humarb Research 05

Forecasting Cross-Venue Funding Cash Flows: A Walk-Forward, Cost-Adjusted Benchmark

Hang Liu · humarb Research, London, United Kingdom · https://humarb.ai · ORCID 0009-0008-1059-3926

Preprint · 3 November 2026 · Not peer reviewed · DOI pending

Abstract

Cross-venue perpetual-funding comparisons require forecasts of holding-period cash flows, not annualizations of the latest quote. We construct a public-data benchmark covering 40 crypto assets, five venues, and 800 directed venue pairs over 90 days. Six causal estimators are evaluated at 24-, 72-, and 168-hour horizons in a frozen 30-day holdout containing 72,800 pair-origins per horizon. A one-day exponentially weighted estimator reduces MAE from \$1.762 to \$1.574 at 24 hours relative to a seven-day median (Holm-adjusted $p = 0.0035$); improvements at longer horizons are not uniformly significant after multiplicity control. The conservative endpoint's coverage falls from 80.13% to 75.86% as the horizon rises from 24 to 168 hours. Online-scaled central intervals approach nominal 90% coverage but retain days below 80%. Among positively forecast routes, median realized funding is \$0.17–\$2.79, far below a \$50 illustrative round-trip cost. The results support horizon-specific forecast evaluation and show why funding accuracy cannot be equated with executable net carry.

Keywords: perpetual futures; funding rates; forecast evaluation; walk-forward validation; conformal calibration; transaction costs

JEL classification: G12; G13; G14; C58

Corresponding author: Hang Liu, humarb Research, London, United Kingdom. Research channel: https://humarb.ai/research/.

Series: humarb Research 05. Methodological and empirical preprint. License: CC BY 4.0.

1 Introduction

Perpetual futures do not expire. Their link to the underlying market is maintained, in part, by periodic transfers between long and short positions. When two venues quote different funding rates for the same asset, the difference appears to offer a simple cash-flow forecast: hold the leg that receives funding, pay funding on the other, and repeat. The apparent simplicity is misleading. A displayed rate is normally a per-period quantity rather than an hourly one; the next rate can differ from the last; settlement grids can be asynchronous; and the cash that remains after entry, exit, transfer, and price-basis effects need not resemble the quoted spread. A research design that begins with an annualized quote therefore starts one step removed from both the statistical object and the economic object of interest.

This paper studies the statistical object directly: the cumulative net funding cash flow of a directed, two-leg perpetual position over a fixed horizon. For a short leg S and a long leg L, each with \$10,000 of notional, the target is the funding received on S minus the funding paid on L between a decision time t and $t+H$. Positive rates mean that longs pay shorts. The forecasting problem is causal by construction. Only settlements available at t may inform an estimate, and the evaluation window begins after the decision-time settlement boundary. This design separates a forecast of funding cash flow from a forecast of total trading profit. The latter additionally requires historical executable prices, depth, transfer paths, margin terms, and basis changes, which are not available in the public funding histories used here.

The distinction is economically important. humarb Research 01 defines executable net carry in terms of position-size-specific costs and depth walks, while humarb Research 02 provides a settled-history, depth-walked, and liquidation-aware accounting framework (Liu, 2026a,b). Neither paper claims to validate a forecasting estimator out of sample. The present paper takes the next step: it freezes a public universe and an evaluation protocol, compares candidate funding estimators on later observations, and then places the resulting forecasts under explicit cost scenarios. The earlier accounting identity is retained, but no accounting identity is treated as evidence of predictive accuracy.

1.1 Research question and main findings

The research question is narrow and testable: *how accurately do simple causal estimators predict realized cross-venue funding cash flow over 24, 72, and 168 hours, and how much economic content remains after stated costs?* The sample comprises 40 crypto assets observed on Hyperliquid, Extended, Pacifica, Lighter, and GRVT. Every asset gives 20 directed venue pairs, hence 800 pairs in total. The fixed data window runs from 2026-07-23 00:00Z through 2026-10-21 00:00Z. Following a 30-day initial lookback, the development segment ends on 2026-09-14; the following 30 days of decision origins form the untouched holdout. Each horizon has 72,800 holdout pair-origins. These observations are numerous but not independent: they overlap through time and share assets and venues. All inference therefore distinguishes row counts from the much smaller number of temporal blocks.

Four findings emerge. First, responsiveness helps, but its value depends on the horizon. The one-day exponentially weighted moving average (EWMA) attains MAE of \$1.574 at 24 hours, compared with \$1.762 for the seven-day median. The Diebold–Mariano comparison remains significant after a Holm correction across all 15 method–horizon tests ($p = 0.0035$). At 72 hours, EWMA with one- and three-day half-lives has lower MAE than the seven-day median, but their Holm-adjusted p-values are both 0.0959. At 168 hours, the one-day EWMA does not improve MAE relative to the seven-day median, while the 14-day median is lower descriptively but not significant after adjustment. It would therefore be incorrect to replace one universal rule with another.

Second, the sign of the funding differential is more predictable than its magnitude. One-day EWMA

achieves sign accuracy of 69.60%, 75.24%, and 72.93% at the three horizons, compared with 65.68%, 63.83%, and 63.55% for the seven-day median. Yet absolute relative errors remain large and unstable near small forecasts. This is the familiar distinction between a useful ranking signal and a reliable cash-flow level. A ranking can be directionally informative while remaining unsuitable for quoting a precise expected receipt.

Third, conservative-endpoint coverage deteriorates with horizon. The three-estimator lower endpoint covers 80.13% of realized outcomes at 24 hours, 76.87% at 72 hours, and 75.86% at 168 hours. Day-block 95% intervals are [78.75%, 81.42%], [75.49%, 78.36%], and [74.38%, 77.35%]. The 24-hour result does not establish coverage above 80%, and the longer-horizon results fall below it. Central 90% intervals based on development residuals also under-cover statically, especially at 168 hours. A frozen online scaling rule brings overall coverage close to 90%, with slightly better interval scores, but the worst daily coverage remains 77.42%, 76.83%, and 79.50%. Average calibration is not uniform calibration.

Fourth, forecast accuracy and economic materiality are different questions. Among routes with a positive seven-day-median forecast, median realized funding is only \$0.169, \$0.980, and \$2.790 per \$10,000 per leg. An illustrative base case of \$50 for four executions across two legs plus transfer therefore overwhelms the median cash flow. The share with positive funding after this cost is 0.00%, 0.06%, and 1.02%. Even the low-cost case of \$20 leaves median net funding negative. These figures do not estimate a strategy's return; they show that a statistically meaningful forecasting comparison can coexist with economically insufficient median cash flow.

1.2 Contributions

The paper makes four contributions. First, it supplies a deterministic public-data benchmark with explicit venue-specific units, per-record settlement intervals, a 60-second decision boundary, complete-UTC-day medians, and 30 independently reconciled origins. The universe covers all 40 exact-name crypto assets shared by the five venue lists under the stated listing rules, rather than a BTC-only illustration. The code preserves raw responses, checksums, failure logs, derived observations, and table-generating scripts.

Second, it evaluates point forecasts, directional forecasts, one-sided conservative endpoints, and central intervals within one frozen protocol. The evaluation includes MAE, RMSE, sign accuracy, relative-error diagnostics, Wilson and day-block coverage intervals, interval scores, Kupiec and Christoffersen diagnostics, Diebold–Mariano tests with Newey–West errors, Holm multiplicity adjustment, and CSCV/PBO ranking-stability checks. The analysis distinguishes descriptive row-level precision from temporal evidence and reports where the latter is weak.

Third, it connects forecast evaluation to cost-adjusted economic quantities without claiming unobserved execution. Costs are scenario assumptions rather than fabricated venue quotes. More importantly, historical mid-spread P&L is not silently set to zero. Funding-after-cost results are labeled as a partial accounting object, and the adverse-basis break-even is left visible. This is the principal difference between a funding benchmark and an investable-performance claim.

Fourth, the study supplies a reproducible bridge between methodology and evaluation. It retains the measurement discipline of humarb Research 01 and 02, but replaces their absence of out-of-sample validation with a defined development/holdout split and a public replication path. The objective is not to certify a production estimator. It is to establish which claims survive later data, which are horizon-dependent, and which require additional evidence before they can be made.

Section 2 places the study in the perpetual-pricing, arbitrage, execution, and forecast-evaluation literatures. Section 3 summarizes institutional mechanics. Section 4 defines the data and sample. Section

5 gives estimators and inference. Section 6 reports results. Section 7 examines robustness. Section 8 states limitations, and Section 9 concludes.

2 Related literature

2.1 Perpetual claims and funding mechanics

The economic design of cash-settled perpetual claims has a longer history than cryptocurrency markets. Shiller (1993) proposes perpetual claims based on repeated-measures indices, establishing the conceptual role of a contract without a fixed expiry. Modern crypto perpetuals differ in leverage, collateral, and market organization, but share the need for a mechanism that prevents an unlimited disconnect from a reference price. Funding is one such mechanism, not a dividend and not, by itself, a risk-free payment.

Recent pricing work makes the feedback between premium and funding explicit. Ackerer et al. (2025) derives perpetual-futures prices under constant funding parameters and interest rates. He et al. (2022) formalizes the relation among the perpetual premium, funding intensity, and no-arbitrage restrictions. These models explain why a funding quote is endogenous to the contract price: a high rate can coexist with a price adjustment that offsets the apparent carry. This paper does not estimate a structural pricing model. Instead, it holds the accounting definition fixed and asks how well reduced-form historical estimators anticipate subsequent settlements. The two approaches are complementary. A structural model disciplines what should be priced; a walk-forward benchmark disciplines what a simple estimator actually delivered.

