Université du Québec en Outaouais

Département des sciences administratives

Working Paper No. 41

CHRONOS-X: Universal Intraday Representation Learning for Probabilistic Financial Forecasting

A Multi-Resolution Patch Transformer Tested Against Matched Baselines on 300 U.S. Equities and ETFs

Simon-Pierre Boucher

Département des sciences administratives
Université du Québec en Outaouais

simon-pierre.boucher@uqo.ca

Gatineau – Pavillon Alexandre-Taché
283, boulevard Alexandre-Taché
Gatineau, Québec, Canada J9A 1L8

November 2026

Version 1.0

Abstract

We ask whether a compact, transferable, multi-resolution model can learn minute-level structure that improves calibrated probabilistic forecasts of the next 5, 15, 30 and 60 minutes of U.S. equity and ETF returns. We build a Multi-Resolution Patch Transformer that reads the last 390 one-minute bars together with five days of 5-minute and twenty days of 30-minute past-only aggregates, emits Student-t marginals for each of the next sixty minutes and, in a second stage, a low-rank Gaussian copula over them. The model is trained and evaluated on 234 million one-minute bars of 300 instruments (2018–2025) from a private FirstRate Data lakehouse, selected point-in-time and audited for timestamp, session and corporate-action conventions, against ten baselines sharing the same information set, with CRPS, interval and calibration scores, trading-day cluster inference and overlap-aware Diebold–Mariano tests on about one million evaluation origins per year, sixty held-out instruments and an untouched final year. Every model beats a trailing-scale random walk by 3.5–4.9 % of CRPS; the gain is a volatility gain with no directional or economic content. Sequence models lead at five minutes; a HAR-type variance regression with a kurtosis-matched Student-t leads and is better calibrated from fifteen minutes on. The multi-resolution context adds at most 0.1 % in sample and nothing out of sample, model size (3.5 vs 20.7 million parameters) does not matter, regime conditioning and coherent path sampling add nothing because future minute returns are conditionally independent given the recent history, and pooled training transfers to unseen instruments without penalty. A model that reads only the trailing volatility and the clock recovers nine tenths of the full model's skill. We release the code, the protocol, synthetic fixtures and these null results.

Keywords: intraday forecasting, probabilistic forecasting, transformers, multi-resolution representations, calibration, proper scoring rules, high-frequency data

JEL Classification: C45, C53, C58, G12, G17

1. Introduction

Every minute of a U.S. trading day leaves behind one bar—open, high, low, close, volume—for each of several thousand instruments. Foundation models for time series promise to learn "universal" representations from such streams and to forecast any series zero-shot (Ansari et al., 2024; Das et al., 2023), and a growing literature applies deep sequence models to intraday prices (Sirignano and Cont, 2019; Zhang et al., 2019; Kolm et al., 2023). Whether a compact model that reads several resolutions of minute-level history improves *calibrated* forecasts of the next hour, relative to the simplest volatility models an econometrician would write down, is however an empirical question that is rarely posed with matched information sets, proper scoring rules and dependence-aware inference. This paper poses it narrowly and answers it on data.

We build CHRONOS-X, a research package around a Multi-Resolution Patch Transformer (MRPT). At a decision time inside the regular session the model reads the last 390 completed one-minute bars, the last five days of completed 5-minute aggregates and the last twenty days of 30-minute aggregates, each tokenised in patches with gated feature projections, resolution and temporal-gap embeddings; a shared pre-norm Transformer encoder and a direct decoder with sixty future-calendar queries emit a Student-t marginal for each of the next sixty minute returns, and a second-stage low-rank Gaussian copula turns the marginals into coherent sample paths. All inputs are strictly causal, scale-free (divided by a trailing volatility computed at the origin) and masked where the source has no bar; the model carries no ticker identity. Five hypotheses were written before any tuning: multi-resolution history helps at equal parameters (H1), pooled training transfers to unseen instruments (H2), history-only regime conditioning improves calibration under stress (H3), coherent paths beat independent marginals on path scores (H4), and—exploratory—the forecasts carry economic content after costs (H5).

The data are 300 U.S. stocks and ETFs selected point-in-time at the end of 2017, stratified by asset class, dollar volume, volatility and sector, with sixty instruments held out entirely; 234 million one-minute bars from a private FirstRate Data lakehouse (FinData OS) over 2018–2025, audited for timestamp conventions, session boundaries, early closes, daylight-saving time, missing minutes, splits and dividends. Training uses 2018–2022, validation 2023, the development test 2024 and an untouched final period 2025; evaluation uses 986468 origins in 2024 alone. Baselines share the information set exactly: random walk, seasonal volatility, EWMA, HAR-RV at the minute scale, GARCH(1,1)-t, ridge, elastic net, LightGBM, an MLP and a TCN with the same Student-t head, and the one-minute-only MRPT as the equal-parameter patch-Transformer control. Scores are the CRPS, the weighted interval score, quantile losses, coverage, PIT diagnostics, QLIKE, tail Brier scores and, for paths, energy and variogram scores; inference clusters by trading day and uses overlap-aware HAC Diebold–Mariano statistics with multiple-testing control, because a million

overlapping windows are not a million observations.

The answer is precise and mostly negative for the architecture, and informative about the data. Every model beats the random walk by 3.5–4.9 % in CRPS at every horizon: the minute stream carries real information about the dispersion of the next hour, and none about its direction (directional accuracy 0.50, median magnitudes below one basis point, no economic content after costs; H5 rejected). The sequence models are the best at five minutes; from fifteen minutes on, a heterogeneous autoregression of realised variance with a kurtosis-matched Student-t is both sharper and better calibrated than any neural model, by 0.3–0.7 %. The multi-resolution context that defines MRPT improves on the one-minute-only twin by at most a tenth of a percent—statistically detectable on a million origins, economically nil (H1 rejected in substance); a 20.7-million-parameter model does no better than a 3.5-million one. The remaining hypotheses are examined in the same spirit in Sections 7 and 8: zero-shot transfer to the sixty held-out instruments, regime conditioning on stress days, the copula paths against independent sampling, and the untouched 2025 period.

The contribution is therefore a replicable measurement rather than a new state of the art: a documented data pipeline and evaluation protocol for minute-level probabilistic forecasting under a private-use license, a compact model that is competitive without being superior, and a set of null results that a reader can reproduce from the released code and synthetic fixtures. Section 2 states the hypotheses, Section 3 places them in the literature, Sections 4–6 describe data, model and design, Sections 7–8 report the results, and Sections 9–11 discuss limitations, reproducibility and what the evidence does and does not say.

2. Research question and falsifiable hypotheses

The question is narrow on purpose. At a decision time t inside the regular session, with every one-minute bar up to t complete, can a compact model that reads the recent minute-level history *and* coarser past-only aggregates produce better-calibrated and sharper predictive distributions for the next $H \in \{5, 15, 30, 60\}$ minutes than models that only read the minute stream, or than well-specified non-neural volatility models? "Better" is defined by strictly proper scoring rules (Gneiting and Raftery, 2007; Matheson and Winkler, 1976), never by price-level errors: on a one-minute grid the random walk is a formidable point forecaster, and a model can look impressive on price-level mean squared error while learning nothing (Cont, 2001). We therefore train and score on return-based targets and evaluate full predictive distributions.

Let r_{t+h} denote the log return of the bar opening h minutes after t and $R_{t,H} = \sum_{h \le H} r_{t+h}$ the cumulative return. The model emits, for each future minute, a Student-t marginal with location, scale and degrees of freedom, and (in a second stage) a dependence structure across the H minutes. Five hypotheses were written before any model was tuned and are tested as stated.

Hypothesis 1 (Multi-resolution history)**.** At equal parameter count and training budget, a model whose encoder reads the last 390 one-minute bars, the last five days of 5-minute aggregates and the last twenty days of 30-minute aggregates achieves a lower CRPS (and WIS) for $R_{t,H}$ than the same architecture reading the one-minute bars only.

Hypothesis 2 (Transfer to unseen instruments)**.** A model trained on a pool of instruments, without any ticker identity, scores better on instruments it has never seen than non-neural baselines fitted with the same information set, and its loss on unseen instruments is not materially worse than on pooled instruments.

Hypothesis 3 (History-only regime conditioning)**.** Conditioning the decoder on history-only regime statistics (trailing volatility and relative volume) improves calibration—measured by coverage gaps and the probability-integral-transform (PIT) dispersion—on days whose cross-sectional volatility is in the top decile, relative to the same model without regime conditioning.

Hypothesis 4 (Coherent joint paths)**.** A dependence-aware decoder (a Gaussian copula with a low-rank factor structure over the H future minutes, with the Stage-A marginals held fixed) achieves lower energy and variogram scores for the joint path, and lower CRPS for path functionals (maximum drawdown, realised variance), than sampling the same marginals independently.

Hypothesis 5 (Economic content, exploratory)**.** Signals derived from the predictive median contain incremental information after an assumed round-trip cost; this hypothesis is exploratory because no quote data are available and costs are assumptions.

Each hypothesis has a pre-specified test: a Diebold–Mariano statistic on the loss differential with an overlap-aware heteroskedasticity- and autocorrelation-consistent variance (Diebold and Mariano, 1995; Harvey et al., 1997; Newey and West, 1987), a trading-day cluster bootstrap for interval estimates (all instruments of one day form one cluster, because millions of overlapping windows are not independent observations), and Holm and Benjamini–Hochberg adjustments across horizons and comparisons (Holm, 1979; Benjamini and Hochberg, 1995). Null findings are reported as such.

3. Related work

Intraday volatility and probabilistic evaluation. The classical description of intraday returns combines a deterministic periodic component with persistent conditional heteroskedasticity (Andersen and Bollerslev, 1997; Bollerslev, 1986; Nelson, 1991; Engle, 2001). Realised measures built from intraday returns turned volatility into an (almost) observable quantity and showed that simple heterogeneous autoregressions forecast it well (Andersen et al., 2003; Barndorff-Nielsen and

Shephard, 2002; Corsi, 2009; Bollerslev et al., 2016); Hansen and Lunde (2005) found that little beats a GARCH(1,1) once the comparison is fair, and Patton (2011) showed that only a few loss functions (QLIKE, MSE) rank variance forecasts consistently under noisy proxies. These results frame our baselines: seasonal profiles, EWMA, HAR-RV, GARCH-t. For distributional forecasts we rely on the theory of proper scoring rules (Matheson and Winkler, 1976; Gneiting and Raftery, 2007), the calibration–sharpness paradigm and PIT diagnostics (Gneiting et al., 2007), the weighted interval score (Bracher et al., 2021), closed-form CRPS for the Student-t family as implemented in `scoringRules` (Jordan et al., 2019; Taillardat et al., 2016), threshold-weighted variants (Gneiting and Ranjan, 2011), and multivariate energy and variogram scores (Scheuerer and Hamill, 2015). Inference on loss differentials follows Diebold and Mariano (1995), Harvey et al. (1997), Giacomini and White (2006) and Hansen et al. (2011).

Deep learning for financial microstructure and time series. Sirignano and Cont (2019) documented that a single neural model trained on pooled limit-order-book data transfers across stocks—the closest antecedent of our "universal" question, although with order-book inputs that we do not have. DeepLOB (Zhang et al., 2019) and deep order-flow imbalance models (Kolm et al., 2023) also use quote-level data. With bars only, the relevant architectures are the generic sequence models: temporal convolutions (Bai et al., 2018), DeepAR-style autoregressive likelihood models (Salinas et al., 2020), N-BEATS (Oreshkin et al., 2019), and Transformer variants (Vaswani et al., 2017; Wen et al., 2022). Patch tokenisation (Nie et al., 2022) and variable-as-token inversion (Liu et al., 2023) are the Transformer designs we adapt; Zeng et al. (2022) is the reminder that linear baselines are competitive. Pretrained foundation models for time series (Ansari et al., 2024; Das et al., 2023) promise zero-shot forecasting but assume a single real-valued series without missingness masks or multi-resolution context; we discuss in Section 6 why they are not evaluated here. Diffusion-based multivariate forecasters (Rasul et al., 2021) are a possible Stage-B alternative that we do not pursue given their compute cost. Surveys of financial machine learning (Gu et al., 2020; Kelly and Xiu, 2023; Lim and Zohren, 2021) stress the evaluation pitfalls we try to avoid: overlapping samples, survivorship, look-ahead, and multiple testing (López de Prado, 2018; Bailey and López de Prado, 2019).

Dependence across horizons. Coherent joint paths require a dependence model. Copulas separate marginals from dependence (Sklar, 1959; Patton, 2006); our Stage B uses a Gaussian copula with a low-rank-plus-diagonal correlation over the sixty future minutes, trained on the probability-integral transforms of the observed targets with the marginals frozen.

4. Data

4.1. Source and connector

All bars come from FinData OS, a private lakehouse of FirstRate Data one-minute bars for U.S. stocks and ETFs (6.70 and 0.91 billion rows respectively at audit time, 2000–2026). The connector discovers the catalog, the published Arrow schema and the OpenAPI document, reads small verification pages through the paginated bars endpoint and performs bulk retrieval through asynchronous Parquet exports pinned to a dataset version. Every call keeps the request identifier, dataset identifier and version; every export file is schema-validated and hashed. The license is private use: no raw bar, grid or per-origin return leaves the compute node, and the release contains code, aggregate results and synthetic fixtures only.

4.2. Conventions verified on real rows

Timestamps are bar-*open* instants in UTC: the bar stamped 09:30 ET covers 09:30:00–09:30:59 and is complete at 09:31. A regular XNYS session is 09:30–16:00 ET, i.e. 391 bars on a normal day (slots $0, \ldots, 390$) and 211 on a 13:00 early close; the last slot is the closing print, whose volume is one to two orders of magnitude larger than a typical minute (6.7 million against a median of 0.12 million shares for AAPL on 2024-03-08). The closing print is excluded from context windows and targets. Extended hours (04:00–20:00 ET) are present at the source and excluded here. Zero-volume minutes are absent at the source and are never filled: missingness is a mask and a feature. Daylight-saving transitions are handled through the exchange calendar (first regular bar at 14:30Z on 2024-03-08 and 13:30Z on 2024-03-11). Stored bars are raw; splits are applied explicitly per session from the instrument master, which reproduces the server-side split-adjusted series exactly (AAPL 4:1 on 2020-08-31, ratio 4.000 on prices and volume). Cash dividends are not applied: intraday log returns within a session are invariant to the dividend factor and any dividend adjustment is window dependent; the overnight-gap feature across an ex-dividend date is left raw. Mergers and spin-offs are unavailable; a delisted series simply ends.

