Decomposing Price Chart Signals in Return Prediction

Jonas Theill Bøjstrup[1,§] Bezirgen Veliyev[1,‡,†] Jesper N. Wulff[1,†]

jobo@econ.au.dk bveliyev@econ.au.dk jwulff@econ.au.dk

Abstract

Image models trained on stock charts predict returns out of sample, though a chart adds no information beyond the OHLCV data it renders. We decompose the image model's advantage over tabular benchmarks from the same daily window into three sources: function-class capacity, chart-specific encoding, and how each model smooths noisy returns. Capacity contributes little. Under a distillation accounting, roughly 60% of the apparent gap is tied to learning from noisy realized returns rather than chart-specific content. The residual localizes to recent bars, intraday extremes, and temporal order, and is largely recoverable with explicit predictors.

Keywords: Asset Pricing, Return Predictability, Machine Learning, Technical Analysis, Computer Vision

JEL Classification: G12, G14, C45, C53

[1]Department of Economics and Business Economics, Aarhus University, Denmark.
[§]Center for Research in Energy: Economics and Markets (CoRE), Aarhus University, Denmark.
[‡]Aarhus Center for Econometrics (ACE), Aarhus University, Denmark.
[†]Research fellow at the Danish Finance Institute (DFI).

1 Introduction

Recent work shows that machine-learning models trained directly on images of stock price charts predict cross-sectional returns out of sample, with economic magnitudes that rival the strongest published return-predictive signals (Jiang et al., 2023; Murray et al., 2024; Bøjstrup et al., 2026). A chart, however, contains no new data. A candlestick image is a deterministic rendering of the same open, high, low, close, and volume (OHLCV) series that any tabular asset-pricing model already reads from standard databases. Because the rendering is a fixed function of that series, it cannot raise the information the window carries about future returns (Cover and Thomas, 2006). Whatever predictive content the image model recovers has been available to tabular models for decades, in numeric form.

When an image model outperforms a tabular model built on the same OHLCV history, what is the source of the advantage? Three sources can produce it. First, *function class*: the image model may benefit from a more flexible function class than its tabular counterparts. Second, *chart encoding*: it may benefit from a representation that exposes structure the numeric predictors do not capture. Third, *target noise*: realized returns are noisy estimates of the conditional expectation (Elton, 1999), and how strongly a model smooths that noise depends on its architecture and features. Only the second source corresponds to new economic information; the first reflects modeling choice and the third estimation efficiency. Both representations are lossy functions of the same window, so the encoding share measures what the chart preserves relative to a given set of numeric predictors, not information absent from the OHLCV series itself. The distinction matters because chart-based scores are increasingly used as inputs to cross-sectional tests, components of factor portfolios, and objects of economic interpretation. If most of the advantage reflects capacity or noise smoothing, the chart score is not a new characteristic but a refined estimate of one already implicit in OHLCV.

The existing literature cannot separate the three. Jiang et al. (2023) and Murray et al. (2024) establish that chart-based and image-like signals are distinct from momentum,

reversal, and standard technical indicators, but their benchmarks lie outside the chart's own data window, so the comparisons confound representation with information. Separating the sources requires a tabular benchmark built on the same OHLCV window as the image, so that any performance gap reflects how each model class exploits a common input set rather than differences in the inputs themselves.

We build that benchmark. For each stock-date observation in a daily panel of the 500 largest U.S. equities over 1993 to 2025, we render a deterministic candlestick chart of the prior 20 trading days of OHLCV and pair it with a matched vector of 220 numeric predictors computed from the same window. We train an image model on the chart and two tabular models, one linear and one flexibly nonlinear, on the numeric predictors; all models are estimated through 2003, tuned on 2004 to 2005, and evaluated out of sample over 2006 to 2025. The three models see identical underlying inputs and differ only in representation and function class. The gap between the linear and nonlinear tabular models isolates the function-class share within the tabular representation, and deep tabular benchmarks that fail to narrow the remaining gap confirm that residual capacity is not the binding constraint. What remains, between the nonlinear tabular model and the image model, is what the analysis decomposes into chart-encoding and target-noise shares rather than imposing the split by assumption.

Out of sample, the image model produces an annualized high-minus-low decile spread of 20.67% with a Sharpe ratio of 1.50, against 17.57% and 1.34 for the nonlinear tabular model and 14.03% and 1.04 for the linear one. A perturbation analysis localizes the encoding share. Holding the trained image model fixed, we selectively remove components of the rendered chart, so that any performance change attaches to the removed component rather than to retraining. Masking the most recent bars cuts the spread nearly in half while masking the oldest bars has no effect; destroying temporal order and removing intraday wicks produce intermediate losses; the volume panel matters modestly and color coding not at all. The load-bearing visual structure is recency, left-to-right ordering, and intrabar extremes. One reading is that the chart acts as an attention prior over an unchanged information set, re-weighting the window toward the bars that attention-constrained

processing attends to first (Barber and Odean, 2008; Da et al., 2011; Bordalo et al., 2013; Cosemans and Frehen, 2021).

We then write those components into explicit numeric form: five predictor blocks capturing recent-bar geometry, wick structure, directional sequencing, price-volume interactions, and path shape. Augmented with the blocks, the linear model's Sharpe ratio rises from 1.04 to 1.42. The nonlinear model's spread becomes statistically indistinguishable from the image model's: the baseline gap of 3.10 percentage points shrinks toward zero or changes sign across translated specifications, and most translated specifications match or exceed the image model's Sharpe ratio of 1.50, though no pairwise difference is significant at conventional levels. In joint cross-sectional tests, each score retains incremental predictive content after controlling for the other; the level comparison and the residual-content comparison answer different questions, and translation closes the former while the image score still carries information the translated set does not span.

A distillation step helps interpret what that residual contains (Hinton et al., 2015). Training the augmented tabular model to reproduce the score, rather than to fit realized returns, holds input representation and learner fixed while changing the target. This raises the cross-sectional R^2 between the two scores from 0.45 to 0.78. The Hiera score is smoother than realized returns but embeds the image model's own estimation error, so the gain measures target mismatch conditional on the encoding blocks rather than in the abstract. Under this accounting, roughly 60% of the unexplained variance gap to Hiera is recovered when the tabular learner is trained on the Hiera score rather than on realized returns. This is a variance-share accounting object, not a decomposition of Sharpe-ratio or portfolio-return differences. Function-class capacity beyond the nonlinear tabular benchmark contributes essentially nothing, and the remaining image-specific component is small and localized: an annualized high-minus-low spread of 4.90%, concentrated in the recent bars and intraday extremes the perturbations identify, and stable across subperiods.

The paper makes three contributions. First, a decomposition of the image model's economic advantage into function-class, chart-encoding, and target-noise shares, operationalized

through a matched-input design that holds the information set fixed. The decomposition adapts knowledge distillation, developed for model compression, to a measurement problem in asset pricing: a noise-free teacher target isolates the noisy-supervision channel empirically rather than imposing it by construction. Second, an empirical characterization of which chart components carry the representational advantage, identifying recency, temporal ordering, and intrabar extremes as load-bearing and assigning negligible roles to color coding and the volume panel. Third, a constructive translation of those components into explicit numeric predictors that closes most of the gap and isolates the remainder as a small, localized encoding residual. The decomposition is portable to any setting in which flexible models on alternative representations of the same data, such as text, graphs, or transformed time series, outperform tabular benchmarks built on the same inputs. The takeaway is that chart-based scores are not, in their own right, a new asset-pricing characteristic. They are a more efficient estimate of structure already implicit in the same OHLCV window.

1.1 Related Literature

This paper relates to three strands of literature. The first studies whether past trading histories forecast returns. In the weak-form efficiency tradition (Fama, 1970, 1991), the evidence documents short-horizon return continuation (Jegadeesh, 1990), intermediate-horizon momentum (Jegadeesh and Titman, 1993), price-level effects such as the 52-week high (George and Hwang, 2004), and forecasting value in moving-average rules (Zhu and Zhou, 2009), cross-sectional technical indicators (Han et al., 2013, 2016), and time-series momentum (Moskowitz et al., 2012). Chart-based patterns have been studied directly: Brock et al. (1992) find support for trading-range break rules, and Lo et al. (2000) detect visual patterns systematically using kernel smoothing. Theoretically, past prices can convey information in rational learning environments (Brown and Jennings, 1989), and volume can supplement prices in revealing private information (Blume et al., 1994). Throughout this literature, the measured predictive content of past prices depends on how the trading record is summarized. Close-to-close returns, intrabar ranges, and ordered bar

sequences capture different aspects of the same record (Lehmann, 1990; Lo and MacKinlay, 1990; Chordia et al., 2008), and OHLC data carries information beyond closing prices (Parkinson, 1980).[1] A candlestick chart makes these distinctions visually explicit: each bar encodes how the day's range relates to its opening and closing levels, and the ordered sequence preserves temporal structure that a vector of summary statistics does not. The literature's concerns about data snooping in the search over technical rules (Sullivan et al., 1999; Bajgrowicz and Scaillet, 2012), multiple testing (Harvey et al., 2016; Hou et al., 2020; Jensen et al., 2023), and weak post-publication performance (McLean and Pontiff, 2016; Schwert, 2003) motivate the common-information design directly: rather than proposing another candidate rule and testing whether it survives, the analysis holds the information set fixed and asks what a flexible model extracts from it.

The second strand is the machine-learning literature in empirical asset pricing. Machine-learning methods have advanced the measurement of expected returns by accommodating nonlinearities and interactions that linear models miss. Gu et al. (2020) conduct a large-scale comparison and find that neural networks and tree-based methods generate out-of-sample R^2 values between 0.33% and 0.40% per month, with the gains traceable to nonlinear predictor interactions rather than to a larger predictor set per se. Kelly et al. (2019) estimate latent factors whose loadings vary with observed characteristics. Chen et al. (2024) combine deep neural networks with no-arbitrage restrictions and find that the gains arise mainly from characteristic interactions. Kelly et al. (2024) establish theoretically that expected out-of-sample performance is increasing in model complexity under appropriate shrinkage.

These results also sharpen a second question: once a flexible model reveals that something predicts returns, what does the signal contain? Fama and French (2008) and Lewellen (2015) show that many proposed characteristics are closely related, and Green et al. (2017) find that a large share become redundant once others are controlled for. Freyberger et al.

[1] Alizadeh et al. (2002) show that range-based volatility estimators outperform those built from closing prices alone. Corwin and Schultz (2012) estimate bid-ask spreads from daily highs and lows, because the daily high (low) is almost always a buy (sell) trade. Kavajecz and Odders-White (2004) relate technical analysis to liquidity provision, with key price levels associated with concentrations of limit orders.

(2020) extract the pricing-relevant content of a large characteristic set nonparametrically and find that only a modest number of characteristics are needed once nonlinearities are accommodated. Giglio et al. (2021) develop formal tests for incremental pricing information, and Giglio et al. (2022) emphasize that the value of machine learning in finance extends beyond prediction to the investigation of economic mechanisms. A separate difficulty is that the estimation target is itself measured with error. Realized returns proxy expected returns imprecisely, and the error does not average out over the sample (Elton, 1999), so part of any model's out-of-sample advantage can reflect how it handles noisy supervision rather than what it recovers about returns. Knowledge distillation, developed to transfer a trained network's outputs to a smaller model (Hinton et al., 2015), provides a way to hold the target fixed at a noise-free function of the same inputs and read off that channel directly. We use it for that purpose. This characterization problem is the one we address, in the specific context of chart-based signals: not whether a flexible model can predict returns, but what it extracts and how far that content can be made explicit.

The third strand, and the most directly related, shows that chart-like representations of past price and volume data predict returns but leaves what the chart captures largely open. Jiang et al. (2023) train convolutional neural networks on stock-level OHLC price charts and find that the learned patterns are largely distinct from momentum, reversal, and previously studied technical signals, and that they transfer across geographies and time scales. Their study establishes the predictive value of chart-based signals but underscores the difficulty of interpretation: even after applying several CNN visualization methods, Jiang et al. (2023) describe their interpretive efforts as "admittedly incomplete." Bøjstrup et al. (2026) replace the convolutional backbone with a hierarchical vision transformer trained on the full CRSP equity cross-section, document that the architectural choice changes the magnitude of the chart-based signal, and push further on interpretation through layer-resolved attention diagnostics, but leave the chart-versus-tabular characterization question open. Murray et al. (2024) study a related but distinct object: machine-learning forecasts built from cumulative-return inputs rather than rendered images. They find that the resulting signals are stable, highly nonlinear, and not subsumed by momentum,

reversal, or previously studied technical indicators, and strong among the largest 500 stocks. Both studies compare their signals against known benchmarks; neither holds the raw information set fixed and asks whether a chart-based and a tabular model built from the same data contain the same predictive content.[2]

The remainder of the paper proceeds as follows. Section 2 introduces the common-information forecasting design, the image and tabular representations, and the empirical decomposition that separates target noise, chart encoding, and function-class capacity. Section 3 describes the data and reports the common-information comparison, perturbation evidence, translated tabular benchmarks, and distillation results. Section 4 concludes.

[2]Several studies in adjacent machine-learning settings also recast financial time series as image inputs and process them with convolutional architectures (Chen et al., 2016; Kim and Kim, 2019; Lee et al., 2019; Hoseinzade and Haratizadeh, 2019). In other financial settings, Obaid and Pukthuanthong (2022) construct an investor-sentiment measure from news photos, and Kaczmarek and Pukthuanthong (2024) use firm images to build a measure of peer similarity.

2 Methodology

2.1 Forecasting Setup

For each stock i and trading day t, the forecasting input is the 20-trading-day window of daily open, high, low, close, and volume (OHLCV) ending on day t,

$$W_{i,t} = \{(O_{i,\tau}, H_{i,\tau}, L_{i,\tau}, C_{i,\tau}, V_{i,\tau})\}_{\tau=t-19}^{t} . \tag{2.1}$$

Following Jiang et al. (2023), the target is the direction of the five-day-ahead cumulative close-to-close return,

$$y_{i,t}^{(5)} = \mathbf{1}\left\{\frac{C_{i,t+5}}{C_{i,t}} - 1 > 0\right\} . \tag{2.2}$$

The 20-day window is roughly one trading month, the horizon over which candlestick charts are typically constructed and the lookback used by Han et al. (2016). The five-day horizon is short enough to isolate short-horizon pattern content and long enough to limit bid-ask bounce.

The design holds the underlying OHLCV window fixed and varies only its representation. Two representations of $W_{i,t}$ are constructed. The first is a candlestick-chart image,

$$X_{i,t} = \mathcal{R}(W_{i,t}) , \tag{2.3}$$

where $\mathcal{R}(\cdot)$ is a deterministic operator that renders the OHLCV sequence as a 224×224 RGB chart with two vertically stacked panels. The upper panel shows candlesticks with 5-day and 20-day trailing moving-average overlays, motivated by evidence that moving-average signals carry cross-sectional information (Han et al., 2016). The lower panel shows volume bars, consistent with evidence that trading volume carries information beyond prices (Blume et al., 1994). Within-window price scaling with 3% padding and a $\log(1 + \text{volume})$ transformation standardize the chart across firms and dates. Exact

rendering settings are reported in Table A1.1.[3]

The second representation is a structured numeric vector,

$$Z_{i,t} = \mathcal{G}(W_{i,t}) \in \mathbb{R}^{220}, \tag{2.4}$$

where $\mathcal{G}(\cdot)$ encodes in explicit numeric form the objects used to construct the chart: normalized OHLC values, moving averages, transformed volume, and indicators recording candlestick and volume-bar orientation. The construction of $Z_{i,t}$ mirrors the chart rendering, so the two inputs are built from the same window and differ in representation rather than source information. The normalization and assembly of the 220 features are described in Subsection A3.

Estimation follows a strictly chronological protocol: models are fit on an initial training period, selected on a subsequent validation period, and evaluated on a later out-of-sample period, as is standard in the machine learning asset-pricing literature (Gu et al., 2020; Jiang et al., 2023). Sample construction and calendar splits are described in the data section.

2.2 Decomposition Framework

The common-information design separates why an image model can outperform a tabular benchmark built from the same OHLCV window. At the population level, the image–tabular gap splits into encoding loss and tabular capacity loss. A separate feasible-training gap arises because models are fit on noisy realized returns; this channel is studied through distillation.

Suppressing stock and date subscripts, let $Y \in \{0, 1\}$ denote the five-day-ahead directional target, and let $X = \mathcal{R}(W)$ and $Z = \mathcal{G}(W)$ denote the image and tabular representations

[3]The moving-average overlays are trailing averages of the daily close and, at the earliest bars, draw on closes immediately preceding the window, reaching back to day $t - 38$ for the 20-day line. These pre-window closes are dated at or before t and enter the tabular representation below identically, so the overlays introduce no information from after the forecast date and no asymmetry between the two representations.

of the same window W.