2.2 Predictability and cross-venue carry

The closest forecasting study is Inan (2026), who reports out-of-sample predictability of next-period funding rates for Bitcoin contracts on Binance and Bybit using double-autoregressive models. The evidence supports predictability of one-step funding levels while indicating time-varying local stability. Our target differs in three respects. We forecast cumulative two-venue cash flow rather than a single venue's next rate; we evaluate 24-, 72-, and 168-hour cumulative outcomes rather than only the next period; and we compare 40 assets across five venues under a common frozen holdout. The simpler estimators in this paper are intentional benchmarks: they reveal how much of the practical forecasting problem is solved by recency and smoothing before a more elaborate model is introduced.

Funding information can also be relevant for other moments. Kim (2026) finds that perpetual-futures information, including funding-related variables, improves forecasts of Bitcoin spot volatility. That result concerns volatility, not cumulative carry, and does not imply that funding levels are easy to forecast. Our results are consistent with the distinction: sign predictability is stronger than level accuracy, and coverage remains time-varying even when point forecasts improve.

The cross-sectional motivation comes from crypto carry and segmented markets. Schmeling et al. (2026) documents the economics of crypto carry, while Makarov and Schoar (2020) shows how arbitrage and segmentation shape cryptocurrency prices across venues. Shleifer and Vishny (1997) and Gromb and Vayanos (2010) provide the broader limits-to-arbitrage framework: capital, intermediation, and risk constraints can allow an apparent discrepancy to persist. A persistent funding spread is therefore not necessarily an unexploited opportunity. It may compensate for execution, margin, transfer, or basis risks, or simply fail to persist long enough to cover them.

2.3 Execution and implementation shortfall

Execution research explains why a mid-quote estimate is not a realized return. Perold (1988) formalizes implementation shortfall as the gap between a decision price and realized execution. Almgren and Chriss (2001) models the trade-off between execution cost and timing risk. In a two-leg position, these frictions occur at entry and exit on both legs. The notional at which a funding forecast is quoted consequently matters: percentage fees scale with size, depth-dependent impact can be nonlinear, and transfer costs may have fixed components.

This paper uses cost scenarios rather than claiming to reconstruct historical order books. That choice follows the evidence, not convenience. Public funding histories identify rates and times but not executable bids, asks, queue position, or available depth at every decision and exit. Scenario costs can answer a bounded question—how large is the gap between funding cash and a stated friction?—but cannot identify capacity or realized slippage. humarb Research 01 and 02 address those measurement problems directly (Liu, 2026a,b); the present paper tests the forecasting component and preserves their separation between measured and unmeasured inputs.

2.4 Forecast comparison, interval evaluation, and selection

Forecast evaluation supplies the inferential framework. Diebold and Mariano (1995) introduces the standard predictive-accuracy comparison, and Harvey et al. (1997) provides a small-sample correction for equal mean squared prediction errors. Because our outcomes overlap, we aggregate absolute-error loss differences by UTC day and use Newey–West standard errors (Newey and West, 1987). Multiple candidate–horizon comparisons require a familywise correction; the paper reports Holm-adjusted p-values alongside raw tests. CSCV/PBO is used only as a secondary ranking-stability diagnostic (Bailey et al., 2016). A low PBO means that an in-sample winner tends to retain a favorable rank; it does not imply positive cash flow, just as a high Sharpe-related selection correction would not repair a weak target definition (Bailey and López de Prado, 2014; Smith and Winkler, 2006).

Coverage evaluation addresses a different failure mode. Wilson intervals summarize binomial endpoint uncertainty under independence (Wilson, 1927), but the assumption is not sufficient for overlapping pair-origins. We therefore make day-block resampling primary and treat Wilson, Kupiec, and Christoffersen tests as secondary diagnostics (Kupiec, 1995; Christoffersen, 1998). Central intervals are assessed with proper interval scores (Gneiting and Raftery, 2007), so that higher coverage obtained through excessively wide bands is not rewarded automatically. Online scaling is related in spirit to adaptive conformal inference under distribution shift (Gibbs and Candès, 2021); our implementation is deliberately simple and is not presented as providing coverage at every date or for every venue pair. Block-resampling ideas follow the broader dependent-data bootstrap literature (Politis and Romano, 1994).

The resulting literature position is practical. Perpetual-pricing theory explains why funding and basis cannot be separated conceptually; forecasting research demonstrates the need for genuine out-of-sample tests; execution research explains why accurate funding need not create executable carry; and calibration research shows why nominal coverage must be earned on later data. The benchmark integrates these requirements without treating any one of them as a substitute for the others.

3 Institutional background

3.1 Funding as a sequence of cash transfers

Let $f_{i,k}$ denote the settled funding rate of leg i at time $t_{i,k}$, expressed as a decimal fraction of position notional. Under the sign convention used throughout, a positive rate transfers cash from longs to shorts. If leg L is long and leg S is short, the signed funding cash flow over $(t, t+H]$ is

$$F(t,H) = A\left(\sum_{k:\, t<t_{S,k}\le t+H} f_{S,k} - \sum_{k:\, t<t_{L,k}\le t+H} f_{L,k}\right), \tag{1}$$

where A is notional per leg. The position is delta-neutral only to a first approximation and only if contract multipliers, currencies, and quantities are matched appropriately. Equation (1) is not total P&L. It excludes price changes, fees, spread, impact, financing, transfer costs, and liquidation effects by design.

Historical records are per settlement, not necessarily per hour. If leg i has interval $I_{i,k}$ hours at record k, the comparable hourly rate is

$$r_{i,k} = \frac{f_{i,k}}{I_{i,k}}. \tag{2}$$

Using a common interval such as eight hours for every venue changes both the forecast and its units. GRVT records include an instrument interval, including four- and eight-hour cases in the present universe. Hyperliquid, Extended, Pacifica, and Lighter records are hourly in this sample. Lighter reports percentage points and a payment direction, so its signed decimal rate is $+rate/100$ for 'long' and $-rate/100$ for 'short'. GRVT also reports percentage points, requiring division by 100. These conversions are applied before any estimator is computed.

3.2 Premium, interest, caps, and dead zones

Venue funding formulas generally combine a premium component with an interest-rate component, then apply clamps or caps. Hyperliquid's official documentation describes an eight-hour formula with hourly payment at one-eighth of the computed rate, premium sampling, an interest component, and a cap (Hyperliquid, 2026). Extended states that funding is calculated more frequently but applied once per hour and that historical records contain rates used for payments (Extended, 2026). Pacifica exposes hourly historical funding records (Pacifica, 2026), while Lighter documents its funding mechanism and direction field (Lighter, 2026). GRVT's instrument and funding endpoints expose funding intervals and historical rates (GRVT, 2026). The paper uses these documents for mechanics and API interpretation, not as evidence of forecast accuracy.

Caps and dead zones create predictable short-run regularities. When the premium remains inside a band, a venue can settle at an unchanged baseline rate. This produces many equal cross-venue differences and exact zeros in the sample. A forecast evaluation must not discard those zeros: they are economically real observations. At the same time, relative error becomes undefined or dominated by tiny denominators. We therefore report MAE and RMSE for every observation, restrict relative-error summaries to absolute forecasts above \$0.01, and analyze sign separately.

3.3 Decision-time and settlement boundaries

A funding record timestamped at or immediately after t is not information available at t. Public APIs also differ in timestamp precision: some records lie a few milliseconds or seconds after the nominal settlement

boundary. To prevent a same-settlement leak, estimators use only records with timestamp no later than $t - 60$ seconds, and realized outcomes include records with timestamps greater than $t + 60$ seconds and no greater than $t + H + 60$ seconds. This symmetric 60-second convention excludes the decision-time settlement while including the corresponding terminal settlement. It also prevents an hourly API's millisecond offset from producing $H + 1$ settlements in an H-hour window.

The convention is an implementation rule, not a claim about a trader's latency. A production system would require a stronger information-availability model, including publication delays and clock uncertainty. The benchmark's rule is sufficient to remove the mechanical lookahead in the public histories and is applied identically to every method.

3.4 From funding to total net income

Total income for a matched position can be written schematically as

$$\Pi = F - C^{\text{entry}} - C^{\text{exit}} - C^{\text{transfer}} - C^{\text{financing}} + \Pi^{\text{basis}} + \Pi^{\text{other}}, \tag{3}$$

where Π^{basis} is the signed P&L from cross-venue price-basis changes. No term in Equation (3) may be presumed zero merely because it is missing. Our public snapshot identifies F but not the other realized components. The empirical sections therefore use two explicit labels: *gross funding* for F, and *funding after cost* for $F - C$ under a stated scenario C. Neither label means total arbitrage P&L. This distinction is particularly important at funding discontinuities, when the leg receiving funding can simultaneously experience an adverse price adjustment.

4 Data

4.1 Sources and fixed snapshot

The study uses official, no-key public funding-history endpoints for Hyperliquid, Extended, Pacifica, Lighter, and GRVT. Instrument lists were retrieved on 2026-11-03, and funding records were then downloaded for the fixed historical window 2026-07-23 00:00Z–2026-10-21 00:00Z. The raw snapshot contains 361,140 records in 200 gzip JSONL files, one for each venue–asset pair. Every file has a SHA-256 digest in the acquisition manifest. The initial log recorded four Pacifica rate-limit failures; a later retry completed all four, and both the initial failure record and successful retry are preserved. No credentials, private endpoints, user data, or authenticated exports are used.

