67.5%	−86.0
atr_pullback_1|time|h48	4,062	−0.0348	40.5%	−0.443	−73.2%	−141.3
atr_pullback_1|time|h24	4,529	−0.0363	41.7%	−0.534	−74.3%	−164.6
atr_pullback_0.5|time|h24	4,611	−0.0358	42.6%	−0.965	−65.0%	−164.9
order_block|time|h96	3,327	−0.1308	19.5%	−1.075	−95.6%	−435.1
atr_pullback_0.5|3R|h48	4,140	−0.0428	42.5%	−1.168	−71.9%	−177.3

Among the 48 cells, zero have `mean_R`> 0 and zero have a positive Sharpe ratio. The deflated Sharpe ratio is −3.41, implied by an $\mathbb{E}[\max \text{Sharpe} \mid \text{no edge}] = +3.35$ derived from the cross-sectional dispersion of Sharpe ratios across the grid. The grid optimum sits on the parameter boundary—longest horizon, no profit target, pure time exit—exhibiting a "trade less, lose less" pattern. Order blocks are the worst entry rule in the grid (Sharpe −1.07, drawdown −95.6%): their tight stop makes cost consume 0.16 of one R, against 0.05–0.07 for the pullback rules.

5.2. A scale-free comparison across entry rules

Stop radii differ across entry rules—0.91 ATR for order blocks against roughly 3 ATR for the pullback rules—so `mean_R` cannot be compared across them. Converting gross return, cost, and net return into scale-free ATR units, holding the exit fixed at `time|h96` and the cost fixed at 14 bp plus 3 bp per day, produces the clearest table in the project (Table 8 and Figure 4): **cost is a horizontal line at roughly 0.146 ATR, not one of the nine rules has a gross edge that reaches it, the closest to breakeven is the random control, and the order block ranks second from last.**

Table 8: Scale-free comparison of nine entry rules (common exit, common cost, units of ATR).

Entry rule	Trades	Gross	Cost	Net	Gross/cost
D random pseudo-zone	118,593	**0.1507**	0.1519	−0.0012	**0.99**
C EMA50	7,679	0.0714	0.1458	−0.0744	0.49
B 1.0 ATR pullback	9,292	0.0543	0.1467	−0.0925	0.37
B 1.5 ATR pullback	9,054	0.0519	0.1467	−0.0948	0.35
A immediate entry	9,453	0.0378	0.1459	−0.1081	0.26
B 0.5 ATR pullback	9,435	0.0363	0.1461	−0.1098	0.25
E displacement without BOS	51,986	0.0291	0.1453	−0.1162	0.20
Order block	6,660	**0.0282**	0.1466	−0.1184	**0.19**
C EMA20	8,805	−0.0077	0.1428	−0.1505	−0.05

Every trade is converted to ATR units and no concurrency constraint is imposed, since the random control is oversampled twentyfold and the comparison is of per-trade expectations. The random control is not a tradable strategy: it is unmatched on distance to the swing extreme, so its edge blends the positional advantage of mid-trend locations with market drift. Its only role is to calibrate how much structural screening adds relative to random entry, and the answer is negative.

5.3. Out-of-sample results and cost stress

The selected configuration stays negative out of sample: in the validate period (2024–2025) `mean_R` $= -0.063$ with a bootstrap $p(\texttt{mean\_R} \leq 0) = 0.943$ against a gate requiring below 0.10, and the test period (2026) is negative as well. The frame is nevertheless *not entirely without information*: at zero cost, pullback entry with the trend has a gross expectation of +0.026 R per trade and a gross Sharpe ratio of +0.39. The problem is what breakeven requires (Table 9): roughly **7 bp round-trip cost and zero funding**.