Assumption 1 (Nested representations)**.** *At the representation level, the tabular variables are nested in the chart representation:*

$$\sigma(Z) \subseteq \sigma(X).$$

The condition is a representation-level idealization. It abstracts from finite-resolution rendering, pixel quantization, and clipping, and should not be read as a claim that the 220 numeric tabular features are exactly recoverable from the stored image pixels. Its role is to formalize the common-information benchmark in which both representations are built from the same OHLCV window and the tabular variables carry no source information beyond the chart representation. If Z retains numeric detail from the OHLCV window that is lost in rasterization, the tabular benchmark is advantaged relative to the ideal nesting case; the decomposition should then be interpreted as applying to the idealized chart representation rather than to exact raster pixels.

Define the image and tabular oracle predictors by

$$p_X^\star(x) = \mathbb{E}[Y \mid X = x], \qquad p_Z^\star(z) = \mathbb{E}[Y \mid Z = z].$$

Under Assumption 1, iterated expectations give $p_Z^\star(Z) = \mathbb{E}[p_X^\star(X) \mid Z]$. Define

$$\epsilon_{\mathcal{G}} = \mathbb{E}\Big[(p_X^\star(X) - p_Z^\star(Z))^2\Big], \quad \kappa_{\mathcal{F}} = \inf_{g \in \mathcal{F}} \mathbb{E}\Big[(p_Z^\star(Z) - g(Z))^2\Big].$$

The first term is encoding loss: return-relevant content preserved by the image oracle but not recovered by the tabular representation. The second is capacity loss: the part of the tabular oracle not approximable within $\mathcal{F}$. Here $\mathcal{F}$ denotes the population analogue of the tabular model class used in the empirical comparison, such as linear or nonlinear tabular predictors on Z.

Proposition 1 (Population representation decomposition)**.** *Under Assumption 1,*

$$\inf_{g\in\mathcal{F}} \mathbb{E}\Big[(Y - g(Z))^2\Big] - \mathbb{E}\Big[(Y - p_X^\star(X))^2\Big] = \epsilon_{\mathcal{G}} + \kappa_{\mathcal{F}}.$$

Proposition 1 is a population accounting identity. If the tabular function class were unrestricted, $\kappa_{\mathcal{F}} = 0$ and the gap would reduce to encoding loss. The identity contains no separate target-noise term: the irreducible variance of Y around the conditional mean is common to both oracle risks and cancels in the difference. Target noise matters only in the empirical problem, where fitted models are trained on noisy realized labels rather than on the latent conditional probability or the smoother image-model score. This motivates the distillation exercise.

Let $T(X) = \hat{p}^H(X)$ be the fitted image-model score used as a teacher target, and let

$$g_T^\star \in \arg\min_{g\in\mathcal{F}} \mathbb{E}[(T(X) - g(Z))^2]$$

be the population-best tabular student for reproducing it. For any tabular score $g(Z)$, define its population score-alignment with the teacher by

$$R^2(g, T) = 1 - \frac{\mathbb{E}[(T(X) - g(Z))^2]}{\mathrm{Var}(T(X))},$$

the population analogue of the out-of-sample R^2 to Hiera reported in the distillation exercise. Finally, let

$$m_T(z) = \mathbb{E}[T(X) \mid Z = z]$$

denote the population projection of the fitted image score on the tabular representation. For comparison with the latent full-window signal, define

$$p_W^\star(w) = \mathbb{E}[Y \mid W = w], \qquad h_Z^\star(z) = \mathbb{E}[p_W^\star(W) \mid Z = z].$$

Proposition 2 (Distillation alignment)**.** *Under the regularity assumptions stated in*

Appendix B,

$$1 - R^2(g_T^\star, T) - \frac{\mathbb{E}[(T(X) - m_T(Z))^2]}{\mathrm{Var}(T(X))} \xrightarrow{p} 0.$$

Moreover,

$$1 - R^2(g_T^\star, T) - \frac{\mathbb{E}[(p_W^\star(W) - h_Z^\star(Z))^2]}{\mathrm{Var}(T(X))} = O_p(\delta_T),$$

where δ_T is the teacher-approximation bound defined in Appendix B.

The regularity conditions require the tabular student to approximate the teacher-score projection on Z, the teacher score to be nondegenerate, and the teacher approximation error to be bounded. The first statement is a fitted-score accounting result: $1 - R^2$ after distillation measures the share of Hiera-score variation not recoverable from the tabular representation. The second statement relates this fitted-score residual to the latent full-window encoding residual up to the teacher-approximation error δ_T. Thus, a small teacher approximation error gives the stronger latent-signal interpretation, while the empirical distillation exercise itself remains a fitted-score recoverability object.

The empirical analysis maps these objects to three exercises. Comparisons between linear, nonlinear, and deeper tabular models assess the capacity component $\kappa_{\mathcal{F}}$. Perturbation and translation exercises diagnose which chart-specific configurations account for the residual encoding component associated with $\epsilon_{\mathcal{G}}$. Distillation asks how much more of the image score the tabular representation recovers once the noisy realized-return target is replaced by the smoother image-model score. The formal results are stated under squared-error loss because the projection and distillation objects are squared-error objects; Appendix B discusses the relation to cross-entropy training.

2.3 Prediction Models

Three models are estimated on the common window: one image model on $X_{i,t}$ and two tabular benchmarks on $Z_{i,t}$. The comparison is deliberately asymmetric in complexity. The question is not whether two flexible models perform similarly on the same window, but whether a simpler tabular benchmark constructed from the same data recovers the

image representation's apparent advantage, following the asset-pricing logic of testing whether one representation retains content beyond another (Freyberger et al., 2020). If it does, the remaining image advantage is unlikely to be driven by tabular capacity alone, which is the distinction formalized in Proposition 1.

Our objective is not to rank image architectures or revisit alternative visual backbones, but to decompose the content of a strong fitted image signal. The image model is a Hiera Tiny hierarchical vision transformer (Ryali et al., 2023). Its hierarchical architecture matches the structure of the chart. Some informative objects are local, such as individual candle bodies, wicks, and volume bars. Others emerge over short stretches, such as moving-average crossovers and reversals. Others are visible only at the level of the full path, such as trend shape and the position of recent price action within the window. A hierarchical model processes the image through stages at progressively coarser spatial resolution, so early stages encode localized configurations and later stages aggregate over wider regions (Liu et al., 2021; Ryali et al., 2023). The model maps $X_{i,t}$ into a probability score $s^H_{i,t} \in (0,1)$, interpreted as the estimated probability of a positive five-day-ahead return. The image is partitioned into non-overlapping patches, embedded as tokens, processed through transformer stages with hierarchical downsampling, pooled, and passed through a linear head to a scalar logit, with the score obtained via the logistic function. Training minimizes binary cross-entropy on $y^{(5)}_{i,t}$.

We fine-tune a pretrained checkpoint from a prior same-task run on the broader CRSP cross-section over 1993–2000, reported in Bøjstrup et al. (2026). We inherit the initialization but train on our own top-500 sample under an independent protocol.[4] The classification head and upper transformer blocks are updated while earlier layers stay fixed. To limit the influence of the more frequent return direction, the model is estimated on a class-balanced training subset, following evidence that skewed class frequencies distort classification under binary cross-entropy (He and Garcia, 2009; Buda et al., 2018); validation and

[4]The pretrain loss evaluation ends in 2003, strictly before this paper's validation (2004–2005) and out-of-sample (2006–2025) windows. No stock-date used in pretrain loss evaluation enters this paper's evaluation grids, which the chronological split guarantees. The shared pre-2001 top-500 coverage across the two training subsamples is therefore immaterial to the out-of-sample comparison.

out-of-sample evaluation retain the natural proportions ($P(y{=}1) \approx 0.539$). Because $\hat{p}^H(X)$ serves both as a cross-sectional ranking and as a continuous distillation target, its empirical calibration is reported in Figure D0.1: per-decile predicted and realized hit rates track the 45° line, with mean predicted probability (0.533) within 0.6 percentage points of the natural prior. Model selection uses validation log loss, and optimization uses AdamW with a scheduled learning rate, weight decay, gradient clipping, and early stopping. Hyperparameters are reported in Table A2.2.

The tabular benchmarks operate on the 220-dimensional vector $Z_{i,t}$ constructed from the same window. Two model classes are estimated. Logistic regression (Logit) provides a transparent linear benchmark. A histogram-based gradient-boosting classifier (HistGB) allows nonlinearities and interactions while remaining a structured tabular learner, following the gradient-boosting framework of Friedman (2001); the histogram-based implementation discretizes continuous predictors into bins to reduce split-search costs. Gradient-boosted trees are among the strongest tabular benchmarks in the machine learning asset-pricing literature (Gu et al., 2020). A multilayer perceptron and a tabular transformer on the same inputs perform similarly and do not narrow the gap further; these deep-tabular alternatives are reported in Appendix C as a robustness check against the reading that the image model wins on capacity rather than representation.

Both tabular benchmarks are estimated on the same class-balanced training subset as the image model and evaluated on the same out-of-sample panel under the common chronological split. Their estimator settings are held fixed across all baseline and translated specifications, so performance differences reflect representational content rather than training data, sample composition, or evaluation protocol.[5] Implementation details are reported in Table A2.1.

[5]The tabular benchmarks are not tuned on the 2004–2005 period: the Logit family runs a fixed number of streaming passes, and the HistGB family fixes its tree and regularization settings while letting the number of boosting iterations be set by the estimator's internal early-stopping holdout, drawn at random from the training rows. That period serves as a common split boundary rather than a model-selection set, so the closure of the gap across stages reflects the predictor set rather than estimator tuning.

2.4 Evaluation Design

All out-of-sample comparisons use a common evaluation panel that retains only stock-date observations for which every model produces a forecast and the realized forward return is identical. This alignment matters because the comparison concerns representations of the same data, not differences in data availability.

The evaluation combines portfolio-space and residual-content evidence. In portfolio space, stocks are sorted on each model's score into deciles on each rebalance date, and equal-weighted decile portfolios are held for five trading days to match the forecast horizon. The primary measure is the annualized high-minus-low (H-L) spread and its Sharpe ratio, following standard cross-sectional designs (Fama and French, 1992; Jegadeesh and Titman, 1993). Portfolios are rebalanced every fifth trading day, turnover is reported descriptively, and net-of-cost diagnostics under a standard one-way cost grid are reported in Subsection H1.

Residual content is assessed three ways. Following Fama and MacBeth (1973), each model's score is standardized within date and five-day-ahead returns are regressed on the standardized scores, asking whether one forecast retains cross-sectional content once competing forecasts are included. The same question is posed in portfolio space through spanning regressions of the Hiera spread on the benchmark spreads, with Newey–West standard errors (Newey and West, 1987); a nonzero intercept indicates return variation in the Hiera spread not spanned by the benchmarks. Conditional double sorts complement the regressions: stocks are grouped by one model's score and re-sorted within groups by a second, and economically meaningful second-stage spreads indicate information not absorbed by the first. This nonparametric design follows Daniel and Titman (1997) and captures nonlinear residual content that the linear Fama-MacBeth design may miss.

2.5 Perturbation and Translation

The evaluation design establishes whether the image model extracts information beyond the tabular benchmarks. It does not reveal where in the chart that information resides, or whether it can be made explicit. Two exercises address these questions in sequence. Perturbations locate the load-bearing chart components. Translations write those components into tabular predictors and test whether the benchmarks then recover the signal.

Perturbations alter selected parts of the rendered image while holding the fitted model fixed, so any change in predictions attaches to the information removed rather than to retraining, following Zeiler and Fergus (2014). For each perturbation π, a deterministic pixel-space operator $\Pi_\pi(\cdot)$ maps $X_{i,t}$ to a perturbed image $X_{i,t}^{\pi}$, and the fitted Hiera model is re-evaluated to yield a perturbed score $s_{i,t}^{H,\pi}$. The stock-date effect is $\Delta^{\pi} s_{i,t} = s_{i,t}^{H} - s_{i,t}^{H,\pi}$. Six perturbations each target a distinct visual component: masking the volume panel, masking the last five or first five candlesticks, converting the chart to grayscale, reversing it horizontally, and removing wick pixels. Each masked region corresponds exactly to a known chart object under the fixed renderer geometry; operator definitions are reported in Subsection A4. Because these perturbations modify rendered inputs away from their training distribution, the resulting performance changes need not equal the causal value of the removed component under model retraining. Deletion-based attribution can conflate the removal of informative content with artifacts introduced by the modified input (Hooker et al., 2019). We therefore interpret the exercises as diagnostics of the fixed fitted signal rather than as estimates of how a model trained without the corresponding chart component would perform. The operators target renderer-defined objects rather than generic image patches, and the first-five-bar and grayscale perturbations provide placebo comparisons for the more consequential edits. To characterize the states in which perturbations matter, the score changes are related to within-window state variables through date-by-date cross-sectional regressions (Fama and MacBeth, 1973). The full evaluation, comprising portfolio sorts, Fama-MacBeth and spanning regressions, and

conditional sorts, is repeated with Hiera's perturbed score in place of its baseline score, which shows whether a visually important component also matters for the part of Hiera's signal the benchmarks do not replicate.

Translation asks whether the located components can be recovered in explicit numeric form. Starting from the baseline vector $Z_{i,t}$, a nested sequence of progressively richer feature sets is constructed,

$$Z_{i,t}^{(0)} \subset Z_{i,t}^{(1)} \subset \cdots \subset Z_{i,t}^{(5)}, \tag{2.5}$$

by adding blocks that correspond to the chart components the perturbations identify: recent-bar geometry, wick structure, ordered directional patterns, price-volume interactions, and higher-level path shape. The sequence starts from the 220-dimensional baseline representation and ends with the 421-dimensional full translated representation; the intermediate dimensions are reported with the block definitions in Section 3.4. Each block is built from the same OHLCV window, so the exercise stays tied to the common-information design rather than becoming a generic expansion of the predictor set; feature definitions are reported in Subsection A3. At each stage $k \in \{0, \ldots, 5\}$, Logit and HistGB are re-estimated on $Z_{i,t}^{(k)}$ under the same training, validation, and evaluation protocol. Because the sets are nested, the sequence provides a controlled test of whether progressively richer translations of chart content narrow or eliminate Hiera's advantage over the tabular benchmarks.

3 Empirical Analysis

3.1 Data and evaluation sample

The analysis uses daily U.S. equity data from CRSP. The evaluation universe is a daily panel of the 500 largest stocks by market capitalization, reconstituted each trading day and spanning 2,173 distinct PERMNOs over the full sample. Restricting attention to this segment limits the scope for thin-trading and other microstructure frictions to contaminate the representation comparison (Chordia et al., 2008; Brandt et al., 2009). Each stock-date observation is a rolling 20-trading-day OHLCV window, mapped deterministically into the image and tabular representations defined in Section 2; observations without a full valid window are dropped. The admissible period runs from January 1, 1993 through December 31, 2025 and is split strictly chronologically: 1993–2003 for training, 2004–2005 for validation, and 2006–2025 for out-of-sample evaluation. The final sample contains 4,102,962 valid observations, of which 1,350,693 fall in the training period, 250,556 in the validation period, and 2,501,713 in the test period.

The forecast target is the direction of the five-day-ahead close-to-close return.[6] The positive-class share is 0.5109 in training, 0.5370 in validation, and 0.5394 in the test period, reflecting the upward drift in equity prices. For the economic tests, portfolios are rebalanced every fifth out-of-sample trading day to match the prediction horizon, yielding 1,006 non-overlapping rebalance dates over 2006–2025. Figure 3.1 shows representative candlestick images from the sample.

[6] Close-to-close returns use CRSP's default daily price series, including its standard treatment of distributions and of days without an observed transaction price.