Table 1 summarizes coverage. The three fully hourly centralized record sets and Lighter contain 2,160 records per asset. Pacifica has 2,156 per asset, a small and documented shortfall. GRVT has 271 records for eight-hour instruments and 541 for four-hour instruments. Record counts therefore measure settlements, not directly comparable clock time. The empirical code converts each record to both a per-period cash-flow rate and an hourly estimator input, using the interval attached to that record.

The choice of five venues reflects reproducibility rather than convenience sampling. Other public endpoints in the original research inventory were tested, but network reachability, access restrictions, or incomplete history prevented a common 90-day reconstruction in the present environment. Two additional venues in that inventory had no public historical funding interface. The paper makes no claim that five venues represent the entire perpetual market. It claims that the selected five can be audited end to end under a common protocol.

Table 1: Public funding-history snapshot by venue

Venue	Assets	Records	Records/asset	Median interval (h)
hyperliquid	40	86,400	2,160–2,160	1
extended	40	86,400	2,160–2,160	1
pacifica	40	86,240	2,156–2,156	1
lighter	40	86,400	2,160–2,160	1
grvt	40	15,700	271–541	4

Note: Fixed window 2026-07-23 00:00Z–2026-10-21 00:00Z. The 40 assets are the exact-name crypto intersection of the five venue lists, with listing-date filters applied where the field has historical meaning. Source: official public venue APIs, retrieved 2026-11-03; authors' calculations.

4.2 Universe construction

The universe is the exact-name intersection of the five venue lists, restricted to crypto perpetuals. Hyperliquid markets flagged as delisted at retrieval are excluded. Extended, Pacifica, and Lighter listing timestamps must precede the window start where those fields have historical meaning. GRVT's current registry creation timestamps do not provide a comparable historical listing field and are not used for that filter. The resulting 40 assets are AAVE, ADA, ARB, ASTER, AVAX, BCH, BNB, BTC, CRV, DOGE, ENA, ETH, FARTCOIN, HYPE, ICP, JUP, LDO, LINK, LIT, LTC, MEGA, MON, NEAR, PAXG, PENGU, PUMP, SOL, STRK, SUI, TAO, TRUMP, UNI, VIRTUAL, WIF, WLD, WLFI, XMR, XPL, XRP, and ZEC.

This rule has two useful properties. It includes the full common-asset set rather than selecting assets because their outcomes are attractive, and it guarantees that every directed pair can be constructed on all five venues. It also creates a clear limitation: current-list intersection is not a point-in-time historical membership database. Assets that disappeared before retrieval are absent. The sample is therefore vulnerable to survivorship bias even though delisted flags available at retrieval are respected. A full historical security master is required to remove that limitation.

Each asset generates $5 \times 4 = 20$ ordered pairs (S, L). Both directions are retained. This symmetry avoids choosing a direction with the benefit of realized outcomes and ensures that forecast errors are evaluated on signed positive and negative differentials. For the separate economic analysis, a route is included only when its reference forecast is positive. That filter uses information available at the decision time and selects the economically relevant direction without inspecting realized funding.

4.3 Decision origins and temporal split

Decision origins occur at 00:00, 08:00, and 16:00 UTC each day. The first origin is 2026-08-22, after 30 days of initial history, and the last is 2026-10-14 00:00Z so that a 168-hour outcome ends by the fixed window boundary. The development segment runs from the first origin to 2026-09-14 00:00Z; subsequent origins form the frozen holdout. There are 165,600 development and 218,400 holdout observations across the three horizons. Every horizon–estimator cell contains 72,800 holdout observations. The holdout spans 30 days of decision-time coverage plus the terminal midnight origin.

The protocol was frozen before computing the new holdout results. Its JSON record specifies the universe, horizons, six estimators, conservative rule, interval update, cost cases, missing-data policy, and inferential hierarchy. No holdout result was used to add an estimator, change a half-life, select a cost case, or choose a favorable horizon. The SHA-256 digest of this record is reported in the replication summary.

4.4 Descriptive properties of the forecast target

The target is strongly non-Gaussian and contains economically meaningful zeros. Exact zero realized differentials account for 8.08% of 24-hour holdout rows, reflecting unchanged baseline rates on the two legs. The proportion falls to 0.42% at 72 hours and 0.43% at 168 hours as additional settlements allow small differences to accumulate. Median absolute realized funding is \$0.560, \$2.193, and \$5.604 at the three horizons, while the corresponding 90th percentiles are \$4.608, \$12.176, and \$24.929. Median absolute seven-day-median forecasts are smaller: \$0.401, \$1.202, and \$2.805.

This distribution has three implications for evaluation. First, squared and absolute errors can rank estimators differently because a small number of large rate reversals dominates squared loss. Second, relative error is not meaningful for the exact zeros and is fragile for the many small nonzero forecasts; the \$0.01 denominator rule prevents an arbitrary numerical stabilization from determining the ranking. Third, the full directed panel is mechanically symmetric. Its mean realized funding is approximately zero, so the mean cannot summarize economic attractiveness. The positive-reference subset used for the cost analysis is therefore a separate, decision-time-defined object.

Rank correlation is informative but modest. The Pearson correlation between realized and reference-forecast funding is 0.361 at 24 hours, 0.382 at 72 hours, and 0.401 at 168 hours. These correlations are descriptive and are not used as the main accuracy criterion because they conceal scale errors. An estimator can rank opportunities while systematically understating or overstating cash flow. The joint reporting of correlation, MAE, sign, and coverage avoids selecting whichever association appears most favorable.

4.5 Inclusion, missingness, and validation

A pair-origin is retained only when all six forecasts and the conservative endpoint are available and each leg has at least 90% expected settlement coverage in its outcome window. Median estimators use the most recent N complete UTC days within an $N + 1$-day lookback, where a day requires at least 90% expected coverage. This avoids treating the two partial boundary days of an eight-hour settlement grid as two full observations. No forward fill is used for realized cash flow.

Three validation layers supplement the snapshot. First, unit tests check sign conventions, timestamp boundaries, interval normalization, Wilson reference values, and deterministic output. Second, 30 holdout origins are drawn with seed 20261021 and independently recomputed in a separate implementation. Reference, conservative, and realized values all agree; the maximum absolute discrepancy is 1.42×10^{-14} USD. Third, every reported table is generated from the frozen result files, so the prose, PDF, and web page share one numerical source.

The final sample is large in rows but limited in temporal breadth. It contains one market window, 40 current-common assets, five venues, and 30 holdout days. Those constraints define the external validity of every result that follows.

5 Methods

5.1 Forecast target

For asset a, directed venue pair (S, L), origin t, and horizon $H \in \{24, 72, 168\}$ hours, the realized target is $y_{a,S,L,t,H} = F_{a,S,L}(t, H)$ from Equation (1), with $A = \$10{,}000$ per leg. A forecast is written $\hat{y}^{(m)}$ for

method m. All methods estimate each leg's hourly funding rate and then project it linearly:

$$\hat{y}^{(m)} = AH\left(\hat{r}_S^{(m)}(t) - \hat{r}_L^{(m)}(t)\right). \tag{4}$$

This is a deliberately simple counterfactual: the estimated hourly rate remains constant over the horizon. More complex dynamics are evaluated only through the choice of $\hat{r}$, not by adding unregistered models after seeing outcomes.

The six estimators are the latest hourly rate; medians of 3, 7, and 14 complete daily mean rates; and EWMAs with one- and three-day half-lives. For leg i,

$$\hat{r}_i^{\mathrm{EWMA}(\lambda)}(t) = \frac{\sum_{k\in\mathcal{W}_{14}(t)} w_{i,k} r_{i,k}}{\sum_{k\in\mathcal{W}_{14}(t)} w_{i,k}}, \qquad w_{i,k} = I_{i,k}\exp\{-\ln(2)\,\mathrm{age}_{i,k}/\lambda\}, \tag{5}$$

where $\mathcal{W}_{14}(t)$ is the trailing 14-day window and age is measured in days. The interval weight prevents an eight-hour record from receiving the same total weight as an hourly record merely because it appears once.

The seven-day median is the reference estimator, not a production recommendation. Its daily values are $d_{i,u} = \sum_{k\in u} f_{i,k}/24$ for complete UTC day u, and its forecast is $\mathrm{median}(d_{i,u})$ over the seven most recent complete days. The other medians change only N.

5.2 Conservative endpoint

The conservative endpoint is a one-sided cash-flow bound based on the last, seven-day, and 30-day estimates. For the short leg, the minimum candidate is least favorable because it reduces expected receipts; for the long leg, the maximum is least favorable because it increases expected payments:

$$\hat{r}_S^c = \min\{\hat{r}_S^{\mathrm{last}}, \hat{r}_S^{\mathrm{med7}}, \hat{r}_S^{\mathrm{med30}}\}, \tag{6}$$

$$\hat{r}_L^c = \max\{\hat{r}_L^{\mathrm{last}}, \hat{r}_L^{\mathrm{med7}}, \hat{r}_L^{\mathrm{med30}}\}, \tag{7}$$

$$\hat{y}^c = AH(\hat{r}_S^c - \hat{r}_L^c). \tag{8}$$

Its empirical hit rate is $P(y \geq \hat{y}^c)$. This is not called a confidence interval: the rule is deterministic, and its coverage must be measured on later observations. It is also unchanged across the six point-estimator comparisons, which isolates differences among reference forecasts.