4.3. Point-in-time universe and coverage gate

The selection date is 2017-12-29. Eligible instruments have a first bar at least six months before that date and a last bar at least six months after the sample start (a documented survival requirement), an average daily dollar volume of at least 5 million USD over December 2017 and at least fifteen daily bars in that window. From 2147 candidates we draw 300 instruments proportionally over asset class (75 % stocks, 25 % ETFs), dollar-volume tercile and daily-volatility tercile, with a round-robin over SIC divisions (stocks, from SEC submissions) or ETF name categories inside each stratum. A

first draw revealed regular-hours minute coverage between 6 % and 100 % over 2018–2025 (median 82 %): a quarter of the instruments traded in fewer than 60 % of regular minutes, which makes per-minute targets mostly masked. Before any model was tuned we audited the coverage of all candidates and imposed an availability filter of at least 80 % coverage over the window (1152 candidates remain; Spearman correlation between dollar volume and coverage 0.76); the second draw keeps the same strata and seed. Sixty instruments (20 %) are held out entirely for zero-shot evaluation. Only five delisted candidates passed the filters because FinData's delisted one-minute archives are only partly ingested: the universe is therefore close to a survivor sample, a limitation we carry through the paper.

4.4. Sample, grid and targets

The main sample spans 2018-01-02 to 2025-12-31 (2011 sessions): 233958776 exported rows, of which 220142835 fall in regular hours (93.7 % coverage on average, minimum 80.3 %; 13.8 million extended-hours rows excluded; zero integrity rejections and duplicates; 242 vendor-flagged rows kept with their flags). Each instrument is stored as a dense session $\times$ 391 grid of log open, high, low, close and log volume with a presence mask, plus past-only 5-minute (78 per day) and 30-minute (13 per day) aggregates with their own masks and a past-20-session seasonal profile of log volume and of squared returns. An origin (s, m) is session s at slot m, with bars $0, \ldots, m$ complete; it is eligible when $m \geq 30$, $m + 60$ does not cross the last tradable slot (389, or 209 on an early close), bar m is present and at least 80 % of the next 60 slots are present. Targets are $y_h = \log C_{m+h} - \log C_{p(m+h-1)}$, where $p(\cdot)$ is the last present slot, with mask $\mathbb{1}[\text{bar } m + h \text{ present}]$, so that the masked sum over $h \leq H$ is exactly the H-minute log return of the last observed bar. Derived targets are the cumulative return $R_{t,H}$, the realised variance $\sum y_h^2$, high/low excursions, and the tail event $\left|R_{t,H}\right| > 2.5\,\sigma_t\sqrt{H}$, where σ_t is the trailing root-mean-square one-minute return in the 390-bar window (floored at 10^{-5}), computed from the past only.

4.5. Features

Every bar carries fifteen causal features: scaled return, range, upper and lower wicks, body, close location, log volume, relative volume against the seasonal profile, missing flag, session-boundary flag, log temporal gap (which encodes overnight and weekend gaps), minute fraction, time to close, weekday and early-close flag. Returns and ranges are divided by σ_t (times $\sqrt{5}$ and $\sqrt{30}$ for the aggregates), so inputs are scale free. Nine global history-only statistics—$\log \sigma_t$, the scaled overnight gap, the calendar position, the ratios of 30-minute and intraday realised variance to σ_t^2, and the mean relative volume over 30 minutes—feed a global token and the regime embedding. A leakage gate in the test suite perturbs every bar after the origin and asserts that no input changes; a quarantined

fixture that deliberately uses the next bar is rejected by the same test.

4.6. Splits

Sessions are split chronologically: training up to 2022-12-31, validation 2023, development test 2024 and an untouched final test 2025 evaluated once, with a one-session embargo at the start of each evaluation block. Held-out instruments are evaluated only with models trained on the pooled instruments. Training origins are two random eligible slots per instrument-session (a 1.3-million-example epoch is not materialised: windows are indexed into memory-mapped grids); evaluation origins are every ten minutes from 10:00, i.e. 36 per normal session.

5. The Multi-Resolution Patch Transformer

5.1. Inputs and tokenisation

Branch A reads the last 390 one-minute bars ending at the origin (crossing session boundaries, with boundary and gap flags), tokenised in patches of five bars (78 tokens). Branch B reads the last 390 completed 5-minute buckets (about five trading days; bucket j of the current session is complete only if slot $5j + 4 \leq m$), in patches of three (130 tokens). Branch C reads the last 260 completed 30-minute buckets (twenty days), in patches of two (130 tokens). No unfinished aggregate ever enters a window. Each patch is flattened, concatenated with its fraction of present bars and its maximal log gap, and projected through a gated linear unit (Dauphin et al., 2016) to d_{model}; a learned resolution embedding, a sinusoidal position measured as the distance to the origin in patch units, and a sinusoidal temporal-gap embedding are added. A global token projects the nine history-only statistics. Patches without any present bar remain tokens but are masked in attention.

5.2. Backbone, fusion and regime conditioning

The default backbone is a shared pre-norm Transformer encoder (Vaswani et al., 2017; Ba et al., 2016) over the concatenated token sequence (339 tokens with the global token), with GELU activations (Hendrycks and Gimpel, 2016) and key-padding masks. The alternative fusion keeps a small encoder per resolution and lets the one-minute stream cross-attend to the two coarser streams before a final encoder. A regime embedding maps the global statistics through a two-layer MLP (“stats”), through a soft codebook of eight learned regimes (“learned”), or is disabled (“none”); it is added to the decoder queries. The core model carries no ticker identity; a ticker embedding exists only as an ablation.

5.3. Direct decoder and Student-t head (Stage A)

Sixty learned future-position queries, each augmented with the known future calendar position (minute fraction, time to close, weekday, early-close flag) and the regime embedding, pass through a two-layer Transformer decoder that cross-attends to the encoder memory. A linear head yields three numbers per future minute, parametrised in a scale-equivariant way:

$$\mu_h = \sigma_t \, a_h/10, \qquad s_h = \sigma_t \, \mathrm{softplus}(b_h + 0.5413) + 10^{-7}, \qquad \nu_h = 2.05 + \mathrm{softplus}(c_h + 1.5), \quad (1)$$

so that the scale starts at σ_t and the degrees of freedom start near 3.75 with finite variance. The loss is the masked Student-t negative log-likelihood averaged over present target minutes, computed in single precision regardless of the device. Gaussian and quantile (pinball) heads are ablations. These are marginal predictions per minute; the cumulative-return distribution at H is obtained by Monte Carlo (256 draws) with independent minutes.

5.4. Dependence-aware paths (Stage B)

With the marginals frozen, a copula head produces, from the same decoder states, factor loadings $\Lambda \in \mathbb{R}^{H \times k}$ ($k = 4$) and a positive diagonal ψ, defining a correlation matrix $R = D^{-1/2}(\Lambda\Lambda^{\mathsf{T}} + \mathrm{diag}\,\psi)D^{-1/2}$ over the H future minutes. Training maximises the Gaussian-copula log-likelihood of $z_h = \Phi^{-1}(F_{t_{\nu_h}}((y_h - \mu_h)/s_h))$, with missing minutes marginalised exactly by replacing their rows and columns of R by the identity. Sampling draws $z = \Lambda\varepsilon + \psi^{1/2}\eta$, standardises, maps through Φ and the Student-t quantile function, and never needs a Cholesky factor. Joint samples give cumulative paths, maximum drawdown, realised volatility and interval coverage, scored against independent sampling of the same marginals.

5.5. Sizes and training

Three presets are measured, not assumed: Small ($d = 192$, 4 layers, 6 heads, FFN 768), the brief's specification ($d = 256$, 6 layers, 8 heads, FFN 1024) and Base ($d = 384$, 8 layers, 8 heads, FFN 1536), each with a two-layer decoder. Trainable parameter counts appear in Table 1. Optimisation uses AdamW (Loshchilov and Hutter, 2017) with $\beta = (0.9, 0.98)$, linear warm-up and cosine decay to 10 % of the peak rate, gradient clipping at 1, early stopping on the validation likelihood, fixed seeds and crash-safe checkpoints with resume. Mixed precision is not used in this version.

Table 1. Measured training throughput and memory on the compute node (Apple M3 Ultra, m3u96b, torch 2.15.0, mps; FP32)

Preset	Batch	Parameters	Steps/s	Ex
small	32	3,455,619	20.81	
small	64	3,455,619	12.21	
small	128	3,455,619	5.93	
spec	32	7,677,955	11.54	
spec	64	7,677,955	6.29	
spec	128	7,677,955	2.97	
base	32	20,718,339	6.11	
base	64	20,718,339	3.24	
base	128	20,718,339	1.51	
Data loader (64/batch)	0 workers: 3971 ex/s; 4 workers: 13016 ex/s; 8 workers: 38486 ex/s			

6. Experimental design

6.1. Baselines with a matched information set

Every baseline sees exactly the completed bars available at the origin and is tuned on the training and validation periods only. Non-neural: a random walk with zero drift and a Gaussian scale from the trailing root-mean-square one-minute return (the reference throughout); a seasonal-volatility model that sums the past-20-session profile of squared returns over the future slots, level-adjusted by the trailing 30-minute realised variance; an EWMA ($\lambda = 0.97$ per minute, Student-t with $\nu = 5$); HAR-RV at the minute scale (log realised variance over the next H minutes on log trailing 30-minute, one-day and five-day realised variances and time of day, by OLS, with ν from the kurtosis of standardised residuals); GARCH(1,1)-t on 5-minute training returns per instrument with the analytic variance term structure (Bollerslev, 1986; Hansen and Lunde, 2005); ridge, elastic net (Zou and Hastie, 2005) and LightGBM (Ke et al., 2017) on nineteen origin features, predicting the scaled cumulative return and the log scaled realised variance, assembled into a Student-t. Neural, on the one-minute branch only: an MLP on the last sixty bars and a dilated temporal convolutional network (Bai et al., 2018), both with the same Student-t head and loss as the main model, and the one-minute-only MRPT itself, which is the equal-parameter patch-Transformer control (Nie et al., 2022). Pretrained foundation models (Ansari et al., 2024; Das et al., 2023) are not evaluated: their weights were not available offline on the compute node, their tokenisers assume a single real-valued series without missingness masks, and their terms were not reviewed within this project; iTransformer (Liu et al., 2023) is not implemented. EGARCH forecasts beyond one step require simulation and were too slow for the number of evaluation origins. Table 10 lists the specifications.

6.2. Scores

The primary scores are the sample-based CRPS of the cumulative return at each horizon (256 draws, energy-form estimator), the weighted interval score (Bracher et al., 2021) over eleven central intervals, the mean pinball loss at seven quantile levels, empirical coverage of the 50/80/95 % central intervals, PIT dispersion and Kolmogorov–Smirnov statistics (Gneiting et al., 2007), and sharpness (90 % interval width). For parametric forecasts the closed-form Student-t CRPS (Jordan et al., 2019) and the log score are also computed; the log score is reported only where comparable. Secondary: absolute and squared errors of the cumulative return, QLIKE on the realised variance with positive forecasts (Patton, 2011), the Brier score of the tail event, directional accuracy against the zero-drift reference, and compute cost. Path-level scores (energy and variogram, $p = 0.5$; Scheuerer and Hamill, 2015) and CRPS of maximum drawdown and realised volatility are computed for the Stage-B comparison. We report scaled CRPS ($\mathrm{CRPS}/(\sigma_t\sqrt{H})$) to make instruments and horizons comparable; unscaled values are in the appendix.

6.3. Inference

All intervals are trading-day cluster bootstraps (all instruments and origins of one session form one cluster, 500–2000 resamples). Pairwise comparisons use Diebold–Mariano statistics on the loss differential ordered chronologically with a Bartlett-kernel HAC variance whose lag equals the overlap of consecutive evaluation windows times the number of instruments, so that overlapping origins and cross-sectional dependence are both absorbed (Diebold and Mariano, 1995; Newey and West, 1987); Holm and Benjamini–Hochberg adjustments are applied across horizons and models. Subgroup results (asset class, dollar-volume and volatility tercile, sector, time of day, year, stress days defined by the top decile of cross-sectional σ_t) use the same machinery. Economic evaluation (Hypothesis 5) trades the sign of the predictive median only when its magnitude exceeds an assumed round-trip cost, executes at the next bar with a one-minute delay, and reports gross and net returns per trade with day-clustered intervals under several cost assumptions; it is a sensitivity analysis, not a strategy.

6.4. Ablations and falsification

At the Small scale on the development test: resolutions (1m only, 1m+5m, 1m+5m+30m, equal parameters), fusion (shared vs cross-attention), head (Gaussian vs Student-t), regime (none, statistics, learned), feature sets (returns only, OHLCV without relative volume, time only), ticker embedding, three seeds, shuffled minute order inside the one-minute window at training and evaluation (a falsification: if the model only uses marginal statistics, shuffling should not hurt), the copula head against independent sampling, and model size (Small, Base). Zero-shot evaluation on the sixty

held-out instruments and the untouched 2025 period complete the protocol.

7. Results

7.1. Pilot

The pilot (20 instruments, September–December 2022, test on the last ten sessions, 4659 origins) validated the pipeline from export to LaTeX table and gave a first, deliberately unweighted, reading: every volatility-aware model improved on the random walk by 2–3 % in CRPS and the multi-resolution MRPT, its one-minute-only variant, a TCN and an MLP were within 0.2 % of one another (Table 20, Appendix). Nothing was tuned on it.

7.2. Development test (2024): all models beat the random walk, by about four percent

Table 2 reports the scaled CRPS of the cumulative return on the 986468 evaluation origins of 2024 (240 pooled instruments, 24 origins per session), with trading-day cluster-bootstrap intervals, and Table 3 the relative change against the random walk with its Diebold–Mariano statistic. Figure 1 summarises both. Three facts stand out.