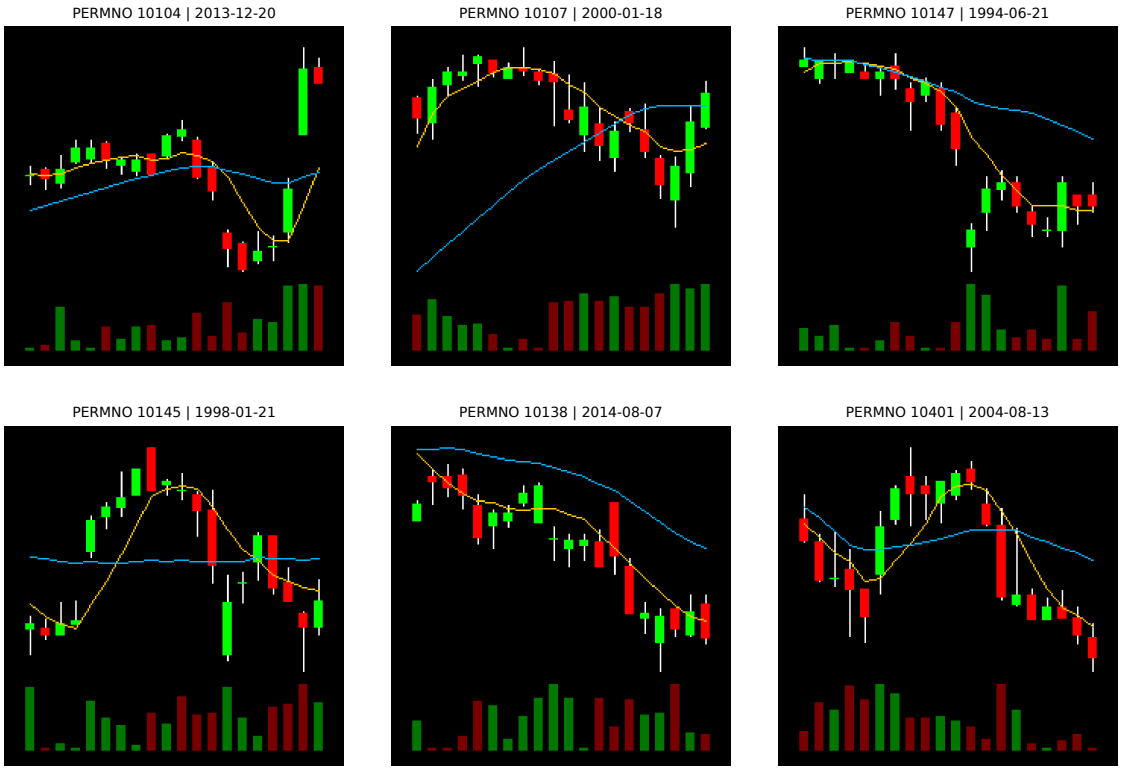


Figure 3.1: Representative candlestick images in the sample. Each panel is a 224×224 RGB candlestick chart rendered from a 20-trading-day open-high-low-close-volume window for one stock-date observation. Within each panel, the upper region shows daily candle bars with green (red) bars indicating up (down) close-to-close returns, the white wicks indicate intraday highs and lows, and the orange and blue overlays are short and long simple moving averages of the closing price. The lower region shows the corresponding daily traded volume, colored with the same up/down convention. The six panels are drawn from different stocks and rebalance dates to illustrate the range of chart morphologies the image model is exposed to in training; they are not selected for predictive content. The rendering specification is detailed in Table A1.1.

Alt text: Six candlestick charts arranged in a two-by-three grid, each showing 20 trading days of stock prices and volume. The examples display varied price patterns, including sharp reversals, sustained declines, rounded peaks, moving-average crossovers, and changes in trading volume; green and red candles indicate positive and negative daily returns, white wicks show intraday ranges, and orange and blue lines show short- and long-term moving averages.

3.2 Common-Information Comparison

We first ask whether the image representation extracts predictive content that the tabular representation does not recover when both are built from the same 20-day OHLCV window. Table 3.1 compares Hiera with the two tabular benchmarks in forecast space and portfolio space over 2006–2025. Hiera is the strongest model on every reported metric. Relative to HistGB, the stronger tabular benchmark, the annualized H–L spread rises from 0.1757 to 0.2067 and the Sharpe ratio from 1.343 to 1.496, at slightly lower turnover (0.818 versus 0.855). The gap relative to Logit is larger still. Because both models receive the same underlying data, the performance gap is attributable to representation and extraction rather than to differences in the information set.[7]

Table 3.1: Hiera versus Baseline Tabular Benchmarks, 2006–2025

	Forecast diagnostics			Portfolio diagnostics		
Model	Tail Spread @ 10%	Pearson IC	Spearman IC	H–L Return	H–L Sharpe	H–L Turnover
Hiera	0.0269***	0.0239***	0.0263***	0.2067	1.496	0.818
HistGB	0.0211***	0.0206***	0.0206***	0.1757	1.343	0.855
Logit	0.0147***	0.0144***	0.0159***	0.1403	1.037	0.870

Note: This table reports out-of-sample forecast and portfolio diagnostics for Hiera and the baseline tabular benchmarks over 2006–2025 on the common evaluation sample. At each rebalance date, stocks are sorted into equal-weighted model-score deciles within the common evaluation universe. Tail Spread @ 10% is the mean date-level difference in the frequency of positive 5-day-ahead outcomes between the top and bottom score deciles, $\Pr(y = 1 \mid \text{top } 10\%) - \Pr(y = 1 \mid \text{bottom } 10\%)$, where $y = 1$ indicates a positive cumulative close-to-close return over the next five trading days. Pearson IC and Spearman IC are the mean date-level cross-sectional Pearson and rank correlations, respectively, between model scores and realized 5-day-ahead returns. H–L Return and H–L Sharpe are the annualized return and Sharpe ratio of the equal-weighted high-minus-low decile spread formed from the same score sorts and rebalanced every five trading days. H–L Turnover is average one-way turnover per rebalance. The common evaluation sample contains 499,409 stock-horizon observations across 1,006 rebalance dates. Significance stars are reported for the forecast diagnostics and are based on Newey–West-adjusted tests with three lags. Split-level classification diagnostics are reported in Table D2.1. Alternative tail-cut decompositions and additional IC diagnostics are reported in Table D1.1 and Table D1.2. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 3.2 shows that the gap is concentrated on the short side. The top deciles of Hiera and HistGB earn similar annualized returns (0.2326 versus 0.2298), while Hiera's low-score decile earns 0.0259 against 0.0541 for HistGB and 0.0663 for Logit. Hiera's advantage lies in identifying weak names at the short end of the cross section; the broader long-short

[7]The headline tests are pre-specified in three blocks: a representation comparison, six frozen perturbations, and a translation-and-distillation set with nested feature blocks and a single pre-specified distillation contrast. Within each block we do not search over alternative perturbations or feature definitions. Cross-sectional Fama–MacBeth slopes, portfolio spanning intercepts, and conditional sorts each address logically distinct margins and use Newey–West-adjusted inference at consistent lag length. We report nominal significance without an additional family-wise correction (Harvey et al., 2016; Jensen et al., 2023).

Table 3.2: Decile Portfolio Comparison, 2006–2025

	Hiera			HistGB			Logit		
Portfolio	Return	Sharpe	Turnover	Return	Sharpe	Turnover	Return	Sharpe	Turnover
Low	0.0259	0.154	0.777	0.0541	0.321	0.842	0.0663	0.399	0.875
2	0.0790	0.456	0.876	0.0575	0.333	0.885	0.0683	0.390	0.890
3	0.0706	0.400	0.890	0.0798	0.449	0.896	0.0955	0.522	0.892
4	0.1064	0.542	0.896	0.1068	0.565	0.900	0.1019	0.565	0.897
5	0.1046	0.546	0.895	0.1257	0.660	0.898	0.1155	0.591	0.897
6	0.1305	0.712	0.895	0.1221	0.666	0.899	0.1366	0.693	0.899
7	0.1457	0.764	0.890	0.1478	0.742	0.896	0.1618	0.684	0.899
8	0.1746	0.764	0.884	0.1517	0.803	0.897	0.1455	0.764	0.895
9	0.1876	0.967	0.877	0.1811	0.782	0.894	0.1587	0.822	0.888
High	0.2326	1.143	0.860	0.2298	1.147	0.869	0.2066	1.061	0.865
H–L	0.2067	1.496	0.818	0.1757	1.343	0.855	0.1403	1.037	0.870

Note: This table reports out-of-sample equal-weighted decile portfolio results for Hiera and the baseline tabular benchmarks over 2006–2025 on the common evaluation sample. At each rebalance date, stocks are sorted into deciles by model score within the common evaluation universe. Low denotes the lowest-score decile and High denotes the highest-score decile. Reported returns are annualized close-to-close arithmetic mean returns. Sharpe ratios are computed from the corresponding decile return series. Turnover is average one-way turnover per rebalance. The H–L row reports the equal-weighted high-minus-low decile spread formed from the same score sorts and rebalanced every five trading days.

decomposition in Table D2.2 confirms this pattern.

Level outperformance does not establish that Hiera's ranking contains independent information: predictors that appear strong in isolation may add little once combined with competing signals (Lewellen, 2015). Table 3.3 addresses this with three tests. Cross-sectional Fama–MacBeth regressions ask whether the Hiera score retains an independent slope once the tabular scores enter directly. Portfolio spanning regressions ask whether the Hiera spread continues to earn alpha after projection onto the tabular spreads. Conditional sorts ask which ranking still discriminates returns within deciles of the competing score. Across all three, the Hiera signal is not subsumed: the Fama–MacBeth slope on the standardized Hiera score is 0.000791 with both tabular controls, and the spanning intercept on the combined tabular spreads is 0.001496, both significant at the 1% level.

Table 3.3: Residual Content of Hiera Conditional on Baseline Tabular Benchmark Scores, 2006–2025

Test	Specification	Reported statistic(s)
Cross-sectional regressions	Hiera + Logit	0.001093***
	Hiera + HistGB	0.000804***
	Hiera + Logit + HistGB	0.000791***
Portfolio spanning regressions	Hiera H–L on Logit + HistGB	0.001496***
	Hiera EW–Low on Logit + HistGB	0.000873***
Conditional sorts	Hiera within Logit deciles	0.1299 / 1.315
	Logit within Hiera deciles	0.0524 / 0.464
	Hiera within HistGB deciles	0.1308 / 1.227
	HistGB within Hiera deciles	0.0736 / 0.754

Note: This table reports residual-content tests for Hiera relative to the baseline tabular benchmarks over 2006–2025 on the common evaluation sample. In the cross-sectional-regression rows, the reported statistic is the average date-level slope on the standardized Hiera score from Fama–MacBeth-style cross-sectional regressions that also include the indicated standardized benchmark scores. In the portfolio-spanning-regression rows, the reported statistic is the intercept from time-series regressions of the Hiera spread return on the indicated baseline benchmark spread returns. EW denotes the equal-weighted market portfolio that holds all stocks in the evaluation universe, so EW–Low is the equal-weighted market-minus-low decile spread. In the conditional-sort rows, stocks are first assigned to deciles on the conditioning score and then to deciles on the secondary score within each conditioning decile; the reported statistic is the annualized return and Sharpe ratio of the equal-weighted conditional high-minus-low spread, reported as return / Sharpe. Reverse conditional sorts are included to assess asymmetry in residual content. Regression-based significance tests use Newey–West-adjusted inference with three lags. The full nested spanning ladder is reported in Table D3.2. Additional overlap and asymmetry diagnostics are reported in Table D3.1. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

The conditional sorts reveal an asymmetry. Conditional on the tabular rankings, Hiera delivers spreads of 0.1299 within Logit deciles and 0.1308 within HistGB deciles; the reverse sorts deliver only 0.0524 and 0.0736. Hiera does not merely differ from the tabular benchmarks. It contains the more consequential residual ordering once the common

component is partialed out. Expanded overlap diagnostics appear in Table D3.1.

3.3 Chart Perturbation Analysis

The previous section establishes that the image representation contains predictive content the tabular benchmarks do not recover. We next ask which components of the chart carry that content, using the perturbation design defined in Section 2.5. We apply targeted perturbations to the rendered chart at evaluation time while holding the trained model fixed, so that any change in performance reflect sensitivity to the disrupted visual component rather than to a change in the training environment.[8] Six perturbations isolate distinct chart components: masking the final five bars (recent price action), masking the first five bars (oldest part of the window), a horizontal flip (preserving local shapes but reversing temporal order), removing wicks (suppressing intrabar highs and lows), masking the volume panel, and converting the image to grayscale. Figure 3.2 illustrates the edits.

Table 3.4 reports the effect of each perturbation on level performance. The results show a sharp recency gradient. Masking the final five bars nearly eliminates tail separation and cuts the annualized H–L spread from 0.2067 to 0.0982; masking the first five bars leaves performance unchanged. This asymmetry is consistent with the importance of recent return realizations at short horizons (Lehmann, 1990; Lo and MacKinlay, 1990). The horizontal flip cuts the spread to 0.1232 even though local shapes are preserved, so Hiera responds to ordered price paths rather than to unordered visual fragments, consistent with the sequential character of chart information in the technical-analysis literature (Lo et al., 2000). Removing wicks lowers the spread more modestly, to 0.1443. Masking the volume panel and grayscale conversion have small effects (0.1806 and 0.2090).

The same gradient carries through to the residual-content diagnostics, with three informative dissociations. Table 3.5 re-estimates the cross-sectional regressions, portfolio spanning tests, and conditional sorts from Subsection 3.2 under each perturbation. Masking

[8]The logic follows the occlusion and perturbation-based interpretability methods developed in the computer-vision literature (Zeiler and Fergus, 2014; Fong and Vedaldi, 2017). Because the model weights are frozen throughout, the exercise characterizes the fitted signal, not the effect of retraining under different inputs.

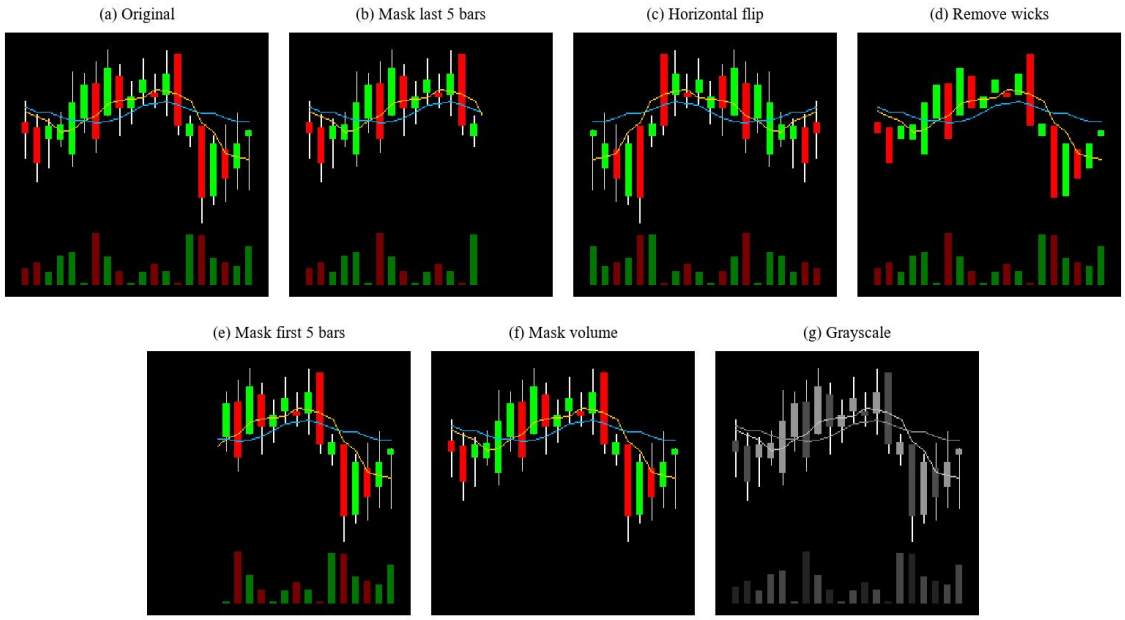


Figure 3.2: Illustrative evaluation-time perturbations of the chart. Panel (a) shows the unmodified rendered chart for a single stock-date observation. Panels (b)–(g) display the six perturbations applied at evaluation time to the same image, with the trained Hiera weights held fixed: masking the last five bars (b), reversing the left-to-right ordering of bars (c), removing the wicks that encode each day's high and low (d), masking the first five bars as a placebo (e), masking the volume panel (f), and converting the image to grayscale (g). The underlying 20-day open-high-low-close-volume window is unchanged; only the visual encoding seen by the model is altered. Performance under each perturbation is reported in Table 3.4.

Alt text: Seven candlestick charts showing the same 20-day stock-price window under different visual transformations. Panel (a) is the original chart. Panel (b) removes the five rightmost price and volume bars, panel (c) reverses the chart horizontally, panel (d) removes the high-low wicks, panel (e) removes the five leftmost price and volume bars, panel (f) removes all volume bars, and panel (g) converts the chart to grayscale. The remaining panels use green and red candles, white wicks, two moving-average lines, and a lower volume panel.

Table 3.4: Hiera Performance under Chart Perturbations, 2006–2025

Perturbation	Tail Spread @ 10%	Δ Tail Spread @ 10%	H–L Return	H–L Sharpe	H–L Turnover
Unperturbed image	0.0269	—	0.2067	1.496	0.818
Mask last 5 bars	0.0025	-0.0244	0.0982	0.722	0.804
Horizontal flip	0.0076	-0.0193	0.1232	0.980	0.742
Remove wicks	0.0226	-0.0043	0.1443	1.061	0.810
Mask volume panel	0.0253	-0.0016	0.1806	1.354	0.841
Grayscale image	0.0267	-0.0002	0.2090	1.505	0.813
Mask first 5 bars	0.0277	+0.0008	0.2068	1.527	0.824

Note: This table reports out-of-sample Hiera performance over 2006–2025 on the common evaluation sample under evaluation-time perturbations to the chart image. The model is held fixed throughout; perturbations are applied to the input image at evaluation only, with no retraining. The Unperturbed image row reports the baseline Hiera results. Tail Spread @ 10% is the mean date-level difference in the frequency of positive 5-day-ahead outcomes between the top and bottom score deciles, $\Pr(y = 1 \mid \text{top } 10\%) - \Pr(y = 1 \mid \text{bottom } 10\%)$, where $y = 1$ indicates a positive cumulative close-to-close return over the next five trading days. Δ Tail Spread @ 10% is the change relative to the unperturbed image. H–L Return and H–L Sharpe are the annualized return and Sharpe ratio of the equal-weighted high-minus-low decile spread formed from the same score sorts and rebalanced every five trading days. H–L Turnover is average one-way turnover per rebalance. Full decile ladders and score-stability diagnostics are reported in Table E2.1 and Table E1.1.

the final five bars weakens both margins: the Fama–MacBeth coefficient on the Hiera score falls from 0.000791 to 0.000371, and the spanning intercept falls from 0.001496 to 0.000813 and loses significance. The horizontal flip roughly halves the Fama–MacBeth coefficient (to 0.000385) and weakens the conditional sorts, yet leaves the spanning intercept essentially unchanged (0.001487) and significant: temporal order drives the cross-sectional part of Hiera's residual ranking, while the tradable spanning component is robust to reversing the bar sequence. Wick removal produces the opposite dissociation: the Fama–MacBeth coefficient remains significant (0.000613), but the spanning intercept falls to 0.000266 and loses significance, so intrabar high-low structure feeds specifically the tradable part of the residual spread. We discuss candidate microstructure interpretations of the wick result in Subsection F5. The placebo, volume-panel, and grayscale perturbations leave the residual content close to its unperturbed values.