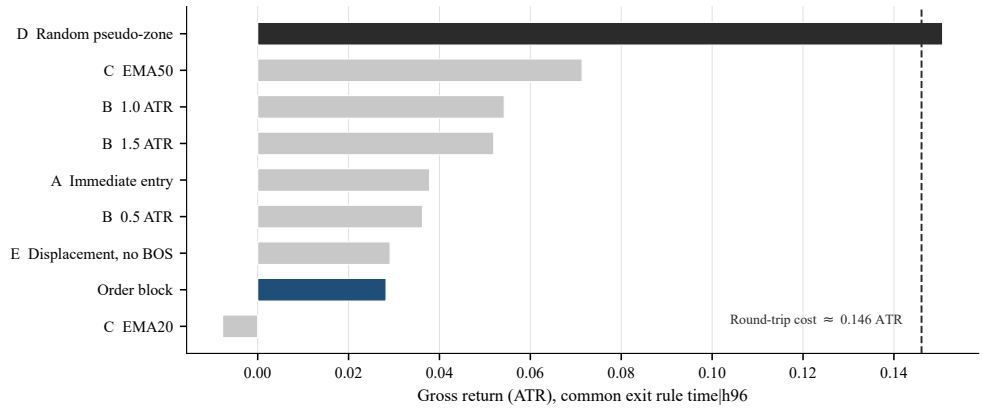

Figure 4: Gross return of nine entry rules in ATR under the common exit `time|h96`, against the horizontal cost line. No rule's gross edge reaches cost; the closest to breakeven is the untradable random control, and the order block ranks second from last.

Table 9: Cost stress test for the selected configuration `atr_pullback_1|time|h96`.

Round trip (bp)	Funding (bp/day)	mean_R	Sharpe	Total R
7	**0**	**−0.0006**	**+0.063**	−3.9
7	3	−0.0047	+0.013	−29.0
7	6	−0.0087	−0.038	−54.2
14 (baseline)	**3**	**−0.0310**	**−0.315**	−193.3
25	3	−0.0724	−0.828	−451.4
40	3	−0.1289	−1.525	−803.4

The zero-cost benchmark has gross `mean_R` = +0.0257 and gross Sharpe +0.390. Breakeven requires about 7 bp round trip with zero funding. Seven basis points implies maker fills on both sides with no slippage, but the configuration exits on time—an unconditional close after 24 hours—so a maker fill cannot be guaranteed. Zero funding is impossible on a perpetual: a 24-hour holding period spans three funding settlements. The escape route "there is information, costs merely eat it" is therefore closed on perpetual futures.

That distinction determines what to do next. If a frame carries no information at all, it should be abandoned; if it carries information whose magnitude falls short of cost, one should look for cheaper execution. Since 7 bp round trip with zero funding is unreachable on perpetuals, the second route is closed too. Of the six G3 gates, G3-0 passes, while G3-1 (deflated Sharpe

−3.41), G3-2 (validate $p = 0.943$), G3-3 (`mean_R` −0.072 at 25 bp), and G3-4 (−0.123 after dropping the top 1%) all fail; G3-5 is formally satisfied but uninformative, since all nine neighboring cells are negative as well.

5.4. Robustness and the equity curve

The robustness checks are *uniformly negative*. Six leave-one-year-out runs: all negative. Nineteen leave-one-symbol-out runs: all negative—and dropping XRP produces the *worst* result of the set, −0.053, so XRP is one of the few symbols contributing positively. The overfitting symptom of "one symbol carries all of the profit" is therefore present, and even with XRP the aggregate is still negative. Dropping the best 1% of trades moves `mean_R` from −0.031 to −0.123, a three- to fourfold widening of the loss. Full-sample equity falls from 1.0 to **0.282** with a maximum drawdown of −82.6% and a CAGR of −20.7% (Figure 5). Over the full sample the profit factor is 0.936, the win rate 38.8%, and the payoff ratio 1.478, while the median trade returns `median_R` = −0.310: wins are indeed large, but the median trade loses money unambiguously.

Figure 5: Full-sample equity curve of the selected configuration `atr_pullback_1|time|h96`, falling from 1.000 to 0.282—roughly 72% of capital lost over 5.5 years—with a maximum drawdown of −82.6%. The absolute shape of the curve depends on the frozen `max_concurrent` and `risk_per_trade` settings, but the statistics used for the verdict (`mean_R`, win rate, payoff ratio) do not.

6. Supplementary experiments

After the main study (P0–P5) was complete we ran three further experiments to confront three residual objections head on: that the bootstrap block

length is a subjective choice; that the conclusion might depend on a single exchange; and that a null result might merely reflect low power. All three were *preregistered before being run*, under the same discipline as the main study, and their artifacts are logged line by line (Appendix C).

