

Adverse Selection Consumes the Touch: A True-Aggressor-Signed Maker-P&L Decomposition Across 13 CME Futures

Daniel Gatto^{*1}

¹Universidade Paulista (UNIP)

June 2026

Abstract

On a price-time-priority futures book a market maker earns a spread for supplying liquidity and pays adverse selection to the informed traders who pick off its quotes (Glosten–Milgrom). The empirical literature lacks a direct, large-scale, *true-aggressor-signed* measurement of how large that cost is relative to the captured spread. We supply one on CME depth-10 data (13 roots, 234 contract-days, 353,387 maker fills): a ground-truth, fill-by-fill maker-P&L decomposition splitting each passive fill into the half-spread captured at the touch (+0.654bp) and the post-fill adverse drift the counterparty takes back (−0.661bp). Adverse selection consumes essentially the whole captured half-spread, leaving net maker P&L indistinguishable from zero (−0.022bp after the 0.015bp fee, every root within ± 0.13 bp). Because passive fills are charged no slippage and every unmodelled friction pushes net more negative, this is a one-sided bound: net $\leq \approx 0$ (Proposition 1). Read through a minimal competitive model the cancellation is *consistent with* the zero-profit condition of liquidity supply, and it carries a derived *sufficient* tick-size survival condition. The one-tick floor clears the fee only for the coarse-tick roots ZN and SI, beyond which the per-root adverse drift decides (the ≈ 0.74 bp frontier separating the survivors is descriptive, not a structural necessity). The sample concentrates in one calm 2023 month (93% of fills), so the cancellation is established cross-sectionally, within a single volatility regime.

Whether the cancellation holds across volatility regimes is confounded by the resolution of the book reconstruction, which we resolve by measuring on a single nanosecond-resolution, true-aggressor-signed trades-and-quotes (TAQ) feed that spans the regimes we test and reproduces the depth-feed decomposition to 0.3pp on matched calm sessions. On that feed adverse selection over-consumes the captured spread in three crisis weeks (adverse-to-captured 105.1% SVB, 111.5% FOMC, 109.8% COVID, CIs excluding 100%) and net maker P&L deepens with volatility

^{*}Correspondence: daniel@daru.finance. Replication code: <https://github.com/DaruFinance/quant-mm-simulator>.

(to -0.139bp), broad across roots in these high- σ snapshots. A continuous 2016–2026 sweep (33,956 root-days, 96% of the universe) sharpens it: a naive pooled σ -curve is composition-confounded, and the within-root σ -gradient of the loss is *root-idiosyncratic*, carried by natural gas, copper, crude and FX, vanishing when those are dropped. So the cross-regime result is matched-feed evidence of regime-dependence (commensurability verified on calm sessions only), not a universal law.

Two corollary nulls follow, since the captured spread and the adverse drift are co-determined: an Avellaneda–Stoikov-style quoting family is uniformly null net of the maker fee (inventory-skew 0 of 234 cells significant), and a true-sign cross-root integrated order-flow-imbalance signal validates its pipeline in-sample ($R^2 = 0.545$) but adds nothing out-of-sample and turns negative at full taker cost. So no inventory-aware quote or linear order-flow signal we test earns a return net of cost in price-time-priority tick-bound futures. The measured adverse drift is, at the role level, the AMM literature’s loss-versus-rebalancing (LVR, $\sigma^2/8$ [7]), which we put on one axis as a consistency check.

Keywords: market making, market microstructure, adverse selection, maker-P&L decomposition, order flow imbalance, loss-versus-rebalancing, trade-sign classification, backtest overfitting, high-frequency trading, CME futures.

JEL classification: G14, G12, C58, D47, G23.

Contents

1	Introduction	3
2	Related Work and Gaps	6
3	Data	7
4	Methodology and Engine	8
5	The Central Result: A Ground-Truth Maker-P&L Decomposition	11
5.1	A competitive zero-profit reading, and a derived survival condition	14
5.2	Conditional Decomposition by Session and Regime	17
6	Cross-Regime Evidence	19
7	Corollary 1: The Quoting Null	23
8	Corollary 2: The Cross-Root Integrated-OFI Lead-Lag Null	24
9	Supporting Measurements	25
10	Discussion	27
11	Conclusion and Limitations	28
A	LVR Cross-Root Volatility Ordering	31
B	Figure Provenance	31

C Results Inventory	32
D Reproducibility	34

1 Introduction

A market maker on a price-time-priority futures book posts a bid and an ask and waits. It earns a spread for standing ready to trade and pays a cost when the counterparty who arrives knows something the price has not yet moved to reflect, the cost Glosten–Milgrom [2] call *adverse selection*: the trades that fill are disproportionately the ones about to be on the wrong side of the next price move. How large is that cost relative to the spread the maker actually captures? Glosten–Milgrom answer it in principle, but the empirical literature has lacked a direct, large-scale, true-sign measurement. The paper supplies one, fill by fill, on true-aggressor-signed CME data, and the answer is that adverse selection consumes essentially the whole captured half-spread, leaving net maker P&L statistically indistinguishable from zero across 13 futures roots. The contribution is that measurement (a direct, large-scale, true-aggressor-signed measurement of the Glosten–Milgrom cost at cross-sectional scale, not the discovery of a surprising trade-off), together with the reading it licenses: through a minimal competitive model the cancellation is the zero-profit condition of liquidity supply (free entry pins the equilibrium half-spread at adverse drift plus fee), which we hold as *consistent with* rather than a test of entry, and which we close with a *derived* tick-size survival condition (Section 5.1). The value is in pinning the magnitude on ground-truth data and giving the net-zero a structural reading rather than treating it as a coincidence. Two corollary nulls about standard quoting and order-flow signals follow from it. The depth-10 sample is four contiguous months inside one calm-to-moderate 2023 volatility regime (July carries 93% of the fills), so the cancellation is checked month to month within that regime; whether it holds *across* regimes is initially confounded by the resolution of the book reconstruction, a confound we resolve by holding the reconstruction fixed on a single nanosecond-resolution TAQ feed, and find the cancellation *does* break down under stress (Section 6, Section 11).

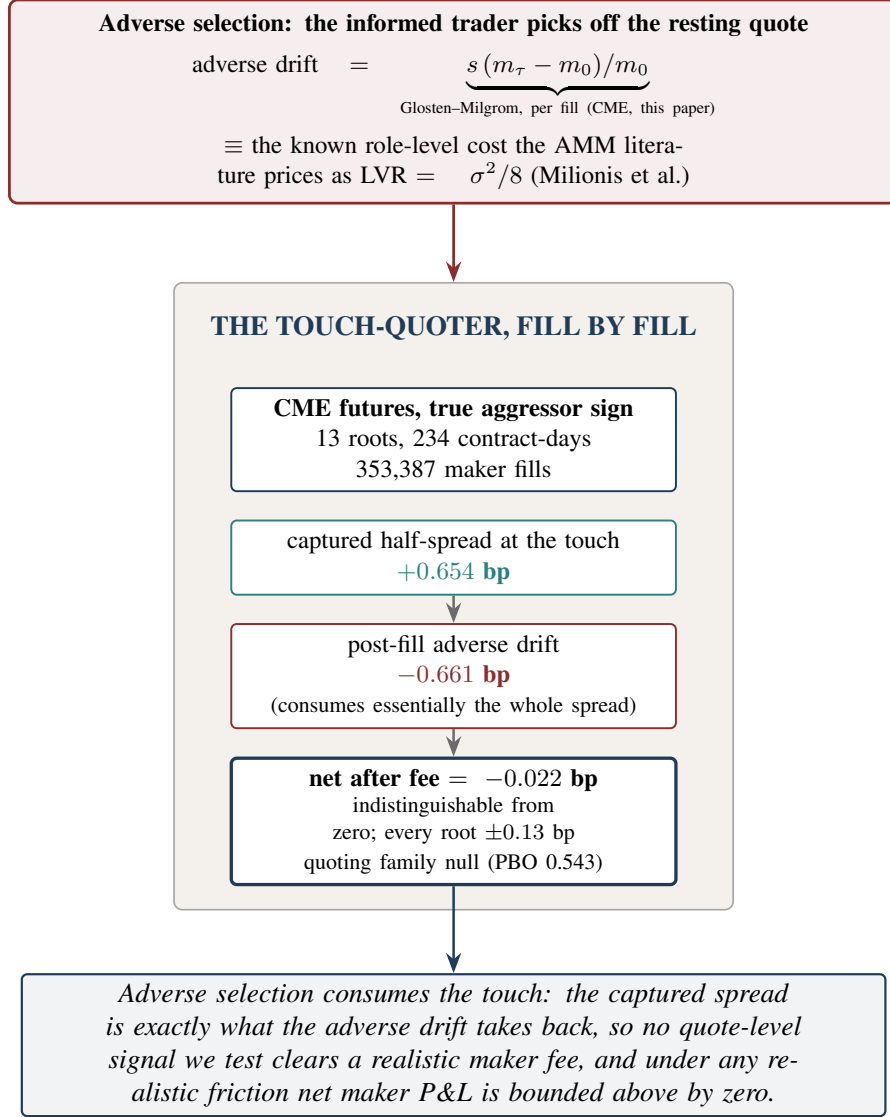


Figure 1: The spine of the paper: a ground-truth maker-P&L decomposition on true-sign CME futures. The half-spread captured at the touch (+0.654bp) is consumed almost exactly by the post-fill adverse drift the informed counterparty takes back (-0.661 bp), leaving net maker P&L indistinguishable from zero (pooled -0.022 bp; the quoting family that tries to beat this is uniformly null, PBO = 0.543). The same adverse drift is, as a known role-level result, the AMM liquidity provider’s LVR ($\sigma^2/8$, [7]); the data are consistent with one shared volatility scale.

The central result: adverse selection consumes the touch. The spine of the paper is a ground-truth maker-P&L decomposition on the true-sign CME data (Section 5). Fill by fill, it splits the touch-quoter’s economics into the half-spread captured at the touch and the post-fill adverse drift the informed counterparty collects, and the adverse drift consumes essentially the entire captured half-spread: +0.654bp captured against -0.661 bp of drift, leaving net maker P&L indistinguishable from zero (pooled -0.022bp after the fee, every one of the 13 roots within ± 0.13 bp). Because passive fills are charged no slippage and every realistic friction pushes net more negative, this is a one-sided bound (net ≤ 0 under any realistic friction), stated once in Section 5. The measured adverse drift is, as a known theoretical result, the same role-level cost the AMM literature calls

loss-versus-rebalancing (LVR, $\sigma^2/8$ [7]): both are the informed counterparty’s pickoff of a liquidity supplier, governed by the volatility scale. We illustrate that known equivalence on one axis (Section 5) rather than assert it, and find the cross-root data consistent with one shared volatility driver. We flag plainly what the illustration can and cannot show: the cross-root data cannot distinguish one σ -driven cost from two unrelated costs that both scale with σ ; the unification rests on the known role-level equivalence, and the σ -scaling we report is consistent-with, not confirmatory.

Two corollary nulls. From that cancellation the rest follows. *Corollary 1 (the quoting null, Section 7):* if the captured spread and the adverse drift are co-determined by the same information, no inventory-aware quote can separate them, and a family of Avellaneda–Stoikov-style strategies (about five structural shapes, evaluated over a 1,126-cell nominal grid) is uniformly null net of a realistic maker fee: the inventory-skew family is 0 significant even at an uncorrected 5%, and $PBO = 0.543$ is the coin-flip value of a true null (BH/BHY add a consistency note, demoted because the permutation floor sits above their thresholds, Section 7). *Corollary 2 (the cross-root OFI null, Section 8):* a true-sign cross-root integrated-OFI signal that fits its own pipeline in-sample (own contemporaneous $R^2 = 0.545$) adds nothing out-of-sample (cross-root incremental $R^2 = -0.005$) and turns sharply negative after taker cost. The reach of these nulls is the constructions we test: standard inventory-aware quoting and linear order-flow signals in a price-time-priority tick-bound venue, not a claim that no edge of any kind exists anywhere. One level up, the same per-position cost is a *covariance* rather than a level, and covariance is reducible even when no single stream carries alpha; preliminary cross-chain DEX-LP evidence suggests it is diversifiable, which we note as a forward pointer (Section 10) and develop elsewhere rather than fold into this paper.

Contributions. (a) A ground-truth maker-P&L decomposition on tick-bound true-sign futures, in the Moallemi–Yuan [6] idiom, that measures spread capture against post-fill adverse drift fill-by-fill at 10s and 60s and shows adverse selection consuming essentially the full captured half-spread on every root, leaving net maker P&L indistinguishable from zero, a one-sided result: under any realistic friction net is bounded above by ≈ 0 . It comes with a tick-size frontier locating the spread-capture threshold (a derived sufficient one-tick floor clears the fee only for the coarse-tick roots, met by ZN and SI; the ≈ 0.74 bp tick width that separates the survivors is descriptive), and an illustration that the measured adverse drift is the known role-level equivalent of the AMM literature’s LVR ($\sigma^2/8$), with the cross-root data consistent with one shared volatility scale (consistent-with, not a test that distinguishes one σ -driven cost from two). (a’) A matched-feed cross-regime extension that shows the calm cancellation is a regime-specific floor, not a universal law: built on a nanosecond-resolution, true-signed TAQ feed through a depth-1 ingester that reproduces the depth-feed decomposition to 0.3pp on matched calm sessions, it finds adverse selection *over*-consuming the captured spread in stress (adverse-to-captured 105–111% across the 2023 SVB, 2022 FOMC and 2020 COVID weeks) and net maker P&L deepening with volatility ($-0.022 \rightarrow -0.139$ bp), broad-based across roots in those high- σ snapshots, once a reconstruction-resolution confound (a 17.4pp same-dates gap between the high- and coarse-resolution reconstructions) is cleared by holding the reconstruction fixed. A continuous 2016–2026 sweep (33,956 root-days, 96% of the universe) then adds the dose-response: a naive pooled σ -curve is composition-confounded (a Simpson’s-paradox sign inversion), and the within-root σ -gradient of the loss is *root-idiosyncratic*

(carried by natural gas, copper, crude and FX, vanishing when they are dropped), not a broad law. (b) A large-scale *true-sign* confirmation that Avellaneda–Stoikov-style quoting fails net of costs in price-time-priority tick-bound futures, a stated gap in a literature whose net-of-cost evidence is mostly equity or simulated, with the overfitting machinery (PBO via CSCV as the resolution-independent lead, BHY FDR as a consistency note, nominal trial counts) applied throughout. (c) A true-sign cross-root integrated-OFI null that extends the Cont–Cucuringu–Zhang [4] equities cross-impact program to CME futures and finds the contemporaneous cross-impact subsumed by own integrated OFI, with no lagged predictive content surviving out-of-sample net of cost. Section 10 notes, as a forward pointer, that one level up the same cost is a covariance an LP can diversify, developed in separate work.