5.3 Point, sign, and endpoint metrics

For each horizon and estimator, we compute

$$\mathrm{MAE} = \frac{1}{n}\sum_j |e_j|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_j e_j^2}, \qquad e_j = y_j - \hat{y}_j. \tag{9}$$

Relative absolute error is summarized by the median of $|e_j/\hat{y}_j|$ for $|\hat{y}_j| > \$0.01$; zeros and near-zeros are not assigned an arbitrary epsilon denominator. Sign accuracy is $n^{-1}\sum_j \mathbf{1}\{\mathrm{sign}(y_j) = \mathrm{sign}(\hat{y}_j)\}$, including exact-zero agreements. Reference-floor coverage is $P(y \geq \hat{y})$, while conservative coverage is $P(y \geq \hat{y}^c)$.

The naive Wilson 95% interval is reported for continuity with the original metric definition (Wilson, 1927). It is not the primary uncertainty statement because observations within a day share market conditions and consecutive horizons overlap. Primary coverage intervals resample UTC days with re-

placement 2,000 times, carrying all pair-origins in a selected day into the same replicate. This preserves cross-sectional dependence within a day but does not model dependence between adjacent days; the resulting intervals remain interpretable as day-block, not fully general, uncertainty estimates.

5.4 Diebold–Mariano tests and multiplicity

Each alternative is compared with the seven-day median using absolute-error loss differences $\Delta_{j,t} = |e^{\text{alt}}_{j,t}| - |e^{\text{med7}}_{j,t}|$. A daily series is formed by averaging Δ across all pair-origins on that day. The test statistic uses Newey–West variance with 2, 4, and 8 daily lags for 24-, 72-, and 168-hour horizons. These lags correspond to the registered horizon overlap plus one day. The 168-hour holdout has few non-overlapping weekly outcomes, so its tests are low-power even with a large row count.

Fifteen comparisons are made: five alternatives across three horizons. Holm-adjusted p-values control the familywise error rate over this family. A result is described as statistically supported only when its adjusted value is below 0.05. Raw p-values remain visible so that borderline evidence is not turned into a binary success narrative.

As a secondary check, CSCV/PBO is computed on holdout daily MAE with six estimators and $S = 8$ or 10 contiguous blocks. For every symmetric split, the estimator with the lowest in-sample MAE is ranked out of sample. PBO is the fraction of splits in which this winner ranks below the out-of-sample median. The statistic evaluates ranking stability, not positive returns or executable performance (Bailey et al., 2016).

5.5 Central interval calibration

Central 90% intervals begin with development residuals $e = y - \hat{y}^{\text{med7}}$. Let $q_{.05}$ and $q_{.95}$ be their empirical 5th and 95th percentiles, computed using the usual linear sample quantile convention (Hyndman and Fan, 1996). The static interval is $[\hat{y} + q_{.05}, \hat{y} + q_{.95}]$. A frozen online rule then scales both residual offsets by m_d:

$$\log m_{d+1} = \log m_d + 0.5(\text{miss}_d - 0.10), \qquad m_0 = 1, \tag{10}$$

with m_d clipped to $[0.25, 10]$ and updated once per UTC day after that day's outcomes are scored. The rule reacts to average miss frequency, never to a future day. It is related to online conformal calibration but is not assigned a finite-sample or per-window guarantee.

Intervals are evaluated by empirical coverage, worst-day coverage, mean width, and the proper interval score

$$S_\alpha(l, u; y) = (u - l) + \frac{2}{\alpha}(l - y)\mathbf{1}\{y < l\} + \frac{2}{\alpha}(y - u)\mathbf{1}\{y > u\}, \tag{11}$$

with $\alpha = 0.10$ (Gneiting and Raftery, 2007). Kupiec unconditional-coverage and Christoffersen independence tests are included as secondary diagnostics. Their independence assumptions are particularly fragile in this panel; day-block intervals and worst-day results are therefore given greater interpretive weight.

5.6 Cost scenarios and economic selection

Economic quantities use only routes with $\hat{y}^{\text{med7}} > 0$. This point-in-time filter selects the direction expected to receive net funding and leaves 33,005 routes per horizon. Cost is expressed per \$10,000 per leg. Each scenario charges a per-execution basis-point rate four times—entry and exit on two legs—and adds transfer cost:

Scenario	Per execution	Four-execution cost	Transfer	Total C
Low	5 bp	\$20	\$0	\$20
Base	10 bp	\$40	\$10	\$50
High	20 bp	\$80	\$25	\$105

These are illustrative assumptions, not venue-specific historical fee measurements. The per-execution rate bundles fees, half-spread, and impact for compact sensitivity analysis; it does not identify their individual sizes or a capacity curve. Funding after cost is $y - C$ for realized funding and $\hat{y} - C$ for forecast funding. Simple annualization multiplies by $8760/H$ without compounding. Historical mid-spread P&L remains unmeasured and is excluded from, rather than zeroed in, this calculation. The resulting quantity can be interpreted as the amount of adverse basis that realized gross funding less scenario cost could withstand before becoming negative.

6 Results

6.1 Recency improves short-horizon point forecasts

Table 2 reports the frozen-holdout results. At 24 hours, the one-day EWMA has the lowest MAE, \$1.574, and RMSE, \$2.753. The seven-day median records \$1.762 and \$3.188, respectively. The 14-day median also improves on the seven-day median descriptively, but less than the EWMA. The latest rate has the best median absolute relative error (0.858) but the worst MAE and RMSE, a warning that a metric conditioned on small denominators can rank a noisy estimator too favorably.

At 72 hours, the three-day EWMA has the lowest MAE (\$4.491) and RMSE (\$7.191), closely followed by the one-day EWMA (\$4.526 and \$7.555). The seven-day median's values are \$4.820 and \$8.159. At 168 hours, the 14-day median has the lowest MAE (\$9.928), while the three-day EWMA has the lowest RMSE among the responsive estimators (\$15.161). The seven-day median reaches \$10.632 and \$16.730. No estimator is uniformly best across both loss functions and all horizons.

Table 2: Frozen-holdout point-forecast performance

H (h)	Estimator	MAE (\$)	RMSE (\$)	Sign accuracy	$P(y \geq \hat{y})$
24	Last	2.035	4.200	66.82%	56.45%
24	Median 3d	1.813	3.247	68.35%	53.48%
24	Median 7d	1.762	3.188	65.68%	53.34%
24	Median 14d	1.718	3.102	63.27%	52.75%
24	EWMA 1d	1.574	2.753	69.60%	50.09%
24	EWMA 3d	1.663	2.819	66.81%	50.10%
72	Last	6.689	13.264	55.89%	50.64%
72	Median 3d	5.105	8.654	65.98%	50.31%
72	Median 7d	4.820	8.159	63.83%	50.34%
72	Median 14d	4.620	7.650	62.39%	50.25%
72	EWMA 1d	4.526	7.555	75.24%	50.00%
72	EWMA 3d	4.491	7.191	73.31%	50.00%
168	Last	16.313	31.937	52.87%	50.22%
168	Median 3d	11.828	19.170	64.77%	50.18%
168	Median 7d	10.632	16.730	63.55%	50.17%
168	Median 14d	9.928	15.149	61.94%	50.15%
168	EWMA 1d	10.716	17.045	72.93%	50.02%
168	EWMA 3d	10.116	15.161	72.97%	50.02%

Note: n = 72,800 in every horizon–estimator cell. USD amounts are per \$10,000 of notional per leg. MAE/RMSE use $y - \hat{y}$; sign accuracy compares exact signs, including zeros. Source: authors' calculations from the public snapshot.

Figure 1 makes the horizon effect visible. Longer windows have larger cash-flow errors in dollar terms, roughly in proportion to the amount of funding that can change before settlement. They also allow more time for a smoothing estimator to become stale. The appropriate conclusion is not that long windows are intrinsically harder in every normalized sense, but that a single short-run estimate cannot be extrapolated for a week without material forecast risk.

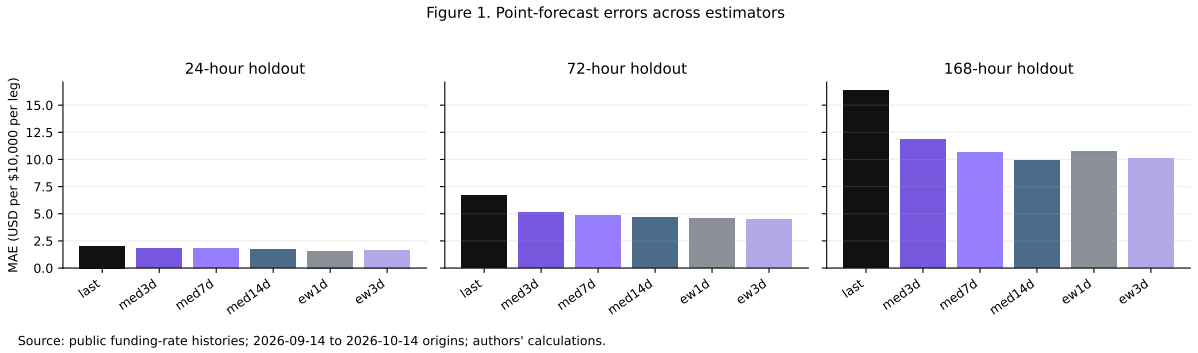


Figure 1: Point-forecast MAE by estimator and horizon in the frozen holdout. Amounts are USD per $10,000 notional per leg. Source: official public funding histories and authors' calculations.