6.1. Sensitivity to bootstrap block length

The headline specification resamples calendar-month blocks, which is a subjective choice; if the conclusion held only at one block length the falsification would be fragile. We resample the same frozen data under six schemes—half-month, the one-month baseline, two months, three months, six months, and a stationary bootstrap—and re-estimate every interval. **All four preregistered checks pass under all six schemes, 24 out of 24**: the interval for Δ_D always excludes zero, the interval for Δ_{simple} always spans zero, the interval for the reversal-rate gap always excludes zero, and the validate-period $p(\texttt{mean\_R} \leq 0)$ always exceeds 0.10. Even in the coarsest six-month scheme, with only 12 blocks, the upper bound of the Δ_D interval is -0.037 and still excludes zero. An independently rewritten resampler reproduces the three core intervals of the headline specification *bit for bit* at the one-month setting, so the two implementations validate each other (Figure 6).

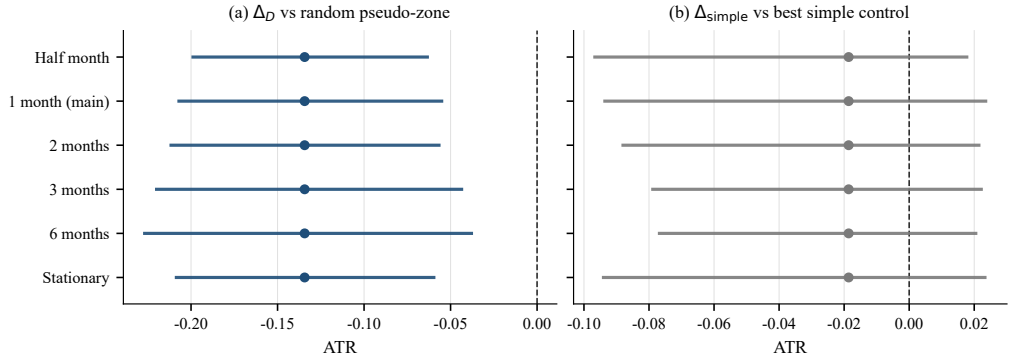

Figure 6: Sensitivity to block length. The 90% intervals for Δ_D and Δ_{simple} deliver the same verdict under all six schemes: Δ_D always excludes zero and Δ_{simple} always spans it. The falsification is not an artifact of monthly-block resampling.

One finding did not match our expectations and is worth reporting. As block length grows, the width of the Δ_D interval increases monotonically, from 0.137 to 0.191, a rise of 39%, whereas the width for μ_{OB} falls slightly,

from 0.186 to 0.163. Serial correlation therefore lives in the *paired differences* rather than in the *levels*: the component common across symbols and across time shows up in the difference, not in the level. This changes no verdict, but it does confirm ex post that the main study was right to designate Δ_{simple}, a paired difference, rather than μ_{OB}, a level, as the headline estimand.

6.2. Frozen replication on a second venue (Binance)

Reliance on a single exchange is a weakness the main study lists against itself. Because the same crypto asset moves nearly identically across venues, we run a *frozen replication*: identical configuration, identical code, identical seed, identical gates, with the price data source as the only variable (OKX replaced by Binance). Bitcoin data were downloaded from Binance Vision (2021-01 through 2026-06, 192,672 15-minute bars, no gaps and no duplicates); BCH, ETC, and TRX are unavailable and, as preregistered, were not substituted, leaving 16 symbols over the common window 2021-01-29 to 2026-07-08. We deliberately do not re-run the 48-cell grid, which would add 48 fresh multiple comparisons while answering no new question, and instead re-derive the complete P2, the reversal test, and one frozen P3 configuration. Results are in Table 10 and Figure 7.

Table 10: Frozen Binance replication against the OKX headline specification.