2 Related Work and Gaps

Three literatures meet at this paper, and it fills a stated white-space gap in each.

Optimal quoting and its net-of-cost test. Avellaneda & Stoikov [1] give the canonical inventory-aware quoting model: skew the two-sided quote against current inventory, recentre on a reservation price, and the optimal half-spread trades inventory risk against fill intensity. The model and its successors (Guéant–Lehalle–Fernandez-Tapia [8], Cartea–Jaimungal [9]) assume a martingale mid and a quote-driven venue. Whether that quoting edge *survives net of cost in a price-time-priority, tick-bound order-driven venue* is a different question. The existing net-of-cost evidence is mostly equity or simulated. We measure it directly on true-sign futures (the decomposition, Section 5, and its quoting-null corollary).

Order-flow imbalance and cross-impact. Cont, Kukanov & Stoikov [3] establish that best-level own-asset OFI explains contemporaneous price changes with high R^2 ; we reproduce this (own contemporaneous $R^2 = 0.545$) as a pipeline check. Cont, Cucuringu & Zhang [4] extend the program to *cross-impact* across equities, finding own-asset OFI dominant and cross terms largely contemporaneous. The only prior cross-asset *futures* flow study, OFR Working Paper 19-04 [5], uses coarse, inferred (not true-signed) flow. The white space is a *true-sign* cross-root integrated-OFI lead-lag test in CME futures; we run it (the second corollary null, Section 8) and it is null, as the Cont–Cucuringu–Zhang prior predicts.

Inventory cost and its decomposition. Glosten & Milgrom [2] model the spread as compensation for adverse selection; Moallemi & Yuan [6] value queue position against exactly the post-fill drift a resting order suffers, the mechanism our decomposition measures. On the AMM side, Milionis, Moallemi, Roughgarden & Zhang [7] derive $LVR = \sigma^2/8$, the LP’s adverse-selection cost against an arbitrageur; at the role level this is the same primitive as the maker’s adverse drift, a known theoretical equivalence. What is missing is a direct, fill-by-fill *measurement* of how much of the captured spread that drift consumes on true-sign futures, which is what this paper supplies (the decomposition, Section 5); we illustrate the LVR equivalence on the same axis as a consistency check rather than re-deriving it.

Trade-sign classification and deflation. The futures trade-sign canon (Chakrabarty–Pascual–Shkilko [10], Easley–López de Prado–O’Hara’s BVC [11]) classifies aggressor side from price and volume because the true sign is rarely available; our data carries the true exchange aggressor sign, so we validate those classifiers against ground truth at scale (Section 9). On overfitting, we lead on the Bailey–López de Prado probability-of-backtest-overfitting [12] (resolution-independent), report the Benjamini–Hochberg–Yekutieli [13] arbitrary-dependence FDR as a consistency note (its rank-1 threshold sits below our permutation floor, Section 4), and use the author’s companion within-strategy permutation framework [15] as the permutation-null backbone; HAC Sharpe inference follows Lo [14]. We never multiply walk-forward window folds into corpus breadth, and report the nominal trial count (the raw slate size, not a deflated effective-trials estimate) so that “strategies $\times$ windows” is not mistaken for a larger search than it is; a genuine deflation is moot for the quoting family, which is null before any correction.

3 Data

One dataset carries the paper: true-aggressor-signed CME futures depth-10 books and tapes, provided by AlgoSeek. We describe it here.

CME futures (the quoting / decomposition / OFI venue). The dataset carries **234 contract-days across 13 roots** and 29 distinct trading dates spanning **July–October 2023** (19 trading days in July, then a weekly cadence through October), four contiguous months that all fall inside one calm-to-moderate 2023 volatility regime, which lets the central decomposition be checked for month-to-month stability *within* that regime (Section 5); a cross-*regime* test is out of scope and untested (Section 11). The 13 roots span four asset blocks: equity index (**ES** E-mini S&P 500, **NQ** E-mini Nasdaq-100, **RTY** E-mini Russell 2000), rates (**ZN** 10Y T-Note, **ZF** 5Y T-Note, **ZB** 30Y T-Bond), energy (**CL** WTI crude, **NG** Henry Hub natural gas), metals (**GC** gold, **SI** silver, **HG** copper), and FX (**6E** EUR/USD, **6J** JPY/USD), a genuinely cross-sectional futures basket. Each contract-day carries a **depth-10** limit order book (10 levels per side) per snapshot and, on every trade, the *true exchange aggressor sign* rather than a tick-rule or Lee–Ready inference, the rare asset that anchors the whole paper. Contract-days per root sum to 234 (ES 22, CL/NQ 19, 6E/6J/GC/HG/NG/SI/ZB/ZF/ZN 18, RTY 12); multiple front contracts appear (e.g. ESU3/ESZ3, CLQ3/CLU3), with the exact contract carried in the symbol column. The total true-signed trade count over the 234 contract-days is **12,426,499**. The cross-root OFI panel uses the **7 fully-synchronized 13-root days** (20230709–0713, 0719, 0730), with two 12-root days (0716, 0723) reported as secondary. Table 1 reports per-root coverage.

Table 1: CME futures universe: 13 roots, 234 contract-days, true exchange aggressor sign on every trade. The basket spans equity index, rates, energy, metals, and FX.

root	block	c-days	root	block	c-days
ES	equity index	22	SI	metals	18
NQ	equity index	19	HG	metals	18
RTY	equity index	12	GC	metals	18
ZN	rates	18	CL	energy	19
ZF	rates	18	NG	energy	18
ZB	rates	18	6E	FX	18
			6J	FX	18
Total			234 contract-days; 12,426,499 signed trades		

Cross-sectional overview · 13 CME futures, 234 contract-days

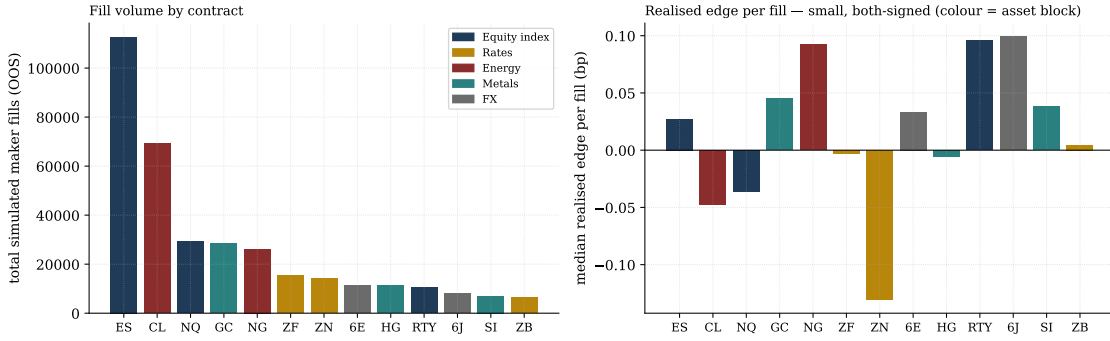


Figure 2: The CME sample is a genuinely cross-sectional futures basket: equity index, rates, energy, metals, and FX roots, each with depth-10 books and true aggressor signs. Left: total simulated maker fills per root, coloured by asset block; right: the small, both-signed median realised edge per fill (ZN -0.13 bp to 6J $+0.10$ bp), previewing the near-break-even quoting result.

4 Methodology and Engine

Every result below obeys the same engine, the same maker economics, and the same significance machinery, stated once here.

The simulator. The engine (`mmsim/sim/fast_sim.py`) is an event-driven L2-book and tape replay with queue-position-aware fills, continuous fractional inventory, a pluggable quoter contract, a hedge engine, and a costed multi-leg trade ledger with post-fill markout. A vectorised numba kernel replaces the per-event Python fill pipeline, running $\approx 80\times$ faster than the pure-Python reference on a full CME root-day (the speedup grows with event count; saved timing artifact and the memory-bound note in Section D). The kernel is faithful, not a shortcut: the runner’s quoters are cancel-on-replace every snapshot, so each resting order lives exactly one snapshot interval, and the kernel reproduces the per-interval queue-consumption semantics (queue position = resting size at our level; each aggressing trade consumes $\min(\text{trade size}, \text{queue position})$, spillover fills us). This is a deliberately *optimistic-fill* assumption: once an interval’s aggressing volume clears the queue resting ahead of us, the next aggressing unit fills our order (we sit at the front of the residual queue),

which biases toward more fills, not fewer, and so only strengthens the one-sided net- ≤ 0 bound of Section 5 (more passive fills can only add adverse selection, never manufacture spread the data does not show). It is verified against the pure-Python reference: the emitted fill sequence (timestamp, side, price, size, maker flag) matches to a 10^{-10} price / 10^{-12} size rounding tolerance (`scripts/verify_fast_parity.py`, `tests/test_fast_sim.py`), and the markout and permutation hot loops are *bit-identical* (array-equal) to their scalar references (`tests/test_markout.py`). Causality is structural: a fill at interval i depends only on the snapshot opening i and trades within i (all timestamps $\leq$ fill time).

Maker economics (the corrected cost model). The cost model encodes the economics of a passive fill correctly, a correction locked in pre-registration before any headline number. A maker fill *captures the quoted spread and fills at its limit price, with no slippage*; slippage applies only to taker/crossing actions. The maker fee is venue-correct: for CME futures the per-contract fee maps to **0.015 bp with no rebate**. Adverse selection (markout) is retained as the maker’s real cost. The Phase-1 specification charged the maker as a taker (0.2 bp slippage on passive fills); fixing it is what makes the baseline approximately break-even rather than badly negative. The null is therefore honest rather than an artifact of over-charging the liquidity supplier. Per-fill economics are streamed to a parquet ledger (one Leg row per fill, framework trade-log v1 schema).

The maker-P&L decomposition. For the touch-quoter, we split each fill’s economics in the Moallemi–Yuan idiom [6], marked at horizon $\tau \in \{10s, 60s\}$, with s the maker side (+1 bid, −1 ask), mid_0 the mid at fill, mid_τ the mid at τ , and fill the fill price:

$$\text{spread capture} = s(\text{mid}_0 - \text{fill})/\text{mid}_0 \ (\geq 0 \text{ at the touch}), \quad (1)$$

$$\text{adverse selection} = s(\text{mid}_\tau - \text{mid}_0)/\text{mid}_0 \ (< 0 \Rightarrow \text{picked off}), \quad (2)$$

$$\text{net markout} = \text{spread capture} + \text{adverse selection} = s(\text{mid}_\tau - \text{fill})/\text{mid}_0, \quad (3)$$

$$\text{net after fee} = \text{net markout} - \text{maker fee}/10^4. \quad (4)$$

Per-root values are fill-weighted across root-days. Equation (1) is the half-spread the liquidity supplier earns; (2) is what the informed counterparty takes back. The central result of the paper is the empirical observation that (1) and $-(2)$ are nearly equal.

Cross-regime measurement on a matched feed. The depth-10 reconstruction underlies the calm headline; to test other regimes under *one* consistent reconstruction we use a nanosecond-resolution, true-aggressor-signed top-of-book (TAQ) feed of CME data, which carries the level-1 quote updates (the best bid/ask book) and the trade events with the exchange-given aggressor sign across the regimes we test. Since the touch decomposition (1)–(4) consumes only the L1 mid at fill, the mid at τ , and the true sign, a depth-1 ingester (`mmsim.ingest.fut_taq`) builds the level-1 book from the quote events and reuses the depth pipeline’s trade parser for the signed tape verbatim, so the *same* engine, cost model, markout, throttle, and front-month rule run on it unchanged; the fast parser is bit-identical to a pure-Python reference (`tests/test_fut_taq.py`). Front contracts are selected per root-day by maximum file size (activity), matching the depth pipeline’s largest-by-activity rule. Because the TAQ and depth reconstructions are distinct, we never compare across them without a *same-dates commensurability control*: on the 52 matched Jul-2023 root-days the TAQ-built decomposition reproduces the depth-feed result to a

0.3pp ratio and 0.001bp net difference (Section 5), which is what licenses extending the TAQ decomposition to the stress episodes. We mark the control’s scope plainly (it is a *calm*-window equivalence) and read the stress ratios as matched-feed evidence accordingly.

Cross-root integrated OFI. Per root, we compute the standard Cont 10-level OFI [3] per book update, then form the *integrated* OFI as the first principal component of the standardised 10-level OFI vector, sign-fixed to load positively on level-1 OFI and scaled to unit variance *on the in-sample window only* (no test-window leakage). The vectorised extraction is bit-identical to the scalar reference. We bin to fixed 10s clock bins (sensitivity at {1s, 5s, 30s} reported but not used to pick the headline); the bin return is the mid log-return, $\text{mid} = (\text{best bid} + \text{best ask})/2$ at bin close, strictly causal. Nested models are (0) own contemporaneous in-sample sanity, (1) own lagged $r_i(t+1) \sim \text{ofi}_i(t)$, and (2) own plus cross lagged $r_i(t+1) \sim \text{ofi}_i(t) + \sum_{j \neq i} \beta_{ij} \text{ofi}_j(t)$. The cross-root *incremental* out-of-sample R^2 is $R^2(2) - R^2(1)$, out-of-sample only. The trading evaluation sets position $i \propto \text{model}(2)$ prediction, vol-scaled per root under an equal-risk-budget gross cap, and charges *taker* costs on turnover at $0 \times / 0.5 \times / 1 \times$ the measured per-root half-spread plus the 0.015 bp CME taker fee.

Pre-registration and significance. Two analyses were pre-registered before any number was produced. A Phase-2 pre-registration (locked 2026-06-05) fixes the quoting and H4 verdicts and the documented maker-economics correction; a second pre-registration (locked 2026-06-09) registers the cross-root OFI hypothesis as new, states the honest prior (contemporaneous cross-impact subsumed by own integrated OFI; any edge is the lagged term, short-horizon, likely cost-fragile), and pre-commits the falsifiable verdicts. On significance we use **Benjamini–Hochberg–Yekutieli** (BHY) rather than plain BH because the per-root-day strategies are structurally correlated; BHY rescales the BH threshold by the harmonic factor $c(m) = \sum_{i=1}^m 1/i = 7.60$ for arbitrary dependence. We lead the overfitting verdict on the **probability of backtest overfitting** (PBO) via CSCV [12], which is resolution-independent, supported by **dual permutation nulls** (block/stationary bootstrap as primary, single-bar shuffle as secondary, $M = 1000$) and Lo HAC standard errors [14] on the portfolio Sharpe, and we report the **nominal trial count** (the raw slate size, not a deflated effective-trials estimate) so that walk-forward window folds are never multiplied into corpus breadth. One caveat governs how we read the FDR step: with $M = 1000$ permutations the smallest attainable p -value is $1/(M+1) \approx 9.99 \times 10^{-4}$, which sits above the Benjamini–Hochberg rank-1 threshold at $m = 1,126 (\approx 4.4 \times 10^{-5})$ and far above the BHY rank-1 threshold ($\approx 5.8 \times 10^{-6}$). No strategy could therefore clear BH or BHY *regardless of true edge*, so a count of 0 significant under those procedures is mechanically guaranteed by the permutation resolution and is not by itself evidence of the null. We accordingly treat BH/BHY as a consistency note and rest the null on PBO and the raw effect sizes, stated where each null is reported.