The residual content that survives beyond the tabular benchmarks is concentrated in recent and correctly ordered price-path information, with intrabar extremal structure contributing specifically to the tradable component of the residual spread. Volume and color encoding play minor roles. This localization narrows the set of chart features a tabular representation must encode explicitly in order to recover the image signal, the question taken up in the next section.

Table 3.5: Perturbations and Hiera's Residual Content, 2006–2025

Perturbation	Conditional sorts				Cross-sectional regression	Portfolio spanning regression
	Within Logit		Within HistGB		FM coefficient	TS intercept
	H–L Return	H–L Sharpe	H–L Return	H–L Sharpe	(Logit + HistGB)	(Logit + HistGB)
Unperturbed image	0.1299	1.315	0.1308	1.227	0.000791***	0.001496***
Mask last 5 bars	0.0805	0.617	0.0473	0.378	0.000371**	0.000813
Horizontal flip	0.0965	0.879	0.1100	0.940	0.000385**	0.001487***
Remove wicks	0.1001	0.958	0.0704	0.633	0.000613***	0.000266
Mask volume panel	0.0954	0.913	0.1084	0.943	0.000634***	0.001133***
Grayscale image	0.1501	1.189	0.1092	1.000	0.000787***	0.001709***
Mask first 5 bars	0.1436	1.398	0.1308	1.205	0.000749***	0.001501***

Note: This table reports how evaluation-time chart perturbations affect the residual content of Hiera relative to the baseline tabular benchmarks over 2006–2025 on the common evaluation sample. In the conditional-sort columns, stocks are first assigned to deciles on the conditioning score and then to deciles on the perturbed Hiera score within each conditioning decile; the reported statistics are the annualized return and Sharpe ratio of the equal-weighted conditional high-minus-low spread. In the cross-sectional-regression column, the reported statistic is the average date-level slope on the standardized Hiera score from a Fama–MacBeth-style cross-sectional regression that controls jointly for the standardized Logit and HistGB scores. In the portfolio-spanning-regression column, the reported statistic is the intercept from a time-series regression of the Hiera high-minus-low spread return on the Logit and HistGB high-minus-low spread returns. This intercept is reported per rebalance period and is not annualized. Perturbations are applied to the input image at evaluation only, with no retraining. Regression-based significance tests use Newey–West-adjusted inference with three lags. The full conditional-sort decomposition and the full nested fragility ladder are reported in Table E2.2 and Table E2.3. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

3.4 Translating the Image Signal

The perturbation analysis in Section 3.3 identifies which chart components carry Hiera's fitted signal, but it does not by itself determine whether those components can be recovered in tabular form. We next implement the translation design defined in Section 2.5. This section constructs explicit tabular predictors, which we call translated feature blocks, and tests how much of the gap to Hiera they close. Each translated predictor is a deterministic function of the same 20-bar OHLCV window used by both Hiera and the baseline tabular representation, so the exercise remains within common information. Each perturbation result motivates one translated block: the right-edge result motivates a *Geometry* block (local price configuration of the most recent bars); the flip result motivates a *Direction* block (short-horizon directional sequencing and reversals) and a *Path* block (trajectory-level features such as efficiency, curvature, and endpoint location); the wick result motivates a *Wick* block (intrabar high-low structure); and the volume-panel result motivates a *Volume* block constructed as price-volume interactions rather than raw volume levels, since raw volume already enters the baseline representation. Carrying all five blocks forward avoids conditioning the tabular design on the perturbation ranking and lets the nested, add-one, and omission comparisons in Table 3.6 determine which blocks are useful in isolation, which matter jointly, and which add little once the others are present. Formal definitions appear in Appendix A3 and in Table A3.1–Table A3.5.[9]

Table 3.6 shows that the spread-level gap to Hiera is largely recoverable once the perturbation-identified chart components are written into explicit tabular predictors, and that the strongest HistGB specifications exceed Hiera's Sharpe ratio. For HistGB the recovery is not monotone in the set of included blocks. *Geometry* is the strongest one-block addition, raising the Sharpe from 1.343 to 1.595, yet the fully translated specification earns only 1.247, and removing *Geometry* or *Volume* from the full model produces the two

[9]As an external robustness check, Appendix Table F4.1 compares the translated models with a Murray et al. (2024)-style cumulative-path block constructed from the same 20-day window. That block is predictive in this sample, but it does not reproduce the same pattern of incremental spread improvements as the perturbation-guided translated models, and residual-content tests continue to reject subsumption of Hiera and the translated models by that external path signal.

Table 3.6: Translated Tabular Benchmarks

Model	Specification	Predictors	Tail Spread @ 10%	H–L Return	H–L Sharpe	Δ H–L Sharpe vs. Base
			Panel A: Cumulative Translation			
HistGB	Base	220	0.0211	0.1757	1.343	0.000
	+ Geometry	280	0.0231	0.1974	1.595	0.252
	+ Wick	320	0.0231	0.1882	1.514	0.171
	+ Direction	360	0.0217	0.1817	1.448	0.105
	+ Volume	399	0.0232	0.1879	1.509	0.166
	+ Path	421	0.0219	0.2131	1.247	-0.096
Logit	Base	220	0.0147	0.1403	1.037	0.000
	+ Geometry	280	0.0150	0.1469	1.112	0.075
	+ Wick	320	0.0150	0.1436	1.079	0.042
	+ Direction	360	0.0181	0.1647	1.281	0.244
	+ Volume	399	0.0156	0.1711	1.401	0.364
	+ Path	421	0.0140	0.1663	1.424	0.387
			Panel B: One Block Additions			
HistGB	Base	220	0.0211	0.1757	1.343	0.000
	Base + Geometry	280	0.0231	0.1974	1.595	0.252
	Base + Wick	260	0.0244	0.1970	1.497	0.154
	Base + Direction	260	0.0260	0.2052	1.548	0.205
	Base + Volume	259	0.0213	0.1768	1.270	-0.073
	Base + Path	242	0.0226	0.2373	1.325	-0.018
Logit	Base	220	0.0147	0.1403	1.037	0.000
	Base + Geometry	280	0.0150	0.1469	1.112	0.075
	Base + Wick	260	0.0132	0.1361	1.010	-0.027
	Base + Direction	260	0.0162	0.1653	1.197	0.160
	Base + Volume	259	0.0147	0.1405	1.120	0.083
	Base + Path	242	0.0164	0.1525	1.173	0.136

(Continued on next page)

Table 3.6: Translated Tabular Benchmarks (Continued)

Model	Specification	Predictors	Tail Spread @ 10%	H–L Return	H–L Sharpe	Δ H–L Sharpe vs. Base
			Panel C: One Block Omissions			
HistGB	Full	421	0.0219	0.2131	1.247	-0.096
	Full − Geometry	361	0.0246	0.2137	1.744	0.401
	Full − Wick	381	0.0214	0.2052	1.229	-0.114
	Full − Direction	381	0.0275	0.2047	1.236	-0.107
	Full − Volume	382	0.0255	0.2082	1.729	0.386
	Full − Path	399	0.0232	0.1879	1.509	0.166
Logit	Full	421	0.0140	0.1663	1.424	0.387
	Full − Geometry	361	0.0163	0.1676	1.413	0.376
	Full − Wick	381	0.0146	0.1671	1.419	0.382
	Full − Direction	381	0.0151	0.1991	1.181	0.144
	Full − Volume	382	0.0197	0.1715	1.360	0.323
	Full − Path	399	0.0156	0.1711	1.401	0.364

Note: This table reports out-of-sample results for the translated tabular benchmarks over 2006–2025 on the common evaluation sample. Panel A adds the translated feature blocks sequentially to the baseline tabular specification within each model family. Panel B adds each translated feature block individually to the baseline specification. Panel C compares the fully translated specification with versions that omit one translated feature block at a time. Base denotes the baseline tabular specification and Full denotes the specification that includes all translated feature blocks. Predictors reports the number of predictors in the corresponding specification. Tail Spread @ 10% is the mean date-level difference in the frequency of positive 5-day-ahead outcomes between the top and bottom score deciles, $\Pr(y = 1 \mid \text{top } 10\%) - \Pr(y = 1 \mid \text{bottom } 10\%)$, where $y = 1$ indicates a positive cumulative close-to-close return over the next five trading days. H–L Return and H–L Sharpe are the annualized return and Sharpe ratio of the equal-weighted high-minus-low decile spread formed from the corresponding score sorts and rebalanced every five trading days. Δ H–L Sharpe vs. Base is measured relative to the baseline specification within the corresponding model family. Formal definitions of the translated feature blocks are provided in Appendix A3 and in Table A3.1, Table A3.2, Table A3.3, Table A3.4, and Table A3.5. Full decile ladders for Panels A–C are reported in Table F2.2, Table F2.3, and Table F2.4. Tail-spread decompositions for Panels A–C are reported in Table F2.1, Table F3.1, and Table F2.5.

best HistGB values in the table (1.744 and 1.729, both above Hiera's 1.496). The marginal effect of a block therefore depends on which other blocks are present, a pattern consistent with correlated translated features and the flexibility of the nonlinear learner. For Logit the recovery is monotone: the fully translated specification raises the Sharpe from 1.037 to 1.424, and the omission tests show the largest drop when *Direction* is removed, consistent with the linear benchmark benefiting most from explicit encoding of temporal order.

Within-family improvements do not by themselves show that the translated models move closer to Hiera: two models can earn similar spreads while ranking the cross section differently. Table 3.7 reports pairwise return and Sharpe differences against Hiera. For HistGB, the baseline return gap of 3.10 percentage points is already insignificant ($p = 0.172$) and narrows further or changes sign under translation. For Logit, the baseline gap is significant (6.64 pp, $p = 0.013$) but contracts as translated blocks are added, and neither the return nor the Sharpe difference is significant at conventional levels in the richer translated specifications. Spread-level convergence, and for HistGB reversal, is however a statement about levels, not about information: the paper's claim concerns whether the two representations carry the same cross-sectional content, which is the subject of the next section.

Table 3.7: Direct Comparison of Hiera and the Translated Tabular Benchmarks, 2006–2025

Model	Specification	Δ H–L Return (pp)	Return p-value	Δ H–L Sharpe	Sharpe p-value
		Panel A: Cumulative Translation			
HistGB	Base	3.10	0.172	0.153	0.234
	+ Geometry	0.93	0.695	-0.099	0.551
	+ Wick	1.85	0.415	-0.018	0.951
	+ Direction	2.50	0.271	0.048	0.676
	+ Volume	1.88	0.419	-0.013	0.983
	+ Path	-0.64	0.864	0.249	0.372
Logit	Base	6.64**	0.013	0.459**	0.021
	+ Geometry	5.98**	0.016	0.384*	0.051
	+ Wick	6.31**	0.012	0.417**	0.030
	+ Direction	4.20	0.102	0.216	0.342
	+ Volume	3.56	0.111	0.095	0.558
	+ Path	4.04*	0.078	0.073	0.665
		Panel B: One Block Additions			
HistGB	Base	3.10	0.172	0.153	0.234
	Base + Geometry	0.93	0.695	-0.099	0.551
	Base + Wick	0.97	0.686	0.000	0.946
	Base + Direction	0.15	0.950	-0.052	0.771
	Base + Volume	2.99	0.179	0.226	0.222
	Base + Path	-3.06	0.436	0.171	0.662
Logit	Base	6.64**	0.013	0.459**	0.021
	Base + Geometry	5.98**	0.016	0.384*	0.051
	Base + Wick	7.06***	0.008	0.486**	0.011
	Base + Direction	4.14	0.113	0.300	0.156
	Base + Volume	6.62***	0.008	0.376**	0.027
	Base + Path	5.42**	0.035	0.323	0.120
		Panel C: One Block Omissions			
HistGB	Full	-0.64	0.864	0.249	0.372
	Full − Geometry	-0.70	0.768	-0.247	0.136
	Full − Wick	0.15	0.968	0.267	0.384
	Full − Direction	0.20	0.958	0.260	0.327
	Full − Volume	-0.15	0.943	-0.233	0.133
	Full − Path	1.88	0.419	-0.013	0.983
Logit	Full	4.04*	0.078	0.073	0.665
	Full − Geometry	3.91	0.102	0.083	0.630
	Full − Wick	3.96*	0.083	0.077	0.632
	Full − Direction	0.76	0.830	0.315	0.196
	Full − Volume	3.52	0.174	0.136	0.552
	Full − Path	3.56	0.111	0.095	0.558

Note: This table reports direct out-of-sample comparisons between Hiera and the translated tabular benchmarks over 2006–2025 on the common evaluation sample. Each entry compares the equal-weighted high-minus-low decile spread formed from Hiera scores with the corresponding equal-weighted high-minus-low decile spread formed from the indicated translated tabular specification. Δ H–L Return (pp) is the annualized difference in spread returns, in percentage points, computed as Hiera minus model. Its reported p-value is based on a Newey–West test of the paired rebalance-return difference series with five lags. Δ H–L Sharpe is the difference in annualized spread Sharpe ratios, again computed as Hiera minus model. Its reported p-value is obtained from a circular block bootstrap applied to the paired spread-return series using block length five. Positive values favor Hiera. Panel A reports cumulative translation results, Panel B reports one-block additions to the baseline specification, and Panel C reports one-block omissions from the fully translated specification. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

3.5 Residual Content after Translation

We reapply the cross-sectional regressions, portfolio spanning tests, and conditional sorts from Section 3.2, with the translated tabular benchmarks in place of the original baselines. Table 3.8 asks whether the translated benchmarks retain residual content after the Hiera score is controlled for. The answer is yes in both model families, and the change is starkest for Logit: the baseline linear benchmark added little once Hiera entered, whereas the richer translated Logit models retain positive and significant cross-sectional slopes, generate significant spanning intercepts, and deliver nontrivial conditional-sort spreads. Translation therefore does more than move the linear benchmark closer to Hiera in spread terms. It converts part of the chart content into tabular information that remains return-relevant after Hiera is taken into account. For HistGB, which retained residual content before translation, that content strengthens along the cumulative ladder: the cross-sectional coefficient rises from 0.000571 to 0.000702 and the spanning intercept from 0.000874 to 0.002107.

Table 3.9 asks the reverse question: does Hiera retain residual content once the translated benchmarks are controlled for? Across both model families and all three panels, it does. Even against the strongest translated HistGB models, the Hiera cross-sectional coefficient remains at least 0.000637 and the Hiera-within-model conditional spreads remain at least 0.1017 annualized; against the fully translated Logit specification, the coefficient is 0.000947 with a conditional spread of 0.1233. All spanning intercepts remain significant.

The two tables jointly characterize the post-translation relationship between the representations as partial overlap rather than nesting. The perturbation-guided translation converts part of the chart signal into explicit tabular predictors, and that converted content is economically meaningful: it improves the tabular benchmarks, narrows their unconditional distance to Hiera, and generates its own residual content after Hiera is controlled for. But the image representation retains return-relevant information beyond those predictors even after the tabular side is rebuilt around the chart components identified in the perturbation analysis. The gap is reduced, not eliminated.