Quantity	OKX (main, 19 symbols)	Binance (replication, 16)
First-retest fills N	6,662	5,583
μ_{OB} (ATR)	$+0.0199$ $[-0.0761, +0.1102]$	$+0.0182$ $[-0.0841, +0.1187]$
Δ_D (ATR)	-0.1343 $[-0.2079, -0.0542]$	-0.1640 $[-0.2465, -0.0727]$
Δ_{simple} (ATR)	-0.0186 $[-0.0941, +0.0240]$	-0.0477 $[-0.1207, -0.0093]$
Reversal rate	55.54% [54.11, 56.97]	55.86% [54.32, 57.43]
Reversal gap (OB − random)	-3.69pp $[-5.08, -2.35]$	-3.88pp $[-5.42, -2.39]$
Full breach (OB / random)	30.30% / 27.96%	29.33% / 27.09%
Gross reversal edge / cost	68%	76%
Frozen P3 config, `mean_R`	-0.0310	-0.0196
Verdict token	`order_block_falsified`	`order_block_falsified`

Binance is less favorable on every metric, yet each venue's Δ_{simple} point estimate falls inside the other's interval, so the two estimates are compatible. The `mean_R` row is the full-sample figure for the frozen P3 configuration. The Δ_{simple} interval excludes zero on Binance, in the less favorable direction, and spans zero on OKX; the main conclusion uses the more conservative OKX version, and the Binance version serves as supplementary evidence that the replication is not merely same-signed but more adverse. One sign flip is disclosed: in the design period the frozen P3 configuration has `mean_R` $= +0.0165$ on Binance, but this is insignificant ($p = 0.390$, interval spanning zero), and validate, test, and the full sample are all negative.

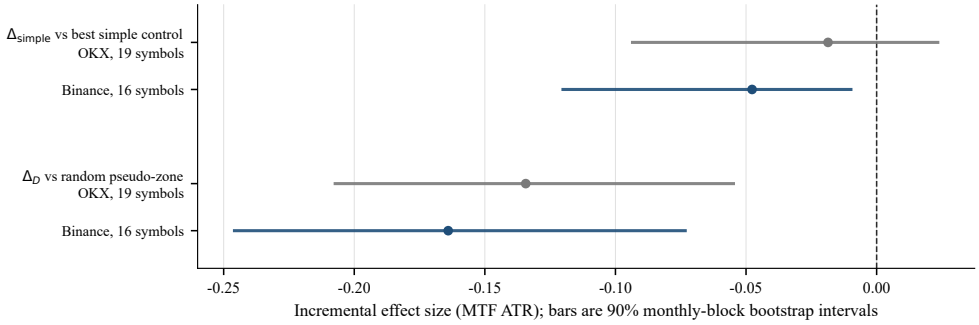

Figure 7: Cross-venue frozen replication. Δ_D and Δ_{simple} carry the same sign on both venues, and both are more negative on Binance. The falsification is not an artifact of OKX microstructure.

The preregistered verdict token on Binance remains `order_block_falsified`, with all four gates resolving identically. The reversal test replicates with particular clarity: the monotone ordering of the eight rules is identical rule for rule, the 50% self-check on immediate entry

reproduces independently at 50.83%, random pseudo-zones still reverse more often (a gap of −3.88 pp with an interval excluding zero), and the gross reversal edge is still only 76% of cost. The one number whose sign differs from OKX—design-period `mean_R` turning positive at +0.0165—is insignificant at $p = 0.390$: a quantity close to zero changing sign under replication is evidence that it is close to zero, not that it works. *The purpose of this experiment is to rule out venue-microstructure dependence as a class of failure, not to double the evidence*: the same asset is market-made across both venues by the same arbitrageurs, and the main study already proxies funding with Binance data, so all new information comes from the price leg.

6.3. SESOI, economic equivalence, and power

The standard objection to a null result is that it may simply be underpowered. This experiment upgrades "not significant" to "statistically equivalent to zero" and quantifies the difference between failing to detect an effect and there being none. The SESOI is set to one round-trip cost of 0.113 ATR, a quantity defined by costs and not by the data: an effect smaller than the cost of getting in and out cannot pay for a single entry and exit even if it is real. Equivalence tests (TOST) and power curves are reported in Table 11 and Figure 8.

Table 11: Economic equivalence tests, power, and minimum detectable effect.