The next three sections carry the spine. We lead with the ground-truth maker-P&L decomposition (Section 5), which measures the captured spread against the post-fill adverse drift fill-by-fill and shows the drift consuming essentially the whole spread; on the same axis we illustrate that this drift is the known role-level equivalent of the $\sigma^2/8$ LVR. From the cancellation it finds, two nulls follow as corollaries on the same true-sign CME data: a quoting family cannot separate a spread the adverse drift co-determines

(Section 7), and a cross-root order-flow signal that fits its own pipeline in-sample adds nothing out-of-sample and dies after cost (Section 8). The three agree.

5 The Central Result: A Ground-Truth Maker-P&L Decomposition

At the spine sits a direct measurement: split a passive maker fill into the spread it captures and the adverse drift it pays, and quantify how much of the spread the drift offsets. We decompose the touch-quoter’s economics fill-by-fill using (1)–(4) over **353,387** true-signed fills across **232** root-days (two root-days dropped). The result is a near-exact cancellation: adverse selection consumes essentially the full captured half-spread. The maker captures **+0.654 bp** at the touch and gives back an adverse post-fill drift of **−0.661 bp** over the next 10 seconds, so net markout is **−0.007 bp** and net after the 0.015 bp maker fee is **−0.022 bp**. A root-day cluster bootstrap (10,000 resamples over the 232 root-days) puts the 95% confidence interval on net-after-fee at $[-0.046, +0.001]$ bp, containing zero, with $P(\text{net} < 0) = 0.97$; the interval on the adverse-to-captured ratio is $[97.7\%, 104.7\%]$, containing 100% (point 101.1%, $0.661/0.654 = 1.0112$). Net maker P&L is thus statistically indistinguishable from zero and adverse selection from the full captured half-spread: the informed counterparty collects essentially the entire spread, and the one-sided friction argument below places the true net no higher. Extending the markout horizon to 60 seconds deepens the adverse term to **−0.687 bp** and net-after-fee to **−0.048 bp**. Waiting longer only confirms the drift is real, not a microstructure blip that mean-reverts. The two nulls of this part, the quoting family (Section 7) and the cross-root OFI signal (Section 8), are *consistent with* this one cancellation rather than deductive consequences of it.

The net-zero result is a one-sided bound. This near-zero net rests on the maker economics of Section 4, which charge a passive fill *no slippage*: it captures its quoted spread and pays only the adverse drift and the fee. That is the most favourable accounting a liquidity supplier can be given. It matters for how to read the result. Every realistic friction we have not modelled pushes net in one direction only, more negative: partial fills (the queue does not always clear to our level, so we capture less spread than booked), cancel/replace latency (a quote that should have been pulled is picked off), and the cost of unwinding accumulated inventory (a taker action with real slippage). None of these can make a passive program *more* profitable than the frictionless-fill accounting. The honest claim is therefore a one-sided bound, stronger than the **−0.022 bp** point estimate: *net maker P&L is bounded above by ≈ 0 under any realistic friction*. We state this once here and rely on it throughout. One scope note on the bound: it constrains the net *level*, not the adverse-to-captured *ratio*. The optimistic fill changes the *mix* of fills, not merely their count (the marginal front-of-residual-queue fills it adds sit at low queue rank, where adverse selection is shallowest, Figure 5b), so we make no one-sided claim for the 101.1% ratio; it is reported as a point estimate with its bootstrap CI $[97.7\%, 104.7\%]$, and only the net carries the directional guarantee.

Proposition 1 (Cost monotonicity of the net). *Fix the touch-quoter and the passive fills the engine generates. Let $\hat{N}$ be their net under the Section 4 accounting (each fill charged the captured spread, the adverse drift, and the fee, but no slippage), and let N be the net*

of the same fills once any real friction omitted from that accounting is charged: effective passive slippage, cancel/replace latency on a quote that should have been pulled, and the slippage of unwinding accumulated inventory. Each is a non-negative per-fill cost $\phi_k \geq 0$, so $N = \hat{N} - \sum_k \phi_k \leq \hat{N}$.

So the no-slippage accounting is *cost-optimistic*: charging any realistic friction only lowers net, and the measured $\hat{N} = -0.022$ bp (CI $[-0.046, +0.001]$) is an upper bound on the net achievable by these fills. This is the operative claim, stronger than the point estimate: realizable net is bounded above by ≈ 0 . One modelling choice sits *outside* the proposition: the optimistic front-of-residual-queue fill (A2, Section 4) changes *which* fills occur, not their per-fill cost. We do not fold it into the bound; we note only that the marginal fills it grants are low-queue-rank, where the adverse drift is shallowest (Figure 5b), so we do not expect a realistic fill model to flip the sign, an expectation rather than part of the guarantee. The bound is on the net *level*; the 101.1% ratio carries no such guarantee.

The cancellation holds month to month within one regime. The 234 contract-days span **July through October 2023**, not a single month, so the decomposition can be checked out-of-month *within* that window. It holds: fill-weighted net-after-fee stays within ± 0.10 bp of zero in every month (July -0.020 bp, August -0.062 , September $+0.094$, October -0.056), with adverse selection consuming 81–114% of the captured spread month to month (101.1% pooled). The near-exact cancellation is therefore a month-to-month regularity, not a single-month artifact. We scope this carefully and do not over-claim it as a regime-robustness test: these are four *contiguous* months that all fall inside one calm-to-moderate 2023 volatility regime, July carries 93% of the fills, and the August–September windows are weekly-sampled and thin (hundreds to a few thousand fills). The precise ratio breathes with the month rather than sitting at a fixed 101%. This is month-to-month stability *within* one regime; whether it extends *across* regimes is taken up in Section 6, where a feed control shows the obvious cross-regime comparison is confounded.

Every root tells the same story. Far from an average over a few dominant roots, the cancellation holds root-by-root (Table 2). Every one of the 13 roots sits within ± 0.13 bp of zero net-after-fee, ranging from $+0.083$ bp (ZN, the 10Y note, the most maker-friendly) to -0.122 bp (NG, natural gas, the least). The roots with the widest captured spreads also suffer the widest adverse drift: NG captures a remarkable $+2.701$ bp and loses -2.808 bp; ZB captures $+1.309$ bp and loses -1.281 bp. The spread a market maker can earn and the adverse selection it pays are co-determined by the same information environment, which is exactly the Glosten–Milgrom [2] prediction and exactly why no quoting cleverness in Section 7 can separate them. The honest caveat is that this null is shown at the realistic CME 0.015 bp, no-rebate level; a rebate venue (some crypto or equity maker programs) would shift the maker economics, which we report rather than assume (Section 11).

Table 2: Maker-P&L decomposition by root (10s horizon, basis points, fill-weighted): captured spread is almost exactly offset by adverse selection, pinning net-after-fee within ± 0.13 bp of zero on every root. Adverse selection consumes essentially the full captured half-spread overall (pooled ratio indistinguishable from 100%).

root	# fills	spread capture	adverse selection	net markout	net after fee
ZN	14,391	+0.730	−0.632	+0.098	+0.083
SI	6,999	+1.087	−1.009	+0.077	+0.062
CL	69,728	+0.914	−0.872	+0.042	+0.027
ZB	6,575	+1.309	−1.281	+0.028	+0.013
ZF	15,657	+0.394	−0.375	+0.019	+0.004
NQ	29,733	+0.112	−0.098	+0.014	−0.001
HG	11,473	+0.826	−0.817	+0.010	−0.005
ES	113,689	+0.280	−0.301	−0.021	−0.036
6E	11,357	+0.277	−0.312	−0.035	−0.050
GC	28,741	+0.310	−0.351	−0.041	−0.056
RTY	10,512	+0.277	−0.358	−0.081	−0.096
6J	8,109	+0.419	−0.520	−0.101	−0.116
NG	26,423	+2.701	−2.808	−0.107	−0.122
ALL	353,387	+0.654	−0.661	−0.007	−0.022

Maker-P&L decomposition: 353,387 true-signed CME fills, 232 root-days

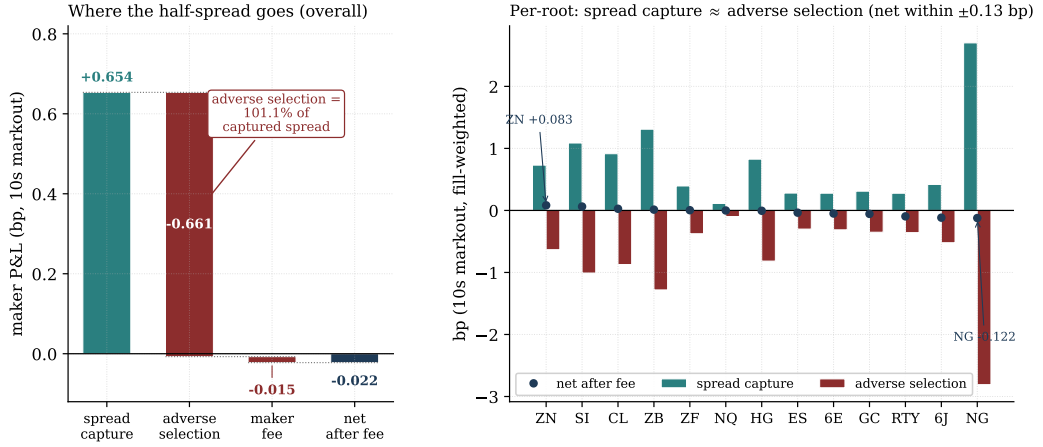


Figure 3: The central result: the half-spread captured at the touch (+0.654bp) is consumed almost exactly by post-fill adverse drift (−0.661 bp), leaving net maker P&L indistinguishable from zero on every root. The quoting null (Section 7) is the corollary.

An illustration: the adverse drift and the $\sigma^2/8$ LVR. At the role level the adverse drift we measure is a *known* theoretical equivalent of the AMM literature’s loss-versus-rebalancing: both are the informed counterparty’s pickoff of a passive liquidity supplier, with Milionis et al. [7] pricing the AMM version at $\sigma^2/8$. We do not re-derive that equivalence; we put the two on one axis purely as a consistency check on the CME measurement (runs/maker_decomp/lvr_identity.json), and the check self-limits. The two do not coincide in level: literal $\sigma^2/8$ over 10s runs some $28,000\times$ below the 0.661 bp adverse drift, because $\sigma^2/8$ is a variance rate (quadratic in σ) while the per-fill adverse drift is a mean signed move (linear in σ), so there is no per-fill identity. What

the cross-root data show is a shared volatility ordering: both the adverse drift and the implied LVR rise with realised σ_τ , but the raw co-movement is tail-driven and, once σ_τ is partialled out, all but vanishes.¹ A cross-root pattern in which both scale with σ is equally the signature of *one* σ -driven cost or of *two* unrelated costs that both scale with σ , and the data cannot distinguish them; we therefore report the scaling as consistent-with, not confirmatory, and rest the unification on the known role-level equivalence rather than on these descriptive correlations.

5.1 A competitive zero-profit reading, and a derived survival condition

The cancellation supports a sharper reading than “the Glosten–Milgrom cost is large.” Write the touch-quoter’s per-fill economics as $\text{net} = s - a - f$, with s the captured half-spread, a the adverse drift, and f the fee. A risk-neutral maker enters only while $s > a + f$; on a competitive book entry bids the captured spread down until $s^* = a + f$ and per-fill net vanishes. The pooled -0.022bp , indistinguishable from zero, is exactly that: captured spread one fee above the adverse drift, where free entry would pin it. We read this as *consistent with* a competitive zero-profit equilibrium, holding the same line we hold on the LVR illustration: we measure $s - a - f \approx 0$ and interpret it, we do not vary entry or identify the channel, and “competition drove net to zero” is not distinguished here from “net is near zero for another reason.” Closing the reading with the volatility scaling $a \approx \kappa \sigma_\tau$ ($\kappa \approx 0.74$, the through-origin fit of the LVR footnote and Appendix A) adds no evidence: it is algebraically downstream of the cancellation, since $\text{capture} \approx \text{adverse}$ and $\text{adverse} \approx \kappa \sigma_\tau$ give $\text{capture} \approx \kappa \sigma_\tau$, the decomposition restated in σ rather than a second test.

The tick grid turns this reading into a structural companion (the same per-fill identity read across roots, not a logical consequence of the pooled cancellation). A book cannot quote inside one tick, so the captured half-spread is bounded below by $\tau/2$, with τ the tick width in basis points, a *lower bound* we use below; realized fill-weighted capture sits at this floor only for the finest-tick root (ES) and well above it for most (NQ captures $1.37\times$ its floor). Combined with the per-fill ledger $\text{net} = s - a_i - f$ evaluated at each root’s *own* measured adverse drift a_i (not the cross-root $\kappa\sigma$ fit, whose $R^2 = 0.77$ scatter would contaminate a per-root statement), one condition is genuinely *derived* and holds root-by-root (Table 3). (i) *A sufficient floor condition*: if $\tau/2 \geq a_i + f$ then, since $\text{capture} \geq \tau/2$, the tick floor alone clears the cost and the maker is net-positive whatever competition does. It is met only by ZN ($\tau/2 = 0.707 \geq a_i + f = 0.647$) and SI ($1.072 \geq 1.024$), both net-positive as guaranteed. (ii) Empirically every surviving root is also coarse-tick (the $\approx 0.74\text{bp}$ crossover of Figure 4, set by the marginal survivor ZF at 0.736bp), but we read this as a *descriptive sort-order boundary, not a structural necessity*: it rests on the single marginal survivor (ZF, 0.736bp), a coarse tick neither guarantees survival (NG and HG clear it yet lose) nor is provably required, and the binding driver beyond the sufficient floor is the adverse drift, not the tick.