Table 3.8: Residual Content of the Translated Tabular Benchmarks after Controlling for Hiera, 2006–2025

		Cross-sectional regression	Portfolio spanning regression	Conditional sorts: Model within Hiera	
Model	Specification	FM $\beta_{\text{Model}\|\text{Hiera}}$	TS $\alpha_{\text{Model}\|\text{Hiera}}$	Ann. H–L return	Sharpe
		Panel A: Cumulative Translation			
HistGB	Base	0.000571***	0.000874*	0.0768	0.456
	+ Geometry	0.000678***	0.001492***	0.1570	0.792
	+ Wick	0.000617***	0.001152***	0.1094	0.698
	+ Direction	0.000690***	0.001067***	0.0404	0.195
	+ Volume	0.000680***	0.001340***	0.1350	0.590
	+ Path	0.000702***	0.002107**	0.1343	0.518
Logit	Base	0.000205	0.000352	0.0790	0.394
	+ Geometry	0.000253*	0.000461	0.0372	0.121
	+ Wick	0.000268**	0.000387	0.0488	0.143
	+ Direction	0.000426***	0.000967**	0.0873	0.460
	+ Volume	0.000505***	0.001056**	0.0957	0.716
	+ Path	0.000528***	0.001220***	0.1033	0.757
		Panel B: One Block Additions			
HistGB	Base	0.000571***	0.000874*	0.0768	0.456
	Base + Geometry	0.000678***	0.001492***	0.1570	0.792
	Base + Wick	0.000484***	0.001306***	0.0493	0.316
	Base + Direction	0.000814***	0.001331***	0.1546	0.795
	Base + Volume	0.000487***	0.000535	-0.0368	-0.080
	Base + Path	0.000862***	0.002227***	0.1570	0.706
Logit	Base	0.000205	0.000352	0.0790	0.394
	Base + Geometry	0.000253*	0.000461	0.0372	0.121
	Base + Wick	0.000224*	0.000274	0.0086	0.029
	Base + Direction	0.000417***	0.000742	0.0870	0.451
	Base + Volume	0.000319**	0.000509	0.0483	0.308
	Base + Path	0.000303**	0.000739	0.0821	0.491

(Continued on next page)

Table 3.8: Residual Content of the Translated Tabular Benchmarks after Controlling for Hiera (Continued)

		Cross-sectional regression	Portfolio spanning regression	Conditional sorts: Model within Hiera	
Model	Specification	FM $\beta_{\text{Model\|Hiera}}$	TS $\alpha_{\text{Model\|Hiera}}$	Ann. H–L return	Sharpe
		Panel C: One Block Omissions			
HistGB	Full	0.000702^{***}	0.002107^{**}	0.1343	0.518
	Full − Geometry	0.000785^{***}	0.001974^{***}	0.1548	0.856
	Full − Wick	0.000681^{***}	0.001973^{**}	0.1777	0.530
	Full − Direction	0.000626^{***}	0.001988^{**}	0.1327	0.558
	Full − Volume	0.000743^{***}	0.001714^{***}	0.1429	0.665
	Full − Path	0.000680^{***}	0.001340^{***}	0.1350	0.590
Logit	Full	0.000528^{***}	0.001220^{***}	0.1033	0.757
	Full − Geometry	0.000501^{***}	0.001176^{***}	0.1215	0.896
	Full − Wick	0.000504^{***}	0.001190^{***}	0.0806	0.606
	Full − Direction	0.000469^{***}	0.001733^{**}	0.1367	0.762
	Full − Volume	0.000449^{***}	0.001312^{***}	0.0970	0.428
	Full − Path	0.000505^{***}	0.001056^{**}	0.0957	0.716

Note: This table reports residual-content tests for the translated tabular benchmarks after controlling for Hiera over 2006–2025 on the common evaluation sample. Panel A reports cumulative translation results, Panel B reports one-block additions to the baseline specification, and Panel C reports one-block omissions from the fully translated specification. FM $\beta_{\text{Model|Hiera}}$ is the average date-level slope on the translated tabular benchmark score from a Fama–MacBeth-style cross-sectional regression that includes both the translated tabular benchmark score and the Hiera score. TS $\alpha_{\text{Model|Hiera}}$ is the intercept from a time-series regression of the translated benchmark high-minus-low spread return on the Hiera high-minus-low spread return. This intercept is reported per rebalance period and is not annualized. In the conditional-sort columns, stocks are first assigned to deciles on the Hiera score and then to deciles on the translated tabular benchmark score within each Hiera decile; the reported statistics are the annualized return and Sharpe ratio of the equal-weighted conditional high-minus-low spread. Regression-based significance tests use Newey–West-adjusted inference with three lags. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 3.9: Residual Content of Hiera after Controlling for the Translated Tabular Benchmarks, 2006–2025

		Cross-sectional regression	Portfolio spanning regression	Conditional sorts: Hiera within Model	
Family	Specification	FM $\beta_{\text{Hiera}\|\text{Model}}$	TS $\alpha_{\text{Hiera}\|\text{Model}}$	Ann. H–L return	Sharpe
		Panel A: Cumulative Translation			
HistGB	Base	0.000803***	0.001625***	0.1909	0.835
	+ Geometry	0.000745***	0.001217***	0.1361	0.629
	+ Wick	0.000800***	0.001199***	0.1271	0.666
	+ Direction	0.000736***	0.001396***	0.1219	0.536
	+ Volume	0.000771***	0.001427***	0.1348	0.577
	+ Path	0.000748***	0.002672***	0.1625	0.544
Logit	Base	0.001092***	0.002381***	0.1287	0.761
	+ Geometry	0.001071***	0.002194***	0.2570	0.534
	+ Wick	0.001066***	0.002258***	0.1577	1.054
	+ Direction	0.000979***	0.001986***	0.1722	0.620
	+ Volume	0.000962***	0.001624***	0.1636	0.760
	+ Path	0.000947***	0.001761***	0.1233	0.731
		Panel B: One Block Additions			
HistGB	Base	0.000803***	0.001625***	0.1909	0.835
	Base + Geometry	0.000745***	0.001217***	0.1361	0.629
	Base + Wick	0.000877***	0.001369***	0.1442	0.745
	Base + Direction	0.000651***	0.001146**	0.1017	0.546
	Base + Volume	0.000898***	0.001596***	0.1252	0.858
	Base + Path	0.000637***	0.002405***	0.2187	0.636
Logit	Base	0.001092***	0.002381***	0.1287	0.761
	Base + Geometry	0.001071***	0.002194***	0.2570	0.534
	Base + Wick	0.001092***	0.002421***	0.1540	0.746
	Base + Direction	0.000994***	0.002071***	0.1509	0.539
	Base + Volume	0.001057***	0.002222***	0.1133	0.649
	Base + Path	0.001052***	0.002196***	0.1394	0.583

(Continued on next page)

Table 3.9: Residual Content of Hiera after Controlling for the Translated Tabular Benchmarks (Continued)

		Cross-sectional regression	Portfolio spanning regression	Conditional sorts: Hiera within Model	
Family	Specification	FM $\beta_{\text{Hiera}\|\text{Model}}$	TS $\alpha_{\text{Hiera}\|\text{Model}}$	Ann. H–L return	Sharpe
		Panel C: One Block Omissions			
HistGB	Full	0.000748^{***}	0.002672^{***}	0.1625	0.544
	Full − Geometry	0.000717^{***}	0.001125^{**}	0.1332	0.495
	Full − Wick	0.000761^{***}	0.002674^{***}	0.1951	0.632
	Full − Direction	0.000790^{***}	0.002671^{***}	0.1609	0.468
	Full − Volume	0.000710^{***}	0.000897^{**}	0.1581	0.451
	Full − Path	0.000771^{***}	0.001427^{***}	0.1348	0.577
Logit	Full	0.000947^{***}	0.001761^{***}	0.1233	0.731
	Full − Geometry	0.000966^{***}	0.001736^{***}	0.1661	0.734
	Full − Wick	0.000956^{***}	0.001736^{***}	0.1354	0.616
	Full − Direction	0.000984^{***}	0.002666^{***}	0.1304	0.838
	Full − Volume	0.000966^{***}	0.002018^{***}	0.1491	0.420
	Full − Path	0.000962^{***}	0.001624^{***}	0.1636	0.760

Note: This table reports residual-content tests for Hiera after controlling for the translated tabular benchmarks over 2006–2025 on the common evaluation sample. Panel A reports cumulative translation results, Panel B reports one-block additions to the baseline specification, and Panel C reports one-block omissions from the fully translated specification. FM $\beta_{\text{Hiera}|\text{Model}}$ is the average date-level slope on the Hiera score from a Fama–MacBeth-style cross-sectional regression that includes both the Hiera score and the translated tabular benchmark score. TS $\alpha_{\text{Hiera}|\text{Model}}$ is the intercept from a time-series regression of the Hiera high-minus-low spread return on the translated tabular benchmark high-minus-low spread return. This intercept is reported per rebalance period and is not annualized. In the conditional-sort columns, stocks are first assigned to deciles on the translated tabular benchmark score and then to deciles on the Hiera score within each translated-benchmark decile; the reported statistics are the annualized return and Sharpe ratio of the equal-weighted conditional high-minus-low spread. Regression-based significance tests use Newey–West-adjusted inference with three lags. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

3.6 Forecast Combination

The partial-overlap result in Section 3.5 implies that combining the two representations should improve on either alone. We test this with a simple 50/50 probability-average ensemble of Hiera and the fully translated HistGB model. The combination uses HistGB because it is the stronger tabular benchmark and the more demanding comparator: any combination gain against a flexible nonlinear learner cannot be attributed to the tabular side lacking functional form. Table 3.10 reports the results.

The ensemble raises the annualized H–L return to 0.2540, against 0.2067 for Hiera and 0.2131 for HistGB (Full); both ICs and the tail spread also increase. The gain is concentrated on the long side: the High–EW return rises to 0.1560 versus 0.1069 for Hiera, while short-side performance matches Hiera (EW–Low Sharpe 1.350 versus 1.353). The ensemble does not dominate on Sharpe (1.398 versus 1.496), reflecting higher spread volatility from combining signals with imperfectly correlated residuals. This contrast is itself informative: a 50/50 average of redundant signals would reproduce the common ordering and lie between its components on both axes, whereas a combination that adds information can improve one axis while giving up ground on the other.

The cross-sectional and portfolio diagnostics in the lower panel of Table 3.10 carry different weight. Because the ensemble score is a fixed linear combination of the component scores, the Fama–MacBeth regressions reparametrize the joint regression on Hiera and HistGB rather than test content beyond the combined signal; the near mirror-image comparator coefficients ($t = 1.55$ and $t = -1.56$) are the mechanical reflection of this, as detailed in Appendix G. The portfolio spanning tests are the informative margin, because the ensemble decile spread sorts on the combined score and is not a linear function of the component spreads. There the result is one-sided. HistGB earns no residual alpha on the ensemble spread (-0.0002 per rebalance, $p = 0.523$), so the ensemble fully spans the translated tabular portfolio. Hiera retains a small but significant margin (0.0019 per rebalance, $p = 0.037$). The ensemble captures nearly all of both component signals at the portfolio level, with a small residual image component surviving, consistent with partial

Table 3.10: Forecast Combination of Hiera and the Fully Translated HistGB Model, 2006–2025

Model	Pearson IC	Spearman IC	Tail Spread @ 10%	High–EW Return	EW–Low Return	H–L Return
Hiera	0.024	0.026	0.0263	0.1069	0.0998	0.2067
HistGB (Full)	0.023	0.023	0.0220	0.1318	0.0813	0.2131
Ensemble (50/50)	0.026	0.027	0.0268	0.1560	0.0980	0.2540

Model	High–EW Sharpe	EW–Low Sharpe	H–L Sharpe
Hiera	1.173	1.353	1.496
HistGB (Full)	0.960	1.166	1.247
Ensemble (50/50)	1.084	1.350	1.398

Incremental-content tests

Comparison	FM coef. on ensemble	FM coef. on comparator	H–L alpha of ensemble on comparator	p-value	H–L alpha of comparator on ensemble	p-value
Ensemble vs. Hiera	0.001 ($t = 4.44$)	0.000 ($t = 1.55$)	0.0019	0.015	0.0019	0.037
Ensemble vs. HistGB (Full)	0.002 ($t = 4.90$)	-0.000 ($t = -1.56$)	0.0008	0.001	-0.0002	0.523

Note: This table reports out-of-sample results for a simple 50/50 probability-average ensemble of Hiera and the fully translated HistGB model over 2006–2025 on the common aligned out-of-sample panel for which both component scores are available. Hiera-only diagnostics in this table (e.g., Tail Spread @ 10% = 0.0263) are computed on this aligned panel and therefore differ marginally from the Hiera-only values in Table 3.1 (0.0269), which are computed on the full Hiera evaluation panel. The ensemble score is the equal-weighted average of the two model probabilities. Pearson IC and Spearman IC are the mean date-level cross-sectional Pearson and rank correlations, respectively, between model scores and realized 5-day-ahead returns. Tail Spread @ 10% is the mean date-level difference in the frequency of positive 5-day-ahead outcomes between the top and bottom score deciles, $\Pr(y = 1 \mid \text{top } 10\%) - \Pr(y = 1 \mid \text{bottom } 10\%)$, where $y = 1$ indicates a positive cumulative close-to-close return over the next five trading days. EW denotes the equal-weighted market portfolio that holds all stocks in the evaluation universe. High–EW Return is the annualized return of the highest-score decile relative to EW, EW–Low Return is the annualized return of EW relative to the lowest-score decile, and H–L Return is the annualized return of the equal-weighted high-minus-low decile spread. The corresponding Sharpe ratios are computed from the same rebalance-return series. In the incremental-content rows, the FM columns report average date-level slopes from Fama–MacBeth-style cross-sectional regressions that include both the ensemble score and the comparator score; reported t-statistics are based on Newey–West-adjusted inference with three lags. The H–L alpha columns report intercepts from time-series regressions of one high-minus-low spread return on the other; these intercepts are reported per rebalance period and are not annualized, and reported p-values are based on Newey–West-adjusted inference with three lags. Full headline classification diagnostics, score-overlap measures, and leg-level spanning results are reported in Appendix G.

overlap. What share of the original gap that residual represents, and how much of the remainder is an artifact of the noisy return target, is the subject of the final exercise.

3.7 Distillation and Decomposition

This section applies the distillation diagnostic introduced in Section 2.2. The population decomposition in Proposition 1 separates encoding loss from tabular capacity loss. Distillation addresses the remaining feasible-training question: how much more of the image score can the tabular representation recover when the noisy realized-return target is replaced by the smoother image-model score?

Operationally, we train HistGB on Hiera's continuous score rather than on the directional target, in the manner of knowledge distillation (Hinton et al., 2015). We then compare the alignment of the return-trained and score-trained tabular models with the Hiera score. The increase in out-of-sample R^2 to Hiera is interpreted as the noisy-supervision, or target-mismatch, component of the apparent image advantage under the distillation accounting. The residual lack of alignment after distillation measures the part of the image-model score not recovered from the tabular representation, under the maintained teacher interpretation in Proposition 2. Deep tabular benchmarks in Appendix C provide a check on whether remaining capacity constraints are binding.[10]

Panel A of Table 3.11 shows that distillation moves the tabular score much closer to Hiera without materially changing raw spread performance. The date-average Pearson correlation with Hiera rises from 0.596 to 0.848 and the out-of-sample R^2 from 0.454 to 0.783. The annualized H–L return is essentially unchanged (0.2131 versus 0.2102), while the Sharpe ratio rises from 1.247 to 1.445, reflecting tighter alignment with Hiera's more efficient ranking. The effect of distillation is not a stronger spread strategy; it is a measurement of how much of Hiera's internal ordering the translated predictor space reproduces once the noisy return target is replaced.

[10]Distillation is restricted to HistGB because the exercise requires a continuous-target student with the same nonlinear tabular architecture used in the strongest benchmark. Logit is retained as the transparent binary-response benchmark.

Table 3.11: Distillation from the Fully Translated Predictor Set, 2006–2025

Panel A: Score Alignment and Spread Performance

Model	Pearson	Spearman	Date-avg. Pearson	Date-avg. Spearman	OOS R^2 to Hiera	Ann. H–L Return	H–L Sharpe
HistGB Full	0.674	0.648	0.596	0.572	0.4539	0.2131	1.247
HistGB distilled	0.885	0.867	0.848	0.825	0.7830	0.2102	1.445

Panel B: Conditional diagnostics

Comparison	FM $\beta_{a\|b}$	FM $\beta_{b\|a}$	TS $\alpha_{a\|b}$	TS $\alpha_{b\|a}$	Residual H–L $a \mid b$	Residual H–L $b \mid a$
HistGB Full vs. Hiera	0.000702***	0.000748***	0.002107**	0.002672***	0.0907***	0.1062***
HistGB distilled vs. Hiera	0.000711***	0.000558***	0.000641*	0.000862**	0.0618**	0.0490**
HistGB distilled vs. HistGB Full	0.000757***	0.000630***	0.002331***	0.001723***	0.1185***	0.1141***

Note: This table reports out-of-sample distillation results for two tabular models built on the fully translated predictor set over 2006–2025 on the common aligned out-of-sample panel for which all compared scores are available. Panel A reports score-alignment and spread-performance diagnostics relative to Hiera. Pearson and Spearman are full-sample correlations between the model score and the Hiera score. Date-avg. Pearson and Date-avg. Spearman are the mean across rebalance dates of the corresponding date-level cross-sectional correlations. OOS R^2 to Hiera is the pooled OLS-calibrated R^2 on the common out-of-sample panel, equal to the squared pooled Pearson correlation between the model score and the Hiera score, so that both rows use the same definition and the residual share in Table 3.12 is the complement of the distilled row. Ann. H–L Return and H–L Sharpe are the annualized return and Sharpe ratio of the equal-weighted high-minus-low decile spread formed from each model score and rebalanced every five trading days. Panel B reports conditional diagnostics for pairwise score comparisons. In each row, a denotes the first score named and b the second. FM $\beta_{a|b}$ is the average date-level slope on score a from a Fama–MacBeth-style cross-sectional regression that includes both score a and score b. TS $\alpha_{a|b}$ is the intercept from a time-series regression of the high-minus-low spread return formed on score a on the corresponding high-minus-low spread return formed on score b; this intercept is reported per rebalance period and is not annualized. Residual H–L $a \mid b$ is the annualized return of the equal-weighted high-minus-low decile spread formed from the residual obtained by projecting score a on score b date by date. Reported significance tests in Panel B use Newey–West-adjusted inference with three lags. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Panel B of Table 3.11 confirms that the tighter alignment is not cosmetic. Against the distilled model, Hiera's spanning alpha falls from 0.002672 to 0.000862 and its residual H–L spread from 0.1062 to 0.0490; the distilled model's own margins on Hiera decline in parallel, from 0.002107 to 0.000641 and from 0.0907 to 0.0618. A substantial part of the pre-distillation distance between the two representations was target mismatch rather than missing information.