The Diebold–Mariano results in Table 3 sharpen this interpretation. The one-day EWMA's daily absolute-loss advantage over the seven-day median is −0.190 at 24 hours, with Newey–West $t = -3.58$ and Holm-adjusted $p = 0.00346$. This is the clearest statistically supported improvement in the study. At 72 hours, one- and three-day EWMA advantages of −0.304 and −0.335 both have adjusted $p = 0.0959$, above the registered 5% threshold. At 168 hours, the 14-day median has raw $p = 0.0448$ but adjusted $p = 0.2689$; the one-day EWMA has adjusted $p = 0.7677$. These results do not support a universal replacement of the seven-day median.

Table 3: Diebold–Mariano comparisons against the seven-day median

H (h)	Alternative	Mean loss diff.	NW t	Raw p	Holm p
24	Last	+0.273	+4.20	2.66e-05	0.00032
24	Median 3d	+0.047	+1.06	0.289	0.579
24	Median 14d	-0.044	-1.83	0.0672	0.27
24	EWMA 1d	-0.190	-3.58	0.000346	0.00346
24	EWMA 3d	-0.100	-2.44	0.0147	0.103
72	Last	+1.858	+7.24	4.35e-13	6.1e-12
72	Median 3d	+0.271	+4.08	4.5e-05	0.000495
72	Median 14d	-0.202	-1.93	0.0539	0.27
72	EWMA 1d	-0.304	-2.55	0.0107	0.0959
72	EWMA 3d	-0.335	-2.53	0.0114	0.0959
168	Last	+5.643	+7.96	1.68e-15	2.52e-14
168	Median 3d	+1.154	+4.95	7.31e-07	9.51e-06
168	Median 14d	-0.710	-2.01	0.0448	0.269
168	EWMA 1d	+0.047	+0.30	0.768	0.768
168	EWMA 3d	-0.538	-1.60	0.109	0.328

Note: Negative loss differences favor the alternative. Loss is absolute error; daily means use Newey–West lags 2, 4, and 8 days for $H = 24, 72, 168$, respectively. Holm adjustment covers all 15 comparisons. Source: authors' calculations.

6.2 Direction is more stable than cash-flow magnitude

Sign accuracy displays a different pattern. The one-day EWMA is correct on 69.60% of 24-hour signs, 75.24% of 72-hour signs, and 72.93% of 168-hour signs. The corresponding seven-day-median values are 65.68%, 63.83%, and 63.55%. The three-day EWMA is close behind at longer horizons. Even the latest rate, which has poor long-horizon MAE, retains 66.82% sign accuracy at 24 hours.

The contrast suggests that recent funding information is useful for orientation but less reliable for level forecasts. In practice, a positive predicted differential often remains positive in realized funding, but its size changes materially. Median absolute relative errors for one-day EWMA are 0.983, 0.945, and 0.967, so even the better estimator has typical relative deviations close to the magnitude of the forecast among non-negligible predictions. This is consistent with time-varying predictability rather than a stable cash-flow law (Inan, 2026).

Reference-floor coverage provides another view of bias. For one-day EWMA it is almost exactly 50% at each horizon, meaning realized funding lies above the forecast about half the time. The seven-day median has 53.34%, 50.34%, and 50.17%. A near-50% value is compatible with an approximately median-unbiased point forecast; it says nothing about absolute error, tail coverage, or cost recovery.

6.3 Conservative coverage weakens with horizon

The conservative endpoint covers 80.13% of 24-hour outcomes. Its day-block 95% interval, [78.75%, 81.42%], includes 80%, so the data do not establish that the underlying rate exceeds that threshold. At 72 and 168 hours, coverage declines to 76.87% and 75.86%, with day-block intervals [75.49%, 78.36%] and [74.38%, 77.35%]. Figure 2 contrasts these rates with reference-floor coverage. The conservative construction shifts the forecast downward, but the shift is not large enough to maintain a common coverage level as the horizon increases.

Table 4: Conservative-endpoint and central-interval coverage

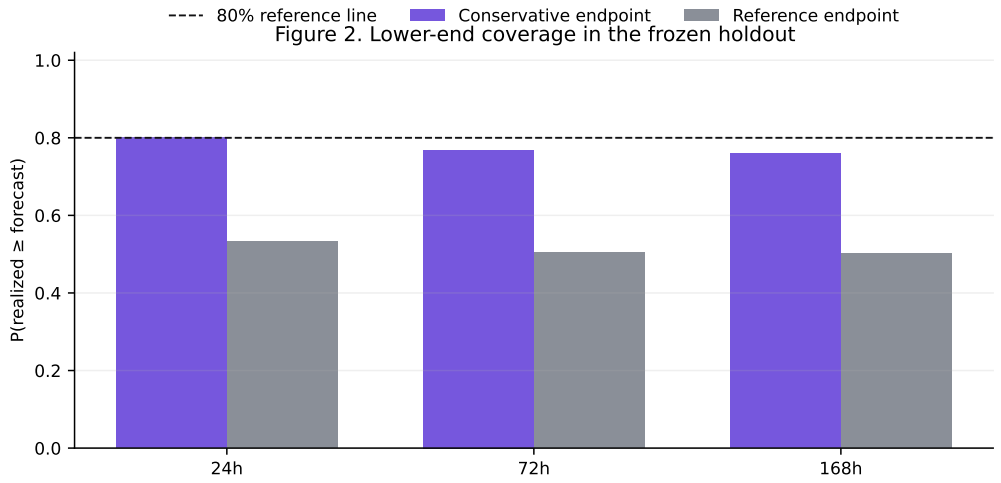
H	Conservative	Day-block 95% CI	Online 90%	Day-block 95% CI	Worst day	Score
24	80.13%	[78.75%, 81.42%]	91.14%	[89.48%, 92.62%]	77.42%	16.02
72	76.87%	[75.49%, 78.36%]	90.86%	[88.81%, 92.51%]	76.83%	40.59
168	75.86%	[74.38%, 77.35%]	90.39%	[88.70%, 91.85%]	79.50%	80.70

Note: The conservative endpoint is the three-estimator lower cash-flow endpoint. Central intervals use the seven-day median with development residuals and the frozen online scaling rule. Day-block intervals use 2,000 resamples; the nominal iid Wilson and Kupiec/Christoffersen results are retained in the replication tables but are not primary evidence. Source: authors' calculations.

This finding qualifies the earlier BTC-only evidence rather than confirming it. A broader asset set, explicit per-record intervals, a later frozen holdout, and both pair directions produce lower coverage at longer horizons. Because the universe and implementation differ, the change should not be interpreted as a pure time effect. It does show that an 85%–87% historical illustration cannot be transported to the present benchmark without re-estimation.

6.4 Average interval calibration hides weak days

Static central intervals cover 90.65%, 89.32%, and 88.05% at the three horizons. The frozen online scaling rule changes these to 91.14%, 90.86%, and 90.39%. Mean widths increase, particularly at 168 hours, but interval scores improve slightly: 16.02 versus 16.12 at 24 hours, 40.59 versus 40.94 at 72

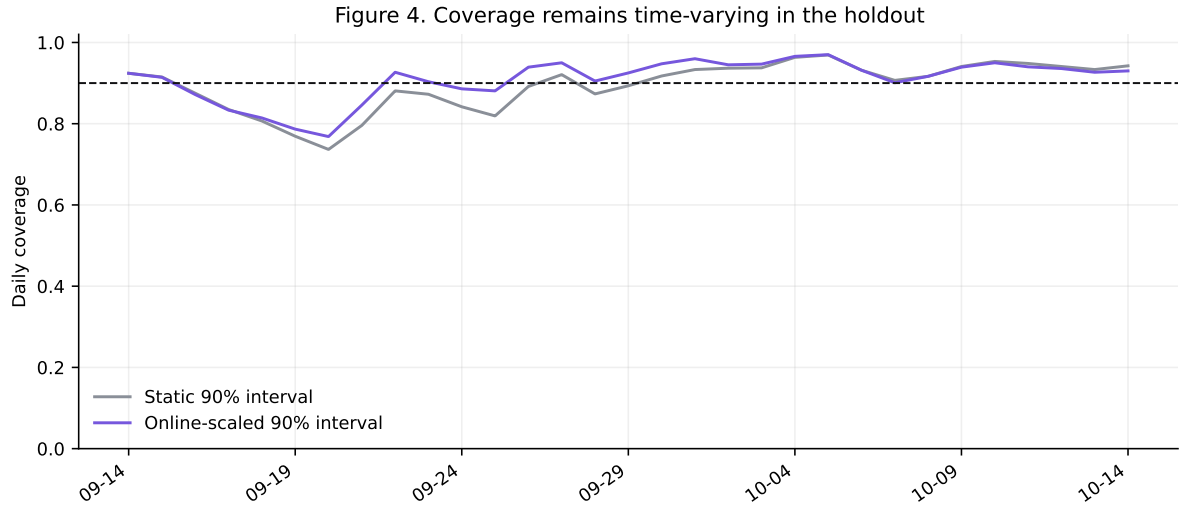


Source: public funding-rate histories; authors' calculations. Overlapping pair-origins; day-block intervals reported in Table 4.

