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

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

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

1. 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.

2. 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, 2008; 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 *The Model Confidence Set* (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 5 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.

3. Data

3.1. 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.

3.2. 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.

3.3. 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.

3.4. 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

$h \le H$ 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.

3.5. 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.

3.6. 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.

4. The Multi-Resolution Patch Transformer

4.1. 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 $5j + 4 \le m$), 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.

4.2. 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.

4.3. 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.

4.4. 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.

4.5. 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 **??**. 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.

5. Experimental design

5.1. 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 1 lists the specifications.

5.2. 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.

5.3. 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.

5.4. 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.

6. 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 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 5); 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.

7. 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`.

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*, 2008, *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.

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.
The Model Confidence Set

***The Model Confidence Set**, Econometrica, 2011, 79 (2), 453–497.*

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 1. 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,

Table 2. 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		

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, 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