Panel A of Table 3.12 summarizes the distillation accounting in two variance-share blocks. The score block decomposes Hiera's cross-sectional score variance using the pooled R^2 from Table 3.11 Panel A; the return block uses the time-series spanning-regression R^2 between each model's H–L return and Hiera's H–L return. In each block, the first row is the share already captured by the translated HistGB model trained on realized Y. The second row is the additional share recovered when the same tabular model is trained on Hiera's score. The third row is the residual share not recovered by the tabular representation after distillation.

The two blocks agree on the qualitative pattern. In score variance, HistGB Full captures 45.4% of Hiera's score under noisy supervision, a further 32.9% becomes recoverable under the image-score target, and 21.7% remains unrecovered. In return-series variance, the corresponding shares are 17.5%, 49.4%, and 33.2%. Under the distillation accounting, the additional recovered share accounts for roughly 60% of the original gap to Hiera (0.329/0.546 in scores and 0.494/0.825 in returns). Thus, much of the apparent image advantage is tied to target mismatch under the distillation accounting rather than to irreducible chart-specific information.

Panel B of Table 3.12 gives the residual score component an economic magnitude. The residual H–L spread of Hiera after date-by-date projection on the distilled student is 0.0490 annualized with a 95% confidence interval of [0.0101, 0.0879]. This is the portfolio-space magnitude of the image-score component not recovered by the tabular representation after distillation. The opposite-direction residual, 0.0618, measures tabular content not present in the image score and should not be interpreted as encoding loss.

Table 3.12: Variance-Share Decomposition of Hiera's Cross-Sectional Score, 2006–2025

Panel A: Variance-share decomposition				
	Score variance ($\hat{p}^H$)		Return-series variance (H–L)	
Share of Hiera variance	R^2	95% CI	R^2	95% CI
Captured by HistGB Full (trained on Y)	0.454	[0.443, 0.466]	0.175	[0.059, 0.462]
Recovered under distillation/target mismatch	0.329	[0.321, 0.337]	0.494	[0.194, 0.661]
Residual not recovered after distillation	0.217	[0.212, 0.223]	0.332	[0.187, 0.532]

Panel B: Annualized residual H–L spread that survives distillation		
	Residual H–L spread	95% CI
Distilled student | Hiera (forward)	0.0618	[0.0143, 0.1093]
Hiera | distilled student (reverse)	0.0490	[0.0101, 0.0879]

Note: This table reports two parallel additive decompositions of Hiera, both following the variance-share accounting described in the text. Each column block sums to 1 by construction. The score-variance block (left) decomposes the variance of Hiera's cross-sectional score $\hat{p}^H$ on the common aligned out-of-sample panel (n = 499,409 stock-rebalance observations across 1,006 rebalance dates) and uses the pooled OLS-calibrated R^2 definition described in Table 3.11 Panel A (squared pooled Pearson correlation between the model score and the Hiera score). The return-series block (right) decomposes the variance of Hiera's high-minus-low decile spread return across the 1,006 five-day rebalance periods using the corresponding time-series spanning-regression R^2 between each model's H–L return and Hiera's H–L return (the value 0.175 in the first row is the R^2 of the time-series spanning regression between HistGB Full's and Hiera's H–L returns—the same regression whose per-period intercepts, the TS α columns, are reported for the HistGB Full vs. Hiera comparison in Table 3.11 Panel B; Table 3.11 tabulates that spanning intercept rather than its R^2). Within each block, the first row is the share already captured by the structured tabular model trained on realized returns; the second row is the additional share recovered when the same model is trained on Hiera's continuous score; the third row is the complementary residual share not recovered after distillation. Score 95% CIs are obtained by a 200-replication block bootstrap that resamples whole rebalance dates with seed 42; return 95% CIs are obtained by a 200-replication non-overlapping period bootstrap on the H–L return series. The wider return-series intervals reflect the much smaller effective sample size (1,006 periods vs. 499,409 cross-sectional observations) and a higher noise-to-signal ratio in aggregated portfolio returns. The deep tabular benchmarks in Appendix C do not narrow the gap to Hiera relative to HistGB on the same feature set, so additional tabular capacity does not appear to be the binding constraint in this setting. Panel B reports the annualized residual H–L spread that survives distillation, with Newey–West 95% confidence intervals at three lags. The Hiera | distilled student row is the portfolio-space magnitude of the residual image-score component not recovered by the tabular representation after distillation, obtained from the residual of Hiera after a date-by-date projection on the distilled student; the distilled student | Hiera row is the complementary opposite-direction residual, capturing tabular score content not present in the image score.

The distillation alignment is stable across the out-of-sample window. Splitting the sample at January 2020, the out-of-sample R^2 between the distilled student and Hiera is 0.78 in both segments to two decimals. The residual H–L magnitude is 0.0511 in 2006–2019 and 0.0443 in 2020–2025, compared with 0.0490 over the full window. It is significant at the 5% level over the full sample ($t = 2.47$) and in the early segment ($t = 2.03$); the shorter post-2020 segment has lower power (301 rebalance periods, $t = 1.46$) but no sign reversal. Thus, the residual image-score component is not localized to a single subperiod, and the distillation-recovered share is not mechanically driven by either window. Full subperiod results for all models are reported in Subsection H2.

4 Conclusion

We ask whether a chart representation of daily OHLCV data contains return-relevant information that explicit numeric summaries of the same data do not recover. The empirical design holds the information set fixed: the image model and the tabular benchmarks receive the same 20-day window for the largest 500 U.S. stocks. The image representation produces a stronger cross-sectional ordering and a larger H–L spread than the strongest tabular benchmark; cross-sectional regressions, portfolio spanning tests, and conditional sorts all reject subsumption. The perturbation analysis localizes the chart advantage to recent price action, temporal ordering, and intrabar extremal structure, with a more limited role for volume and little role for color encoding. Writing those components back into explicit tabular predictors closes much of the gap, but not all of it.

Our interpretive contribution is a three-way decomposition of the apparent image–tabular gap. Distillation of the tabular model on Hiera's continuous score raises the pooled OLS-calibrated out-of-sample R^2 between the two scores from 0.454 to 0.783, so roughly 60% of the apparent gap to Hiera is tied to target mismatch under the distillation accounting. The remaining image-specific component is measured by the residual H–L spread that survives distillation: 0.0490 annualized, localized by perturbation to features the renderer preserves natively but the tabular encoding partially elides. The image signal is therefore substantially recoverable in tabular form but not fully exhausted by it, and the residual that survives is statistically nonzero, economically modest, and concentrated in visual features the renderer makes explicit. Future work may extend the decomposition to smaller-cap segments, alternative forecast horizons, and non-equity asset classes, and apply the framework to canonical machine-learning benchmarks in finance where similar complex-versus-simpler tradeoffs arise.

References

Alizadeh, S., Brandt, M. W., and Diebold, F. X. (2002). Range-Based Estimation of Stochastic Volatility Models. *The Journal of Finance*, 57(3):1047–1091.

Andrews, D. W. K. (1993). Tests for Parameter Instability and Structural Change with Unknown Change Point. *Econometrica*, 61(4):821–856.

Bajgrowicz, P. and Scaillet, O. (2012). Technical Trading Revisited: False Discoveries, Persistence Tests, and Transaction Costs. *Journal of Financial Economics*, 106(3):473–491.

Barber, B. M. and Odean, T. (2008). All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *The Review of Financial Studies*, 21(2):785–818.

Blume, L., Easley, D., and O'Hara, M. (1994). Market Statistics and Technical Analysis: The Role of Volume. *The Journal of Finance*, 49(1):153–181.

Bøjstrup, J. T., Veliyev, B., and Wulff, J. N. (2026). From Pixels to Signals: Hierarchical Vision Transformers for Return Prediction. SSRN working paper.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2013). Salience and Asset Prices. *American Economic Review*, 103(3):623–628.

Brandt, M. W., Santa-Clara, P., and Valkanov, R. (2009). Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns. *The Review of Financial Studies*, 22(9):3411–3447.

Brock, W., Lakonishok, J., and LeBaron, B. (1992). Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *The Journal of Finance*, 47(5):1731–1764.

Brown, D. P. and Jennings, R. H. (1989). On Technical Analysis. *The Review of Financial Studies*, 2(4):527–551.

Buda, M., Maki, A., and Mazurowski, M. A. (2018). A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. *Neural Networks*, 106:249–259.

Chen, J.-F., Chen, W.-L., Huang, C.-P., Huang, S.-H., and Chen, A.-P. (2016). Financial Time-Series Data Analysis Using Deep Convolutional Neural Networks. In *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, pages 87–92. IEEE.

Chen, L., Pelger, M., and Zhu, J. (2024). Deep Learning in Asset Pricing. *Management Science*, 70(2):714–750.

Chordia, T., Roll, R., and Subrahmanyam, A. (2008). Liquidity and Market Efficiency. *Journal of Financial Economics*, 87(2):249–268.

Corwin, S. A. and Schultz, P. (2012). A Simple Way to Estimate Bid-Ask Spreads from Daily High and Low Prices. *The Journal of Finance*, 67(2):719–760.

Cosemans, M. and Frehen, R. (2021). Salience Theory and Stock Prices: Empirical Evidence. *Journal of Financial Economics*, 140(2):460–483.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition.

Da, Z., Engelberg, J., and Gao, P. (2011). In Search of Attention. *The Journal of Finance*, 66(5):1461–1499.

Daniel, K. and Titman, S. (1997). Evidence on the Characteristics of Cross-Sectional Variation in Stock Returns. *The Journal of Finance*, 52(1):1–33.

Elton, E. J. (1999). Expected Return, Realized Return, and Asset Pricing Tests. *Journal of Finance*, 54(4):1199–1220.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, 25(2):383–417.

Fama, E. F. (1991). Efficient Capital Markets: II. *The Journal of Finance*, 46(5):1575–1617.

Fama, E. F. and French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*, 47(2):427–465.

Fama, E. F. and French, K. R. (2008). Dissecting Anomalies. *The Journal of Finance*, 63(4):1653–1678.

Fama, E. F. and MacBeth, J. D. (1973). Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy*, 81(3):607–636.

Fong, R. C. and Vedaldi, A. (2017). Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457. IEEE.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting Characteristics Nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232.

George, T. J. and Hwang, C.-Y. (2004). The 52-Week High and Momentum Investing. *The Journal of Finance*, 59(5):2145–2176.

Giglio, S., Kelly, B., and Xiu, D. (2022). Factor Models, Machine Learning, and Asset Pricing. *Annual Review of Financial Economics*, 14:337–368.

Giglio, S., Liao, Y., and Xiu, D. (2021). Thousands of Alpha Tests. *The Review of Financial Studies*, 34(7):3456–3496.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems*, volume 34, pages 18932–18943.

Green, J., Hand, J. R. M., and Zhang, X. F. (2017). The Characteristics That Provide Independent Information about Average U.S. Monthly Stock Returns. *The Review of Financial Studies*, 30(12):4389–4436.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273.

Han, Y., Yang, K., and Zhou, G. (2013). A New Anomaly: The Cross-Sectional Profitability of Technical Analysis. *Journal of Financial and Quantitative Analysis*, 48(5):1433–1461.

Han, Y., Zhou, G., and Zhu, Y. (2016). A Trend Factor: Any Economic Gains from Using Information over Investment Horizons? *Journal of Financial Economics*, 122(2):352–375.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the Cross-Section of Expected Returns. *The Review of Financial Studies*, 29(1):5–68.

He, H. and Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. (2019). A Benchmark for Interpretability Methods in Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32.

Hoseinzade, E. and Haratizadeh, S. (2019). CNNpred: CNN-Based Stock Market Prediction Using a Diverse Set of Variables. *Expert Systems with Applications*, 129:273–285.

Hou, K., Xue, C., and Zhang, L. (2020). Replicating Anomalies. *The Review of Financial Studies*, 33(5):2019–2133.

Jegadeesh, N. (1990). Evidence of Predictable Behavior of Security Returns. *The Journal of Finance*, 45(3):881–898.

Jegadeesh, N. and Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance*, 48(1):65–91.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is There a Replication Crisis in Finance? *The Journal of Finance*, 78(5):2465–2518.

Jiang, J., Kelly, B. T., and Xiu, D. (2023). (Re-)Imag(in)ing Price Trends. *The Journal of Finance*, 78(6):3193–3249.

Kaczmarek, T. and Pukthuanthong, K. (2024). Just Look: Knowing Peers with Image Representation. SSRN working paper.

Kavajecz, K. A. and Odders-White, E. R. (2004). Technical Analysis and Liquidity Provision. *The Review of Financial Studies*, 17(4):1043–1071.

Kelly, B., Malamud, S., and Zhou, K. (2024). The Virtue of Complexity in Return Prediction. *The Journal of Finance*, 79(1):459–503.

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics Are Covariances: A Unified Model of Risk and Return. *Journal of Financial Economics*, 134(3):501–524.

Kim, T. and Kim, H. Y. (2019). Forecasting Stock Prices with a Feature Fusion LSTM-CNN Model Using Different Representations of the Same Data. *PLOS ONE*, 14(2):e0212320.

Lee, J., Kim, R., Koh, Y., and Kang, J. (2019). Global Stock Market Prediction Based on Stock Chart Images Using Deep Q-Network. *IEEE Access*, 7:167260–167277.

Lehmann, B. N. (1990). Fads, Martingales, and Market Efficiency. *The Quarterly Journal of Economics*, 105(1):1–28.

Lewellen, J. (2015). The Cross-Section of Expected Stock Returns. *Critical Finance Review*, 4(1):1–44.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Lo, A. W. and MacKinlay, A. C. (1990). When Are Contrarian Profits Due to Stock Market Overreaction? *The Review of Financial Studies*, 3(2):175–205.

Lo, A. W., Mamaysky, H., and Wang, J. (2000). Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation. *The Journal of Finance*, 55(4):1705–1765.

McLean, R. D. and Pontiff, J. (2016). Does Academic Research Destroy Stock Return Predictability? *The Journal of Finance*, 71(1):5–32.

Moskowitz, T. J., Ooi, Y. H., and Pedersen, L. H. (2012). Time Series Momentum. *Journal of Financial Economics*, 104(2):228–250.

Murray, S., Xia, Y., and Xiao, H. (2024). Charting by Machines. *Journal of Financial Economics*, 153:103791.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703–708.

Obaid, K. and Pukthuanthong, K. (2022). A Picture Is Worth a Thousand Words: Measuring Investor Sentiment by Combining Machine Learning and Photos from News. *Journal of Financial Economics*, 144(1):273–297.

Parkinson, M. (1980). The Extreme Value Method for Estimating the Variance of the Rate of Return. *The Journal of Business*, 53(1):61–65.

Ryali, C., Hu, Y.-T., Bolya, D., Wei, C., Fan, H., Huang, P.-Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., Malik, J., Li, Y., and Feichtenhofer, C. (2023). Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29441–29454. PMLR.

Schwert, G. W. (2003). Anomalies and Market Efficiency. In Constantinides, G. M., Harris, M., and Stulz, R. M., editors, *Handbook of the Economics of Finance*, volume 1, chapter 15, pages 939–974. Elsevier.

Sullivan, R., Timmermann, A., and White, H. (1999). Data-Snooping, Technical Trading Rule Performance, and the Bootstrap. *The Journal of Finance*, 54(5):1647–1691.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer, Cham.

Zhu, Y. and Zhou, G. (2009). Technical Analysis: An Asset Allocation Perspective on the Use of Moving Averages. *Journal of Financial Economics*, 92(3):519–544.

Appendix

A Implementation and Construction Details

This section documents the fixed implementation choices and deterministic construction rules that make the common-information design operational. Its role is not to introduce new economic content, but to record how the image representation, the tabular representation, and the translated feature blocks are constructed from the same 20-trading-day OHLCV window, and how the corresponding estimators are held fixed across comparisons. Read together, the subsections below provide the reproducibility details that underlie the main-text comparison between Hiera and the tabular benchmarks.

A1 Chart Construction

This subsection reports the deterministic renderer used to map each 20-trading-day OHLCV window into the image representation. Because the central comparison is conducted under common information, the chart should be interpreted as a fixed visual encoding of the same source window that also underlies the tabular representation, not as a source of additional raw information. Table A1.1 therefore records the chart geometry, scaling rules, and visual conventions applied uniformly across observations.