Figure 2: Empirical probability that realized funding is at least the stated endpoint. The dashed line marks 80%, not a certified target. Source: authors' calculations.

hours, and 80.70 versus 81.63 at 168 hours. The adjustment therefore provides a modest coverage–width improvement in this holdout, not merely wider intervals at any cost.

The daily path is less reassuring. Worst-day online coverage is 77.42%, 76.83%, and 79.50%. Figure 4 shows that misses cluster in time rather than arriving as independent rows. Christoffersen independence diagnostics reject independence throughout, while Kupiec tests reject exact 90% unconditional coverage in several cells despite large row counts. These tests are not treated as definitive panel inference, but they agree with the visual evidence: average calibration conceals regime-specific failure.



Source: public funding-rate histories; 72-hour horizon; authors' calculations.

Figure 3: Daily holdout coverage of static and online-scaled central intervals at the 72-hour horizon. Source: authors' calculations.

An online rule also reacts after a miss has occurred. It cannot guarantee coverage in an initial shock window. This limitation is the interval analogue of forecast staleness: the data needed to widen the band are observed only after the relevant regime has begun. Any operational interpretation must therefore focus on the worst day and block uncertainty, not only the full-sample average.

6.5 Scenario costs dominate median funding

Among routes with a positive reference forecast, the median forecast itself is \$0.550, \$1.650, and \$3.850. Median realized funding is lower: \$0.169, \$0.980, and \$2.790. The median realized-to-predicted ratios are 0.553, 0.571, and 0.519. Realized funding is positive on 67.79%, 70.20%, and 69.70% of these selected routes, so direction remains useful, but the median cash amount is small.

Cost scenarios change the economic conclusion completely. In the low case, median funding after cost is −\$19.83, −\$19.02, and −\$17.21. In the base case it is −\$49.83, −\$49.02, and −\$47.21. Only 0.17%, 2.92%, and 11.57% of selected routes have positive funding after the low cost; the base-case shares are 0.00%, 0.06%, and 1.02%. High-cost positive shares are effectively zero. Table 5 reports the distribution and simple annualized median, which remains negative in every scenario.

Table 5: Funding after cost for routes with a positive reference forecast

H	Cost case	Gross med. (\$)	Cost (\$)	Net med. (\$)	Net > 0	Annualized med. (\$)
24	low	0.169	20	-19.831	0.17%	-7,238
24	base	0.169	50	-49.831	0.00%	-18,188
24	high	0.169	105	-104.831	0.00%	-38,263
72	low	0.980	20	-19.020	2.92%	-2,314
72	base	0.980	50	-49.020	0.06%	-5,964
72	high	0.980	105	-104.020	0.00%	-12,656
168	low	2.790	20	-17.210	11.57%	-897
168	base	2.790	50	-47.210	1.02%	-2,462
168	high	2.790	105	-102.210	0.03%	-5,330

Note: n = 33,006 per horizon. Cost scenarios are illustrative, not venue-specific fee quotes: four executions across two legs, plus transfer. Annualization is simple $8760/H$, with no compounding. Historical mid-spread P&L is unavailable, is not set to zero, and therefore these are funding-after-cost results rather than total arbitrage P&L. Source: authors' calculations.

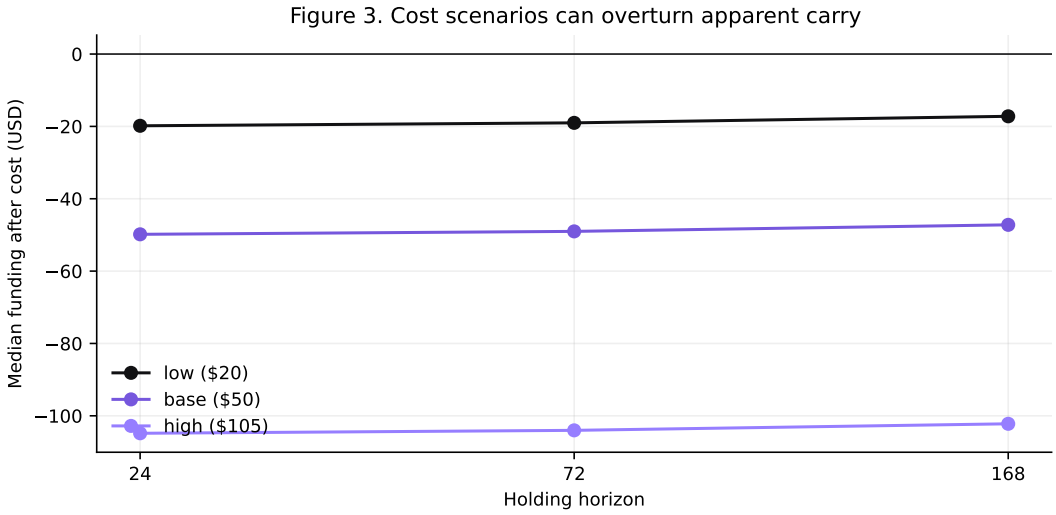


Figure 4: Median realized funding after scenario costs among positively forecast routes. Historical basis P&L is not measured and is not set to zero. Source: authors' calculations.

Means are higher than medians—\$0.908, \$2.406, and \$4.794 before cost—indicating a right tail of larger realized funding. The 90th percentiles are \$4.03, \$10.56, and \$21.79. Thus the conclusion is not that every route lacks cash flow. It is that the economic outcome is concentrated in a minority of

observations, and median funding is an inadequate basis for covering two-leg execution. Identifying whether that tail is persistent, predictable without outcome leakage, and large enough after measured depth-dependent costs requires a separate selection study. The present benchmark does not claim it.

7 Robustness

The robustness analysis changes one dimension at a time and retains unfavorable slices. Five dimensions are considered: forecast horizon, subperiod, asset group, venue composition, and estimator/cost specification. The objective is not to find the most flattering cell but to identify where the main interpretation fails.

7.1 Subperiods and asset groups

At 72 hours, MAE falls from \$5.902 in the first 15 holdout days to \$3.679 in the last 15. Conservative coverage moves in the opposite direction, from 79.30% to 74.09%. Lower average error therefore does not imply better tail protection. This divergence is precisely why point and coverage metrics must be reported together. It also indicates temporal non-stationarity within a holdout of only one month.

The five major assets—BTC, ETH, SOL, XRP, and BNB—have lower MAE (\$3.626) and higher conservative coverage (81.64%) than the remaining 35 assets (\$4.991 and 76.19%). Sign accuracy is slightly lower for majors in this window (61.67% versus 64.14%), so liquidity or baseline-rate stability should not be equated with uniformly better forecasts. Long-tail outcomes drive both higher error and weaker lower-end coverage. The full-sample result is therefore not solely a BTC result, but it is materially affected by the inclusion of long-tail assets.

Table 6: Robustness slices at the 72-hour horizon

Slice	n	MAE (\$)	Sign accuracy	Conservative coverage
first 15 holdout days	36,000	5.902	62.11%	79.30%
last 15 holdout days	34,400	3.679	65.35%	74.09%
majors	9,100	3.626	61.67%	81.64%
long tail	63,700	4.991	64.14%	76.19%
exclude hyperliquid	43,680	4.526	66.08%	77.02%
exclude extended	43,680	4.836	61.71%	75.77%
exclude pacifica	43,680	5.198	66.48%	77.97%
exclude lighter	43,680	4.396	58.96%	76.35%
exclude grvt	43,680	5.146	65.92%	77.25%

Note: Venue exclusions remove every directed pair containing that venue. Majors are BTC, ETH, SOL, XRP, and BNB.
Source: authors' calculations.

7.2 Venue exclusions

Leaving out Hyperliquid, Extended, Pacifica, Lighter, or GRVT changes both the sample and the outcome. MAE ranges from \$4.396 without Lighter to \$5.198 without Pacifica. Sign accuracy ranges from 58.96% without Lighter to 66.48% without Pacifica. Conservative coverage is highest without Pacifica (77.97%) and lowest without Extended (75.77%), but no exclusion restores 80% at 72 hours. The central coverage conclusion is thus not attributable to a single venue.

These are compositional transport tests, not venue fixed effects. Removing a venue deletes every pair containing it and changes the relative weight of the others. Differences cannot be interpreted as causal

estimates of a venue's forecastability. They show that the aggregate result is conditional on the five-venue panel and that pairwise heterogeneity remains important.

7.3 Estimator and ranking sensitivity

Alternative estimators produce the largest change in point-forecast performance, as Section 6 documents. The responsive EWMA family is strongest on direction and short-horizon MAE, while the 14-day median is competitive at 168 hours. This disagreement across horizons is itself a robustness finding: a model chosen solely from the 24-hour table would not dominate at one week.

CSCV/PBO gives low values for ranking stability. With $S = 8$, PBO is 0.000 at 24 and 72 hours and 0.014 at 168 hours; with $S = 10$, it is 0.000, 0.000, and 0.004. Median out-of-sample ranks are one at 24 hours and two at longer horizons. The best in-sample estimator therefore tends to remain near the top out of sample in this frozen holdout. However, PBO does not measure whether the associated cash flow is large or positive, and the best-MAE level changes from development to evaluation. It must not be quoted as evidence of strategy robustness.