Among the net-negative roots the loss splits cleanly in two by whether the captured

¹Per root, adverse drift $\approx 0.74 \sigma_\tau$ (through-origin $R^2 = 0.77$) and the $\sigma^2/8$ LVR co-varies with the adverse drift at Pearson $r = 0.92$; but r falls to 0.57 dropping NG and 0.39 dropping the high-vol tail (NG, SI, CL), bootstrap 95% CI $[0.33, 0.99]$, and the partial correlation given σ_τ collapses from 0.56 to 0.03 – 0.13 . What is stable is that both quantities track σ_τ (correlations 0.88 and 0.97) and the rank association survives drop-NG (Spearman $0.64/0.54$). Full breakdown in Appendix A (Table 8); source `lvr_robustness.json`.

spread clears the adverse drift. *Fee-dominated near-breakevens* (NQ and HG) capture marginally *more* than their adverse drift (NQ +0.112 against −0.098, HG +0.826 against −0.817) and go negative only by the 0.015 bp fee; they sit on the zero-profit line, corroborating the reading above. The other six (ES, RTY, 6E, GC, 6J, NG) are *adverse-dominated*: the adverse drift exceeds the captured spread, by a margin that grows with the information environment (ES 0.301 against 0.280, rising to NG 2.808 against 2.701). For ES alone the captured spread sits right at its one-tick floor (0.280 at a 0.282 bp floor), the one root where the tick binds capture directly; elsewhere the ordering within the loser set tracks adverse selection, not tick width: 6J carries a coarser tick than ES yet loses far more. The frontier is therefore the same per-fill identity (net = capture − a_i − f) read across roots rather than a competing mechanism: the tick sets a *sufficient* floor (condition i); beyond it, each root’s own adverse drift a_i (the fee at the margin for the near-breakevens) decides survival.

A tick-size frontier for spread-capture survival. The per-root spread the maker can earn is bounded below by structure: a root’s tick is the minimum price increment, so its tick width in basis points (tick/price $\times 10^4$) is the finest spread the book can quote. A maker keeps the spread net of the fee only where the captured half-spread clears the adverse drift plus the fee, and the per-root net-after-fee from the decomposition is the realised outcome of that contest. Ordering the 13 roots by tick width makes the structural reading visible (Table 3, Figure 4): at the paper’s 0.015 bp maker fee, the roots that survive with a positive net-after-fee are the coarse-tick ones, **ZF, CL, ZN, SI, ZB**, whose tick widths run from 0.74 bp (ZF) to 2.53 bp (ZB); the fine-tick equity-index roots go negative (ES −0.036 bp at a 0.56 bp tick, RTY −0.096 bp, NQ a hair under zero at −0.001 bp). The crossover sits at a tick width of about **0.74 bp**, with every surviving root coarse-tick (the marginal one ZF at 0.736 bp); we read this as a *descriptive sort-order boundary*, not a structural requirement (Section 5.1). The genuinely derived condition is the *sufficient* tick floor $\tau/2 \geq a_i + f$, met by ZN and SI; beyond it the adverse drift decides, and the figure shows a coarse tick is no guarantee: NG has the widest tick (3.89 bp) yet the deepest loss (−0.122 bp), because its adverse drift (−2.808 bp) overwhelms even a wide captured spread, and HG (1.32 bp tick) sits just negative for the same reason. A coarse tick raises the floor on the spread a maker can quote, but a root with a violent information environment gives the spread straight back regardless. On fees, the headline uses the locked 0.015 bp; a per-product band derived from the CME per-side dollar fee over each contract’s notional (about \$0.10 to \$0.40 per side, e.g. ES $\approx$ 0.011 bp, ZN $\approx$ 0.018 bp, both sub-0.015 bp, against NG’s 0.097 bp on its small notional) leaves the survivor set unchanged for every root except at the very margin, so the frontier is not an artifact of a single fee assumption.

Table 3: Tick-size frontier (per root, 0.015bp maker fee, 10s, fill-weighted). Tick width is the minimum quotable spread in basis points; net-after-fee is captured spread minus adverse selection minus fee. Only the coarse-tick roots (tick width above the ≈ 0.74 bp crossover) keep the spread net of the fee; the coarse-tick frontier is descriptive (NG/HG are coarse-tick yet lose), the genuinely derived condition being the sufficient floor $\tau/2 \geq a_i + f$ (met by ZN, SI). Sorted by tick width.

root	tick	notional (\$)	tick width (bp)	captured (bp)	net-after-fee (bp)	survives
NQ	0.25	303,928	0.165	+0.112	−0.001	no
6E	0.00005	137,572	0.454	+0.277	−0.050	no
GC	0.10	193,080	0.518	+0.310	−0.056	no
RTY	0.10	93,850	0.533	+0.277	−0.096	no
ES	0.25	221,881	0.563	+0.280	−0.036	no
6J	0.0000005	88,906	0.703	+0.419	−0.116	no
ZF	0.0078125	106,168	0.736	+0.394	+0.004	yes
HG	0.0005	94,512	1.323	+0.826	−0.005	no
CL	0.01	73,870	1.354	+0.914	+0.027	yes
ZN	0.015625	110,523	1.414	+0.730	+0.083	yes
SI	0.005	116,600	2.144	+1.087	+0.062	yes
ZB	0.03125	123,516	2.530	+1.309	+0.013	yes
NG	0.001	25,690	3.893	+2.701	−0.122	no

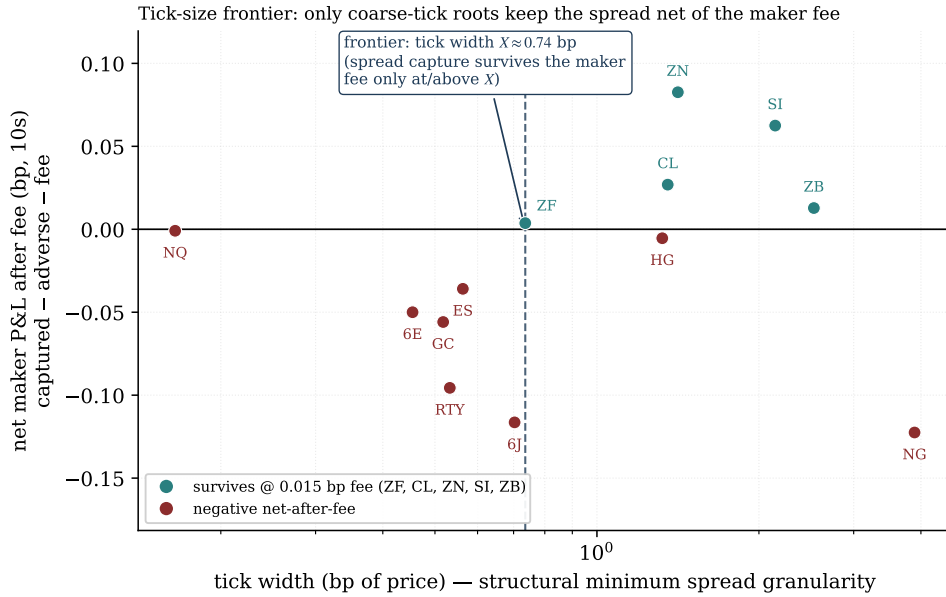


Figure 4: The tick-size frontier. Net maker P&L after the 0.015bp fee against tick width (log scale) for the 13 roots. Spread capture survives the fee only at and above a tick width of ≈ 0.74 bp (ZF, CL, ZN, SI, ZB), where the minimum quotable spread is wide enough to clear the adverse drift; the fine-tick equity-index roots fall below zero. NG, widest tick but deepest loss, shows a coarse tick is no guarantee and not a structural requirement: every survivor is coarse-tick (a descriptive frontier), but beyond the sufficient tick floor the binding driver is the adverse drift, which for NG surrenders the spread regardless.

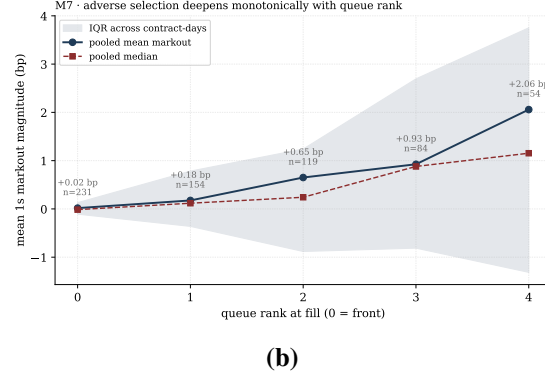
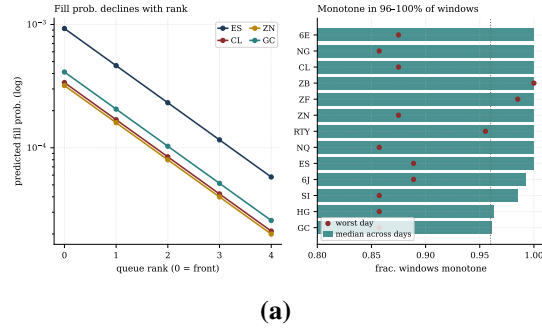


Figure 5: The two halves of the mechanism, in the queue. (a) The modelled geometric fill probability declines monotonically with queue rank, and empirical fill rates are rank-monotone in 96–100% of walk-forward windows across all 13 roots, validating the queue-position fill model. (b) Adverse selection deepens monotonically with that same rank: pooled mean 1s markout magnitude grows from ≈ 0.02 bp at the front of book to > 2.0 bp at deep ranks, the microstructural source of the adverse-selection cost that consumes the captured spread.

5.2 Conditional Decomposition by Session and Regime

The pooled cancellation could in principle hide a regime where the maker does keep the spread. A desk would condition: quote only in the calm regime, or only away from the volatile session open, where adverse selection might be milder. We test the sharpest version of that idea on a pre-committed axis, root by session by regime, attaching to each of the 232 root-days’ fills both a session tag (the first 60 minutes of the day’s fills, the open auction and early regular hours, versus the rest of the session) and an online causal volatility-regime label. The regime label reuses the H4 pipeline (Section 9): one-minute mid bars, a $K = 4$ Gaussian HMM fit on each day’s in-day warmup span, vol-ordered so the top state is high-vol, then forward-filtered so the label at a fill uses only bars at or before it (no look-ahead, the same causality the rest of the paper enforces). We aggregate fill-weighted spread, adverse selection, and net-after-fee in the four session-by-regime cells, pooled across roots and per root. One caveat on the labels: the HMM is fit on the thin in-day warmup (the first ~ 36 –60 one-minute bars), where a $K = 4$ Gaussian model is weakly identified and the run logs show non-convergence or degenerate fits on some days, so the regime tags carry real label noise; the labels remain strictly causal (no leakage), and the paper’s headline null does not lean on them (it rests on the pooled and per-root decomposition of Section 5), but the conditional cells below should be read as sitting on noisy regime labels rather than clean states.

Those pooled cells confirm the mechanism rather than break it (Table 4). The high-vol session open is the most adverse slice, net-after-fee -0.056 bp on 2,739 fills, where the adverse drift (-0.764 bp) runs widest of the four cells against a spread ($+0.723$ bp) no larger than elsewhere; the calm open is mildly positive ($+0.047$ bp), and the two mid-session cells sit at -0.029 bp (calm) and 0.000 bp (high-vol). Conditioning on regime does not find a cell where the maker is paid for volatility; it finds that the volatile open is exactly where the inventory cost is worst, which is the Glosten–Milgrom reading.

On the locked exploratory axis, root by high-vol by session open, five of the thirteen per-root cells have a positive net-after-fee: NG $+0.400$ bp ($n = 111$ fills), CL $+0.155$ bp ($n = 248$), NQ $+0.069$ bp ($n = 374$), 6E $+0.066$ bp ($n = 55$), and GC $+0.0005$ bp ($n = 128$). We report these as *exploratory sensitivities with their fill counts, not as a result*.

They are thin slices (55 to 374 fills) of a family the paper shows is null at scale (the inventory-skew family 0 significant before any correction, PBO 0.543, over a 1,126-cell nominal grid, Section 7), they carry no multiplicity correction, the pooled high-vol-open cell across all roots is *negative* (-0.056 bp), and the GC cell is a rounding error from zero. A single conditional positive on an axis chosen after a null does not overturn the null; it is the kind of cell a search would surface by chance, and we flag it as such rather than promote it. The honest reading of the conditional sweep is the pooled one: the captured-spread-equals-adverse-drift cancellation holds across sessions and regimes, and the worst cell is the volatile open.

Table 4: Conditional maker-P&L decomposition (10s, fill-weighted, net-after-fee at 0.015 bp): the four pooled session-by-regime cells (top), and the locked exploratory axis root $\times$ high-vol $\times$ session-open (bottom), over 232 root-days. The high-vol open is the most adverse pooled cell. The five per-root cells with a positive net-after-fee are exploratory sensitivities with their fill counts, not multiplicity-corrected passes; the pooled high-vol-open cell is negative.

pooled cell	# fills	spread (bp)	adverse (bp)	net-after-fee (bp)
calm, session open	3,751	+0.729	−0.667	+0.047
calm, mid-session	271,903	+0.656	−0.670	−0.029
high-vol, session open	2,739	+0.723	−0.764	−0.056
high-vol, mid-session	74,994	+0.638	−0.623	+0.000
<i>Exploratory</i> (root $\times$ high-vol $\times$ open), net-after-fee > 0 only:				
NG	111	+4.441	−4.027	+0.400
CL	248	+1.503	−1.333	+0.155
NQ	374	+0.167	−0.084	+0.069
6E	55	+0.278	−0.196	+0.066
GC	128	+0.298	−0.283	+0.001

Conditional maker-P&L decomposition: 232 root-days, online causal regime (no lookahead), 0.015 bp fee

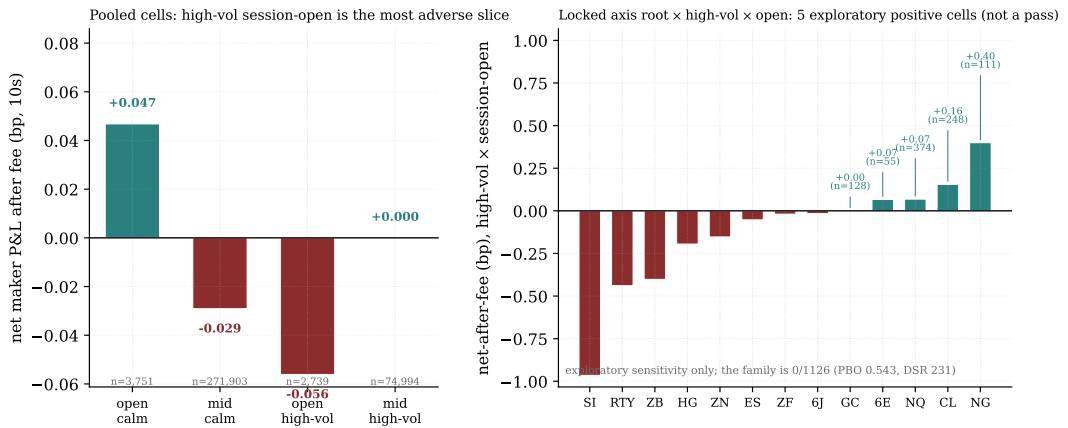


Figure 6: Conditional maker-P&L decomposition over 232 root-days with an online causal regime label (no look-ahead). Left: the four pooled session-by-regime cells; the high-vol session open is the most adverse (-0.056 bp net-after-fee). Right: the locked exploratory axis, per-root net-after-fee in the high-vol session-open cell with fill counts; the five positive cells (NG, CL, NQ, 6E, GC) are thin exploratory sensitivities of an already-null family (inventory-skew 0 significant before correction, PBO 0.543), not multiplicity-corrected passes, and the pooled cell is negative.