First, the random walk with a trailing one-minute scale is beaten by every model at every horizon, by 3.5–4.9 %, with DM statistics between −25 and −72: minute-level history carries real information about the *dispersion* of the next hour, and the improvement is essentially a volatility improvement (directional accuracy of every location forecast stays at 0.50).

Second, the neural family (MRPT-S, MRPT-B, TCN, MLP) is the best at five minutes (scaled CRPS 0.4531 for MRPT-S against 0.4615 for HAR-RV, a difference of -1.91 % with $t = -58.1$), whereas the feature-based volatility regressions—HAR-RV, ridge, elastic net and LightGBM—are the best from fifteen minutes on (0.4166 against 0.4187 at sixty minutes; MRPT-S is worse by +0.65 %, $t = 10.4$). MRPT-S beats GARCH(1,1)-t, EWMA, the TCN and the MLP at every horizon (by 0.1–0.9 %, all Holm-adjusted $p < 10^{-6}$).

Third, and central to the paper, **Hypothesis 1 is rejected in economic terms**: the multi-resolution MRPT-S improves on its one-minute-only twin by -0.04 % at five minutes and -0.10 % at sixty minutes. The differences are statistically detectable ($t = -5.2$ at $H = 60$, 986468 origins over 252 days) but they are an order of magnitude smaller than the gap between any neural model and the random walk, and smaller than the gap between the neural models and HAR-RV at sixty minutes. Five days of 5-minute aggregates and twenty days of 30-minute aggregates add almost nothing that the last 390 one-minute bars do not already contain. Size does not matter either: the 20.7-million-parameter Base model and the 3.5-million-parameter Small model are indistinguishable (+0.04 % at $H = 60$, Holm $p = 0.1$).

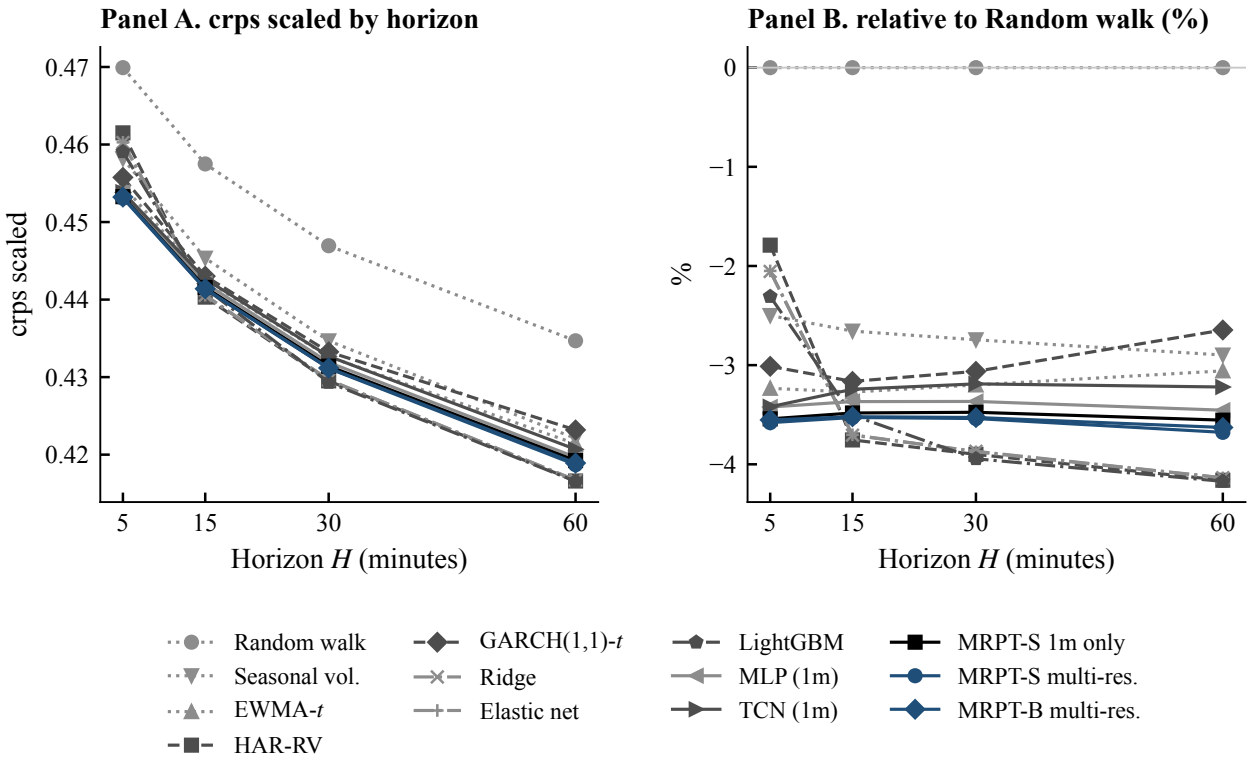


Figure 1. Development test 2024 (240 instruments, 986468 origins). Panel A: scaled CRPS of the cumulative-return forecast by horizon. Panel B: relative change versus the random walk. Source: `results/metrics/main/leaderboard_test.csv`.

Table 2. Main: mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.4699 [0.4648, 0.4754]	0.4575 [0.4515, 0.4639]	0.4469 [0.4412, 0.4538]	0.4347 [0.4271, 0.4441]
Seasonal volatility	0.4582 [0.4527, 0.4641]	0.4453 [0.4390, 0.4519]	0.4347 [0.4285, 0.4418]	0.4221 [0.4139, 0.4320]
EWMA-*t*	0.4547 [0.4492, 0.4606]	0.4425 [0.4363, 0.4491]	0.4326 [0.4264, 0.4396]	0.4214 [0.4133, 0.4314]
HAR-RV	0.4615 [0.4563, 0.4671]	0.4403 [0.4338, 0.4471]	0.4295 [0.4230, 0.4369]	0.4166 [0.4080, 0.4270]
GARCH(1,1)-*t*	0.4558 [0.4502, 0.4616]	0.4430 [0.4367, 0.4497]	0.4333 [0.4271, 0.4403]	0.4232 [0.4153, 0.4328]
Ridge	0.4603 [0.4542, 0.4667]	0.4405 [0.4339, 0.4476]	0.4297 [0.4231, 0.4371]	0.4167 [0.4081, 0.4272]
Elastic net	0.4603 [0.4543, 0.4668]	0.4406 [0.4339, 0.4476]	0.4296 [0.4231, 0.4370]	0.4167 [0.4081, 0.4272]
LightGBM	0.4591 [0.4531, 0.4655]	0.4415 [0.4347, 0.4484]	0.4293 [0.4228, 0.4366]	0.4166 [0.4080, 0.4269]
MLP (1m)	0.4538 [0.4484, 0.4598]	0.4421 [0.4357, 0.4486]	0.4319 [0.4258, 0.4388]	0.4197 [0.4116, 0.4291]
TCN (1m)	0.4538 [0.4484, 0.4597]	0.4427 [0.4364, 0.4491]	0.4327 [0.4265, 0.4396]	0.4207 [0.4126, 0.4303]
MRPT-S (1m+5m+30m)	0.4531 [0.4476, 0.4590]	0.4414 [0.4351, 0.4479]	0.4311 [0.4249, 0.4381]	0.4187 [0.4105, 0.4284]
MRPT-S (1m only)	0.4533 [0.4478, 0.4592]	0.4416 [0.4352, 0.4481]	0.4314 [0.4251, 0.4382]	0.4192 [0.4110, 0.4290]
MRPT-B (1m+5m+30m)	0.4532 [0.4478, 0.4591]	0.4414 [0.4351, 0.4479]	0.4312 [0.4249, 0.4381]	0.4189 [0.4107, 0.4287]

Cluster-bootstrap (trading-day) 95% intervals in brackets.

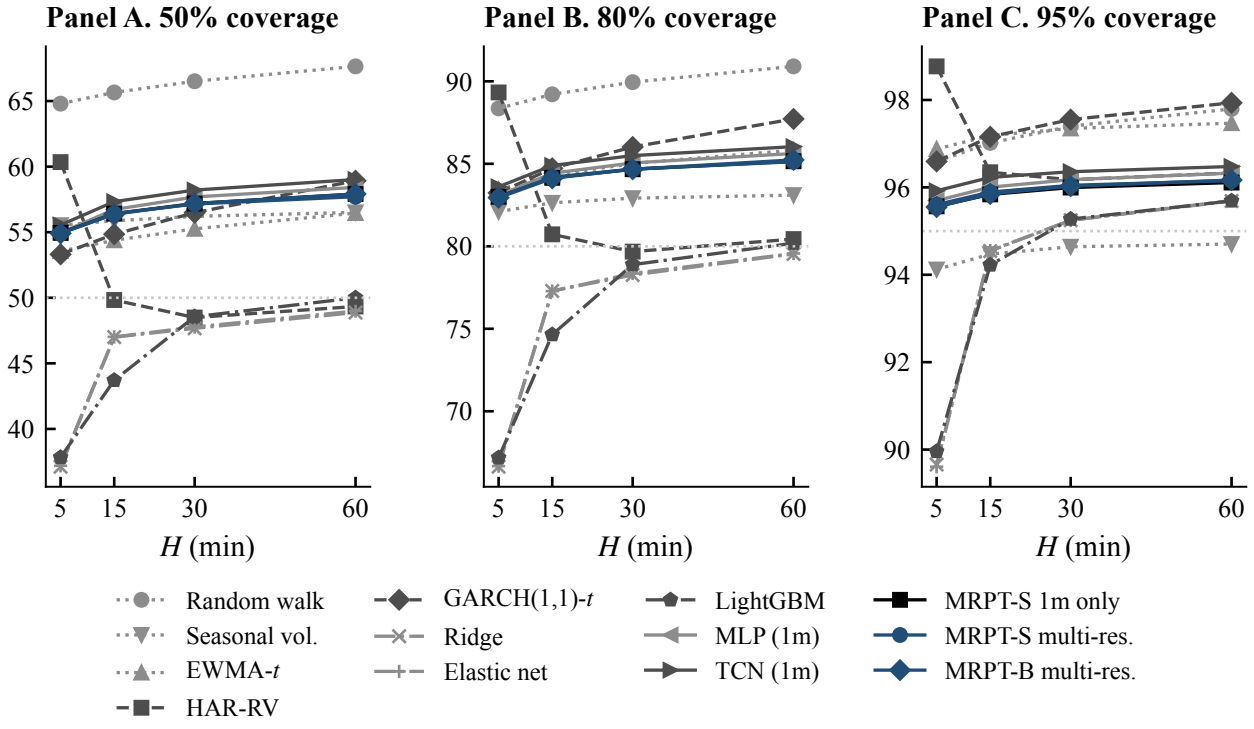


Figure 2. Empirical coverage of the central 50 %, 80 % and 95 % prediction intervals by horizon (dotted line = nominal), development test 2024.

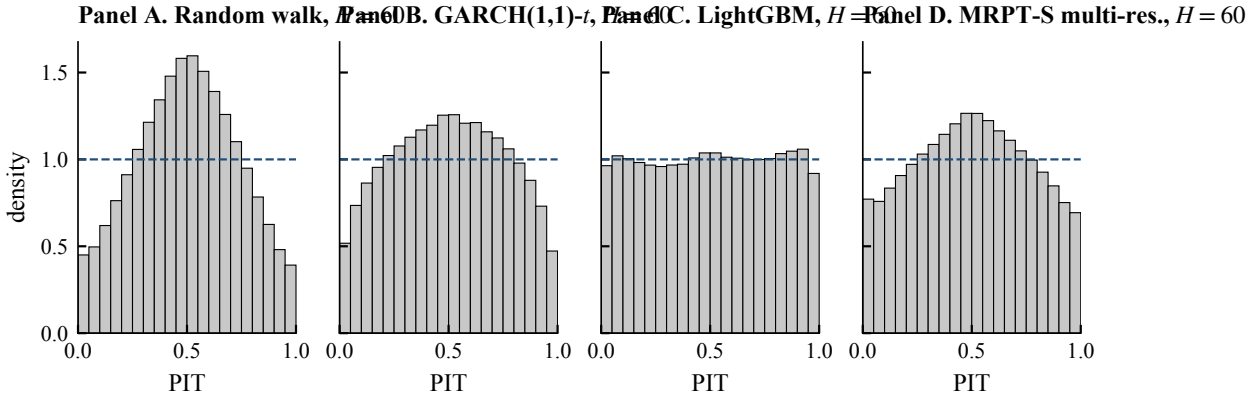


Figure 3. PIT histograms at $H = 60$ for the random walk, GARCH(1,1)-t, LightGBM and MRPT-S (dashed line = uniform density).

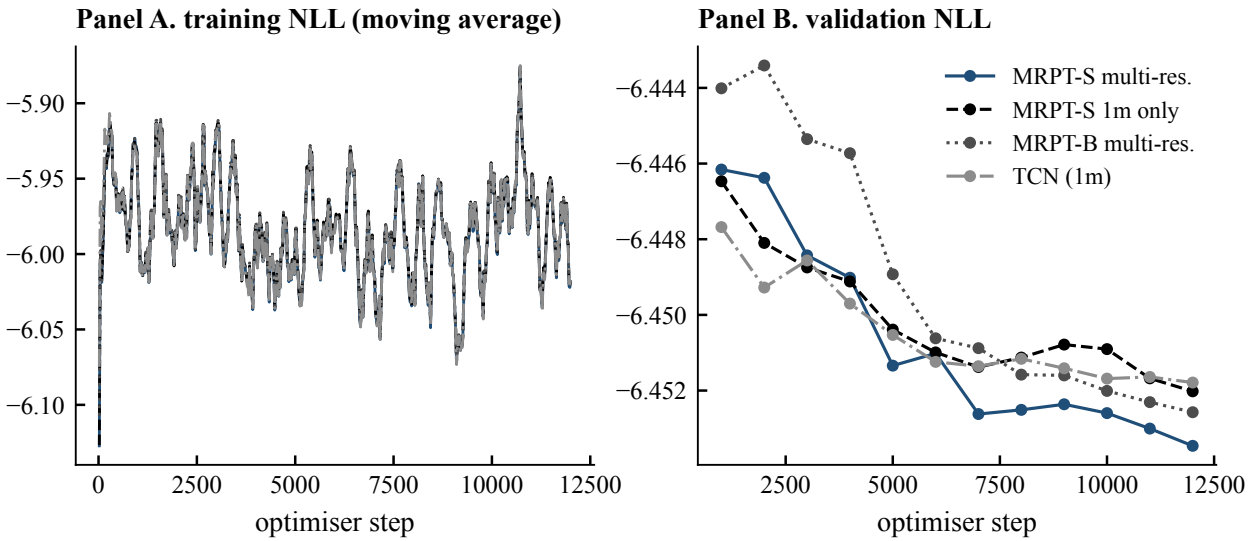


Figure 4. Training (moving average) and validation negative log-likelihood per optimiser step, main study.