Table 7: CSCV/PBO ranking-stability diagnostics

H	S	Splits	PBO	Median OOS rank	IS MAE	OOS MAE
24.0	8.0	70.0	0.00%	1	1.557	1.557
24.0	10.0	252.0	0.00%	1	1.559	1.559
72.0	8.0	70.0	0.00%	2	4.420	4.547
72.0	10.0	252.0	0.00%	2	4.429	4.522
168.0	8.0	70.0	1.43%	2	9.811	10.079
168.0	10.0	252.0	0.40%	2	9.828	10.005

Note: Six estimators; daily mean absolute error; holdout only. Low PBO concerns ranking stability, not economic profitability.
Source: authors' calculations.

The multiplicity results provide a second caution. Descriptively best models and statistically distinguishable models are not always the same. The 14-day median at 168 hours has the lowest MAE but does not beat the seven-day median after Holm correction. Conversely, the one-day EWMA at 24 hours clears the adjusted threshold. The evidence supports selective, horizon-specific statements rather than a global estimator ranking.

7.4 Cost sensitivity

Cost is the most consequential robustness dimension for economic interpretation. The base case is not cherry-picked to make outcomes negative; the low case is also reported, and even it leaves median net funding negative. The high case merely strengthens the same conclusion. Because the actual cost parameters are scenario assumptions, these tables should be read as thresholds. For each horizon, the median gross funding is far below even $20 of round-trip cost, so a large class of plausible cost specifications produces the same median sign.

The right tail is less robust to simple generalization. A subset of routes exceeds cost, especially at 168 hours, but this study has not tested whether a rule can identify that subset on later data. Ranking all routes by their forecast and selecting positive ones is not equivalent to selecting the eventual high-funding tail. A capacity or tail-selection claim would require its own preregistered candidate family and multiplicity control.

7.5 Overall stability judgment

The evidence is **conditional**, not robust in an unrestricted sense. Short-horizon recency gains are the most stable statistical result. Directional usefulness is visible across several slices, but magnitude errors remain large. Conservative and interval coverage vary with horizon, day, asset group, and venue composition. Finally, economic significance is fragile to modest costs at the median. The most damaging single perturbation for the conservative claim is extending the horizon from 24 to 168 hours; for the economic claim, it is imposing even the low round-trip cost scenario.

8 Limitations

Universe and survivorship. The universe is an exact-name intersection of venue lists retrieved on 2026-11-03, not a point-in-time historical security master. Contracts that left a venue before retrieval can be absent, and GRVT does not expose a comparable listing-date field in the current registry. The 40-asset sample is consequently subject to survivorship and identifier-mapping risk. It also excludes all assets not shared by every venue, which can tilt the panel toward more established contracts. Results should not be extrapolated to the full market or to venues without a reconstructed common window.

Temporal coverage. The snapshot contains 90 days, with 30 days of holdout decision origins. This satisfies the minimum one-month calibration window but does not span multiple market regimes. The 168-hour analysis has only a small number of non-overlapping weekly outcomes. Row counts of 72,800 per horizon should not be mistaken for independent sample size. Day-block resampling addresses within-day clustering but cannot create missing temporal regimes or fully resolve dependence across adjacent weeks.

Linear projection. Every estimator holds an hourly funding estimate constant over the forecast horizon. This is an interpretable counterfactual, not a model of future premium dynamics, cap changes, settlement-rule changes, or endogenous arbitrage flows. Historical per-record intervals normalize the inputs, but future intervals and caps are not forecast. Structural funding models or regime-switching approaches may improve the projection; they require a separately frozen comparison.

Timestamp and information availability. The 60-second settlement boundary removes mechanical same-settlement leakage in the public records. It does not measure API publication latency, clock disagreement, or the time required to observe and process a settlement. Real-time implementation would need a stricter availability model. Public APIs can also revise historical records; a later download may not reproduce an earlier vendor vintage bit for bit, even with the same query window.

Execution and total P&L. The study does not possess historical order books, executable depth, queue positions, account-specific fees, collateral-financing terms, transfer paths, or liquidation outcomes. The three cost cases are transparent scenarios, not measured costs. They cannot establish capacity, implementation shortfall, or a break-even participation rate. Spot borrow is not applicable to the perp–perp structure modeled here, but collateral and financing costs remain unmeasured rather than irrelevant.

Basis risk. Historical cross-venue mid-spread P&L is unavailable and is not set to zero. Funding-after-cost tables therefore describe $F - C$, not total net income. Adverse basis can exceed funding, particularly around settlement discontinuities. The median funding-after-cost value is only a break-even reference for that omitted component. No conclusion about realized arbitrage profit, Sharpe ratio, drawdown, or liquidation-adjusted return is warranted.

Contract comparability. Exact base-asset names improve contract matching but do not eliminate differences in quote currency, collateral, contract multiplier, index composition, or funding-notional convention. The analysis uses a constant \$10,000 notional per leg for comparability. Actual funding transfers

can depend on changing oracle or mark values. A complete implementation would need contract-specific quantities and historical conversion rules.

Inference and selection. Diebold–Mariano tests rely on daily aggregation and chosen Newey–West lags; the longest horizon remains low-power. Holm adjustment covers the registered 15 comparisons, not every descriptive slice in the paper. CSCV/PBO is a secondary ranking diagnostic computed on the holdout and is not a substitute for temporal out-of-sample evidence. Cost and subgroup tables are descriptive unless explicitly tested.

Replication and redistribution. All results can be recomputed from public endpoints with the supplied code. This technical reproducibility is not a legal determination that every raw record may be redistributed. Venue terms and database rights must be checked before publishing a raw-data archive. The safe fallback is to distribute scripts, manifests, derived non-raw tables, and retrieval instructions while leaving raw downloads to each researcher.

External review. The numerical pipeline and rendered artifacts receive local checks, but independent academic peer review and an external similarity service are not substitutes for those checks and have not been completed. The preprint should be read as a reproducible working paper, not as a journal-accepted study or a certified production model.

9 Conclusion

This paper turns cross-venue funding comparison into a reproducible forecasting exercise. A fixed 90-day snapshot, 40 common crypto assets, five venues, 800 directed pairs, and a frozen 30-day holdout show that simple historical estimators have useful but limited information. Responsive EWMA forecasts improve directional accuracy across horizons and deliver the clearest short-horizon MAE improvement. Their longer-horizon advantages are not uniformly significant after multiplicity control, so estimator choice should remain horizon-specific.

The coverage results impose a second constraint. A conservative lower endpoint that appears adequate at 24 hours weakens at 72 and 168 hours. Online-scaled central intervals can approach nominal average coverage, but weak days remain and misses cluster in time. Forecast evaluation must therefore include tails and temporal dependence, not just average point error.

The economic results impose a third constraint. Positively forecast routes often retain a positive funding sign, yet median realized funding is small relative to explicit two-leg cost scenarios. This does not rule out a profitable tail or a more effective selection rule. It does rule out interpreting funding predictability as executable net carry without additional evidence.

The appropriate next step is not a larger collection of tuned estimators on the same holdout. It is a preregistered extension with point-in-time contract membership, additional market regimes, historical executable books, and measured basis changes. Until then, the supported claim is deliberately bounded: recent funding history contains forecast information, but accurate total cash-flow prediction and implementable cross-venue income remain separate, unresolved problems.

Declarations

Authorship and affiliation. Hang Liu is the named author. The research was conducted by humarb Research, London, United Kingdom; `https://humarb.ai` is its research website and `https://humarb.ai/research/` is the research channel. The author is responsible for the final content.

Competing interests. No competing-interest disclosure was supplied in the research package. The author must confirm the final statement before Zenodo publication.

Funding. No external funding statement was supplied in the research package. The author must confirm the funding disclosure before Zenodo publication.

Data availability. The empirical data are historical funding records and instrument metadata retrieved from the official public APIs of Hyperliquid, Extended, Pacifica, Lighter, and GRVT on 2026-11-03. The fixed study window is 2026-07-23 00:00Z–2026-10-21 00:00Z. Retrieval scripts, endpoint descriptions, file checksums, and derived result tables accompany the paper. Raw-data redistribution is conditional on a final review of venue terms; researchers can regenerate the snapshot directly from the public endpoints.

Code availability. The reproduction package contains acquisition, normalization, estimation, inference, figure, table, and independent-reconciliation code. The preregistered configuration and deterministic seeds are included. Every table in the paper is generated from the same frozen result files.

Generative AI use. Generative AI was used for literature retrieval assistance, code drafting and checking, language editing, and document/layout assistance. No AI system was used to invent observations, references, or numerical results. All numerical claims were computed from the public snapshot and reconciled against generated outputs; the named author retains responsibility for verification and publication.

Research status. This is a preprint and has not been peer reviewed. External similarity checking and the author's final conflict/funding confirmations remain publication steps.

Not investment advice. This paper is methodological and empirical research. Historical forecasts and scenario calculations are hypothetical evidence, not forecasts of future performance, trading instructions, or investment advice. No representation is made that any route is executable, profitable, or suitable for a particular reader.

References

Damien Ackerer, Julien Hugonnier, and Urban Jermann. Perpetual futures pricing. *Mathematical Finance*, 36(3):481–499, 2025. doi: 10.1111/mafi.70018.

Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3(2): 5–39, 2001. doi: 10.21314/JOR.2001.041.

David H. Bailey and Marcos López de Prado. The deflated sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management*, 40(5):94–107, 2014. doi: 10.3905/jpm.2014.40.5.094.

David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. The probability of backtest overfitting. *The Journal of Computational Finance*, 2016. doi: 10.21314/JCF.2016.322.