6 Cross-Regime Evidence

The cancellation of Section 5 is established within a single calm-to-moderate volatility regime. Whether it survives *across* regimes is a separate question, and answering it cleanly requires first clearing a data-feed confound; once that is resolved on a single nanosecond feed, a continuous decade-long sweep traces the full volatility dose-response.

Is the cancellation regime-dependent? On a matched feed, yes. The natural test is whether the cancellation holds in a *distinct* volatility regime, and it is treacherous because the obvious comparison crosses reconstructions of differing time resolution. The nanosecond-resolution depth-10 reconstruction underlies the calm Jul–Oct 2023 headline; a coarser-resolution reconstruction of the same CME data reconstructs the book differently. A same-dates control proves the hazard: on the *identical* five Jul-2023 sessions the high-resolution reconstruction reads 99.5% and the coarser-resolution one 82.1%: a 17.4pp gap with dates, roots, and regime all held fixed, attributable to the reconstruction alone (it persists at the 60s horizon, 14.9pp; artificially coarsening the high-resolution markout timeline does *not* reproduce it, staying near 98%, so it is structural to the book reconstruction, not a clock-precision artifact). An earlier draft read a calm-vs-stress difference across the two reconstructions (101% against 87.6%) as regime dependence; since the reconstruction alone moves the ratio 17.4pp on identical dates, that reading could not stand and we withdrew it (Section 11). The resolution is to hold the reconstruction fixed across all regimes on a single nanosecond-resolution, true-aggressor-signed TAQ feed. The touch decomposition needs only the level-1 mid at fill, the mid at the markout horizon, and the true aggressor sign (all carried by the TAQ tape), so a depth-1 ingester (Section 4, `mmsim.ingest.fut_taq`) runs the *identical* decomposition on it across regimes.

The TAQ feed is commensurable with depth, and the cancellation breaks down in stress. Before trusting the TAQ feed across regimes we tie it to the depth headline where both exist. Re-running the decomposition from the TAQ-built level-1 book over the 52 matched Jul-2023 calm root-days reproduces the depth-feed numbers to within measurement noise: adverse-to-captured 99.6% (TAQ) against 99.3% (depth), net -0.012 against -0.011 bp, with a per-root-day median absolute difference of 0.010 – 0.017 bp on each component, against the coarser-resolution reconstruction’s 17.4pp gap on the same dates. The nanosecond TAQ and depth reconstructions are the same measurement; the coarser-resolution reconstruction is the outlier. We then run the TAQ decomposition across three stress episodes spanning the volatility range, each 13 roots and 65 root-days under the same engine, cost model, markout, and front-month rule: the March-2023 SVB banking week, the June-2022 FOMC/post-CPI week, and the March-2020 COVID crash week (Table 5, Figure 8). On this one feed the calm cancellation gives way. The adverse-to-captured ratio rises from $\approx 100\%$ in the calm regime to 105.1% at $\sigma_\tau = 2.1$ bp (SVB), 111.5% at 2.5 bp (FOMC) and 109.8% at 6.8 bp (COVID), and net-after-fee deepens with volatility: -0.046 , -0.087 and -0.139 bp against the calm -0.022 bp. Adverse selection does not merely consume the captured spread in stress; it *over*-consumes it, and the maker’s net loss grows with σ .

The stress breakdown is broad, not tail-driven. We hold the stress result to the same standard as the headline, which is the root-day cluster bootstrap. Re-running that bootstrap per episode (the same $B = 10,000$ root-day resampling that gives the calm

CI of [97.7, 104.7]%) puts the adverse-to-captured ratio CI at [100.7, 109.6]% (SVB), [102.8, 120.4]% (FOMC) and [102.7, 116.8]% (COVID), each excluding 100%, and net-after-fee at $[-0.070, -0.020]$, $[-0.177, -0.030]$ and $[-0.223, -0.049]$ bp, each excluding zero (`stress_bootstrap_ci.json`). So the over-consumption clears its own within-feed sampling noise: narrowly at SVB (the lower bound sits just above 100%) and comfortably at FOMC and COVID. The point estimates corroborate root-by-root. The ratio exceeds 100% at the *median* root in every episode (105.9% SVB, 104.4% FOMC, 104.0% COVID), not only fill-weighted, and the median root is net-negative throughout ($-0.038, -0.047, -0.088$ bp); 10, 9 and 11 of the 13 roots respectively lose money net. Dropping the three highest-fill root-days per episode leaves net negative and similar (COVID $-0.139 \rightarrow -0.122$ bp, SVB $-0.046 \rightarrow -0.033$ bp), so the breakdown is not an artifact of one busy contract: the single extreme name (NG in the FOMC week, -1.18 bp) inflates the fill-weighted FOMC ratio, but the median-by-root figure shows the effect is general. One caveat scopes the inference, stated plainly: commensurability is established on the *calm* sessions used for the headline; the stress extension assumes the TAQ-vs-depth agreement carries into stress, which we do not separately verify against a depth-10 reconstruction in the stress episodes. The mechanism is unchanged across that bridge (the touch economics are a function of the level-1 mid and the true sign, which the TAQ tape measures identically in calm and stress), but a referee should read the stress ratios as matched-feed evidence, not a depth-verified identity. What is clean and feed-internal is the *direction*: under one consistent measurement, the cancellation is regime-dependent, and adverse selection wins by more as volatility rises.

A continuous cross-regime view: the σ -gradient is root-idiosyncratic. The three stress weeks are three high- σ points. To trace the full σ dose-response we extend the same TAQ decomposition to a continuous 2016–2026 sweep, a frozen 33,956 root-day snapshot across the 13 roots (96% of the available universe; every available trading day, the same engine, cost model, markout and front-month rule). Two findings sharpen the regime picture (Figure 7). *First, a methodological one.* A naive pooled σ -curve is *composition-confounded*: high- σ days are a structurally different root set than low- σ days (the calm $\sigma < 1$ bp set is rates, FX and equity, with energy and metals nearly absent), so the pooled adverse-to-captured rank correlation against σ is -0.111 , near zero and sign-inverted, a Simpson’s-paradox artifact. The cross-regime relationship is interpretable only *within root* (root fixed effects, inference clustered over roots). *Second, conditioned within root, the σ -gradient of net maker P&L is real but root-idiosyncratic.* The leverage-weighted aggregate slope is -0.036 bp per bp of σ (cluster-bootstrap CI $[-0.072, -0.005]$; sign and per-root concentration stable under a σ -rank recoding that discards extreme-day leverage), but it is *carried by a handful of contracts* (natural gas, copper, crude and the two FX, with per-root slopes -0.135 to -0.031 , each individually significant in a day-level within-root bootstrap, crude and JPY only marginally), and is not robust to removing them: the aggregate already loses significance on dropping just the three highest-leverage roots (CL, SI, NG $\rightarrow -0.038$ bp, CI containing zero) and vanishes on dropping the five drivers (-0.003 bp, CI containing zero); gold, silver, the equity indices and the rates complex have flat slopes. It is neither a clean asset-class law (copper steepens; gold and silver do not) nor explained by realised- σ range (FX steepens at the *narrowest* range). The mechanism is the decomposition read across σ , and it is the same in *every* root: a rising σ widens the captured half-spread *and* deepens the adverse drift (Glosten–Milgrom: more volatility brings both wider quotes and more informed pickoff). What differs is the *balance*. In

the driver roots the adverse-deepening outpaces the spread-widening (natural gas $+0.27$ capture against -0.36 adverse per bp of σ), so net deepens; in the σ -insensitive roots the two move together and cancel (gold $+0.07/-0.07$, silver $+0.08/-0.08$, the 10Y note $+0.07/-0.06$), leaving net flat. So the root-idiosyncrasy is not whether σ reaches the maker (it does, everywhere) but whether the spread it can widen keeps pace with the information σ carries; it fails to only in the high-information commodity and FX roots. The concentration is otherwise robust on this snapshot: the per-root σ -gradient is horizon-invariant (the 60s markout gives the same within-root slope, drivers -0.061 against -0.065 at 10s) and time-stable (every driver negative in both chronological halves, split April 2021). We report the aggregate slope as suggestive (a 13-cluster bootstrap is anti-conservative) and the per-root concentration and its mechanism as the substantive finding; the composition-confound caution is the robust, reproducible contribution.

This refines, rather than contradicts, the “broad-based” stress result above. The three crisis weeks are cross-sectional *snapshots* at extreme σ , where most roots are net-negative, a *level* statement (9–11 of 13 lose money in a given crisis week). The continuous sweep measures the σ -gradient (a *dose-response*, how fast a root’s net deepens as its own σ rises), and that gradient is concentrated in a few roots. A contract can be net-negative at crisis σ yet have a near-flat σ -slope (its loss roughly σ -independent), and the two sets are empirically distinct: seven non-driver roots (silver, the three equity indices, and the rates complex) are net-negative in two or more of the three crisis weeks yet carry flat σ -slopes; they lose in a crisis but their loss does not deepen with σ . Both therefore hold: stress is broadly costly as a snapshot, while the *volatility-sensitivity* of the maker’s loss is carried by energy, copper and FX. The desk reading is the sharper of the two: do not infer a cross-regime law from a single pooled σ -curve, and expect σ -conditional maker degradation specifically in natural gas, copper, crude and FX.

No quote or flow signal harvests a return on any episode. The small positive net-after-fee the coarser-resolution episodes show is a one-sided upper bound (Section 4) and, per the reconstruction control, partly a reconstruction artifact; either way the central message holds across every episode: no inventory-aware quote or order-flow signal earns a return net of cost. Re-running both corollary tests on the June-2022 coarser-resolution data, the quoting family is null (PBO = 0.500, BH/BHY 0/325; the inventory-skew family adds nothing over symmetric quoting: median skew-minus-symmetric P&L 0, 0/65 permutation-significant), and the cross-root integrated-OFI signal validates its pipeline in-sample ($R^2 = 0.559$) but adds nothing out-of-sample (incremental OOS $R^2 = -0.006$, 0/13 roots BHY-significant, block-null $p = 0.80$) and turns from $+0.10$ bp gross to -1.43 bp at full taker cost. The two corollary nulls are thus robust across feeds and episodes; no tested cleverness improves on quoting at the touch. Banked to `harvestability_jun2022.json`.

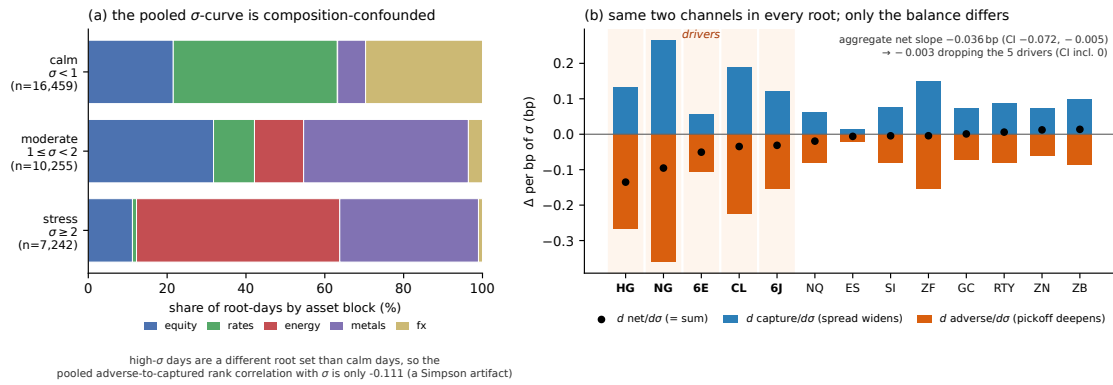


Figure 7: The continuous cross-regime sweep (33,956 root-days, 96% of the universe, 2016–2026; same engine, cost model, markout and front-month rule). (a) *The pooled σ -curve is composition-confounded.* High- σ days are a structurally different root set than calm days (the calm $\sigma < 1$ bp set is rates, FX and equity, while the $\sigma \geq 2$ bp tail is energy and metals), so the pooled adverse-to-captured rank correlation with σ is only -0.111 , near zero and sign-inverted (a Simpson's-paradox artifact). The cross-regime relationship is interpretable only within root. (b) *Conditioned within root, the same two channels move in every root; only the balance differs.* A rising σ widens the captured half-spread (blue) and deepens the post-fill adverse drift (orange) in all 13 roots (Glosten–Milgrom); the net σ -slope (black dot) is their sum. In the five driver roots (shaded: natural gas, copper, crude, EUR and JPY) the adverse-deepening outpaces the spread-widening, so net deepens; in the others the channels cancel and net is flat. The leverage-weighted aggregate slope is -0.036 bp per bp of σ but is carried by those five; dropping them leaves -0.003 bp, CI containing zero. Source `frozen3_results.json`; reproduce with `frozen_analysis_fast.py`.