Table 3. Main: CRPS relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Seasonal volatility	-2.89% (-48.8)	-3.09% (-49.9)	-3.22% (-39.3)	-3.44% (-30.7)
EWMA-*t*	-3.65% (-57.2)	-3.69% (-53.4)	-3.60% (-37.3)	-3.47% (-25.5)
HAR-RV	-2.19% (-35.7)	-4.36% (-64.1)	-4.56% (-47.4)	-4.89% (-36.6)
GARCH(1,1)-*t*	-3.60% (-57.1)	-3.77% (-56.0)	-3.70% (-40.5)	-3.40% (-27.9)
Ridge	-2.88% (-33.2)	-4.35% (-58.3)	-4.55% (-45.9)	-4.89% (-36.2)
Elastic net	-2.88% (-33.2)	-4.35% (-58.2)	-4.56% (-46.1)	-4.89% (-36.2)
LightGBM	-2.88% (-33.4)	-4.12% (-51.1)	-4.56% (-47.4)	-4.85% (-37.6)
MLP (1m)	-3.94% (-71.6)	-3.93% (-68.6)	-3.94% (-51.5)	-4.07% (-38.7)
TCN (1m)	-3.93% (-69.2)	-3.75% (-65.1)	-3.68% (-48.7)	-3.73% (-37.3)
MRPT-S (1m+5m+30m)	-4.05% (-69.3)	-4.03% (-67.0)	-4.05% (-50.6)	-4.27% (-39.7)
MRPT-S (1m only)	-4.02% (-69.2)	-4.00% (-66.9)	-4.00% (-50.4)	-4.17% (-39.6)
MRPT-B (1m+5m+30m)	-4.04% (-69.5)	-4.04% (-67.2)	-4.06% (-51.3)	-4.23% (-39.8)

Relative change in mean CRPS versus the reference (negative = better); Diebold–Mariano t-statistic with HAC variance in parentheses.

7.3. Calibration

Table 4 and Figure 2 show empirical coverage of the 50/80/95 % central intervals and the PIT Kolmogorov–Smirnov statistic. The random walk is over-dispersed everywhere (50 % intervals cover 68 % of outcomes at sixty minutes). The Student-t neural models and GARCH/EWMA are nearly nominal at 95 % but over-dispersed in the centre (50 % coverage 0.55–0.59): their degrees of freedom and scale produce tails that are right and a body that is too wide. HAR-RV, ridge and LightGBM are the best-calibrated models from fifteen minutes on (0.49/0.80/0.96 at sixty minutes, KS 0.008) and the worst at five minutes (50 % coverage 0.37, 95 % coverage 0.90): a lognormal variance regression on 390-minute features cannot react to the last five minutes. PIT histograms (Figure 3) make the same point visually. Scores and calibration thus tell one consistent story: at the shortest horizon the sequence models' ability to read the very last bars wins; at longer horizons a well-specified heterogeneous variance regression with a kurtosis-matched Student-t is both sharper and better calibrated.

Table 4. Main: empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
Random walk (trailing RMS)	0.65	0.88	0.97	0.078	0.66	0.89	0.97	0.079	0.67	0.90	0.97	0.083	0.68	0.91	0.98	0.089
Seasonal volatility	0.56	0.82	0.94	0.031	0.56	0.83	0.94	0.031	0.56	0.83	0.95	0.033	0.57	0.83	0.95	0.034
EWMA-t	0.54	0.83	0.97	0.023	0.54	0.84	0.97	0.024	0.55	0.85	0.97	0.028	0.56	0.86	0.97	0.034
HAR-RV	0.60	0.89	0.99	0.058	0.50	0.81	0.96	0.008	0.48	0.80	0.96	0.009	0.49	0.80	0.96	0.008
GARCH(1,1)-t	0.53	0.83	0.97	0.021	0.55	0.85	0.97	0.027	0.56	0.86	0.98	0.035	0.59	0.88	0.98	0.047
Ridge	0.37	0.67	0.90	0.076	0.47	0.77	0.95	0.018	0.48	0.78	0.95	0.015	0.49	0.80	0.96	0.009
Elastic net	0.37	0.67	0.90	0.076	0.47	0.77	0.95	0.018	0.48	0.78	0.95	0.014	0.49	0.80	0.96	0.009
LightGBM	0.38	0.67	0.90	0.073	0.44	0.75	0.94	0.036	0.49	0.79	0.95	0.010	0.50	0.80	0.96	0.008
MLP (1m)	0.55	0.83	0.96	0.029	0.57	0.84	0.96	0.034	0.58	0.85	0.96	0.039	0.58	0.86	0.96	0.043
TCN (1m)	0.56	0.84	0.96	0.032	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.042	0.59	0.86	0.96	0.046
MRPT-S (1m+5m+30m)	0.55	0.83	0.96	0.029	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.040
MRPT-S (1m only)	0.55	0.83	0.96	0.030	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.041
MRPT-B (1m+5m+30m)	0.55	0.83	0.96	0.029	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.041

7.4. Subgroups

Splitting the MRPT-S versus HAR-RV contrast at sixty minutes by subgroup, HAR-RV's advantage is concentrated in low-dollar-volume stocks (+1.3 %), low-volatility instruments (+1.0 %), transport and utilities (+1.9 %), fixed-income ETFs (+1.8 %) and the stocks without a SIC code (+2.8 %); MRPT-S is better on broad-equity, sector and volatility-linked ETFs (−0.3 to −1.1 %) and statistically tied on high-dollar-volume instruments. Afternoon origins favour HAR-RV (+1.0 %), morning origins much less (+0.2 %). On stress days (top decile of cross-sectional σ_t, 100846 origins) the ordering is unchanged and every model's coverage moves by at most one percentage point: none of the models breaks under stress, and none gains a special advantage there.

7.5. Economic content (Hypothesis 5)

The predictive medians carry no usable directional information. Their correlation with the realised sixty-minute return is 0.003 (MRPT-S), 0.025 (MRPT-B), 0.014 (TCN) and 0.007 (MLP), the typical magnitude of the median is below one basis point, and a sign rule that trades only when the median exceeds an assumed round-trip cost of 1 basis point trades 17 % of the origins for a net return of 0.0 bp $[-1.7, +1.8]$; at 2.5–10 bp the traded share falls below 4 % and every interval contains zero. We report this null without further sensitivity analysis.

7.6. Zero-shot transfer and the untouched 2025 period

Hypothesis 2 asks whether pooled training transfers. On the sixty instruments never seen in training (250322 origins in 2024, Table 5), MRPT-S improves on the random walk by -4.12 % at five minutes and -4.34 % at sixty, against -4.05 % and -4.27 % on the pooled instruments: there is no transfer penalty. Against the baselines fitted on those same held-out instruments' own histories, MRPT-S beats GARCH(1,1)-t by 0.9–1.6 %, EWMA by 0.6–1.2 %, the TCN and the MLP by 0.1–0.4 %, HAR-RV by -1.16 % at five minutes, and is statistically tied with HAR-RV, ridge and LightGBM from thirty minutes on (Holm-adjusted $p = 1$). The multi-resolution gain over the one-minute-only twin is again a tenth of a percent at most. Transfer therefore holds in the only sense the data support: a pooled model without ticker identity forecasts unseen liquid instruments as well as it forecasts the instruments it was trained on, and as well as—not better than—a heterogeneous variance regression refitted on the new instrument.

The untouched 2025 period (Table 6; 1008425 pooled origins and 254334 held-out origins, models and baselines frozen from the main study) repeats the 2024 picture almost line by line. All models beat the random walk by 3.8–4.4 %; the neural models lead at five minutes (0.4594 against 0.4674 for HAR-RV, -1.79 %); HAR-RV, ridge and LightGBM lead from fifteen minutes on (0.4232 against 0.4256, +0.48 %, $t = 6.2$). The multi-resolution effect vanishes entirely out of sample: MRPT-S versus its one-minute-only twin is +0.01 % at five minutes and +0.01 % at sixty minutes (Holm $p = 1$), and on the held-out instruments in 2025 every one of the four contrasts has $p = 1$. Base and Small remain indistinguishable. One nuance appears on 2025 stress days (top decile of cross-sectional volatility, 101561 origins): HAR-RV's 50/80/95 % intervals cover 46/77/95 % of outcomes while MRPT-S covers 55/84/95 %; the regression's variance forecasts lag on the days when volatility jumps, and the sequence model's wider body is, on those days, closer to the truth at 80 %, with the same CRPS (0.4559 versus 0.4557 scaled).

Table 5. Main (held-out instruments): mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.4744 [0.4686, 0.4803]	0.4653 [0.4584, 0.4716]	0.4541 [0.4468, 0.4617]	0.4415 [0.4316, 0.4521]
Seasonal volatility	0.4623 [0.4562, 0.4686]	0.4531 [0.4458, 0.4602]	0.4419 [0.4342, 0.4500]	0.4289 [0.4184, 0.4399]
EWMA-t	0.4593 [0.4533, 0.4655]	0.4508 [0.4435, 0.4579]	0.4404 [0.4327, 0.4485]	0.4287 [0.4184, 0.4398]
HAR-RV	0.4616 [0.4556, 0.4677]	0.4490 [0.4414, 0.4563]	0.4378 [0.4299, 0.4461]	0.4248 [0.4142, 0.4360]
GARCH(1,1)-t	0.4607 [0.4545, 0.4670]	0.4517 [0.4443, 0.4591]	0.4415 [0.4339, 0.4496]	0.4306 [0.4203, 0.4414]
Ridge	0.4621 [0.4554, 0.4689]	0.4494 [0.4416, 0.4568]	0.4376 [0.4296, 0.4461]	0.4242 [0.4136, 0.4355]
Elastic net	0.4621 [0.4554, 0.4689]	0.4494 [0.4417, 0.4569]	0.4376 [0.4296, 0.4461]	0.4242 [0.4135, 0.4354]
LightGBM	0.4647 [0.4581, 0.4715]	0.4496 [0.4419, 0.4571]	0.4382 [0.4302, 0.4467]	0.4248 [0.4142, 0.4359]
MLP (1m)	0.4582 [0.4520, 0.4646]	0.4499 [0.4428, 0.4569]	0.4393 [0.4317, 0.4470]	0.4264 [0.4162, 0.4371]
TCN (1m)	0.4578 [0.4516, 0.4641]	0.4497 [0.4425, 0.4567]	0.4392 [0.4316, 0.4471]	0.4265 [0.4164, 0.4373]
MRPT-S (1m+5m+30m)	0.4571 [0.4510, 0.4634]	0.4488 [0.4415, 0.4559]	0.4381 [0.4304, 0.4460]	0.4250 [0.4150, 0.4357]
MRPT-S (1m only)	0.4573 [0.4511, 0.4636]	0.4490 [0.4417, 0.4561]	0.4384 [0.4307, 0.4463]	0.4255 [0.4154, 0.4364]
MRPT-B (1m+5m+30m)	0.4573 [0.4512, 0.4635]	0.4490 [0.4417, 0.4561]	0.4383 [0.4306, 0.4462]	0.4253 [0.4151, 0.4362]

Cluster-bootstrap (trading-day) 95% intervals in brackets.

Table 6. Final: mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.4759 [0.4683, 0.4828]	0.4621 [0.4543, 0.4697]	0.4521 [0.4439, 0.4602]	0.4409 [0.4316, 0.4512]
Seasonal volatility	0.4655 [0.4575, 0.4732]	0.4505 [0.4427, 0.4585]	0.4406 [0.4319, 0.4497]	0.4291 [0.4188, 0.4401]
EWMA-t	0.4609 [0.4529, 0.4682]	0.4470 [0.4391, 0.4548]	0.4380 [0.4293, 0.4469]	0.4278 [0.4175, 0.4387]
HAR-RV	0.4674 [0.4597, 0.4744]	0.4450 [0.4368, 0.4532]	0.4352 [0.4259, 0.4444]	0.4232 [0.4125, 0.4347]
GARCH(1,1)-t	0.4616 [0.4537, 0.4690]	0.4468 [0.4388, 0.4547]	0.4380 [0.4292, 0.4468]	0.4286 [0.4183, 0.4397]
Ridge	0.4663 [0.4577, 0.4743]	0.4449 [0.4365, 0.4532]	0.4350 [0.4257, 0.4443]	0.4232 [0.4126, 0.4346]
Elastic net	0.4663 [0.4577, 0.4743]	0.4449 [0.4366, 0.4532]	0.4351 [0.4258, 0.4443]	0.4232 [0.4126, 0.4346]
LightGBM	0.4655 [0.4570, 0.4734]	0.4457 [0.4374, 0.4541]	0.4347 [0.4254, 0.4439]	0.4232 [0.4125, 0.4345]
MLP (1m)	0.4600 [0.4519, 0.4672]	0.4465 [0.4385, 0.4545]	0.4371 [0.4282, 0.4458]	0.4260 [0.4158, 0.4369]
TCN (1m)	0.4600 [0.4520, 0.4671]	0.4471 [0.4391, 0.4550]	0.4379 [0.4291, 0.4467]	0.4271 [0.4169, 0.4380]
MRPT-S (1m+5m+30m)	0.4594 [0.4514, 0.4666]	0.4462 [0.4382, 0.4540]	0.4368 [0.4278, 0.4456]	0.4256 [0.4154, 0.4366]
MRPT-S (1m only)	0.4594 [0.4514, 0.4666]	0.4461 [0.4381, 0.4540]	0.4367 [0.4278, 0.4456]	0.4257 [0.4154, 0.4367]
MRPT-B (1m+5m+30m)	0.4595 [0.4516, 0.4667]	0.4462 [0.4383, 0.4541]	0.4368 [0.4279, 0.4456]	0.4257 [0.4155, 0.4366]

Cluster-bootstrap (trading-day) 95% intervals in brackets.