A2 Estimation Settings

This subsection reports the fixed estimation settings for the tabular benchmark families and for Hiera. The purpose is to separate changes in predictor content from changes in model tuning: across the nested translated specifications, estimator settings are held fixed, so stage-to-stage differences reflect the information added by the translated feature blocks rather than re-optimization of the learning algorithm. For the tabular benchmark families, the same fitting rules are used throughout the ladder; for Hiera, the reported configuration is the fixed training specification used in the main results.

Table A1.1: Fixed Renderer Specification for the Image Representation

Element	Value
Image size	224×224 pixels
Price-panel height	156 pixels
Volume-panel height	44 pixels
Gap between panels	4 pixels
Margins	Left/right: 14 pixels (bars centered); top/bottom: 10 pixels
Bar width	6 pixels
Gap between bars	4 pixels
5-day moving average	5 trading days
20-day moving average	20 trading days
Price scaling	Within-window range with 3% padding
Price range basis	Includes moving-average lines
Volume transform	$\log(1 + \text{volume})$
Volume clipping	Within-window 5th to 95th percentiles
Color convention	Separate colors for up days, down days, wicks, and moving averages

Note: This table reports the fixed rendering parameters used to map each 20-trading-day OHLCV window into the chart image used as the image representation. The renderer is deterministic and identical across observations. The output image contains an upper price panel with candlesticks and 5-day and 20-day moving-average overlays and a lower volume panel. The two overlays are trailing simple moving averages of the daily close computed from each stock's full available price history. For windows away from the start of a stock's sample the earliest displayed bars take their average from closes immediately preceding the window, the 5-day line reaching back to day $t-23$ and the 20-day line to day $t-38$. All such closes are dated at or before t, and the same moving-average series enter the tabular representation, so the chart introduces no information beyond, and none later than, the data shared with the tabular encoder. Near the start of a stock's sample the average is defined only once enough history has accrued, so the earliest windows in a stock's history carry a partial or absent moving-average line; this affects a negligible share of windows. Price values are scaled within window using the stated range basis and padding rule. Volume is transformed as $\log(1 + \text{volume})$ and clipped using the stated within-window percentile rule.

The image-based forecasts are produced by a Hiera Tiny hierarchical vision transformer. The purpose of this appendix entry is only to record the compact model mapping used in the empirical design. Optimization settings, trainable layers, and model-selection rules are reported separately in Table A2.2.

Let $\mathcal{P}(\cdot)$ denote the patch-embedding map applied to the rendered chart $X_{i,t}$. The initial token representation is

$$h_{i,t}^{(0)} = \mathcal{P}(X_{i,t})\,. \tag{A2.1}$$

Hiera then applies a sequence of multistage transformer blocks,

$$h_{i,t}^{(\ell+1)} = \mathcal{T}_\ell\Big(h_{i,t}^{(\ell)}\Big)\,, \qquad \ell = 0, \ldots, L-1, \tag{A2.2}$$

where stage transitions reduce spatial resolution and increase channel capacity, yielding a multiscale representation of the chart.

Let

$$\bar{h}_{i,t} = \mathrm{Pool}\Big(h_{i,t}^{(L)}\Big) \tag{A2.3}$$

Table A2.1: Estimator Configuration for the Tabular Benchmark Families

Component	Specification
	Panel A: Logit benchmark family
Estimator	`SGDClassifier` with logistic loss
Loss	`loss="log_loss"`
Penalty	ℓ_2 penalty, `alpha`$= 10^{-4}$
Intercept	`fit_intercept=True`
Learning rate	Constant, `learning_rate="constant"`, `eta0`$= 0.01$
Averaging	`average=True`
Epochs / passes	3 streaming passes over the selected training rows
Standardization	Enabled
Feature clipping	Values clipped to $[-8, 8]$
Missing / infinite values	Replaced by 0 before fitting
Random seed	`random_state`$= 47$
	Panel B: HistGB benchmark family
Estimator	`HistGradientBoostingClassifier`
Loss	`loss="log_loss"`
Learning rate	`learning_rate`$= 0.05$
Maximum iterations	`max_iter`$= 300$
Tree size	`max_leaf_nodes`$= 31$
Minimum leaf size	`min_samples_leaf`$= 100$
ℓ_2 regularization	`l2_regularization`$= 0.0$
Early stopping	Enabled, with `validation_fraction`$= 0.1$ and `n_iter_no_change`$= 20$
Standardization	Not used
Feature clipping	Not used
Missing / infinite values	Replaced by 0 before fitting
Random seed	`random_state`$= 42$

Note: This table reports the fixed estimator settings for the two tabular benchmark families. These settings are held constant across all baseline and translated tabular specifications, so differences across stages reflect changes in the predictor set rather than changes in estimator tuning. For the Logit benchmark family, a single training-set standardizer is computed once on the union of candidate features and then reused for all stage-specific fits; the relevant stage-specific feature subset is extracted after standardization and clipped to $[-8, 8]$. For the HistGB benchmark family, the engineered features are used directly without standardization or clipping. In both families, missing or infinite values are replaced by 0 before fitting, and reported scores are positive-class probabilities.

denote the pooled terminal representation. The classification head maps this representation into a scalar logit,

$$f_\theta(X_{i,t}) = w' \bar{h}_{i,t} + b, \tag{A2.4}$$

and the corresponding forecast score is

$$s_{i,t}^H = \sigma(f_\theta(X_{i,t})) = \frac{1}{1 + \exp(-f_\theta(X_{i,t}))}. \tag{A2.5}$$

The score $s_{i,t}^H$ is interpreted as the estimated probability that the five-day-ahead cumulative return is positive.

For a training sample of size N, estimation minimizes the binary cross-entropy objective

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{(i,t)} \left[y_{i,t}^{(5)} \log s_{i,t}^{H} + \left(1 - y_{i,t}^{(5)}\right) \log\left(1 - s_{i,t}^{H}\right) \right]. \tag{A2.6}$$

Table A2.2: Training Configuration for Hiera

Component	Specification
Architecture	Hiera Tiny with a single-logit binary classification head and dropout rate 0.25 before the classifier
Input	224×224 RGB chart images constructed from 20-trading-day OHLCV windows
Prediction target	$y_{i,t}^{(5)} = \mathbf{1}\{R_{i,t,t+5} > 0\}$, the sign of the 5-day cumulative close-to-close return
Training period	January 1, 1993 to December 31, 2003
Validation period	January 1, 2004 to December 31, 2005
Training sample rule	Class-balanced subset with 300,000 observations from each class
Validation sample rule	Full validation sample
Initialization	Transfer from a prior same-task Hiera run
Fine-tuning regime	Classification head plus the final three transformer blocks are trainable; earlier layers are frozen
Loss	Binary cross-entropy
Optimizer	AdamW
Base learning rate	10^{-4}
Head learning-rate multiplier	10
Layer-wise decay	0.75
Weight decay	5×10^{-3}
Gradient clipping	Global norm clipped at 1.0
Schedule	1 epoch of warmup, then cosine decay to a minimum learning-rate ratio of 0.05
Maximum epochs	12
Early stopping	Patience 2, activated after at least 3 epochs
Batch size	1,024
Mixed precision	bfloat16 when supported; otherwise float16 automatic mixed precision
Teacher–student consistency	Supported in code but disabled in the reported specification
Model selection criterion	Validation log loss

Note: This table reports the training configuration for the Hiera specification used in the main results. Hiera is trained on the image representation using the common target $y_{i,t}^{(5)}$, with model selection based on validation log loss over the 2004–2005 validation period. The reported specification uses transfer initialization from a prior same-task Hiera run and fine-tunes the classification head together with the final three transformer blocks while keeping earlier layers frozen.

A3 Construction of Translated Feature Blocks

This subsection defines the translated feature blocks that augment the baseline tabular representation in the translated benchmark ladder introduced in the methodology section. Consistent with the common-information design, every added feature is a deterministic function of the same dated input record used to construct both the image representation and the baseline tabular representation: the 20-trading-day OHLCV window ending at t, together with the earlier closing prices required to initialize the 5-day and 20-day moving-average series. The earliest displayed 20-day moving average uses closing prices

back to $t-38$, so the complete raw input required by the two representations is

$$\{O_{i,\tau}, H_{i,\tau}, L_{i,\tau}, C_{i,\tau}, V_{i,\tau}\}_{\tau=t-19}^{t} \cup \{C_{i,\tau}\}_{\tau=t-38}^{t-20}, \tag{A3.1}$$

all dated at or before t. The purpose of the notation below is therefore not to introduce a new data source, but to formalize how the chart components isolated in the perturbation analysis are written back into explicit tabular predictors. Tables A3.1–A3.5 report the feature definitions for the Geometry, Wick, Direction, Volume, and Path translated feature blocks. Within each block, features are concatenated in the row order shown in the corresponding table; within a lagged feature group, lags are ordered by $\ell = 0, \ldots, 4$, so the most recent observation appears first.

Renderer-aligned baseline coordinates

For a given stock-date observation (i, t), let

$$\tau_j = t - 19 + j, \qquad j = 0, \ldots, 19, \tag{A3.2}$$

so that $j = 0$ denotes the earliest displayed bar and $j = 19$ denotes the most recent displayed bar. For $k \in \{5, 20\}$, define the trailing simple moving average

$$M_{i,\tau}^{(k)} = \frac{1}{k} \sum_{q=0}^{k-1} C_{i,\tau-q}, \tag{A3.3}$$

and its availability indicator

$$\delta_{i,\tau}^{(k)} = \mathbf{1}\left\{C_{i,\tau}, C_{i,\tau-1}, \ldots, C_{i,\tau-k+1} \text{ are all observed}\right\}. \tag{A3.4}$$

When $\delta_{i,\tau}^{(k)} = 0$, the corresponding moving-average coordinate is treated as missing before the final zero-imputation step described below.

Define the within-window price-panel extrema by

$$P_{i,t}^{\min} = \min \left\{ \min_{0\le j\le 19} L_{i,\tau_j}, \min_{\substack{0\le j\le 19 \\ \delta_{i,\tau_j}^{(5)}=1}} M_{i,\tau_j}^{(5)}, \min_{\substack{0\le j\le 19 \\ \delta_{i,\tau_j}^{(20)}=1}} M_{i,\tau_j}^{(20)} \right\}, \tag{A3.5}$$

and

$$P_{i,t}^{\max} = \max \left\{ \max_{0\le j\le 19} H_{i,\tau_j}, \max_{\substack{0\le j\le 19 \\ \delta_{i,\tau_j}^{(5)}=1}} M_{i,\tau_j}^{(5)}, \max_{\substack{0\le j\le 19 \\ \delta_{i,\tau_j}^{(20)}=1}} M_{i,\tau_j}^{(20)} \right\}. \tag{A3.6}$$

If one of the moving-average sets is empty, its minimum or maximum is omitted from the corresponding outer operator. Open and close prices need not enter separately because valid OHLC observations satisfy $L_{i,\tau} \le O_{i,\tau}, C_{i,\tau} \le H_{i,\tau}$.

Let

$$R_{i,t}^{P} = \begin{cases} P_{i,t}^{\max} - P_{i,t}^{\min}, & P_{i,t}^{\max} - P_{i,t}^{\min} > 0 \text{ and finite}, \\ 0.01 \max\left\{ |P_{i,t}^{\max}|, 1 \right\}, & \text{otherwise}, \end{cases} \tag{A3.7}$$

and apply the renderer's 3% vertical padding:

$$P_{i,t}^{-} = P_{i,t}^{\min} - 0.03 R_{i,t}^{P}, \qquad P_{i,t}^{+} = P_{i,t}^{\max} + 0.03 R_{i,t}^{P}. \tag{A3.8}$$

For later use, define the clipping operator

$$\mathrm{clip}_{[a,b]}(x) = \min\{\max\{x, a\}, b\}. \tag{A3.9}$$

For $X \in \{O, H, L, C\}$, the renderer-aligned price coordinate is

$$\widetilde{X}_{i,\tau_j} = \mathrm{clip}_{[0,1]} \left(\frac{X_{i,\tau_j} - P_{i,t}^{-}}{P_{i,t}^{+} - P_{i,t}^{-}} \right). \tag{A3.10}$$

The moving-average coordinates use the same transformation:

$$\widetilde{M}_{i,\tau_j}^{(k)} = \begin{cases} \mathrm{clip}_{[0,1]}\left(\dfrac{M_{i,\tau_j}^{(k)} - P_{i,t}^{-}}{P_{i,t}^{+} - P_{i,t}^{-}}\right), & \delta_{i,\tau_j}^{(k)} = 1, \\ 0, & \delta_{i,\tau_j}^{(k)} = 0, \end{cases} \qquad k \in \{5, 20\}. \tag{A3.11}$$

For volume, define

$$x_{i,\tau_j}^{V} = \log\bigl(1 + V_{i,\tau_j}\bigr). \tag{A3.12}$$

Let

$$x_{i,t,[0]}^{V} \le x_{i,t,[1]}^{V} \le \cdots \le x_{i,t,[19]}^{V}$$

denote the ordered values of $\{x_{i,\tau_j}^{V}\}_{j=0}^{19}$. The continuous empirical quantile used in the implementation is defined, for $p \in [0,1]$, by

$$h_p = 19p, \qquad r_p = \lfloor h_p \rfloor, \qquad \lambda_p = h_p - r_p, \tag{A3.13}$$

$$Q_{i,t}^{V}(p) = \begin{cases} (1-\lambda_p)x_{i,t,[r_p]}^{V} + \lambda_p x_{i,t,[r_p+1]}^{V}, & r_p < 19, \\ x_{i,t,[19]}^{V}, & r_p = 19. \end{cases} \tag{A3.14}$$

Set

$$q_{i,t,-}^{V} = Q_{i,t}^{V}(0.05), \qquad q_{i,t,+}^{V} = Q_{i,t}^{V}(0.95), \qquad x_{i,t,\min}^{V} = \min_{0\le j\le 19} x_{i,\tau_j}^{V}, \qquad x_{i,t,\max}^{V} = \max_{0\le j\le 19} x_{i,\tau_j}^{V}. \tag{A3.15}$$

The effective lower and upper volume bounds are

$$\bigl(\underline{q}_{i,t}^{V}, \overline{q}_{i,t}^{V}\bigr) = \begin{cases} \bigl(q_{i,t,-}^{V}, q_{i,t,+}^{V}\bigr), & q_{i,t,+}^{V} > q_{i,t,-}^{V}, \\ \bigl(x_{i,t,\min}^{V}, x_{i,t,\max}^{V}\bigr), & q_{i,t,+}^{V} \le q_{i,t,-}^{V} \text{ and } x_{i,t,\max}^{V} > x_{i,t,\min}^{V}, \\ (0,1), & \text{otherwise.} \end{cases} \tag{A3.16}$$

The mapped volume coordinate is

$$\widetilde{V}_{i,\tau_j} = \begin{cases} \mathrm{clip}_{[0,1]}\left(\dfrac{x^V_{i,\tau_j} - \underline{q}^V_{i,t}}{\overline{q}^V_{i,t} - \underline{q}^V_{i,t} + 10^{-12}}\right), & \overline{q}^V_{i,t} > \underline{q}^V_{i,t}, \\ 0.5, & \overline{q}^V_{i,t} \le \underline{q}^V_{i,t}. \end{cases} \tag{A3.17}$$

The baseline representation also records the binary orientations used by the renderer. The candle orientation is

$$u^P_{i,\tau_j} = \mathbf{1}\left\{C_{i,\tau_j} \ge O_{i,\tau_j}\right\}. \tag{A3.18}$$

The volume-bar orientation follows the candle orientation except for a zero-body bar, for which the implementation uses the previous close as a tie-break:

$$u^V_{i,\tau_j} = \begin{cases} 1, & C_{i,\tau_j} > O_{i,\tau_j}, \\ 0, & C_{i,\tau_j} < O_{i,\tau_j}, \\ 1, & C_{i,\tau_j} = O_{i,\tau_j} \text{ and } j = 0, \\ \mathbf{1}\left\{C_{i,\tau_j} \ge C_{i,\tau_{j-1}}\right\}, & C_{i,\tau_j} = O_{i,\tau_j} \text{ and } j \ge 1. \end{cases} \tag{A3.19}$$

To keep the remaining notation compact, define

$$o_j = \widetilde{O}_{i,\tau_j}, \quad h_j = \widetilde{H}_{i,\tau_j}, \quad l_j = \widetilde{L}_{i,\tau_j}, \quad c_j = \widetilde{C}_{i,\tau_j}, \quad v_j = \widetilde{V}_{i,\tau_j}, \tag{A3.20}$$

together with

$$m^{(5)}_j = \widetilde{M}^{(5)}_{i,\tau_j}, \qquad m^{(20)}_j = \widetilde{M}^{(20)}_{i,\tau_j}, \qquad \delta^{(5)}_j = \delta^{(5)}_{i,\tau_j}, \qquad \delta^{(20)}_j = \delta^{(20)}_{i,\tau_j}, \tag{A3.21}$$

and

$$u^P_j = u^P_{i,\tau_j}, \qquad u^V_j = u^V_{i,\tau_j}. \tag{A3.22}$$

For each displayed bar, define the 11-dimensional baseline subvector

$$z_{i,t,j}^{(0)} = \left(o_j, h_j, l_j, c_j, m_j^{(5)}, m_j^{(20)}, \delta_j^{(5)}, \delta_j^{(20)}, v_j, u_j^P, u_j^V\right). \tag{A3.23}$$

The baseline tabular stage concatenates these subvectors in chronological order:

$$Z_{i,t}^{(0)} = Z_{i,t} = \left(z_{i,t,0}^{(0)}, z_{i,t,1}^{(0)}, \ldots, z_{i,t,19}^{(0)}\right), \qquad \dim Z_{i,t}^{(0)} = 20 \times 11 = 220. \tag{A3.24}$$

Before model estimation, any remaining missing, positive-infinite, or negative-infinite coordinate is replaced by zero. The availability indicators preserve whether a zero-valued moving-average coordinate arose from this imputation.