Quantity	Estimate	90% CI	TOST	p	Power	MDE
Δ_{simple} (headline)	−0.0186	$[-0.0941, +0.0240]$	Equivalent	0.021	**0.939**	**0.087**
μ_{OB} (level)	+0.0199	$[-0.0761, +0.1102]$	Equivalent	0.046	0.654	0.139
Δ_D	−0.1343	$[-0.2079, -0.0542]$	Not equiv.	—	0.757	0.120

SESOI = one round-trip cost = 0.113 ATR, and the TOST column reports equivalence against the bound ±0.113 ATR with its p value in the next column. "Power" is power at the SESOI and "MDE" the minimum effect detectable with 80% power. TOST is a bootstrap version of two one-sided tests. Under the stricter bound of ±0.05 ATR, neither Δ_{simple} nor μ_{OB} is equivalent ($p = 0.318$ and 0.299), which we report as it stands. Non-superiority test: the 95th percentile of the bootstrap distribution is $0.1102 <$ SESOI, so at the 5% level we reject the hypothesis that order blocks earn back one round-trip cost ($p = 0.046$, a marginal rejection). Power is estimated nonparametrically from the centered bootstrap distribution, cross-checked against a normal approximation.

Two of three preregistered checks pass and one explicitly fails; all are reported. **S3-1 passes**: Δ_{simple} and μ_{OB} are both statistically equivalent to

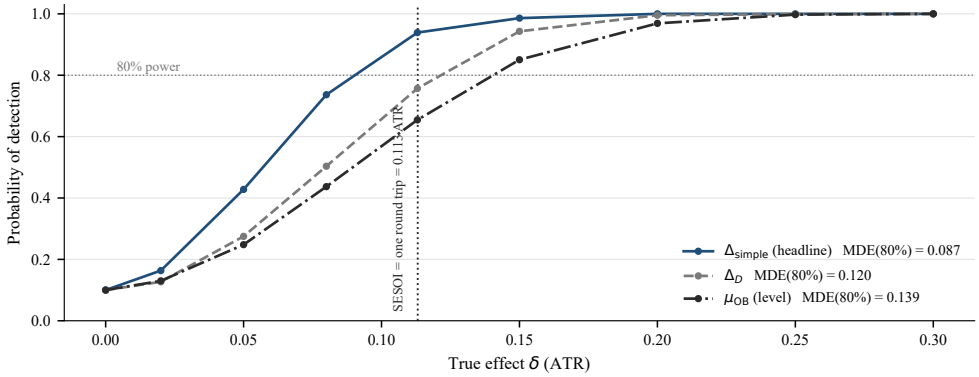

Figure 8: Power curves against the SESOI. The headline estimand Δ_{simple} has 93.9% power at the SESOI, with an MDE of 0.087 below it; the absolute level μ_{OB} has only 65.4% power, with an MDE of 0.139 above it. For the headline estimand the rebuttal "the null result is just low power" is quantitatively excluded.

zero within a bound of one round-trip cost, so the wording can be upgraded from "indistinguishable from zero" to "equivalent to zero". **S3-3 passes**: at the 5% level we reject the hypothesis that order blocks earn back a round trip (95th percentile 0.1102 against 0.1131, $p = 0.046$)—though the rejection is marginal, by 0.003 ATR, so the text does not describe it as "far below". **S3-2 fails**: the check requires both quantities to have an MDE below the SESOI, and while Δ_{simple} qualifies (0.087, with 93.9% power at the SESOI), μ_{OB} does not (0.139, with only 65.4% power). The reason is structural. μ_{OB} is a single-group level and absorbs the entire common variation of the market, whereas Δ_{simple} is a paired difference within the same set of BOS events, so the common component is differenced away and the standard error is 37% smaller—the same phenomenon, seen from the other side, as the finding in Section 6.1 that serial correlation is strong in differences and weak in levels. This *bounds how the falsification may be stated*: we can say that order blocks provide no increment beyond the cost scale relative to the strongest simple control and that we reject their earning back one round trip; we *cannot* say that the absolute return to order blocks has been proven smaller than cost. The main conclusion rests on the former, which is a further reason why the main study designated Δ_{simple} rather than μ_{OB} as the headline estimand in advance.