Table 7. Final: empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
Random walk (trailing RMS)	0.64	0.88	0.96	0.074	0.66	0.89	0.97	0.080	0.66	0.90	0.97	0.083	0.68	0.91	0.98	0.089
Seasonal volatility	0.56	0.82	0.94	0.032	0.56	0.83	0.95	0.034	0.56	0.83	0.95	0.035	0.57	0.83	0.95	0.037
EWMA-t	0.53	0.83	0.97	0.021	0.54	0.84	0.97	0.025	0.55	0.85	0.97	0.029	0.56	0.86	0.97	0.035
HAR-RV	0.60	0.89	0.99	0.055	0.50	0.81	0.96	0.008	0.48	0.80	0.96	0.013	0.49	0.80	0.96	0.010
GARCH(1,1)-t	0.52	0.82	0.96	0.020	0.54	0.84	0.97	0.021	0.55	0.85	0.97	0.028	0.57	0.86	0.97	0.039
Ridge	0.37	0.67	0.90	0.076	0.47	0.77	0.95	0.018	0.48	0.78	0.95	0.018	0.49	0.79	0.96	0.013
Elastic net	0.37	0.67	0.90	0.076	0.47	0.77	0.95	0.019	0.48	0.78	0.95	0.018	0.49	0.80	0.96	0.013
LightGBM	0.38	0.67	0.90	0.075	0.44	0.75	0.94	0.037	0.48	0.79	0.95	0.014	0.50	0.80	0.96	0.012
MLP (1m)	0.55	0.83	0.96	0.027	0.57	0.85	0.96	0.036	0.58	0.85	0.96	0.041	0.58	0.86	0.96	0.044
TCN (1m)	0.55	0.84	0.96	0.030	0.58	0.85	0.96	0.039	0.58	0.86	0.96	0.042	0.59	0.86	0.96	0.047
MRPT-S (1m+5m+30m)	0.55	0.83	0.96	0.028	0.57	0.85	0.96	0.035	0.57	0.85	0.96	0.038	0.58	0.85	0.96	0.041
MRPT-S (1m only)	0.55	0.83	0.96	0.027	0.57	0.84	0.96	0.034	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.040
MRPT-B (1m+5m+30m)	0.55	0.83	0.96	0.028	0.57	0.84	0.96	0.035	0.57	0.85	0.96	0.039	0.58	0.85	0.96	0.041

8. Ablations and falsification

Every variant in Table 8 is a Small model trained for 4000 steps with the same seed and compared, on the 2024 development test, with a reference trained under exactly the same budget (MRPT-S, 4k); two further seeds of the reference give the noise floor. Figure 5 shows the relative CRPS changes at sixty minutes with their day-cluster intervals.

Table 8. Ablations: mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.4699 [0.4648, 0.4754]	0.4575 [0.4515, 0.4639]	0.4469 [0.4412, 0.4538]	0.4347 [0.4271, 0.4441]
Seasonal volatility	0.4582 [0.4527, 0.4641]	0.4453 [0.4390, 0.4519]	0.4347 [0.4285, 0.4418]	0.4221 [0.4139, 0.4320]
EWMA-t	0.4547 [0.4492, 0.4606]	0.4425 [0.4363, 0.4491]	0.4326 [0.4264, 0.4396]	0.4214 [0.4133, 0.4314]
MRPT-S (1m+5m)	0.4537 [0.4482, 0.4596]	0.4418 [0.4355, 0.4483]	0.4316 [0.4253, 0.4384]	0.4191 [0.4110, 0.4289]
MRPT-S cross-attention fusion	0.4535 [0.4480, 0.4594]	0.4416 [0.4353, 0.4481]	0.4313 [0.4251, 0.4382]	0.4190 [0.4109, 0.4288]
MRPT-S Gaussian head	nan	nan	nan	nan
MRPT-S no regime	0.4534 [0.4479, 0.4593]	0.4414 [0.4352, 0.4480]	0.4312 [0.4249, 0.4380]	0.4188 [0.4107, 0.4286]
MRPT-S learned regime	0.4531 [0.4476, 0.4591]	0.4410 [0.4347, 0.4476]	0.4306 [0.4243, 0.4376]	0.4182 [0.4100, 0.4281]
MRPT-S returns only	0.4542 [0.4487, 0.4601]	0.4424 [0.4362, 0.4489]	0.4322 [0.4260, 0.4392]	0.4201 [0.4119, 0.4299]
MRPT-S time only	0.4542 [0.4488, 0.4601]	0.4424 [0.4362, 0.4489]	0.4323 [0.4260, 0.4392]	0.4200 [0.4119, 0.4299]
MRPT-S OHLCV (no rel. volume)	0.4536 [0.4481, 0.4595]	0.4418 [0.4355, 0.4482]	0.4316 [0.4253, 0.4385]	0.4193 [0.4113, 0.4291]
MRPT-S shuffled minutes	0.4538 [0.4483, 0.4598]	0.4420 [0.4357, 0.4484]	0.4318 [0.4255, 0.4388]	0.4196 [0.4115, 0.4294]
MRPT-S + ticker id	0.4534 [0.4479, 0.4593]	0.4418 [0.4356, 0.4483]	0.4318 [0.4256, 0.4387]	0.4199 [0.4118, 0.4297]
MRPT-S + copula (joint sampling)	0.4534 [0.4479, 0.4593]	0.4417 [0.4355, 0.4481]	0.4316 [0.4254, 0.4385]	0.4194 [0.4113, 0.4291]
MRPT-S reference (4k steps)	0.4533 [0.4478, 0.4592]	0.4415 [0.4352, 0.4480]	0.4312 [0.4249, 0.4381]	0.4188 [0.4107, 0.4286]
MRPT-S 1m only (4k)	0.4538 [0.4483, 0.4597]	0.4420 [0.4358, 0.4485]	0.4318 [0.4256, 0.4386]	0.4195 [0.4113, 0.4293]
MRPT-S reference, seed 2	0.4531 [0.4476, 0.4591]	0.4411 [0.4347, 0.4476]	0.4308 [0.4245, 0.4379]	0.4187 [0.4103, 0.4287]
MRPT-S reference, seed 3	0.4531 [0.4476, 0.4590]	0.4413 [0.4350, 0.4479]	0.4312 [0.4250, 0.4380]	0.4194 [0.4115, 0.4290]
MRPT-S + copula, independent sampling	0.4535 [0.4480, 0.4593]	0.4416 [0.4354, 0.4481]	0.4314 [0.4252, 0.4383]	0.4191 [0.4110, 0.4288]

Cluster-bootstrap (trading-day) 95% intervals in brackets.

Resolutions (H1 again). At this shorter budget removing the 5- and 30-minute branches costs +0.12 to +0.16 % of CRPS, removing only the 30-minute branch +0.10 %; the seed-to-seed spread of the reference is −0.03 to +0.16 %. Multi-resolution context is therefore a real but marginal regulariser at small budgets and, as Section 7 showed, nothing at all at the full budget and out of sample.

What the model actually uses. The most informative ablations are the destructive ones. A model that sees *only* the calendar and missingness flags—no returns, no ranges, no volumes, just the trailing scale σ_t in its global token and the time of day—is worse than the full model by only +0.20 to +0.29 %. Returns-only inputs cost +0.19 to +0.28 %; dropping relative volume +0.08 to +0.14 %; shuffling the order of the 390 one-minute bars at training and evaluation time +0.13 to +0.21 %. In other words, of the ≈ 4.3 % improvement over the random walk at sixty minutes, roughly nine tenths come from the trailing volatility level and the intraday seasonal, which every baseline also has, and at most three tenths of a percent from any minute-level *structure*—order, shape of bars,

volume—that a sequence model can extract. The falsification test passes in the sense intended: shuffling does hurt, so the model reads order; it just does not have much to read.

Fusion, head, regime (H3), identity. Cross-attention fusion of separate per-scale encoders is slightly worse than the shared backbone (+0.05 %). The Gaussian head is worse than the Student-t head (see Table 8). Removing the regime embedding changes nothing (+0.02 to +0.03 %, Holm $p > 0.3$ at 30 and 60 minutes), and coverage on stress days is identical with and without it (0.57/0.85/0.97 at $H = 60$): Hypothesis 3 is rejected—the global token already carries the regime statistics, and conditioning the decoder on them again adds nothing. A learned eight-code regime codebook improves CRPS by 0.03–0.15 % and lowers central coverage toward nominal (0.56 at 50 %), the only architectural change with a consistent, if small, benefit. A ticker-identity embedding hurts (+0.05 to +0.31 %): instrument identity over-fits at this scale, which is also why the core model has none.

Joint paths (H4). The copula head trained on the probability-integral transforms of the Stage-A marginals converges to a copula log-likelihood of -2.5×10^{-5} per dimension, i.e. to independence, and its loadings stay near zero. Correlated sampling changes the energy score by −0.003 %, the variogram score by −0.002 %, the CRPS of maximum drawdown by −0.01 % and of realised variance by +0.10 %, all with intervals containing zero (Table 9); the CRPS of the cumulative return is +0.08 % worse at sixty minutes. Hypothesis 4 is rejected: once the marginals are conditioned on the recent history, the next sixty minute returns are, to the resolution of these scores, conditionally independent, and a dependence-aware decoder has nothing to model. We did not pursue diffusion decoders.

Table 9. Joint paths versus independent sampling of the same Student-t marginals (MRPT-S + copula, 2024 development test, 986468 origins, 60-minute paths, 200 draws; day-cluster bootstrap 95 % intervals of the difference).

Score	Joint (copula)	Independent	Joint − independent
Energy score	3.490e-3	3.490e-3	-9.2×10^{-8} [-2.3×10^{-7}, $+5.0 \times 10^{-8}$]
Variogram score ($p = 0.5$)	1.5113e-4	1.5114e-4	-3.8×10^{-9} [-1.0×10^{-8}, $+2.2 \times 10^{-9}$]
CRPS, maximum drawdown	1.4632e-3	1.4633e-3	-1.8×10^{-7} [-6.2×10^{-7}, $+2.8 \times 10^{-7}$]
CRPS, realised variance	1.1881e-5	1.1869e-5	$+1.2 \times 10^{-8}$ [-2.8×10^{-9}, $+3.0 \times 10^{-8}$]

9. Limitations

Survivorship. The universe requires six months of presence after the sample start and, more importantly, 80 % regular-hours minute coverage over 2018–2025; only five delisted candidates

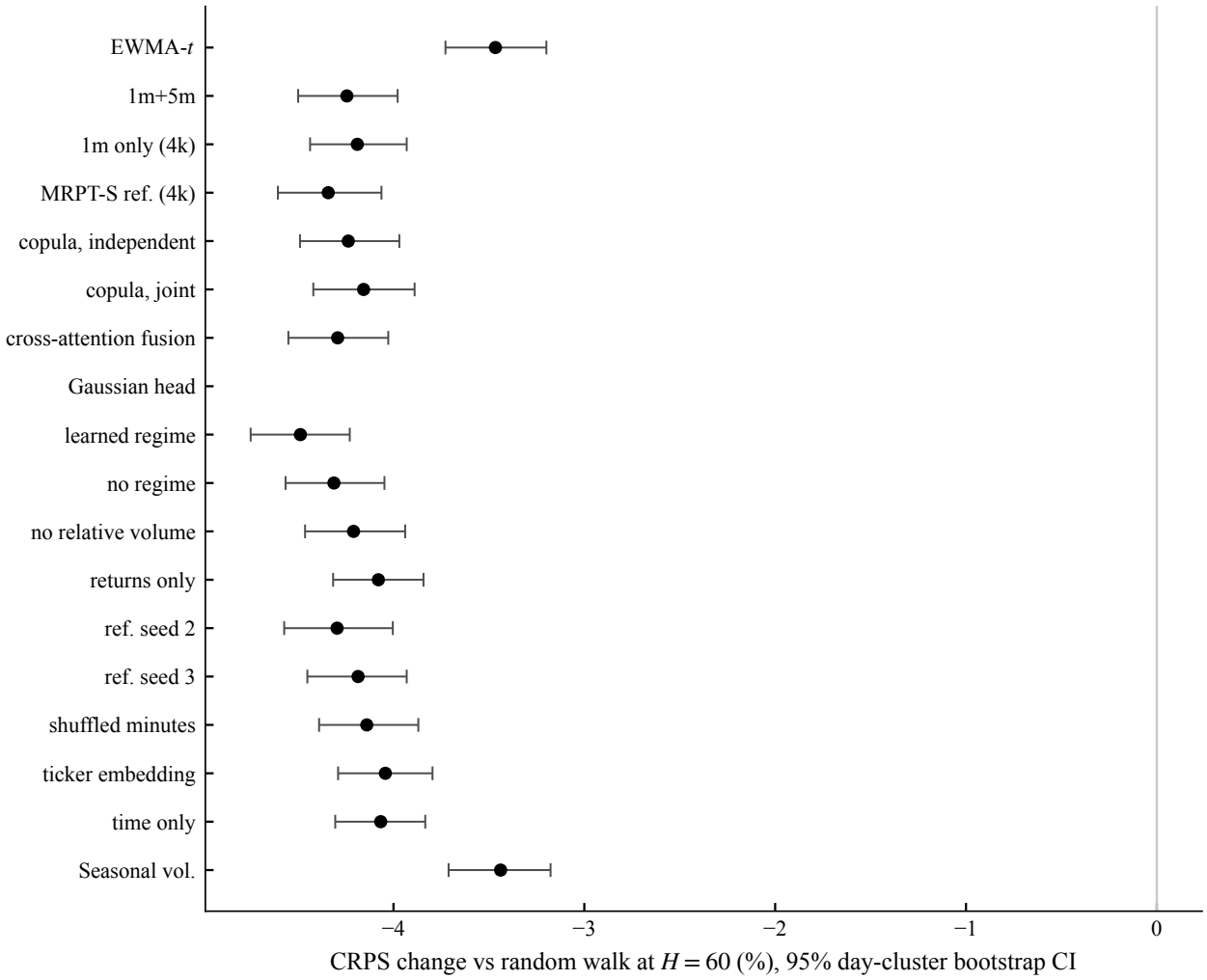


Figure 5. Ablations at the Small scale (4 000 steps each): relative CRPS change at $H = 60$ versus the random walk, with 95 % day-cluster bootstrap intervals; the equal-budget reference and its two extra seeds are included.

passed because the vendor's delisted one-minute archives are only partly ingested in FinData OS. The results therefore describe liquid instruments that survived the sample window. Nothing in the architecture depends on this choice, but the transfer claim (Hypothesis 2) is a claim about unseen *liquid survivors*, not about the whole market.