Peter F. Christoffersen. Evaluating interval forecasts. *International Economic Review*, 39(4):841–862, 1998. doi: 10.2307/2527341.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. doi: 10.1080/07350015.1995.10524599.

Extended. Extended api documentation: Funding rates, 2026. URL `https://api.docs.extended.exchange/`. Official documentation, accessed 2026-11-03.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.00170.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Denis Gromb and Dimitri Vayanos. Limits of arbitrage. *Annual Review of Financial Economics*, 2(1): 251–275, 2010. doi: 10.1146/annurev-financial-073009-104107.

GRVT. Market data api: Instruments and funding rate, 2026. URL https://api-docs.grvt.io/market_data_api/. Official API documentation, accessed 2026-11-03.

David Harvey, Stephen Leybourne, and Paul Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291, 1997. doi: 10.1016/S0169-2070(96)00719-4.

Songrun He, Asaf Manela, Omri Ross, and Victor von Wachter. Fundamentals of perpetual futures, 2022. URL https://arxiv.org/abs/2212.06888.

Rob J. Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50 (4):361–365, 1996. doi: 10.1080/00031305.1996.10473566.

Hyperliquid. Funding, 2026. URL https://hyperliquid.gitbook.io/hyperliquid-docs/trading/funding. Official documentation, accessed 2026-11-03.

Emre Inan. Predictability of funding rates. *International Review of Financial Analysis*, 119:105464, 2026. doi: 10.1016/j.irfa.2026.105464.

Da-Hea Kim. Perpetual futures as predictors of bitcoin volatility. *Journal of Empirical Finance*, 89: 101779, 2026. doi: 10.1016/j.jempfin.2026.101779.

Paul H. Kupiec. Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2):73–84, 1995. doi: 10.3905/jod.1995.407942.

Lighter. Funding, 2026. URL https://docs.lighter.xyz/trading/funding. Official documentation, accessed 2026-11-03.

Hang Liu. Measuring executable net carry in cross-venue perpetual funding-rate arbitrage. humarb Research 01, Zenodo preprint, 2026a. URL https://doi.org/10.5281/zenodo.23868361.

Hang Liu. Framing net carry for cross-venue perpetual funding positions: A settled-history, depth-walked, and liquidation-aware method. humarb Research 02, Zenodo preprint, 2026b. URL https://doi.org/10.5281/zenodo.23897013.

Igor Makarov and Antoinette Schoar. Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, 135(2):293–319, 2020. doi: 10.1016/j.jfineco.2019.07.001.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610.

Pacifica. Get historical funding, 2026. URL `https://docs.pacifica.fi/api-documentation/api/rest-api/markets/get-historical-funding`. Official API documentation, accessed 2026-11-03.

André F. Perold. The implementation shortfall: Paper versus reality. *The Journal of Portfolio Management*, 14(3):4–9, 1988. doi: 10.3905/jpm.1988.409150.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.

Maik Schmeling, Andreas Schrimpf, and Karamfil Todorov. Crypto carry. *Management Science*, 2026. doi: 10.1287/mnsc.2024.05069.

Robert J. Shiller. Measuring asset values for cash settlement in derivative markets: Hedonic repeated measures indices and perpetual claims. *The Journal of Finance*, 48(3):911–931, 1993. doi: 10.1111/j.1540-6261.1993.tb04024.x.

Andrei Shleifer and Robert W. Vishny. The limits of arbitrage. *The Journal of Finance*, 52(1):35–55, 1997. doi: 10.1111/j.1540-6261.1997.tb03807.x.

James E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006. doi: 10.1287/mnsc.1050.0451.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

A Notation and estimator definitions

Table 8: Principal symbols

Symbol	Definition	Unit
A	Constant notional per leg	USD
S, L	Short and long legs of a directed venue pair	–
t	Decision origin, on the 00/08/16Z grid	UTC timestamp
H	Forecast and realization horizon	hours
$f_{i,k}$	Settled funding rate for leg i, record k	decimal per period
$I_{i,k}$	Settlement interval attached to record k	hours
$r_{i,k}$	Comparable hourly rate, $f_{i,k}/I_{i,k}$	decimal per hour
y	Realized two-leg net funding in $(t, t+H]$	USD
$\hat{y}^{(m)}$	Forecast from estimator m	USD
$\hat{y}^{c}$	Conservative lower cash-flow endpoint	USD
e	Forecast error, $y - \hat{y}$	USD
m_d	Online residual-band multiplier for day d	–
C	Scenario round-trip and transfer cost	USD

The latest estimator uses the most recent known $r_{i,k}$. Daily-median estimators first sum per-period rates within each complete UTC day and divide by 24, yielding a daily mean hourly rate; they then take the median over the most recent N complete days. EWMA estimators use a 14-day lookback, interval-weighted exponential decay, and one- or three-day half-lives. The 30-day median is used only in the conservative endpoint. All are example research parameters, not production settings.

B Data dictionary

B.1 Raw snapshot

Each raw JSONL record is retained in the schema returned by its public endpoint. The normalization layer creates the following common fields.

Field	Meaning	Validation
`ts_ms`	Funding settlement timestamp in Unix milliseconds	Integer; fixed window enforced
`period`	Signed funding rate as decimal per settlement	Finite; venue unit conversion applied
`interval`	Settlement interval in hours	Positive; per-record GRVT field preserved
`hourly`	`period/interval`	Comparable estimator input
`venue`	Hyperliquid, Extended, Pacifica, Lighter, or GRVT	Exact controlled vocabulary
`asset`	Common base-asset identifier	One of the registered 40 names

Lighter's sign is negative when its payment direction is 'short' and positive when it is 'long'; its rate and GRVT's rate are divided by 100. The other three venues use decimal fractions directly. Duplicate timestamps, if present, are resolved deterministically after sorting; no missing realized rate is filled with zero.

B.2 Derived pair-origin table

Column	Meaning	Unit/domain
`asset, short, long`	Asset and directed venue identifiers	Categorical
`t_start, day`	Origin and UTC calendar day	Unix seconds; YYYY-MM-DD
`horizon`	Outcome horizon	24, 72, or 168 hours
`actual`	Realized short-minus-long funding	USD per $10,000/leg
`conservative`	Three-estimator lower endpoint	USD
`sample`	Temporal segment	development or holdout
`last`	Latest-rate forecast	USD
`median_3d`, `median_7d`, `median_14d`	Complete-day median forecasts	USD
`ewma_1d`, `ewma_3d`	Interval-weighted EWMA forecasts	USD

A row exists only if all six forecasts and the conservative endpoint are finite and both outcome legs meet the 90% settlement-coverage rule. The table contains 384,000 rows. The holdout has 218,400 rows, evenly divided among the three horizons.

C Inference and reproducibility details

C.1 Boundary and missing-data algorithm

For an origin t, the estimator includes records with timestamp at most $t - 60$ seconds. For an outcome of length H, it includes records with timestamp greater than $t + 60$ seconds and no greater than $t + H + 60$ seconds. Expected

outcome coverage is the sum of record intervals divided by H. For daily estimator inputs, a UTC day is complete when its interval sum is at least 21.6 hours. Medians use an $N + 1$-day lookback and the most recent N complete days, preventing the partial first and last days of an eight-hour grid from shortening the effective sample.

C.2 Independent 30-origin reconciliation

The independent audit uses seed 20261021 to draw 30 holdout rows. A second implementation reads the raw JSONL files directly and recomputes the seven-day reference, three-candidate conservative endpoint, and realized funding without calling the main estimator classes. The reconciliation table reports stored and recomputed values. All 30 rows pass a 10^{-8} USD absolute tolerance; the largest discrepancy is 1.42×10^{-14} USD, attributable to floating-point operation order.

C.3 Run log and negative results

The package retains the initial API failure log, successful retry manifest, analysis log, complete metric tables, and all subgroup results. No failed candidate is removed from the comparison. The latest-rate estimator has the weakest long-horizon MAE; static intervals under-cover at 168 hours; the last 15 holdout days have lower conservative coverage than the first 15; and every cost scenario has a negative median funding-after-cost value. These outcomes are included in the paper rather than treated as failed runs.

C.4 Computational environment

The reference implementation uses Python 3.12 with NumPy, pandas, SciPy, and Matplotlib. The deterministic random seed is 20261103 for block resampling and 20261021 for the independent audit. The paper is generated from the same CSV result files as the web page. SHA-256 digests cover the preregistered protocol, raw files, derived observation table, PDF, metadata, and webpage package.

D Additional interpretation safeguards

The full directed-pair panel is symmetric. Its mean gross funding is mechanically close to zero because reversing S and L reverses the signed cash flow. Economic interpretation therefore uses the preregistered positive-reference subset rather than the full-panel mean. This subset has 33,005 routes per horizon. It is not independent of the point-forecast evaluation and is not presented as a second holdout.

Simple annualized values in Table 5 are arithmetic scale conversions of a per-route median. They do not assume that a position can be repeated, that capital remains continuously deployed, or that overlapping positions can be financed. They are included only to make the order of magnitude comparable across horizons. The negative values should not be read as an estimate of an annual strategy loss.

The low, base, and high cost cases are not calibrated to any one venue. A venue with zero explicit trading fees can still incur spread, impact, transfer, financing, and basis costs. Conversely, a maker execution can have a different fee and fill probability from the bundled scenario. Replacing the scenarios with measured depth walks is a necessary extension before capacity or executable net carry can be assessed.