7. Discussion

7.1. Mechanism: why a real reversal does not pay

One mechanism accounts for all three layers of the verdict. The reversal rate is real because of the geometry of the *origin of measurement*: the deeper one waits, the closer +1 ATR is to where price has just been, and the easier it is to reach—which has nothing to do with whether the price level has memory. What does the work is *waiting*, not *information*. Waiting does not pay because, in a competitive market, a better price and a lower fill probability are two sides of the same trade (Section 4.5): 29.5% of zones are never retested, 87% of those moves had already run away with the trend, and the opportunity cost of waiting for the retest exactly offsets the improvement in entry price. Finally, even if that real, depth-driven reversal edge were captured in full, its magnitude is only 68% of one round-trip cost. Between "there is a reversal" and "one can make money" lies an order of magnitude of cost. That a method taught this widely should have negative net expectation is unsurprising in light of the evidence that individual investors as a group underperform[14].

7.2. Methodological lessons

Three lessons generalize to technical-strategy backtests. First, the *fill-bar convention* (Section 2.5): whenever only the tightest-stop cell looks exceptional, check the fill-bar convention before anything else—without that check, this project would have published a spurious Sharpe ratio of 3.41 and the opposite conclusion. Second, *the identification of controls*: because Δ_D is unmatched on distance to the swing extreme, it absorbs a positional effect, which is why the headline estimand must be Δ_{simple}, anchored to the same events; a random control is always a reference benchmark and never a strategy. Third, *ex post variables cannot be used as filters*: the pullback-depth table—fully breached zones lead to deep losses—is measured after the fill, "fully breached" is nearly equivalent to "the stop was hit", and neither is knowable at the moment of trading.

7.3. A three-layer verdict on SMC

On the method itself, a three-layer statement is more accurate than calling it a scam. *Literally true*: price really is more likely to reverse after returning to an order block (55.54%, with an interval excluding 50), which is not an illusion but a measurable phenomenon. *Wrongly attributed*: that rate is

explained by the depth of the pullback waited for, not by memory attached to the level—immediate entry at 50.28% and size-matched random zones at 59.24% both make the point. *Too small*: the gross expectation of the reversal is 68% of one round-trip cost, so it could not cover costs even if the attribution were correct. "Precision entry" is falsified as well: even among the observations that ended in profit, half had first moved more than 0.75 ATR against the position.

7.4. Scope and limitations

We list the limitations in order of importance. First, the primary evidence comes from a single venue, OKX. The frozen Binance replication in Section 6.2 rules out microstructure dependence as a class of failure, but prices on the two venues are nearly identical, so the replication eliminates a failure mode rather than supplying independent evidence. Second, we model neither capacity nor market impact; the 14 bp figure includes 2 bp of slippage per side, which is broadly reasonable for small size in 19 large-cap perpetuals but is not verified. Third, `max_concurrent` = 6 and `risk_per_trade` = 0.005 are set by hand rather than calibrated to data—they were frozen precisely so that leverage could not be used to manufacture a Sharpe ratio—and while the absolute shape of the equity curve depends on them, the statistics used for the verdict (`mean_R`, win rate, payoff ratio) do not. Fourth, the P3 preregistration was written after the fact and is weaker than P2, and the body convention is a second convention run on the same data. Fifth, any discretionary "zone quality" judgment—for instance a joint condition involving some form of liquidity gap—is not implemented; if real SMC traders apply a decisive condition we have not implemented, and if it happens to reduce 6,662 events to a few hundred high-quality ones, the conclusion could be overturned. Our specification is recomputable, however (configuration in Appendix D), so such an objection can be settled by a concrete test rather than left as an argument.

8. Conclusion

The central claim of the order-block method is *literally true*: price is indeed more likely to reverse after a retest, 55.5% of the time. But the attribution is *wrong*—the effect belongs to pullback depth, not to memory attached to a price level—and the magnitude is *insufficient*, the reversal edge amounting to 68% of one round-trip cost. A preregistered event study (Δ_D =

−0.134 ATR, interval excluding zero), a 48-cell full-cost backtest (uniformly negative, deflated Sharpe −3.41), a frozen replication on a second venue (same sign and more adverse), and an economic equivalence test (equivalent to zero within one round-trip cost) deliver one verdict: this smart-money-concepts trend-continuation and retest strategy is **not tradable, and no live capital should be committed to it**. The more general lesson concerns method: had we built the backtest first and searched for the signal second, without checking the fill-bar convention, this paper would have reported a spurious Sharpe ratio of 3.41 and the opposite conclusion, and would have stopped there.