Table 5: Maker-P&L decomposition across volatility regimes on the *matched* nanosecond feed (10s, fill-weighted, bp). *Top block:* under one consistent reconstruction (depth-10 for the calm headline, the commensurable nanosecond TAQ for the stress episodes), the adverse-to-captured ratio rises from $\approx 100\%$ calm to 105–111% in stress and net-after-fee deepens with σ_τ , so the cancellation is regime-dependent. *Middle block (commensurability):* on the 52 matched Jul-2023 root-days the TAQ-built L1 book reproduces the depth-10 decomposition to 0.3pp (99.6% vs 99.3%), while a coarser-resolution reconstruction reads 82.1%, a 17.4pp *reconstruction* artifact, not a regime effect. *Bottom block:* the four coarser-resolution episodes (confounded, shown for reference) read a flat 79–88% across the same σ_τ span; that flatness is the artifact that sank an earlier draft's reading. Stress ratios exceed 100% at the median root too (104–106%), with 9–11 of 13 roots net-negative per episode, and a per-episode root-day cluster bootstrap puts each stress ratio's 95% CI above 100% and each net CI below zero (Section 5, `stress_bootstrap_ci.json`). Same engine/cost/markout/front-month rule throughout.

episode	feed	σ_τ	# fills	capture	adverse	adv/cap	net
<i>Nanosecond feed (the regime test):</i>							
2023 calm (Jul–Oct)	depth-10	1.0	353,387	+0.654	−0.661	101.1%	−0.022
2023 SVB (Mar)	nanosec TAQ	2.1	383,088	+0.616	−0.647	105.1%	−0.046
2022 FOMC (Jun)	nanosec TAQ	2.5	289,248	+0.629	−0.701	111.5%	−0.087
2020 COVID (Mar)	nanosec TAQ	6.8	362,014	+1.265	−1.389	109.8%	−0.139
<i>Commensurability control, 52 matched Jul-2023 root-days:</i>							
same dates	depth-10	1.0	142,505	+0.666	−0.662	99.3%	−0.011
same dates	nanosec TAQ	1.0	142,278	+0.676	−0.674	99.6%	−0.012
same dates	coarse recon	1.1	214,597	+0.485	−0.398	82.1%	+0.072
<i>Coarser-resolution reconstruction (confounded, for reference):</i>							
2021 calm	coarse recon	1.1	200,014	+0.438	−0.366	83.5%	+0.057
2023 SVB (Mar)	coarse recon	2.1	418,712	+0.521	−0.412	79.1%	+0.094
2022 FOMC (Jun)	coarse recon	2.5	338,194	+0.516	−0.452	87.6%	+0.049
2020 COVID (Mar)	coarse recon	7.3	552,910	+1.104	−0.907	82.2%	+0.182

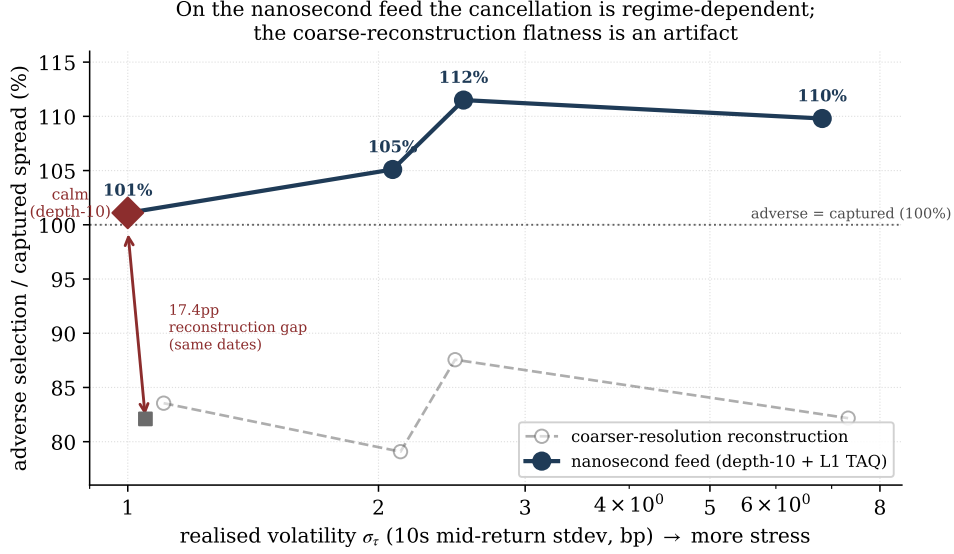


Figure 8: The adverse-to-captured ratio against realised volatility σ_τ (log scale). On the *matched* nanosecond feed (navy; calm point from depth-10, the diamond, the three stress points from the commensurable nanosecond TAQ) the ratio rises from $\approx 101\%$ in calm markets to 105–111% under stress, so the cancellation is regime-dependent and adverse selection over-consumes the captured spread as volatility rises. The coarser-resolution reconstruction (grey, dashed) is flat at 79–88% across the same $7\times$ span, but the vertical arrow marks the same-dates control: on identical Jul-2023 sessions the high- and coarse-resolution reconstructions differ by 17.4pp, so that flatness is a reconstruction artifact, not the absence of a regime effect. The TAQ and depth reconstructions agree to 0.3pp on the matched calm dates.

7 Corollary 1: The Quoting Null

If the captured spread and the adverse drift are co-determined by the same information (Section 5), then no inventory-aware quote can hold one without paying the other, and a search over quoting cleverness should come up empty. It does. We enumerate **about five structural shapes** a desk would reach for, inventory-skew quoting (the Avellaneda–Stoikov move [1]), realised-spread and zero-fee-skill diagnostics, micro-price quoting, and integrated-OFI quoting, each shape swept over per-root-day configurations into a **1,126-cell** nominal grid across the 13 roots and 29 dates, evaluated walk-forward with the corrected maker economics of Section 4. This is a five-shape family on a 1,126-cell grid, not 1,126 independent strategies; we report the count nominally and do not deflate it (the family is null before any correction, so a deflation would be moot). The family is uniformly null, and we read that off the raw effect sizes and the overfitting probability, not off the FDR count, which at our permutation resolution cannot reject regardless of true edge.² *The inventory-skew family that is the Avellaneda–Stoikov object proper has 0 of*

²The family decomposes into five structural shapes: inventory skew (M1, the A–S object proper, 234 cells), realised-spread and zero-fee-skill diagnostics (M6, 231 cells each), micro-price (M3, 217) and integrated-OFI (M3, 213) quoting, for the 1,126-cell total. We report the Benjamini–Hochberg / Benjamini–Hochberg–Yekutieli FDR step (0/1,126 under both, $c(m) = 7.60$) only for completeness, because at $M = 1,000$ permutations it is mechanically incapable of rejecting: the smallest attainable p -value, $1/(M+1) \approx 9.99 \times 10^{-4}$, sits above the BH rank-1 threshold ($\approx 4.4 \times 10^{-5}$ at $m = 1,126$) and the BHY rank-1 threshold ($\approx 5.8 \times 10^{-6}$), so no cell could clear either bar regardless of edge and the 0/1,126 carries no evidential weight on its own. Raising M to $\sim 2.3 \times 10^4$ (BH) or $\sim 1.7 \times 10^5$ (BHY) would make the bars

its 234 cells significant even at an uncorrected 5%, before any multiplicity correction at all. Across the whole grid the only cells significant at an uncorrected 5% (103 of 1,126) are either zero-magnitude realised-spread diagnostics (median $|\text{effect}| \approx 1 \times 10^{-5}$ bp, the spread the maker quotes, mechanically nonzero and economically nothing) or *gross*, pre-cost quantities, the zero-fee-skill and micro-price / integrated-OFI gross signals, which the cost-frontier analysis of Section 8 turns from a +0.091 bp gross edge to −1.30bp net of taker cost. Corroborating the raw count, the probability of backtest overfitting from CSCV (36 configurations $\times$ 29 dates, 8 blocks, 70 splits) is **0.543**, the coin-flip value expected under a true null: the in-sample-best policy has no out-of-sample predictive power, and PBO does not depend on the permutation resolution. We read it as corroboration rather than the lead, because PBO ≈ 0.5 on its own is also what low power would produce; the load-bearing fact is the raw zero-of-234. So the Avellaneda–Stoikov quoting edge, tested at scale on true-sign futures, does not survive a realistic 0.015 bp maker fee in this venue.

This is the corollary the decomposition predicts, not a separate finding: the cancellation of Section 5 is the mechanism, and the uniform null is what it implies for any quote that tries to beat it.

8 Corollary 2: The Cross-Root Integrated-OFI Lead-Lag Null

The quoting null closes the single-book routes; a desk’s next move is to look across books, reading order flow across correlated roots in case the lead-lag the single quote cannot see lives there. If ES flow leads ZN by a few hundred milliseconds, an integrated cross-root OFI signal might recover what the quote alone cannot. We test it, pre-registered (Section 4), on the 7-day synchronized 13-root panel of **30,915** 10s bins (17 walk-forward windows). The pipeline works: own contemporaneous integrated OFI explains the contemporaneous return with in-sample $R^2 = \mathbf{0.545}$, reproducing Cont–Kukanov–Stoikov [3]. The predictive content is the problem. The own *lagged* in-sample R^2 is only 0.0066, and the cross-root *incremental* out-of-sample R^2 is −0.005, positive in just **1 of 17** windows (5.88%). No root benefits: **0 of 13** are BHY-significant for cross benefit. The block-bootstrap null (which preserves vol clustering and destroys cross-root lead-lag alignment) gives $p = \mathbf{0.091}$; the single-bar shuffle (which destroys all serial structure) gives $p = 0.081$; the result is null and robust to scheme. PBO from CSCV is 0.40.

Costs bury what little is there. Even the sign of the gross signal does not survive execution (Table 6). At zero cost the portfolio earns +0.091 bp per 10s with a Lo HAC Sharpe of +0.14 ($t_{\text{HAC}} = +23.56$) and a positive-window fraction of 1.0; at half the measured per-root half-spread it is already −0.61 bp (Sharpe −0.91, positive-window fraction 0.0); at the full spread it is −1.30bp (Sharpe −1.67, $t_{\text{HAC}} = -272.09$). The tail-guard reward-to-risk ratio collapses from ∞ to 0.0. The per-root incremental out-of-sample R^2 is negative for all 13 roots, ranging from −0.0013 (ZF) to −0.0070 (RTY). This is the rigorous true-sign negative the literature lacked: the Cont–Cucuringu–Zhang [4] prior, that contemporaneous cross-impact is subsumed by own integrated OFI and the lagged term is thin, holds in CME futures with the true aggressor sign, and what thin lagged content exists does not survive taker costs.

attainable; we did not, because the verdict rests on the resolution-independent PBO and the raw effect sizes.

Table 6: Cross-root OFI cost frontier (per 10s, 30,915 bins, 17 windows): the gross signal is marginal and turns sharply negative under any realistic taker cost. Own contemporaneous in-sample $R^2 = 0.545$; cross-root incremental OOS $R^2 = -0.005$; 0/13 BHY-significant; block-null $p = 0.091$.

cost	net cross (bp/10s)	net own (bp/10s)	pos-window frac	tail-guard RRR	Lo HAC Sharpe
0 $\times$ (gross)	+0.091	+0.116	1.0	∞	+0.145
0.5 $\times$ ($\frac{1}{2}$ spread)	-0.607	-0.589	0.0	0.0	-0.911
1 $\times$ (full spread)	-1.304	-1.295	0.0	0.0	-1.674

Cross-root integrated-OFI lead-lag null (true-sign CME, 30,915 10s bins, 17 WFO windows)

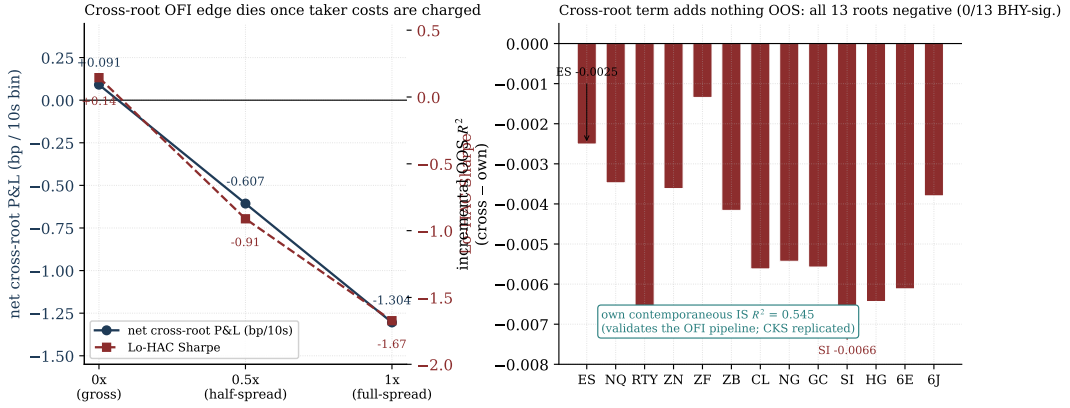


Figure 9: Cross-root integrated OFI adds nothing net: own contemporaneous $R^2 = 0.545$ validates the pipeline (not shown), but cross-root incremental OOS R^2 is negative for every root (right) and the costed portfolio turns from +0.091 bp gross to -1.30 bp at full spread (left).

Spine verdict. The two nulls are corollaries of the decomposition, not separate findings: because the captured spread and the adverse drift are co-determined, neither the inventory-aware quote (Section 7) nor the linear cross-root order-flow signal (Section 8) can separate them, so neither earns a return net of cost in this price-time-priority tick-bound venue. Whether the same cost behaves differently one level up, as a covariance rather than a level, is a forward pointer we sketch in the discussion (Section 10).

9 Supporting Measurements

Beyond the headline results, a few further measurements bear on the data's credibility (the trade sign), on the inventory-cost story (regime conditioning of inventory risk), and on robustness (latency). We report them as support, not as headline findings.

True-vs-inferred sign accuracy (M8). Because the CME data carries the true exchange aggressor sign, we can grade the trade-sign classifiers the futures literature must rely on, over all **12,426,499** trades (Table 7). The tick rule is weak, ranging from **0.607** (ZB, the 30Y bond) to **0.834** (NQ), fill-weighted, far below the $\approx 90\%$ equity benchmark and confirming the futures result that the tick rule degrades in coarse-tick instruments. The quote rule, which our ground-truth sign lets us validate directly, is strong: **0.917** (CL) to **0.993** (NQ). This both certifies that the true-sign anchor is doing real work (the classifiers

a non-true-sign study would use are materially wrong on these roots) and warns that any futures microstructure result built on tick-rule signs inherits a 17–39% sign-error rate on the worst roots.