Bars are not quotes. FirstRate bars carry no bid–ask spread, trades or order flow. Microstructure information that is known to matter for minute-level prediction (Sirignano and Cont, 2019; Kolm et al., 2023) is absent, and the economic evaluation of Hypothesis 5 rests on assumed costs.

Regular hours only. Origins whose horizon crosses the close are excluded, extended hours are not studied, and the closing print is excluded from context and targets. The model has never seen the first thirty minutes as an origin (the first origin is at 10:00), although those minutes are in its context.

Missingness and targets. Missing minutes are masked rather than filled; the target of a minute that follows a gap is the return accrued since the last observed bar, so its scale is slightly larger than a one-minute return. With coverage above 80 % and the 80 % future-presence rule this affects a small fraction of targets, and it affects all models identically, but it is a modelling convention, not a property of the market.

Corporate actions. Splits are applied explicitly; dividends are not, and the overnight gap across an ex-dividend date is left raw. Mergers and spin-offs end a series without notice.

Compute and scale. A single Apple Silicon node without mixed precision bounds the number of optimiser steps and seeds. The brief's upper range (50–100 million parameters) was not explored; the Base model is the largest size trained. Training-time estimates are measured, not extrapolated, but the comparison between sizes is at matched steps rather than matched wall-clock.

Baselines not run. Pretrained time-series foundation models, iTransformer and EGARCH were not evaluated (Section 6); their absence is a limitation of this version, not evidence about them.

Multiplicity and overlap. Evaluation origins overlap within a session and are cross-sectionally dependent; all inference uses day clusters and HAC variances, and multiple-testing adjustments are reported, but the effective sample size is far smaller than the number of origins suggests. Readers should weigh interval estimates, not point estimates.

10. Reproducibility and release

The repository `wp41_uqo` contains the package `chronos_x` (data connector, calendar, quality checks, corporate actions, grids and windows; models; training; scoring and inference; baselines), the numbered scripts that produced every artifact cited here, the configurations that freeze the protocol (`configs/pilot`, `main`, `ablations`, `baselines`, `final`), the test suite (calendar and timestamp conventions, grid construction on synthetic bars, leakage gate with a quarantined future-feature fixture, model shapes and parameter counts, numerical checks of the Student-t density and CRPS against Monte Carlo, copula sampling and likelihood, live API contract tests), the universe manifest with its hash, the provenance logs of every API call, the phase reports, and the LaTeX source whose tables and macros are generated from the saved results. A synthetic cache in the exact FinData schema lets anyone run the entire pipeline without a license (`configs/synthetic_smoke.yaml`). Raw bars, grids and per-origin returns never leave the compute node; the released artifacts are code, aggregate scores and figures of scores. The FinData dataset versions used are 17375 (stocks) and 5162 (ETFs); the API audit and the coverage audit are saved with the results. Seeds, checkpoints and the exact commands are listed in `docs/REPRODUCIBILITY.md`.

11. Conclusion

We asked whether a compact, transferable, multi-resolution Transformer can learn minute-level structure that improves calibrated forecasts of the next 5 to 60 minutes, and we tested the question against matched baselines, with proper scoring rules, on 234 million one-minute bars of 300 U.S. stocks and ETFs, with sixty unseen instruments and an untouched year. The answers are precise. Minute-level history improves distributional forecasts over a trailing-scale random walk by about four percent of CRPS at every horizon, for every model, and the improvement is a volatility improvement: no model has directional skill or economic content after costs. The sequence models win at five minutes; a heterogeneous autoregression of realised variance with a Student-t matched to residual kurtosis wins, and is better calibrated, from fifteen minutes on. The multi-resolution context that defines the architecture adds at most a tenth of a percent in sample and nothing out of sample; model size does not matter between 3.5 and 20.7 million parameters; regime conditioning adds nothing; coherent paths add nothing because the next hour's minute returns are conditionally independent given the recent history. Pooled training transfers to unseen instruments without penalty. And a model that reads only the trailing volatility and the clock recovers nine tenths of what the full model knows.

The central empirical statement of this paper is therefore not that a Transformer predicts financial markets. It is that, for liquid U.S. instruments at the one-minute scale and with bars as the only input,

the universal intraday representation worth learning is small: a scale, a clock, a heavy tail, and a few tenths of a percent of residual structure that sequence models and variance regressions extract in different proportions at different horizons. Two directions follow from the limitations rather than from the results: order-book or quote data, which the literature suggests carry the microstructure signal that bars lack, and the delisted instruments that this universe could not include. The package, the protocol and the null results are released so that both can be tested the same way.

References

Andersen, Torben G. and Tim Bollerslev, "Intraday periodicity and volatility persistence in financial markets," *Journal of Empirical Finance*, 1997, *4* (2-3), 115–158.

_ , _ , Francis X. Diebold, and Paul Labys, "Modeling and Forecasting Realized Volatility," *Econometrica*, 2003, *71* (2), 579–625.

Ansari, Abdul Fatir, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang, "Chronos: Learning the Language of Time Series," arXiv:2403.07815 2024.

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton, "Layer Normalization," arXiv:1607.06450 2016.

Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," arXiv:1803.01271 2018.

Bailey, David H. and Marcos López de Prado, "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality," 2019.

Barndorff-Nielsen, Ole E. and Neil Shephard, "Econometric Analysis of Realized Volatility and its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2002, *64* (2), 253–280.

Benjamini, Yoav and Yosef Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1995, *57* (1), 289–300.

Bollerslev, Tim, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, 1986, *31* (3), 307–327.

_ , Andrew J. Patton, and Rogier Quaedvlieg, "Exploiting the errors: A simple approach for improved volatility forecasting," *Journal of Econometrics*, 2016, *192* (1), 1–18.

Bracher, Johannes, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich, "Evaluating epidemic forecasts in an interval format," *PLOS Computational Biology*, 2021, *17* (2), e1008618.

Cont, R., "Empirical properties of asset returns: stylized facts and statistical issues," *Quantitative Finance*, 2001, *1* (2), 223–236.

Corsi, F., "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 2009, *7* (2), 174–196.

Das, Abhimanyu, Weihao Kong, Rajat Sen, and Yichen Zhou, "A decoder-only foundation model for time-series forecasting," arXiv:2310.10688 2023.

Dauphin, Yann N., Angela Fan, Michael Auli, and David Grangier, "Language Modeling with Gated Convolutional Networks," arXiv:1612.08083 2016.

de Prado, Marcos López, *Advances in Financial Machine Learning*, Hoboken, NJ: Wiley, 2018.

Diebold, Francis X. and Roberto S. Mariano, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 1995, *13* (3), 253–263.

Engle, Robert, "GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics," *Journal of Economic Perspectives*, 2001, *15* (4), 157–168.

Giacomini, Raffaella and Halbert White, "Tests of Conditional Predictive Ability," *Econometrica*, 2006, *74* (6), 1545–1578.

Gneiting, Tilmann and Adrian E Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 2007, *102* (477), 359–378.

_ **and Roopesh Ranjan**, "Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules," *Journal of Business & Economic Statistics*, 2011, *29* (3), 411–422.

_ **, Fadoua Balabdaoui, and Adrian E. Raftery**, "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2007, *69* (2), 243–268.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, "Empirical Asset Pricing via Machine Learning," *The Review of Financial Studies*, 2020, *33* (5), 2223–2273.

Hansen, Peter R. and Asger Lunde, "A forecast comparison of volatility models: does anything beat a GARCH(1,1)?," *Journal of Applied Econometrics*, 2005, *20* (7), 873–889.

_ **,** _ **, and James M. Nason**, "The Model Confidence Set," *Econometrica*, 2011, *79* (2), 453–497.

Harvey, David, Stephen Leybourne, and Paul Newbold, "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 1997, *13* (2), 281–291.

Hendrycks, Dan and Kevin Gimpel, "Gaussian Error Linear Units (GELUs)," arXiv:1606.08415 2016.

Holm, Sture, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 1979, *6* (2), 65–70.

Jordan, Alexander, Fabian Krüger, and Sebastian Lerch, "Evaluating Probabilistic Forecasts with scoringRules," *Journal of Statistical Software*, 2019, *90* (12).

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in "Advances in Neural Information Processing Systems 30" 2017, pp. 3146–3154.

Kelly, Bryan and Dacheng Xiu, "Financial Machine Learning," *Foundations and Trends® in Finance*, 2023, *13* (3-4), 205–363.

Kolm, Petter N., Jeremy Turiel, and Nicholas Westray, "Deep order flow imbalance: Extracting alpha at multiple horizons from the limit order book," *Mathematical Finance*, 2023, *33* (4), 1044–1081.

Lim, Bryan and Stefan Zohren, "Time-series forecasting with deep learning: a survey," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2021, *379* (2194), 20200209.

Liu, Yong, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long, "iTransformer: Inverted Transformers Are Effective for Time Series Forecasting," arXiv:2310.06625 2023.

Loshchilov, Ilya and Frank Hutter, "Decoupled Weight Decay Regularization," arXiv:1711.05101 2017.

Matheson, James E. and Robert L. Winkler, "Scoring Rules for Continuous Probability Distributions," *Management Science*, 1976, *22* (10), 1087–1096.

Nelson, Daniel B., "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, 1991, *59* (2), 347.

Newey, Whitney K. and Kenneth D. West, "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 1987, *55* (3), 703.

Nie, Yuqi, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," arXiv:2211.14730 2022.

Oreshkin, Boris N., Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," arXiv:1905.10437 2019.

Patton, Andrew J., "MODELLING ASYMMETRIC EXCHANGE RATE DEPENDENCE*," *International Economic Review*, 2006, *47* (2), 527–556.

_ , "Volatility forecast comparison using imperfect volatility proxies," *Journal of Econometrics*, 2011, *160* (1), 246–256.

Rasul, Kashif, Calvin Seward, Ingmar Schuster, and Roland Vollgraf, "Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting," arXiv:2101.12072 2021.

Salinas, David, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, 2020, *36* (3), 1181–1191.

Scheuerer, Michael and Thomas M. Hamill, "Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities*," *Monthly Weather Review*, 2015, *143* (4), 1321–1334.

Sirignano, Justin and Rama Cont, "Universal features of price formation in financial markets: perspectives from deep learning," *Quantitative Finance*, 2019, *19* (9), 1449–1459.

Sklar, Abe, "Fonctions de répartition à *n* dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, 1959, *8*, 229–231.

Taillardat, Maxime, Olivier Mestre, Michaël Zamo, and Philippe Naveau, "Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics," *Monthly Weather Review*, 2016, *144* (6), 2375–2393.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," arXiv:1706.03762 2017.

Wen, Qingsong, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun, "Transformers in Time Series: A Survey," arXiv:2202.07125 2022.

Zeng, Ailing, Muxi Chen, Lei Zhang, and Qiang Xu, "Are Transformers Effective for Time Series Forecasting?," arXiv:2205.13504 2022.

Zhang, Zihao, Stefan Zohren, and Stephen Roberts, "DeepLOB: Deep Convolutional Neural Networks for Limit Order Books," *IEEE Transactions on Signal Processing*, 2019, *67* (11), 3001–3012.

Zou, Hui and Trevor Hastie, "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2005, *67* (2), 301–320.