Appendix A. Parameter values

P2 runs a single fixed configuration—an event study makes no configuration choice, because its evidentiary standard is the effect size—and every parameter takes the midpoint rather than an endpoint of the candidate range from the upstream design, frozen before any data were run (Table A.1).

Table A.1: Parameter values, frozen, with reasoning.

Parameter	Value	Reasoning
L_{htf} / L_{mtf}	5 / 5	About 280 swings per year on 4h and 1,100 on 1h; ample sample
k_{pivot}	0.5 ATR	Low end of the candidate range; a higher value filters out too many structure points
β (BOS buffer)	0	Pure close break, the weakest assumption
$d_{\min}$ / $\text{eff}_{\min}$	1.0 ATR / 0.5	Low end of the range, deliberately undemanding
M (forward search)	5	Upper end, to reduce attrition from bars with no opposite candle
W (displacement leg cap)	3	As in the upstream design
$w_{\max}$	1.5 ATR	Midpoint of the range
E (expiry)	40 MTF bars	Midpoint of the range ($\approx$ 1.7 days at 1h)
Stop buffer	0.25 ATR	Midpoint of the range
Zone convention	wick / body	The main specification uses wicks; the body convention appears as a robustness appendix

The P3 parameter budget is capped at 48 configurations; the grid was written into the P3 preregistration before any data were run and is pinned by an `assert`.

Appendix B. Look-ahead test suite

The engine passes 31 automated tests, grouped into seven families (Table B.1). T1, truncated-history replay, is the single most effective check on a state machine against leakage, and T4's deliberately leaking negative control guarantees that the causality tests are not vacuously true.

Table B.1: Look-ahead test suite: seven families, 31 test functions.

ID	Test	Description
T1	Truncated-history replay	Feed only `data[:k]` for increasing k and assert that every event row with `avail_at` $\leq t_k$ is unchanged field by field (tolerance 10^{-9}).
T2	Pivot confirmation delay	On synthetic data, a pivot must not appear in any output before the L bars to its right have closed.
T3	Multi-timeframe alignment	On synthetic data, modify the second half of an unfinished 4h bar and assert that all prior LTF decisions are unchanged.
T4	Deliberate-leak negative control	Perturb the future of the zone-activation boolean series to verify that the causality tests are not vacuously true.
T5	Table-level invariants	Identities such as `activated_at` > `ob_bar_ts`.
T6	No full-sample statistics	Assert that every quantile and every standardization uses an expanding or rolling window.
T7	Conservative fill ordering	When entry and stop are touched within the same bar, take the least favorable ordering; 1-minute Bitcoin data calibrate the cost of this assumption.

Appendix C. Supplementary experiments: full tables

Table C.1 gives the complete set of intervals for the block-length sensitivity analysis of Section 6.1. Full artifacts, preregistrations, and result documents for all three supplementary experiments are in the replication package.

Table C.1: Block-length sensitivity: 90% intervals under six schemes.

Scheme	Blocks	μ_{OB}	Δ_D	Δ_{simple}	p_{val} [a]
Half month	132	$[-0.0705, +0.1126]$	$[-0.1998, -0.0625]$	$[-0.0971, +0.0182]$	0.921
1 month (main)	**66**	$[\mathbf{-0.0761}, \mathbf{+0.1102}]$	$[\mathbf{-0.2079}, \mathbf{-0.0542}]$	$[\mathbf{-0.0941}, \mathbf{+0.0240}]$	**0.938**
2 months	34	$[-0.0620, +0.1024]$	$[-0.2125, -0.0558]$	$[-0.0885, +0.0219]$	0.956
3 months	23	$[-0.0669, +0.1023]$	$[-0.2208, -0.0427]$	$[-0.0793, +0.0227]$	0.947
6 months	12	$[-0.0595, +0.1039]$	$[-0.2277, -0.0370]$	$[-0.0773, +0.0210]$	0.966
Stationary (mean 1m)	66	$[-0.0739, +0.1118]$	$[-0.2094, -0.0588]$	$[-0.0945, +0.0238]$	0.943