Table 7: True-vs-inferred sign accuracy by root (fill-weighted, 12,426,499 trades). The tick rule is far below the equity $\approx 90\%$ benchmark (weakest ZB 0.607); the ground-truth-validated quote rule is strong (strongest NQ 0.993, weakest CL 0.917).

root	# trades	tick	quote	root	# trades	tick	quote
ES	3,382,752	0.707	0.952	NG	455,589	0.716	0.939
NQ	2,205,549	0.834	0.993	RTY	444,138	0.734	0.972
CL	1,482,896	0.706	0.917	ZN	444,713	0.666	0.986
GC	1,010,265	0.723	0.965	HG	438,222	0.733	0.933
6E	801,186	0.678	0.932	SI	350,671	0.709	0.983
6J	706,989	0.659	0.942	ZB	240,393	0.607	0.978
ZF	463,136	0.615	0.975	All	12,426,499	0.61–0.83	0.92–0.99

Regime-aware inventory CVaR (H4). The inventory cost is regime-dependent: a maker holding inventory into a high-volatility burst suffers a fatter tail. We test whether a regime-aware spread/inventory rule cuts the high-vol 95% conditional value-at-risk (CVaR) of the held inventory position (in contracts; the CVaR is the mean of the worst-5% tail of $|\text{inventory}|$ after each fill) while *retaining* at least 0.10 of high-vol fills (a CVaR cut bought by simply not trading in high vol is hollow). Of the 13 roots, **10 are evaluable** (3 are undefined-thin: NG, ZB, NQ), and on all **10 of 10** the conditional rule reduces the high-vol inventory CVaR, a one-sided sign test giving $p = \mathbf{0.002}$. The median reduction is **−2.19 contracts** of high-vol inventory CVaR. Of the 10 evaluable, **4 PASS** the full bar (GC, HG, ZN, ZF, a significant cut keeping the fills), three are hollow-overwiden (ES, 6E, 6J), and three are null (SI, RTY, CL). The reading is modest and consistent with the rest of the paper: conditioning on regime trims the worst-case inventory tail on the deepest, most liquid roots, but it does not manufacture a quote-level edge, and on the thin roots it cannot even be evaluated without collapsing fills. We disclose one cost-model wrinkle for this supporting result: the H4 run was executed under the Phase-1 cost model (maker and slippage at 0.2bp), not the corrected maker economics (0.015bp maker fee, no slippage on passive fills) the rest of the paper uses. It does not affect the H4 conclusion, because the reported statistic is the 95% CVaR of $|\text{inventory}|$ measured *in contracts* (the mean of the worst-5% tail of the held position after each fill), a quantity that depends only on the fill and inventory paths and not on the per-fill cost charged; the cost model enters the realised-spread retention check, where it would only make the corrected economics more favourable, not the contract-space CVaR the PASS rests on.

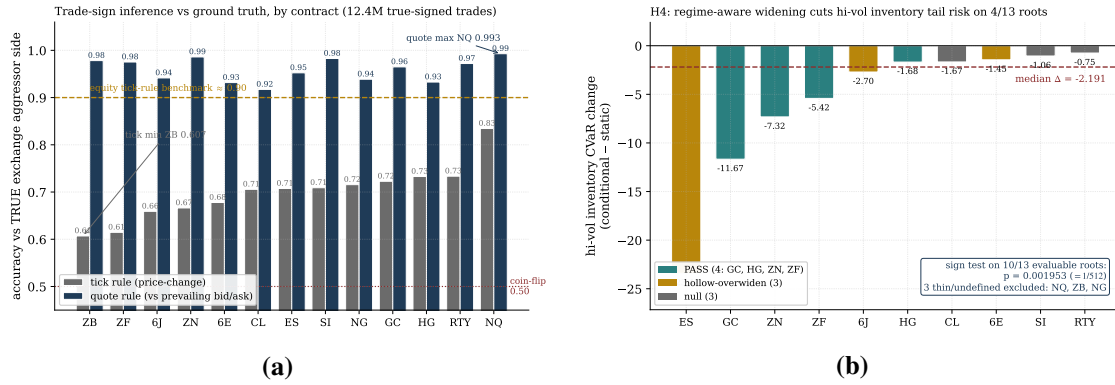


Figure 10: Supporting measurements: the tick rule is unreliable on coarse-tick futures while the ground-truth-validated quote rule is strong (left); regime conditioning reduces the high-vol inventory CVaR (in contracts) on all 10 evaluable roots (sign-test $p = 0.002$), 4 of them passing the full bar, but does not create a quote-level edge (right).

Ruling out a latency artifact (M9). One more robustness check rules out the obvious mechanical objection, that the near-break-even result is an artifact of simulated quote-to-fill latency (Figure 11). The pooled mean realised edge per fill stays in a narrow near-zero band (≈ -0.011 to -0.004 bp) across the whole 0–500ms range of added latency, with no degradation even at 500ms (the mean at 500ms is, if anything, marginally less negative than at 0ms), so the cancellation in Section 5 is a property of the spread/adverse-selection economics, not of the latency assumption.

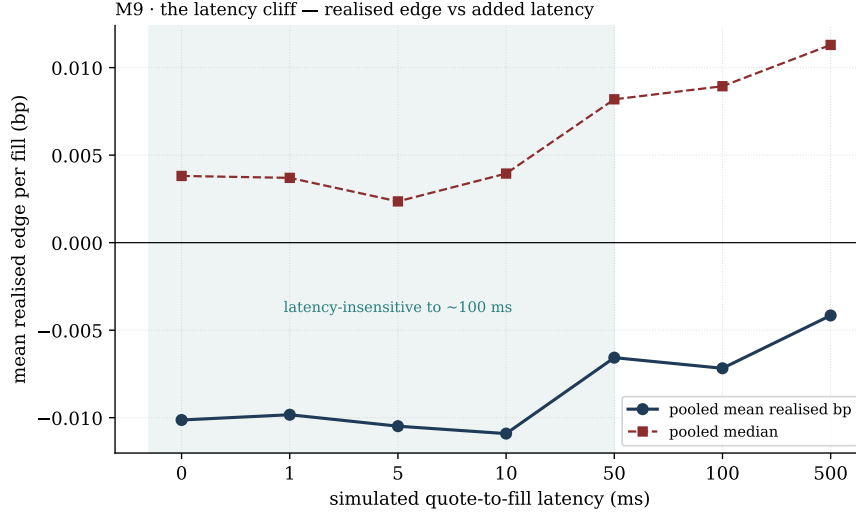


Figure 11: Latency robustness (M9): pooled mean realised edge per fill is latency-insensitive, staying in a narrow near-zero band across the whole 0–500ms range of added quote-to-fill latency with no degradation at 500ms (if anything marginally less negative than at 0ms), indicating the near-break-even quoting result is not a latency artifact.

10 Discussion

This paper’s result is the decomposition, and it backs the title directly: the half-spread a maker earns is consumed by the adverse drift it pays, a near-exact cancellation on every

one of the 13 roots that leaves net maker P&L indistinguishable from zero (Section 5). Because passive fills are charged no slippage, that null is a one-sided bound rather than a point estimate: every realistic friction we have left out pushes net more negative. The defensible claim is that net maker P&L is bounded above by ≈ 0 . The two corollary nulls fall out of it: because the captured spread and the adverse drift are co-determined, the Avellaneda–Stoikov quoting family (Section 7) and the cross-root integrated-OFI signal (Section 8) cannot separate them, so neither construction we test earns a return net of cost in this price-time-priority tick-bound venue.

The known LVR equivalence, as an illustration. At the role level the adverse drift the decomposition measures is the same primitive the AMM literature calls loss-versus-rebalancing (Milionis et al. [7] price the AMM version at $\sigma^2/8$): a known theoretical equivalence, not a finding of ours, which we illustrate on one axis (Section 5) only as a consistency check, finding the cross-root data consistent (the costs paying the widest adverse drift carry the largest volatility) but unable to distinguish *one* σ -driven cost from *two* unrelated ones that both scale with σ , so we report the σ -scaling as consistent-with, not confirmatory.

Practical reading, and a forward pointer. For a market-making or HFT desk the actionable half is the negative: on price-time-priority tick-bound futures, do not expect a quoting signal, however clever, to clear a realistic maker fee, because the captured spread and the adverse drift are co-determined by the same information and net to zero (Section 5), a one-sided bound rather than a knife-edge. The trade-sign result (Section 9) sharpens the warning, since any such signal built on tick-rule signs inherits a large sign-error rate on coarse-tick roots. One direction does stay open, and we flag it only as a forward pointer rather than develop it here. A quote is a bet on the *level* of the next price move, the part of which adverse selection erases; but across many liquidity-supply positions the same cost is a *covariance*, and a covariance is reducible by diversification even when no single stream carries alpha (it cannot, by the same token, manufacture a return). Preliminary cross-chain DEX-LP evidence is consistent with that reading, where robust minimum-variance allocation cuts out-of-sample variance materially net of fees, LVR, gas, and MEV; because it is a different market, instrument, and year, and a single-regime variance reduction rather than a return premium, we develop it in separate work rather than fold it in here.

11 Conclusion and Limitations

In one sentence: on true-sign CME data the adverse drift a maker pays consumes essentially the whole half-spread it captures (+0.654 bp against -0.661 bp, every root within ± 0.13 bp of zero), so net maker P&L is indistinguishable from zero and, under any realistic friction, bounded above by it, and the two corollary nulls of standard quoting and linear order-flow signals follow from that cancellation. On true-sign CME data (13 roots, 234 contract-days, 353,387 maker fills, 12.4M signed trades) the decomposition is the result; the Avellaneda–Stoikov quoting family, about five structural shapes on a 1,126-cell nominal grid, is uniformly null (inventory-skew 0 significant before any correction, PBO 0.543; BH/BHY 0/1126 a consistency note only, below the permutation floor), and a true-sign cross-root integrated-OFI signal adds nothing out-of-sample (-0.005) and dies after cost ($+0.091 \rightarrow -1.30$ bp). The measured adverse drift is the known role-level

equivalent of the AMM literature’s LVR ($\sigma^2/8$); we illustrate that equivalence on one axis as a consistency check, and note as a forward pointer that one level up the same cost is a covariance an LP can diversify, developed in separate work. The reach of the nulls is the constructions we test, not edge in general. The principal threats to validity, which scope the claim, are:

- (i) *The net-zero result is a one-sided bound, not a point estimate.* It rests on charging passive fills no slippage; every realistic friction we omit (partial fills, cancel/replace latency, inventory-unwind cost) pushes net more negative. The honest claim is therefore that net maker P&L is bounded above by ≈ 0 , a stronger statement than the -0.022 bp point.
- (ii) *Cross-regime: established on a matched feed, with one commensurability caveat.* The headline calm cancellation is four contiguous months (July–October 2023, July = 93% of fills) on the depth-10 feed. Comparing it across regimes is confounded if reconstructions of differing resolution are mixed: a same-dates control shows the high- and coarse-resolution reconstructions read 99.5% vs 82.1% on identical Jul-2023 sessions, a 17.4pp gap from the reconstruction alone (structural to the book reconstruction, not a timestamp artifact), and we withdrew an earlier draft’s reading that mistook this for regime dependence. We resolve the confound by holding the reconstruction fixed on the nanosecond-resolution true-signed TAQ feed, which spans all regimes and reproduces the depth-10 decomposition to 0.3pp on the matched calm sessions (99.6% vs 99.3%, 52 root-days). Measured on that one feed the cancellation *is* regime-dependent: adverse selection over-consumes the captured spread in stress (ratio 105–111%, median root 104–106%, per-episode bootstrap CI excluding 100%) and net-after-fee deepens with σ_τ (-0.022 bp calm to -0.139 bp COVID, CI excluding zero), broad-based as a high- σ snapshot (9–11 of 13 roots net-negative per episode). A continuous 2016–2026 sweep (33,956 root-days, 96% of the universe, Section 6) refines this: the per-episode breadth is a cross-sectional level at extreme σ , whereas the within-root σ -gradient of the loss is root-idiosyncratic (carried by natural gas, copper, crude and FX) and a naive pooled σ -curve is composition-confounded; the two are complementary, not contradictory. The commensurability caveat (Section 6) scopes this: it is established on calm sessions only, so the stress ratios are matched-feed evidence of regime-dependence rather than a depth-verified identity. The corollary nulls are unaffected: re-running both on the June-2022 coarser-resolution data leaves them null (quoting PBO = 0.500, 0/325; cross-root OFI incremental OOS $R^2 = -0.006$, -1.43 bp after cost), robust across reconstructions and episodes.
- (iii) *Cross-root OFI is null, not a positive.* The lagged cross-root term is real only in the pipeline-validating sense (own $R^2 = 0.545$); it adds nothing out-of-sample and dies after taker cost. It is the rigorous true-sign negative that fills the stated gap, presented as such.
- (iv) *The LVR equivalence is illustrative, not a test.* The adverse drift is the known role-level equivalent of LVR; the cross-root data are consistent with one shared volatility scale but cannot distinguish one σ -driven cost from two unrelated ones that both scale with σ , so we report the scaling as consistent-with rather than confirmatory.

- (v) *Maker-fee level is venue-specific.* The null is shown at the realistic CME 0.015 bp, no-rebate level; rebate venues (some crypto/equity) would shift the maker economics, which we report rather than assume.
- (vi) *Avellaneda–Stoikov generalisation, and the reactive-quoter scope.* The result is that A–S-style *quoting* fails in price-time-priority tick-bound futures net of cost; it is not a claim that A–S is wrong in its native quote-driven (OTC) setting. More precisely, the tested family is *reactive, snapshot-lived* quoters (cancel-on-replace every snapshot, Section 4): each shape re-skews against current inventory at each snapshot but does not work queue position or run an inter-temporal inventory-control policy across snapshots. The null therefore bears directly on reactive inventory-skew quoting; a fully dynamic A–S controller that holds and works queue position over time is outside the tested family. We expect the decomposition’s cancellation to bind on it too (a dynamic quote still captures spread it pays back as adverse drift), but that is an expectation from the mechanism, not a result we test here.
- (vii) *Shared engine, not independent replication.* The CME analyses share one simulator, cost module, and significance machinery, so their agreement across the quoting family, the decomposition, and the OFI test is convergent evidence under a common pipeline, not independent replication.
- (viii) *The portfolio forward pointer is separate work.* The cross-chain DEX-LP variance reduction we point to in Section 10 is a different market, instrument, and year, and a single-regime variance reduction rather than a return premium; it is developed and scoped elsewhere, not established by this paper.

We read the result as a strong, large-scale true-sign measurement of the Glosten–Milgrom / Moallemi–Yuan prediction in tick-bound futures: in the calm regime adverse selection consumes essentially the whole captured half-spread on every root and net maker P&L is bounded above by zero, with the quoting and OFI nulls following as its corollaries. The cancellation is regime-dependent, and we establish it on a single matched feed after first clearing a confound: the high-resolution and coarser-resolution reconstructions read 99.5% vs 82.1% on identical Jul-2023 sessions (17.4pp from the reconstruction alone), which is why an earlier draft’s cross-reconstruction “regime” reading could not stand and was withdrawn. Resolving it by holding the reconstruction fixed on the nanosecond-resolution TAQ feed (which spans all regimes and matches the depth-10 decomposition to 0.3pp where both exist), the calm $\approx 100\%$ cancellation rises to 105–111% in the SVB, FOMC and COVID episodes and net-after-fee deepens with volatility ($-0.022 \rightarrow -0.139$ bp), broad-based across roots as a high- σ snapshot. A continuous 2016–2026 sweep (33,956 root-days, 96% of the universe) then resolves the dose-response: the within-root σ -gradient of the loss is root-idiosyncratic (natural gas, copper, crude and FX), and a naive pooled σ -curve is composition-confounded. So adverse selection consumes the captured spread in calm markets and over-consumes it in stress; the maker’s loss is not a knife-edge but widens with σ , broadly as a cross-sectional snapshot, with a volatility-sensitivity concentrated in a few contracts. The one open caveat is that the TAQ-vs-depth commensurability is verified on the calm sessions used for the headline; a depth-10 reconstruction in a stress regime would convert the matched-feed evidence into a depth-verified identity. The corollary nulls are robust across reconstructions and episodes.