Appendices

A. Baseline specifications

Table 10. Baseline specifications (information set = completed bars at the origin; tuned on training/validation periods only)

Model	Specification
Random walk	window: 390 one-minute bars; distribution: Gaussian, zero mean, scale = trailing RMS one-minute return x sqrt(H)
Seasonal volatility	profile: past-20-session mean squared return per slot (past only); level_adjustment: trailing 30-minute RV over the seasonal counterpart, clipped to [0.25, 4]
EWMA-*t*	lambda: 0.97 per minute; nu: 5; history: last 3 sessions of one-minute returns
HAR-RV	regressors: log RV 30 min, log RV 390 min, log RV 5 days, time of day, time of day squared, log H; fit: OLS on training origins, one regression per horizon; nu: method of moments on standardised training residuals, clipped to [2.5, 50]
GARCH(1,1)-*t*	model: GARCH(1,1) with Student-t innovations on 5-minute training returns, one fit per instrument (arch 8); forecast: analytic variance term structure over ceil(H/5) steps; own filter over the last 400 five-minute returns
Ridge	alpha: 1.0; features: 19 origin features (src/chronos_x/baselines/features.py); targets: scaled cumulative return (location) and log scaled realised variance (scale); nu from residual kurtosis
Elastic net	alpha: 0.001; l1_ratio: 0.5
LightGBM	n_estimators: 400; learning_rate: 0.03; num_leaves: 31; subsample: 0.8; colsample_bytree: 0.8
MLP	inputs: last 60 one-minute bars (15 features each) + 9 global features + future calendar; hidden: 256 x 2 (GELU, dropout 0.1); head: Student-t per future minute, masked NLL
TCN	channels: 96; blocks: 6 dilated causal convolutions (kernel 3, dilation 1..32) over 390 one-minute bars; head: Student-t per future minute, masked NLL
Not run	*chronos_timesfm*: pretrained foundation models not evaluated: weights not available offline on the compute node; tokenisers assume a single real-valued series without missingness masks; licenses/terms not reviewed within this project; *egarch*: multi-step forecasts require simulation; too slow for the number of evaluation origins; GARCH(1,1)-t reported; *patchtst_itransformer*: the one-minute-only MRPT variant is the equal-parameter patch-transformer control; iTransformer not implemented

B. Additional tables

C. Universe

Pooled instruments (240). ABT, ACH, ACM, ADP, AG, AGG, ALB, ALK, AM, AMLP, APA, ARCC, ARKK, ARMK, AU, AUPH, AXTA, BB, BCE, BDN, BF.B, BHP, BKD, BLDR, BLMN, BND, BOX, BTU, CAH, CB, CG, CHD, CHTR, CI, CMCSA, CME, CNC, CNDT, CNO, CNX, COR, CPRT, CVBF, CVE, CX, CZR, D, DB, DBC, DDD, DGX, DHI, DIS, DKS, DLR, DXC, EBS, EFX, EHC, EMLC, EMN, EOG, EQNR, EQT, ERIC, EWA, EWT, EWY, FAST, FAZ, FEZ, FHN, FLR, FPE, FSLR, FXI, GDDY, GEO, GIL, GLD,

Table 11. Main: mean WIS of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.0007 [0.0007, 0.0008]	0.0013 [0.0012, 0.0013]	0.0017 [0.0017, 0.0018]	0.0024 [0.0023, 0.0024]
Seasonal volatility	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0024]
EWMA-*t*	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0024]
HAR-RV	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
GARCH(1,1)-*t*	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0024]
Ridge	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
Elastic net	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
LightGBM	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MLP (1m)	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
TCN (1m)	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S (1m+5m+30m)	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S (1m only)	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-B (1m+5m+30m)	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]

Cluster-bootstrap (trading-day) 95% intervals in brackets.

Table 12. Main: mean QLIKE of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	-12.3011 [-12.3357, -12.2647]	-11.2438 [-11.2774, -11.2053]	-10.5819 [-10.6160, -10.5427]	-9.9314 [-9.9665, -9.8897]
Seasonal volatility	-12.3858 [-12.4188, -12.3489]	-11.3268 [-11.3639, -11.2834]	-10.6551 [-10.6952, -10.6061]	-9.9916 [-10.0361, -9.9323]
EWMA-*t*	-12.4291 [-12.4603, -12.3927]	-11.3537 [-11.3907, -11.3114]	-10.6715 [-10.7094, -10.6265]	-9.9930 [-10.0339, -9.9413]
HAR-RV	-12.2819 [-12.3137, -12.2475]	-11.3605 [-11.4055, -11.3071]	-10.6947 [-10.7386, -10.6459]	-10.0465 [-10.0884, -9.9945]
GARCH(1,1)-*t*	-12.3390 [-12.3766, -12.2971]	-11.2731 [-11.3120, -11.2289]	-10.5920 [-10.6309, -10.5470]	-9.9059 [-9.9462, -9.8580]
Ridge	-12.2468 [-12.2897, -12.2033]	-11.3655 [-11.4057, -11.3170]	-10.7071 [-10.7479, -10.6570]	-10.0542 [-10.0979, -9.9984]
Elastic net	-12.2460 [-12.2890, -12.2026]	-11.3651 [-11.4051, -11.3165]	-10.7066 [-10.7475, -10.6563]	-10.0537 [-10.0974, -9.9978]
LightGBM	5.8416 [-1.1419, 14.3811]	-11.1290 [-11.3556, -10.8016]	-10.7147 [-10.7553, -10.6689]	-10.0661 [-10.1063, -10.0165]
MLP (1m)	-12.4593 [-12.4939, -12.4208]	-11.3889 [-11.4249, -11.3434]	-10.7168 [-10.7542, -10.6698]	-10.0558 [-10.0975, -10.0073]
TCN (1m)	-12.4704 [-12.5018, -12.4332]	-11.3940 [-11.4310, -11.3510]	-10.7198 [-10.7598, -10.6741]	-10.0619 [-10.1020, -10.0151]
MRPT-S (1m+5m+30m)	-12.4763 [-12.5080, -12.4395]	-11.4077 [-11.4437, -11.3669]	-10.7366 [-10.7742, -10.6941]	-10.0795 [-10.1178, -10.0350]
MRPT-S (1m only)	-12.4730 [-12.5048, -12.4358]	-11.4028 [-11.4405, -11.3594]	-10.7305 [-10.7695, -10.6850]	-10.0715 [-10.1112, -10.0212]
MRPT-B (1m+5m+30m)	-12.4727 [-12.5044, -12.4354]	-11.4041 [-11.4413, -11.3610]	-10.7327 [-10.7716, -10.6880]	-10.0748 [-10.1147, -10.0255]

Cluster-bootstrap (trading-day) 95% intervals in brackets.

Table 13. Main: WIS relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Seasonal volatility	-3.00% (-46.4)	-3.20% (-47.8)	-3.33% (-37.8)	-3.54% (-29.1)
EWMA-*t*	-3.75% (-54.5)	-3.75% (-50.4)	-3.60% (-34.8)	-3.40% (-23.3)
HAR-RV	-2.00% (-29.7)	-4.49% (-62.7)	-4.67% (-46.4)	-4.99% (-35.5)
GARCH(1,1)-*t*	-3.66% (-54.1)	-3.79% (-52.8)	-3.67% (-37.7)	-3.29% (-25.3)
Ridge	-2.83% (-29.9)	-4.50% (-56.7)	-4.68% (-44.7)	-5.02% (-35.0)
Elastic net	-2.83% (-29.8)	-4.49% (-56.5)	-4.69% (-44.9)	-5.02% (-35.0)
LightGBM	-2.79% (-30.1)	-4.26% (-50.5)	-4.72% (-46.8)	-5.01% (-36.9)
MLP (1m)	-4.17% (-71.5)	-4.15% (-68.5)	-4.15% (-51.3)	-4.28% (-38.1)
TCN (1m)	-4.16% (-68.4)	-3.96% (-64.5)	-3.88% (-48.4)	-3.92% (-36.8)
MRPT-S (1m+5m+30m)	-4.30% (-68.4)	-4.28% (-66.6)	-4.28% (-50.4)	-4.50% (-39.2)
MRPT-S (1m only)	-4.26% (-68.4)	-4.23% (-66.3)	-4.22% (-50.0)	-4.38% (-38.8)
MRPT-B (1m+5m+30m)	-4.28% (-68.8)	-4.28% (-66.7)	-4.29% (-51.0)	-4.44% (-39.3)

Relative change in mean WIS versus the reference (negative = better); Diebold–Mariano t-statistic with HAC variance in parentheses.

Table 14. Main: QLIKE relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Seasonal volatility	-8472593190631.71% (-26.5)	-8305269577503.39% (-23.8)	-7325382995932.99% (-14.1)	-6019653909459.67% (-7.4)
EWMA-t	-12797135612615.71% (-41.9)	-10996225269347.05% (-30.9)	-8956891977333.71% (-18.6)	-6156613654049.44% (-9.7)
HAR-RV	+1922904424583.55% (4.0)	-11677912373599.43% (-9.2)	-11280891155603.60% (-11.2)	-11508236846123.03% (-17.7)
GARCH(1,1)-t	-3792274777371.91% (-4.2)	-2936683613506.50% (-4.8)	-1014981889598.51% (-1.7)	+2553431567866.42% (4.3)
Ridge	+5434479946610.18% (7.4)	-12177565285940.69% (-21.4)	-12517051234684.73% (-20.7)	-12281264246708.48% (-17.5)
Elastic net	+5514663565784.96% (7.5)	-12129129818422.78% (-21.2)	-12468574596082.83% (-20.4)	-12234566499031.67% (-17.2)
LightGBM	+1814267205302873.75% (4.8)	+11471871421175.42% (0.8)	-13278787159454.87% (-25.8)	-13466213447329.12% (-26.1)
MLP (1m)	-15818212313424.51% (-41.8)	-14508869311135.06% (-36.7)	-13489944655719.03% (-28.8)	-12445594328523.75% (-19.9)
TCN (1m)	-16924042856514.05% (-63.3)	-15020597397889.10% (-35.3)	-13794157215647.81% (-23.8)	-13055303736382.31% (-25.8)
MRPT-S (1m+5m+30m)	-17518582839370.88% (-61.1)	-16395858212306.31% (-59.2)	-15469510096077.54% (-47.9)	-14813941021870.59% (-38.0)
MRPT-S (1m only)	-17184931027568.06% (-60.5)	-15902869282831.41% (-45.2)	-14857414727344.76% (-34.7)	-14011901777049.01% (-28.0)
MRPT-B (1m+5m+30m)	-17153971110963.66% (-60.7)	-16037277205437.76% (-48.1)	-15081771195336.03% (-37.9)	-14340084012486.90% (-30.2)

Relative change in mean QLIKE versus the reference (negative = better); Diebold–Mariano t-statistic with HAC variance in parentheses.

Table 15. Main (held-out): empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
Random walk (trailing RMS)	0.64	0.88	0.96	0.075	0.65	0.89	0.97	0.073	0.65	0.90	0.97	0.078	0.67	0.91	0.98	0.084
Seasonal volatility	0.55	0.82	0.94	0.026	0.54	0.82	0.94	0.026	0.55	0.82	0.94	0.028	0.55	0.82	0.95	0.029
EWMA-t	0.53	0.83	0.97	0.018	0.53	0.84	0.97	0.019	0.54	0.84	0.97	0.024	0.55	0.85	0.97	0.030
HAR-RV	0.57	0.87	0.98	0.041	0.49	0.80	0.96	0.007	0.50	0.81	0.96	0.008	0.52	0.82	0.96	0.014
GARCH(1,1)-t	0.53	0.83	0.96	0.018	0.54	0.84	0.97	0.022	0.55	0.85	0.97	0.029	0.58	0.87	0.98	0.042
Ridge	0.41	0.70	0.90	0.057	0.46	0.76	0.94	0.026	0.48	0.78	0.95	0.013	0.50	0.80	0.96	0.008
Elastic net	0.41	0.70	0.90	0.057	0.46	0.76	0.94	0.026	0.48	0.78	0.95	0.013	0.50	0.80	0.96	0.008
LightGBM	0.38	0.68	0.91	0.072	0.46	0.77	0.94	0.025	0.48	0.78	0.95	0.012	0.50	0.80	0.95	0.006
MLP (1m)	0.54	0.82	0.96	0.023	0.55	0.84	0.96	0.029	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.037
TCN (1m)	0.55	0.83	0.96	0.027	0.56	0.84	0.96	0.031	0.57	0.85	0.96	0.035	0.58	0.85	0.96	0.039
MRPT-S (1m+5m+30m)	0.54	0.82	0.95	0.024	0.55	0.83	0.96	0.026	0.56	0.84	0.96	0.030	0.56	0.84	0.96	0.033
MRPT-S (1m only)	0.54	0.82	0.95	0.023	0.55	0.83	0.96	0.026	0.56	0.84	0.96	0.029	0.56	0.85	0.96	0.034
MRPT-B (1m+5m+30m)	0.54	0.82	0.95	0.023	0.55	0.83	0.96	0.026	0.56	0.84	0.96	0.030	0.57	0.85	0.96	0.035

Table 16. Final: CRPS relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Seasonal volatility	-2.52% (-27.7)	-2.85% (-31.9)	-2.89% (-26.3)	-3.07% (-20.0)
EWMA-t	-3.51% (-39.0)	-3.59% (-35.8)	-3.39% (-29.7)	-3.14% (-19.7)
HAR-RV	-2.20% (-27.4)	-4.06% (-44.9)	-4.11% (-34.1)	-4.37% (-24.8)
GARCH(1,1)-t	-3.38% (-32.4)	-3.68% (-29.1)	-3.44% (-28.9)	-3.09% (-17.8)
Ridge	-2.68% (-23.5)	-4.10% (-41.9)	-4.17% (-34.0)	-4.42% (-25.4)
Elastic net	-2.68% (-23.5)	-4.10% (-41.8)	-4.17% (-34.0)	-4.42% (-25.4)
LightGBM	-2.57% (-22.2)	-3.87% (-34.1)	-4.22% (-34.6)	-4.41% (-24.9)
MLP (1m)	-3.84% (-52.8)	-3.84% (-49.8)	-3.81% (-41.0)	-3.85% (-29.9)
TCN (1m)	-3.89% (-47.5)	-3.74% (-46.3)	-3.60% (-39.0)	-3.55% (-28.6)
MRPT-S (1m+5m+30m)	-3.96% (-42.3)	-3.89% (-46.2)	-3.85% (-39.6)	-3.91% (-28.7)
MRPT-S (1m only)	-3.97% (-42.4)	-3.91% (-46.1)	-3.88% (-39.5)	-3.92% (-28.5)
MRPT-B (1m+5m+30m)	-3.94% (-42.7)	-3.87% (-49.2)	-3.84% (-40.3)	-3.89% (-30.1)

Relative change in mean CRPS versus the reference (negative = better); Diebold–Mariano t-statistic with HAC variance in parentheses.