Point estimates are identical to the last digit under every scheme, since the experiment replaces only the uncertainty estimator and never touches the data. The half-month scheme is marked n/a for the reversal-rate gap: the detail table carries no timestamps, so half-month blocks cannot be derived, and the preregistration required marking this n/a rather than silently falling back. The one-month scheme reproduces the headline intervals bit for bit under an independently rewritten resampler.

[a] $p_{\mathrm{val}} = p(\texttt{mean\_R} \leq 0)$ in the validate period (2024–2025).

Appendix D. Artifacts and replication

The replication package accompanying the paper contains the frozen design document, the two preregistrations for P2 and P3, the P1–P3 result documents, three supplementary preregistrations and their result documents, the source of the multi-timeframe structure engine (`src/nemesis/artemis/`), the 31 tests, the supplementary experiment scripts, and the line-by-line trial ledger. Headline artifacts are in `outputs/artemis/`, body-convention appendix artifacts in `outputs/artemis_body/`, Binance replication artifacts in `outputs/artemis_binance/`, and supplementary statistics in `outputs/artemis/paper/`; the frozen configuration is `configs/artemis.canonical.yaml`. Every data figure in this paper is generated by `make_figs_en.py`. Key ledger rows: `artemis_p2_event_study` (`n_configs=2`, including the body convention), `artemis_p3_grid` (`n_configs=48`), and `artemis_paper_s1_blocklen` / `_s2_binance_repl` / `_s3_sesoi_power` (`n_configs=1` each, being frozen replications or pure post-processing rather than new searches).

Appendix E. Timeline

Stage	Content
P0	Design freeze: mechanize the SMC method, establish the causality constraint and the control design.
P1	Structure acceptance: 31 tests pass; the fill-bar convention error is found and corrected; the tautology of order-block existence is established.
P2	Event study, genuinely preregistered in advance: gates resolve to `order_block_falsified`.
P3	Full-cost backtest: 48 of 48 cells negative, deflated Sharpe -3.41 (preregistration written after the fact and disclosed as such).
P4	Robustness: leave-one-year-out, leave-one-symbol-out, concentration, and cost stress, all uniformly negative.
P5	Final report: the three-layer verdict and the recommendation against committing live capital.
S1–S3	Supplementary experiments, each preregistered first: block-length sensitivity, frozen Binance replication, equivalence testing and power.

References

[1] Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25, 383–417.

[2] Brock, W., Lakonishok, J., LeBaron, B., 1992. Simple technical trading rules and the stochastic properties of stock returns. *Journal of Finance* 47, 1731–1764.

[3] Lo, A.W., Mamaysky, H., Wang, J., 2000. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *Journal of Finance* 55, 1705–1765.

[4] Urquhart, A., 2016. The inefficiency of Bitcoin. *Economics Letters* 148, 80–82.

[5] Sullivan, R., Timmermann, A., White, H., 1999. Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance* 54, 1647–1691.

[6] White, H., 2000. A reality check for data snooping. *Econometrica* 68, 1097–1126.

[7] Harvey, C.R., Liu, Y., 2015. Backtesting. *Journal of Portfolio Management* 42, 13–28.

[8] Harvey, C.R., Liu, Y., Zhu, H., 2016. …and the cross-section of expected returns. *Review of Financial Studies* 29, 5–68.

[9] Bailey, D.H., López de Prado, M., 2014. The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality. *Journal of Portfolio Management* 40, 94–107.

[10] McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? *Journal of Finance* 71, 5–32.

[11] Lakens, D., 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science* 8, 355–362.

[12] Lakens, D., Scheel, A.M., Isager, P.M., 2018. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science* 1, 259–269.

[13] Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. *Journal of the American Statistical Association* 89, 1303–1313.

[14] Barber, B.M., Odean, T., 2000. Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance* 55, 773–806.