A LVR Cross-Root Volatility Ordering

The $\sigma^2/8$ LVR equivalence is illustrative (Section 5), not a test. Table 8 gives the full robustness breakdown of the cross-root volatility ordering between the implied $\sigma^2/8$ LVR and the measured per-fill adverse drift; the body paragraph summarises it. The raw Pearson $r = 0.92$ is tail-sensitive (drop-NG, drop high-vol tail); what is stable is that both quantities track σ_τ and the rank association survives drop-NG. This is consistent-with one σ -driven cost but cannot distinguish that from two unrelated costs that both scale with σ ; the partial-correlation figures are descriptive only.

Table 8: Cross-root volatility ordering between the implied $\sigma^2/8$ LVR and the measured per-fill adverse drift (13 roots, 10s, bp; from `lvr_robustness.json`), an illustrative consistency check. Raw $r = 0.92$ is tail-sensitive; the rank association and the shared σ_τ scaling are what survive.

statistic	value	reading
Pearson r , full 13 roots	0.92	raw cross-root association
Pearson r , drop NG	0.57	tail-sensitive to the high-vol extreme
Pearson r , drop tail (NG, SI, CL)	0.39	raw r is substantially tail-driven
Spearman ρ , full / drop-NG	0.64 / 0.54	rank association survives the drop
bootstrap 95% CI on r	[0.33, 0.99]	wide, consistent with tail-dependence
adverse, LVR each vs σ_τ	0.88 / 0.97	both quantities track volatility
partial r (adv, LVR σ_τ), full	0.56	co-movement before controlling vol (descriptive)
drop-NG / drop-tail	0.03 / 0.13	little co-movement beyond volatility (descriptive)
<i>Reading:</i> volatility ordering robust on the rank statistics; raw $r = 0.92$ tail-sensitive. Consistent-with one σ -driven cost, not a test that distinguishes one from two; the equivalence itself is the known role-level result.		

B Figure Provenance

For navigation and reproduction, Table 9 maps each figure in the paper to the committed result artifact (under runs/) from which its numbers are rendered. Those artifacts are plain JSON/CSV; no figure depends on network access or a metered API.

Table 9: Figure-to-source provenance. Each figure is rendered from the cited committed result artifact under runs/.

figure	source artifact (committed under runs/)
2 overview	runs/mm_full/*.json (oos_pooled: n_fills, mean_realised_bp)
1 spine diagram	mechanism (thesis); numbers from runs/maker_decomp/
3 P&L waterfall	runs/maker_decomp/decomp_summary.json (+ by_root)
8 feed confound	runs/maker_decomp/regime_comparison_multi.json (+ feed_control_compare.json)
7 cross-regime σ -gradient	continuous/frozen3_results.json (per_root_slope, channel_per_root); rendered by frozen_analysis_fast.py + make_xregime_fig.py
4 tick frontier	runs/maker_decomp/tick_frontier.csv (+ decomp_by_root.csv)
6 conditional sweep	runs/maker_decomp/decomp_by_regime_summary.json (+ decomp_by_regime.csv)
5 queue economics	runs/mm_full/*.json (M4_fill_calibration; M7_queue_rank_economics)
9 OFI cost frontier	runs/xroot_ofi/verdict_xroot_ofi.json (cost_frontier, per_root_cross_benefit)
10 M8 / H4	runs/mm_full/*.json (M8_true_vs_inferred_sign); runs/h4_regime/verdict.json
11 latency cliff	runs/mm_full/*.json (M9_latency_cliff)

C Results Inventory

Table 10 lists the analyses and their one-line verdicts, so the argument can be read at a glance: the decomposition, its two corollary nulls, and the supporting measurements.

Table 10: The analyses and their verdicts: the maker-P&L decomposition (central), its two corollary nulls, and the supporting measurements.

analysis	verdict
decomposition (central)	captured +0.654bp consumed by adverse −0.661bp (ratio 101.1%, indistinguishable from 100%); net-after-fee −0.022bp, all roots ± 0.13 bp; net bounded above by 0 under any realistic friction (calm 2023 regime)
cross-regime (matched nanosecond TAQ feed)	regime-dependent: calm $\approx 100\%$ rises to 105.1% (SVB), 111.5% (FOMC), 109.8% (COVID), net deepening $-0.022 \rightarrow -0.139$ bp with σ_τ ; broad as a high- σ snapshot (9–11/13 roots net<0), but a continuous 2016–2026 sweep (33,956 root-days, 96%) shows the within-root σ -gradient is root-idiosyncratic (energy/copper/crude/FX) and the pooled σ -curve composition-confounded; per-episode root-day bootstrap CI excludes 100% (ratio) and 0 (net). TAQ $\equiv$ depth to 0.3pp on matched calm dates (99.6% vs 99.3%); a coarser-resolution reconstruction’s flat 79–88% was a 17.4pp reconstruction artifact (earlier “regime” reading withdrawn). Commensurability verified on calm sessions
corollary nulls (June-2022 coarser-resolution)	robust across reconstructions/episodes: quoting grid PBO = 0.500, 0/325; cross-root OFI incr OOS $R^2 = -0.006$, −1.43bp after cost
LVR equivalence (illustrative)	adverse drift is the known role-level equivalent of LVR $\sigma^2/8$; cross-root data consistent with one shared σ scale (consistent-with, cannot distinguish one σ -driven cost from two)
tick-size frontier	spread capture survives the 0.015bp fee only above a ≈ 0.74 bp tick width (ZF, CL, ZN, SI, ZB)
conditional sweep	high-vol session-open most adverse (−0.056bp); 5 thin exploratory cells, no multiplicity-corrected pass
quoting null (corollary 1)	inventory-skew family 0 significant at uncorrected 5%; PBO 0.543; 1,126-cell nominal grid, 5 shapes; BH/BHY 0/1126 a consistency note (below the $M=1000$ permutation floor)
cross-root OFI (corollary 2)	own $R^2 = 0.545$; cross incr OOS $R^2 = -0.005$; 0/13; +0.091 $\rightarrow -1.30$ bp; null
M8 sign accuracy	tick 0.61–0.83, quote 0.92–0.99 over 12.4M trades; tick rule unreliable
H4 regime CVaR	10/10 evaluable roots reduce hi-vol inventory CVaR (sign-test $p = 0.002$); 4 PASS the full bar; median −2.19 contracts (run under Phase-1 costs; statistic is contract-space CVaR, cost-independent)

D Reproducibility

This section gathers everything needed to reproduce the study; no reproducibility details are stated elsewhere in the paper.

Code. The simulator, the corrected per-fill cost/ledger model, the markout and decomposition engines, and the cross-root OFI feature/model scripts are released at `https://github.com/DaruFinance/quant-mm-simulator` (Python) with a Rust port at `quant-mm-simulator-rs` and the trade-log v1 schema (one Leg row per fill) in `quant-research-framework`. MIT-licensed. The paper’s figures are rendered from the committed result artifacts under `runs/` (Appendix B). The replay CLI is `python -m mmsim.cli.replay -snapshots ... -trades ... -horizon 10s -json`.

Data availability and what the public repo reproduces. We state the scope plainly so a reader knows what runs from the repository alone and what does not. The CME true-signed tapes that carry every headline number are *not* redistributable: they are CME-originated data obtained from a third-party provider under its terms, and the engines read them from a local data directory with provider credentials taken from environment variables. They are not a public product and cannot be regenerated from raw using the public repository alone. The repository therefore ships two things in lieu of the raw CME data. First, the engine’s correctness guarantees, the cross-language fill/markout parity and the no-lookahead (pollution) tests, run on *bundled* fixtures (a small equity-TAQ fixture in a flat-file TAQ format, ticker IBM, and a 60-minute BTC depth book/tape), so a reader can verify the simulator, the markout kernels, and the causality contract end-to-end without the licensed CME feed. Second, the CME headline numbers (the decomposition, the quoting null, the cross-root OFI frontier, the significance battery) are reproduced from the committed per-run result JSONs under `runs/`, which the figure code renders directly; given the licensed CME data, the same engines rebuild those JSONs from scratch in sequence. A reader should accordingly not expect to regenerate the CME pipeline from raw data using the public repository alone; the bundled fixtures exercise the code paths, and the committed JSONs carry the reported CME figures. The additional regime episodes (2021 calm, 2022 FOMC, 2023 SVB, 2020 COVID) and the same-dates reconstruction control are committed as their decomposition-result JSONs under `runs/`. The decomposition is sensitive to the time resolution of the book reconstruction: a coarser-resolution reconstruction of the same sessions yields a 17.4pp lower adverse-to-captured ratio on identical dates (Section 5, Figure 8), which is why all cross-regime measurement holds the reconstruction fixed on the nanosecond-resolution TAQ feed and cross-reconstruction comparisons are not made. That feed is read by the depth-1 ingester `mmsim.ingest.fut_taq` (parity test `tests/test_fut_taq.py` on a bundled synthetic TAQ fixture, no licensed data); the per-root-day decompositions for the three stress episodes (March-2020 COVID, June-2022 FOMC, March-2023 SVB; 195 root-days) and the 52-root-day TAQ-vs-depth commensurability control are committed under `runs/`, from which Table 5 and Figure 8 render directly. The continuous 2016–2026 sweep behind Section 6 and Figure 7 runs the *same* depth-1 TAQ ingester over every available trading day. Because the licensed TAQ tapes are not redistributable, the sweep ships as its *aggregate* per-root-day decompositions only: one row per root-day (captured spread, adverse drift, net-after-fee, realised σ_τ , fill count), carrying no underlying tape and therefore leak-safe to commit, frozen as a 33,956 root-day snapshot (96% of the available universe) in `continuous/froze`

`n3_results.json` and `continuous/frozen3_robustness.json`. The within-root fixed-effect slopes, the per-root capture/adverse channel decomposition, the composition-confound diagnostics, and the cluster-bootstrap CIs reproduce deterministically (seed 7) via `frozen_analysis_fast.py`; the horizon-invariance, per-driver day-level significance, and time-stability robustness via `robustness_frozen3.py`; and Figure 7 renders from those JSONs via `make_xregime_fig.py`. Given the licensed TAQ feed the same ingester rebuilds the aggregate rows from raw.

Determinism. The numba fast path reproduces the pure-Python reference fill sequence to a 10^{-10} price / 10^{-12} size rounding tolerance (`scripts/verify_fast_parity.py`, `tests/test_fast_sim.py`); the markout and permutation hot loops are *bit-identical* (array-equal) to their scalar references (`tests/test_markout.py`). Causality is structural: a fill consults no event past its own timestamp. RLIMIT_AS caps guard RAM on the 9M-trade days. Headline numbers reproduce via `scripts/run_mm_full.py` then `methodology_retrofit.py` (quoting null), `scripts/maker_pnl_decomp.py` (decomposition and tick frontier), `scripts/lvr_identity.py` (the illustrative LVR/adverse-drift one-axis comparison), `scripts/xroot_ofi_{features,model}.py` (OFI, pre-registration-locked), `scripts/maker_decomp_conditional.py` (the conditional root $\times$ session $\times$ regime sweep), and `scripts/stress_bootstrap_ci.py` (the per-episode root-day cluster bootstrap CIs of Section 5, which first reproduces the calm headline CI as a parity check, $\rightarrow$ `runs/maker_decomp_xregime/stress_bootstrap_ci.json`).

Engine performance. The engine speedup is recorded to a saved timing artifact (`runs/engine_timing.json`, `engine_timing_fullscale.json`): the numba fast path matches the pure-Python reference fill-for-fill and runs $\approx 5\times$ faster on the bundled 60-min BTC fixture (535k events, 1.78s vs 0.35s) and $\approx 83\times$ faster on a full CME root-day (1.25M events, 327.8s vs 3.97s, single-thread). The speedup grows with event count, and the pure-Python reference exhausts memory on the largest multi-million-event root-days (it cannot complete a ~ 4 M-event day in the box’s RAM), which is the operational reason the kernel exists.

Acknowledgements. The author thanks AlgoSeek for supporting this research.

References

- [1] Avellaneda, M., & Stoikov, S. (2008). High-Frequency Trading in a Limit Order Book. *Quantitative Finance* 8(3), 217–224.
- [2] Glosten, L. R., & Milgrom, P. R. (1985). Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. *Journal of Financial Economics* 14(1), 71–100.
- [3] Cont, R., Kukanov, A., & Stoikov, S. (2014). The Price Impact of Order Book Events. *Journal of Financial Econometrics* 12(1), 47–88.
- [4] Cont, R., Cucuringu, M., & Zhang, C. (2023). Cross-Impact of Order Flow Imbalance in Equity Markets. *Quantitative Finance* 23(10), 1373–1393.

- [5] Office of Financial Research (2019). Cross-Asset Market Order Flow, Liquidity, and Price Discovery. OFR Working Paper 19-04.
- [6] Moallemi, C. C., & Yuan, K. (2016). A Model for Queue Position Valuation in a Limit Order Book. Working Paper, Columbia University.
- [7] Milionis, J., Moallemi, C. C., Roughgarden, T., & Zhang, A. L. (2022). Automated Market Making and Loss-Versus-Rebalancing. Working Paper, Columbia University.
- [8] Guéant, O., Lehalle, C.-A., & Fernandez-Tapia, J. (2013). Dealing with the Inventory Risk: A Solution to the Market Making Problem. *Mathematics and Financial Economics* 7(4), 477–507.
- [9] Cartea, Á., Jaimungal, S., & Penalva, J. (2015). *Algorithmic and High-Frequency Trading*. Cambridge University Press.
- [10] Chakrabarty, B., Pascual, R., & Shkilko, A. (2015). Evaluating Trade Classification Algorithms: Bulk Volume Classification versus the Tick Rule and the Lee-Ready Algorithm. *Journal of Financial Markets* 25, 52–79.
- [11] Easley, D., López de Prado, M., & O’Hara, M. (2016). Discerning Information from Trade Data. *Journal of Financial Economics* 120(2), 269–285.
- [12] 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(5), 94–107.
- [13] Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *Annals of Statistics* 29(4), 1165–1188.
- [14] Lo, A. W. (2002). The Statistics of Sharpe Ratios. *Financial Analysts Journal* 58(4), 36–52.
- [15] Gatto, D. V. (2026). Predictive Value of Within-Strategy Permutation Tests for Forward Selection: Evidence from Over 6 Billion Strategy-Level Permutations Across Three Asset Classes. SSRN Working Paper 6636018, March 2026. <https://ssrn.com/abstract=6636018>, doi:10.2139/ssrn.6636018.