Table 17. Final (held-out instruments): mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.4813 [0.4728, 0.4897]	0.4682 [0.4598, 0.4771]	0.4586 [0.4495, 0.4679]	0.4471 [0.4369, 0.4580]
Seasonal volatility	0.4705 [0.4614, 0.4794]	0.4563 [0.4478, 0.4658]	0.4468 [0.4367, 0.4569]	0.4351 [0.4236, 0.4474]
EWMA-*t*	0.4666 [0.4576, 0.4754]	0.4533 [0.4447, 0.4626]	0.4449 [0.4351, 0.4551]	0.4346 [0.4235, 0.4465]
HAR-RV	0.4688 [0.4600, 0.4774]	0.4518 [0.4428, 0.4614]	0.4424 [0.4323, 0.4529]	0.4308 [0.4192, 0.4432]
GARCH(1,1)-*t*	0.4677 [0.4588, 0.4766]	0.4538 [0.4451, 0.4632]	0.4455 [0.4358, 0.4555]	0.4360 [0.4250, 0.4481]
Ridge	0.4691 [0.4594, 0.4786]	0.4517 [0.4426, 0.4614]	0.4420 [0.4318, 0.4526]	0.4303 [0.4187, 0.4428]
Elastic net	0.4691 [0.4595, 0.4786]	0.4517 [0.4427, 0.4614]	0.4421 [0.4319, 0.4526]	0.4302 [0.4186, 0.4427]
LightGBM	0.4722 [0.4623, 0.4817]	0.4521 [0.4430, 0.4619]	0.4428 [0.4324, 0.4533]	0.4311 [0.4196, 0.4438]
MLP (1m)	0.4651 [0.4562, 0.4740]	0.4524 [0.4438, 0.4618]	0.4433 [0.4335, 0.4533]	0.4321 [0.4207, 0.4443]
TCN (1m)	0.4650 [0.4562, 0.4738]	0.4526 [0.4441, 0.4619]	0.4436 [0.4340, 0.4536]	0.4326 [0.4213, 0.4447]
MRPT-S (1m+5m+30m)	0.4644 [0.4556, 0.4732]	0.4518 [0.4432, 0.4611]	0.4428 [0.4330, 0.4528]	0.4316 [0.4202, 0.4437]
MRPT-S (1m only)	0.4644 [0.4556, 0.4733]	0.4518 [0.4433, 0.4611]	0.4428 [0.4330, 0.4529]	0.4317 [0.4203, 0.4438]
MRPT-B (1m+5m+30m)	0.4646 [0.4558, 0.4735]	0.4519 [0.4434, 0.4613]	0.4429 [0.4331, 0.4529]	0.4317 [0.4203, 0.4438]

Cluster-bootstrap (trading-day) 95% intervals in brackets.

Table 18. Ablations: empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
Random walk (trailing RMS)	0.65	0.88	0.97	0.078	0.66	0.89	0.97	0.079	0.67	0.90	0.97	0.083	0.68	0.91	0.98	0.089
Seasonal volatility	0.56	0.82	0.94	0.031	0.56	0.83	0.94	0.031	0.56	0.83	0.95	0.033	0.57	0.83	0.95	0.034
EWMA-*t*	0.54	0.83	0.97	0.023	0.54	0.84	0.97	0.024	0.55	0.85	0.97	0.028	0.56	0.86	0.97	0.034
MRPT-S (1m+5m)	0.55	0.83	0.96	0.030	0.57	0.84	0.96	0.034	0.57	0.85	0.96	0.039	0.58	0.85	0.96	0.041
MRPT-S cross-attention fusion	0.55	0.83	0.96	0.029	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.038	0.58	0.85	0.96	0.042
MRPT-S Gaussian head	0.00	0.00	0.00	0.995	0.00	0.00	0.00	0.995	0.00	0.00	0.00	0.995	0.00	0.00	0.00	0.995
MRPT-S no regime	0.54	0.83	0.95	0.027	0.56	0.84	0.96	0.031	0.57	0.84	0.96	0.035	0.58	0.85	0.96	0.039
MRPT-S learned regime	0.54	0.82	0.95	0.021	0.55	0.83	0.95	0.029	0.56	0.84	0.96	0.035	0.57	0.84	0.96	0.040
MRPT-S returns only	0.56	0.84	0.96	0.034	0.57	0.85	0.96	0.041	0.58	0.86	0.96	0.045	0.59	0.86	0.96	0.049
MRPT-S time only	0.56	0.84	0.96	0.034	0.57	0.85	0.96	0.040	0.58	0.86	0.96	0.045	0.59	0.86	0.96	0.049
MRPT-S OHLCV (no rel. volume)	0.55	0.83	0.96	0.031	0.57	0.85	0.96	0.036	0.58	0.85	0.96	0.041	0.58	0.86	0.96	0.044
MRPT-S shuffled minutes	0.55	0.83	0.96	0.031	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.042	0.58	0.86	0.96	0.046
MRPT-S + ticker id	0.55	0.83	0.96	0.029	0.57	0.84	0.96	0.034	0.58	0.85	0.96	0.039	0.59	0.86	0.96	0.044
MRPT-S + copula (joint sampling)	0.55	0.83	0.96	0.030	0.57	0.84	0.96	0.036	0.58	0.85	0.96	0.041	0.58	0.86	0.96	0.045
MRPT-S reference (4k steps)	0.54	0.83	0.95	0.027	0.56	0.84	0.96	0.032	0.57	0.84	0.96	0.037	0.58	0.85	0.96	0.042
MRPT-S 1m only (4k)	0.55	0.83	0.96	0.030	0.57	0.84	0.96	0.035	0.57	0.85	0.96	0.039	0.58	0.85	0.96	0.043
MRPT-S reference, seed 2	0.54	0.82	0.95	0.029	0.56	0.84	0.96	0.034	0.57	0.84	0.96	0.041	0.58	0.85	0.96	0.048
MRPT-S reference, seed 3	0.54	0.82	0.95	0.020	0.56	0.84	0.96	0.037	0.56	0.84	0.96	0.045	0.57	0.85	0.96	0.056
MRPT-S + copula, independent sampling	0.55	0.83	0.96	0.030	0.57	0.84	0.96	0.036	0.57	0.85	0.96	0.040	0.58	0.85	0.96	0.044

Table 19. Ablations: mean WIS of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.0007 [0.0007, 0.0008]	0.0013 [0.0012, 0.0013]	0.0017 [0.0017, 0.0018]	0.0024 [0.0023, 0.0024]
Seasonal volatility	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0024]
EWMA-t	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0024]
MRPT-S (1m+5m)	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S cross-attention fusion	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S Gaussian head	nan	nan	nan	nan
MRPT-S no regime	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S learned regime	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S returns only	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S time only	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S OHLCV (no rel. volume)	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S shuffled minutes	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S + ticker id	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S + copula (joint sampling)	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S reference (4k steps)	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S 1m only (4k)	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S reference, seed 2	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S reference, seed 3	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]
MRPT-S + copula, independent sampling	0.0007 [0.0007, 0.0007]	0.0012 [0.0012, 0.0012]	0.0017 [0.0016, 0.0017]	0.0023 [0.0022, 0.0023]

Cluster-bootstrap (trading-day) 95% intervals in brackets.

Table 20. Pilot: mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.4803 [0.4618, 0.5012]	0.4874 [0.4641, 0.5093]	0.4952 [0.4688, 0.5225]	0.4819 [0.4381, 0.5395]
Seasonal volatility	0.4746 [0.4545, 0.4957]	0.4793 [0.4526, 0.5047]	0.4870 [0.4569, 0.5190]	0.4767 [0.4268, 0.5419]
EWMA-t	0.4699 [0.4502, 0.4912]	0.4775 [0.4517, 0.5023]	0.4868 [0.4561, 0.5196]	0.4773 [0.4276, 0.5428]
HAR-RV	0.4833 [0.4648, 0.5032]	0.4761 [0.4492, 0.5019]	0.4845 [0.4540, 0.5168]	0.4728 [0.4223, 0.5391]
GARCH(1,1)-t	0.4714 [0.4522, 0.4932]	0.4783 [0.4545, 0.5018]	0.4876 [0.4602, 0.5167]	0.4789 [0.4331, 0.5388]
Ridge	0.4734 [0.4518, 0.4968]	0.4765 [0.4494, 0.5028]	0.4842 [0.4533, 0.5170]	0.4723 [0.4216, 0.5368]
Elastic net	0.4732 [0.4516, 0.4967]	0.4764 [0.4494, 0.5026]	0.4841 [0.4532, 0.5169]	0.4725 [0.4217, 0.5371]
LightGBM	0.5090 [0.4871, 0.5325]	0.5034 [0.4756, 0.5308]	0.5116 [0.4812, 0.5471]	0.5019 [0.4484, 0.5765]
MLP (1m)	0.4689 [0.4497, 0.4908]	0.4769 [0.4522, 0.5015]	0.4847 [0.4559, 0.5157]	0.4743 [0.4250, 0.5390]
TCN (1m)	0.4687 [0.4497, 0.4896]	0.4767 [0.4526, 0.5008]	0.4838 [0.4552, 0.5144]	0.4743 [0.4252, 0.5388]
MRPT-S (1m+5m+30m)	0.4697 [0.4509, 0.4903]	0.4763 [0.4526, 0.5001]	0.4831 [0.4553, 0.5135]	0.4736 [0.4238, 0.5381]
MRPT-S (1m only)	0.4693 [0.4506, 0.4899]	0.4765 [0.4527, 0.5006]	0.4842 [0.4558, 0.5148]	0.4745 [0.4245, 0.5395]

Cluster-bootstrap (trading-day) 95% intervals in brackets.

Table 21. Pilot: empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
Random walk (trailing RMS)	0.62	0.87	0.96	0.070	0.61	0.87	0.96	0.063	0.60	0.86	0.96	0.068	0.61	0.87	0.97	0.075
Seasonal volatility	0.53	0.80	0.94	0.029	0.53	0.80	0.94	0.025	0.52	0.79	0.93	0.026	0.52	0.79	0.93	0.039
EWMA-t	0.51	0.81	0.97	0.017	0.50	0.80	0.96	0.012	0.49	0.80	0.96	0.027	0.51	0.80	0.96	0.034
HAR-RV	0.63	0.92	0.99	0.082	0.48	0.78	0.96	0.021	0.46	0.77	0.95	0.039	0.48	0.77	0.95	0.033
GARCH(1,1)-t	0.54	0.85	0.97	0.035	0.56	0.85	0.98	0.037	0.56	0.85	0.98	0.043	0.59	0.87	0.98	0.065
Ridge	0.41	0.70	0.91	0.074	0.44	0.74	0.94	0.042	0.44	0.75	0.94	0.052	0.46	0.75	0.94	0.040
Elastic net	0.41	0.71	0.91	0.075	0.44	0.74	0.94	0.041	0.44	0.74	0.94	0.053	0.46	0.75	0.94	0.038
LightGBM	0.35	0.64	0.89	0.110	0.41	0.71	0.92	0.063	0.46	0.74	0.92	0.044	0.45	0.75	0.92	0.042
MLP (1m)	0.51	0.80	0.95	0.021	0.53	0.81	0.95	0.024	0.51	0.80	0.94	0.030	0.53	0.80	0.94	0.043
TCN (1m)	0.54	0.82	0.96	0.027	0.55	0.82	0.96	0.034	0.52	0.81	0.95	0.032	0.54	0.81	0.95	0.050
MRPT-S (1m+5m+30m)	0.54	0.83	0.96	0.035	0.55	0.83	0.96	0.034	0.53	0.82	0.95	0.037	0.54	0.81	0.95	0.053
MRPT-S (1m only)	0.55	0.83	0.96	0.039	0.56	0.83	0.96	0.038	0.53	0.82	0.95	0.040	0.54	0.81	0.95	0.053

Table 22. Sector (SIC division) and ETF category counts in the universe

Asset class	Group	*n*
etf	broad_equity	12
etf	commodity	5
etf	fixed_income	8
etf	international	13
etf	leveraged_inverse	14
etf	other	10
etf	sector	10
etf	volatility_alt	3
stock	construction	13
stock	finance_insurance_realestate	30
stock	manufacturing	34
stock	mining	19
stock	retail	26
stock	services	32
stock	transport_utilities	32
stock	unknown	25
stock	wholesale	14

Table 23. Universe allocation: instruments per asset class × dollar-volume tercile × daily-volatility tercile (selection date 2017-12-29; 60 of the 300 instruments are held out)

Asset class	ADV tercile	Vol. low	Vol. mid	Vol. high
etf	high	8	8	8
etf	low	9	8	9
etf	mid	9	8	8
stock	high	25	25	25
stock	low	25	25	25
stock	mid	25	25	25
held out		stocks 42, ETFs 18		

GLNG, GNTX, GOOGL, GOOS, GPC, GSAT, GTN, HAL, HMC, HOMB, HOPE, HSBC, HSIC, HXL, IEF, IGSB, ILF, INDA, IRBT, IRDM, ITUB, IVE, IVZ, IWF, IXUS, IYE, J, JPM, KBE, KBR, KBWB, KIM, KLAC, KMI, KSS, LBTYA, LDOS, LEN, LNT, LOW, LQD, LRCX, LYG, M, MAA, MAC, MINT, MOMO, MPC, MRSH, MUR, MUX, NAVI, NEM, NI, NOW, NRG, NTNX, ODFL, OIH, OKE, OPLN, OPTU, OXY, PAGP, PANW, PBA, PBR, PBR.A, PDBC, PEP, PFF, PFG, PFGC, PHM, PK, POR, PWR, RCI, RIO, RIOT, ROL, ROST, RPM, RUN, SABR, SAN, SBH, SBSW, SBUX, SCHG, SCO, SDOW, SDS, SFM, SH, SHEL, SHV, SHW, SJNK, SLF, SMFG, SPDW, SPEM, SPGI, SPWR, SPY, SPYD, SRPT, SSNC, STAG, TAP, TDC, TEL, TIP, TJX, TMO, TNA, TPR, TRI, TTE, TU, TXRH, UNFI, UPRO, URBN, USFD, USMV, VALE, VDE, VEU, VFH, VGK, VIAV, VIG, VIV, VNQ, VTEB, VTR, VWO, VXX, VZ, WAB, WDAY, WEN, WERN, WES, WIT, WTRG, WU, WWW, WY, XBI, XLP, XLU, XME, XPO, XYL, ZION, ZTO

Held-out instruments (60). ACAD, AEE, AEP, AVT, BDX, BHC, BOTZ, BRK.B, CAKE, CAT, CNI, CVNA, ENB, FNV, FTI, GPK, H, HD, HDB, HLF, HTGC, IAU, ING, ITB, KBH, KRG, LKQ, MCK, MSFT, MTZ, ORCL, PDM, PSQ, PTEN, RBA, SCHA, SCHF, SCHH, SOXX, SPHD, SPSB, SSO, SYY, TLT, TOL, TROW, TSCO, TZA, UAL, UDOW, UE, VEA, VMBS, VMC, W, WMT, WRB, XLRE, XRT, YUM