The cumulative translated ladder is

$$Z_{i,t}^{(1)} = \left(Z_{i,t}^{(0)}, G_{i,t}\right), \qquad Z_{i,t}^{(2)} = \left(Z_{i,t}^{(1)}, \mathcal{W}_{i,t}\right), \qquad Z_{i,t}^{(3)} = \left(Z_{i,t}^{(2)}, D_{i,t}\right), \tag{A3.25}$$

$$Z_{i,t}^{(4)} = \left(Z_{i,t}^{(3)}, \mathcal{V}_{i,t}\right), \qquad Z_{i,t}^{(5)} = \left(Z_{i,t}^{(4)}, P_{i,t}\right). \tag{A3.26}$$

Here $G_{i,t}$ denotes the Geometry block, $\mathcal{W}_{i,t}$ the Wick block, $D_{i,t}$ the Direction block, $\mathcal{V}_{i,t}$ the Volume block, and $P_{i,t}$ the Path block. Raw daily volume retains the notation $V_{i,\tau}$, whereas the translated Volume block is denoted by $\mathcal{V}_{i,t}$. The block sizes reported below imply

$$\dim Z_{i,t}^{(1)} = 280, \qquad \dim Z_{i,t}^{(2)} = 320, \qquad \dim Z_{i,t}^{(3)} = 360, \qquad \dim Z_{i,t}^{(4)} = 399, \qquad \dim Z_{i,t}^{(5)} = 421. \tag{A3.27}$$

Bar-level primitives and operators

For the translated feature construction, set

$$\varepsilon = 10^{-12}. \tag{A3.28}$$

For each $j = 0, \ldots, 19$, define the safeguarded full range, signed body, absolute body, upper wick, and lower wick by

$$r_j = \max\{h_j - l_j, \varepsilon\}, \qquad \beta_j = c_j - o_j, \qquad \gamma_j = |\beta_j|, \tag{A3.29}$$

$$w_j^+ = \max\{h_j - \max(o_j, c_j), 0\}\,, \qquad w_j^- = \max\{\min(o_j, c_j) - l_j, 0\}\,. \tag{A3.30}$$

From these, define the normalized bar-shape objects

$$\mathrm{bf}_j = \frac{\beta_j}{r_j}, \qquad \mathrm{abf}_j = \frac{\gamma_j}{r_j}, \qquad \mathrm{cl}_j = \frac{c_j - l_j}{r_j}, \qquad \mathrm{ol}_j = \frac{o_j - l_j}{r_j}, \tag{A3.31}$$

$$\mathrm{uf}_j = \frac{w_j^+}{r_j}, \qquad \mathrm{lf}_j = \frac{w_j^-}{r_j}, \qquad \mathrm{wf}_j = \frac{w_j^+ + w_j^-}{r_j}, \qquad \mathrm{wa}_j = \frac{w_j^+ - w_j^-}{r_j}, \tag{A3.32}$$

and the clipped wick-to-body ratios

$$\mathrm{utb}_j = \min\left\{\frac{w_j^+}{\max\{\gamma_j, \varepsilon\}}, 5\right\}, \qquad \mathrm{ltb}_j = \min\left\{\frac{w_j^-}{\max\{\gamma_j, \varepsilon\}}, 5\right\}. \tag{A3.33}$$

At the full-window level, define

$$H^{20} = \max_{0 \le j \le 19} h_j, \qquad L^{20} = \min_{0 \le j \le 19} l_j, \qquad R^{20} = \max\{H^{20} - L^{20}, \varepsilon\}, \tag{A3.34}$$

and the bar range relative to the full-window span,

$$\mathrm{rr}_j = \frac{h_j - l_j}{R^{20}}. \tag{A3.35}$$

The auxiliary series used below are

$$g_j = \begin{cases} 0, & j = 0, \\ o_j - c_{j-1}, & j \ge 1, \end{cases} \qquad \Delta c_j = \begin{cases} 0, & j = 0, \\ c_j - c_{j-1}, & j \ge 1, \end{cases} \tag{A3.36}$$

$$\Delta^2 c_j = \begin{cases} 0, & j \le 1, \\ c_j - 2c_{j-1} + c_{j-2}, & j \ge 2, \end{cases} \qquad \mathrm{tp}_j = \frac{h_j + l_j + c_j}{3}, \tag{A3.37}$$

and

$$\Delta^2 \mathrm{tp}_j = \begin{cases} 0, & j \le 1, \\ \mathrm{tp}_j - 2\mathrm{tp}_{j-1} + \mathrm{tp}_{j-2}, & j \ge 2. \end{cases} \tag{A3.38}$$

Define the sign operator by

$$\mathrm{sgn}(x) = \begin{cases} -1, & x < 0, \\ 0, & x = 0, \\ 1, & x > 0, \end{cases} \qquad s_j = \mathrm{sgn}(\Delta c_j). \tag{A3.39}$$

For $k \in \{2, \dots, 20\}$, define the last-k index set

$$\mathcal{I}_k = \{20 - k, \dots, 19\}, \tag{A3.40}$$

and define the previous-five set excluding the current bar by

$$\mathcal{I}_5^- = \{14, \dots, 18\}. \tag{A3.41}$$

The corresponding previous-five high, low, and average range, together with the previous-four average mapped volume, are

$$H_5^- = \max_{j \in \mathcal{I}_5^-} h_j, \qquad L_5^- = \min_{j \in \mathcal{I}_5^-} l_j, \tag{A3.42}$$

$$\bar{r}_5^- = \frac{1}{5} \sum_{j=14}^{18} r_j, \qquad \bar{v}_4^- = \frac{1}{4} \sum_{j=15}^{18} v_j. \tag{A3.43}$$

For any scalar sequence $x = (x_0, \ldots, x_{19})$, define

$$\bar{x}_k = \frac{1}{k} \sum_{j \in \mathcal{I}_k} x_j. \tag{A3.44}$$

The slope operator is the OLS slope of x_j on a centered linear time trend over the last k bars:

$$\mathrm{slope}_k(x) = \frac{\sum_{m=0}^{k-1} (m - \bar{m}_k) x_{20-k+m}}{\sum_{m=0}^{k-1} (m - \bar{m}_k)^2}, \qquad \bar{m}_k = \frac{k-1}{2}. \tag{A3.45}$$

The path-length and efficiency operators are

$$\mathrm{pathlen}_k(x) = \sum_{m=20-k+1}^{19} |x_m - x_{m-1}|, \tag{A3.46}$$

$$\mathrm{eff}_k(x) = \frac{|x_{19} - x_{20-k}|}{\max\{\mathrm{pathlen}_k(x), \varepsilon\}}. \tag{A3.47}$$

The sign-switch count is

$$\mathrm{switch}_k(x) = \sum_{m=20-k+1}^{19} \mathbf{1}\left\{\mathrm{sgn}(x_m)\,\mathrm{sgn}(x_{m-1}) < 0\right\}. \tag{A3.48}$$

A zero-valued observation therefore does not itself create a sign switch.

The endpoint rank is

$$\mathrm{ranklast}_k(x) = \frac{\#\{m \in \mathcal{I}_k : x_m \le x_{19}\} - 1}{k-1}. \tag{A3.49}$$

Ties with the endpoint are counted as weakly below it. The mean absolute deviation from the straight line joining the first and last observations is

$$\mathrm{areadev}_k(x) = \frac{1}{k} \sum_{m=0}^{k-1} \left| x_{20-k+m} - \left[x_{20-k} + \frac{m}{k-1} \left(x_{19} - x_{20-k} \right) \right] \right|, \tag{A3.50}$$

and the turning-count operator is

$$\mathrm{turn}_k(x) = \sum_{m=20-k+2}^{19} \mathbf{1}\left\{\mathrm{sgn}(x_m - x_{m-1})\,\mathrm{sgn}(x_{m-1} - x_{m-2}) < 0\right\}. \tag{A3.51}$$

For positional features, let

$$m_k^{\max}(x) = \min \operatorname*{arg\,max}_{m\in\{0,\dots,k-1\}} x_{20-k+m}, \qquad m_k^{\min}(x) = \min \operatorname*{arg\,min}_{m\in\{0,\dots,k-1\}} x_{20-k+m}, \tag{A3.52}$$

so ties are resolved by the first occurrence. Define

$$\operatorname{posmax}_k(x) = \frac{m_k^{\max}(x)}{k-1}, \qquad \operatorname{posmin}_k(x) = \frac{m_k^{\min}(x)}{k-1}. \tag{A3.53}$$

Both operators lie in $[0,1]$, with zero denoting the first observation and one the final observation in the relevant subwindow.

For the Volume block, define the Pearson correlation over the last k bars by

$$\operatorname{corr}_k(x,y) = \frac{\sum_{j\in\mathcal{I}_k}(x_j-\bar{x}_k)(y_j-\bar{y}_k)}{\sqrt{\sum_{j\in\mathcal{I}_k}(x_j-\bar{x}_k)^2}\sqrt{\sum_{j\in\mathcal{I}_k}(y_j-\bar{y}_k)^2}}, \tag{A3.54}$$

when both denominator terms are strictly positive, and set $\operatorname{corr}_k(x,y)=0$ otherwise.

Finally, define the terminal positive and negative run lengths over at most five close changes by

$$\operatorname{run}_5^{+}(\Delta c) = \max\left\{m\in\{0,\dots,5\} : \Delta c_{19-q} > 0 \text{ for every } q=0,\dots,m-1\right\}, \tag{A3.55}$$

$$\operatorname{run}_5^{-}(\Delta c) = \max\left\{m\in\{0,\dots,5\} : \Delta c_{19-q} < 0 \text{ for every } q=0,\dots,m-1\right\}. \tag{A3.56}$$

For $m=0$, the condition over the empty index set is understood to hold, so both operators are always defined.

Geometry block $G_{i,t}$

Table A3.1: Geometry Features in the Translated Feature Block $G_{i,t}$

Feature group	Count	Definition
Lagged body fraction	5	$\{\mathrm{bf}_{19-\ell}\}_{\ell=0}^{4}$, where $\mathrm{bf}_j = \beta_j / r_j$, that is, the signed candle body relative to the full bar range for the current bar and the four immediately preceding bars.
Lagged absolute body fraction	5	$\{\mathrm{abf}_{19-\ell}\}_{\ell=0}^{4}$, where $\mathrm{abf}_j = \gamma_j / r_j$, that is, the absolute candle body relative to the full bar range over the same five bars.
Lagged close location	5	$\{\mathrm{cl}_{19-\ell}\}_{\ell=0}^{4}$, where $\mathrm{cl}_j = (c_j - l_j)/r_j$, that is, the close location within the daily range for the most recent five bars.
Lagged open location	5	$\{\mathrm{ol}_{19-\ell}\}_{\ell=0}^{4}$, where $\mathrm{ol}_j = (o_j - l_j)/r_j$, that is, the open location within the daily range for the most recent five bars.
Lagged relative range	5	$\{\mathrm{rr}_{19-\ell}\}_{\ell=0}^{4}$, where $\mathrm{rr}_j = (h_j - l_j)/R^{20}$, that is, the bar range relative to the full-window range for the most recent five bars.
Lagged gap	5	$\{g_{19-\ell}\}_{\ell=0}^{4}$, where $g_0 = 0$ and $g_j = o_j - c_{j-1}$ for $j \geq 1$, that is, the gap between the current open and the previous close.
Lagged close change	5	$\{\Delta c_{19-\ell}\}_{\ell=0}^{4}$, that is, close-to-close price changes over the most recent five bars.
Lagged distance to window high	5	$\{H^{20} - c_{19-\ell}\}_{\ell=0}^{4}$, that is, the distance from each of the last five closes to the highest high in the full 20-bar window.
Lagged distance to window low	5	$\{c_{19-\ell} - L^{20}\}_{\ell=0}^{4}$, that is, the distance from each of the last five closes to the lowest low in the full 20-bar window.
Up-count	1	$\sum_{j \in \mathcal{I}_5} u_j^P$, that is, the number of up bars in $\mathcal{I}_5$.
Body-sign switches	1	$\mathrm{switch}_5(\beta)$, that is, the number of sign changes in the signed-body sequence over the last five bars.
Mean body fraction	1	$\bar{\mathrm{bf}}_5 = \frac{1}{5}\sum_{j \in \mathcal{I}_5} \mathrm{bf}_j$, that is, the mean signed body fraction over the last five bars.
Mean absolute body fraction	1	$\bar{\mathrm{abf}}_5 = \frac{1}{5}\sum_{j \in \mathcal{I}_5} \mathrm{abf}_j$, that is, the mean absolute body fraction over the last five bars.
Mean close location	1	$\bar{\mathrm{cl}}_5 = \frac{1}{5}\sum_{j \in \mathcal{I}_5} \mathrm{cl}_j$, that is, the mean close location within the daily range over the last five bars.
Mean relative range	1	$\bar{\mathrm{rr}}_5 = \frac{1}{5}\sum_{j \in \mathcal{I}_5} \mathrm{rr}_j$, that is, the mean relative bar range over the last five bars.
Close slope	1	$\mathrm{slope}_5(c)$, that is, the OLS slope of the close series on a centered linear time trend over $\mathcal{I}_5$.
Range slope	1	$\mathrm{slope}_5(r)$, that is, the OLS slope of the full bar range over $\mathcal{I}_5$.
Net move	1	$c_{19} - c_{15}$, that is, the end-to-start change in the close over the last five bars.
Path length	1	$\mathrm{pathlen}_5(c)$, that is, the total variation of the close path over $\mathcal{I}_5$.
Efficiency	1	$\mathrm{eff}_5(c)$, that is, the ratio of net close displacement to total close-path length over $\mathcal{I}_5$.
Upside breakout	1	$c_{19} - H_5^-$, that is, the current close relative to the highest high over the previous five bars $\mathcal{I}_5^- = \{14, \ldots, 18\}$.
Downside breakout	1	$L_5^- - c_{19}$, that is, the previous-five low relative to the current close.
Close position in previous-five range	1	$\dfrac{c_{19} - L_5^-}{\max(H_5^- - L_5^-, \varepsilon)}$, that is, the position of the current close within the high-low range of the previous five bars.
Range expansion	1	$r_{19} - \bar{r}_5^-$, where $\bar{r}_5^- = \frac{1}{5}\sum_{j=14}^{18} r_j$, that is, the current bar range relative to the average range over the previous five bars.
Total	60	Geometry translated feature block.

Note: This table defines the Geometry translated feature block $G_{i,t}$. All features are deterministic functions of the renderer-aligned baseline coordinates defined above, which are constructed from the common dated input record in Equation (A3.1). The block is added as the first translated feature block in the cumulative translation and combines lagged bar-shape variables for the current bar and the four immediately preceding bars with short-horizon summary measures over the last five bars. The resulting block contains 60 features.

Wick block $\mathcal{W}_{i,t}$

Table A3.2: Wick Features in the Translated Feature Block $\mathcal{W}_{i,t}$

Feature group	Count	Definition
Lagged upper wick fraction	5	$\{\mathrm{uf}_{19-\ell}\}_{\ell=0}^{4}$, where $\mathrm{uf}_j = w_j^+/r_j$, that is, the upper wick relative to the full bar range for the current bar and the four immediately preceding bars.
Lagged lower wick fraction	5	$\{\mathrm{lf}_{19-\ell}\}_{\ell=0}^{4}$, where $\mathrm{lf}_j = w_j^-/r_j$, that is, the lower wick relative to the full bar range over the same five bars.
Lagged total wick fraction	5	$\{\mathrm{wf}_{19-\ell}\}_{\ell=0}^{4}$, where $\mathrm{wf}_j = (w_j^+ + w_j^-)/r_j$, that is, the total wick share over the most recent five bars.
Lagged wick asymmetry	5	$\{\mathrm{wa}_{19-\ell}\}_{\ell=0}^{4}$, where $\mathrm{wa}_j = (w_j^+ - w_j^-)/r_j$, that is, wick asymmetry, with positive values indicating relatively larger upper wicks.
Lagged upper-to-body ratio	5	$\{\mathrm{utb}_{19-\ell}\}_{\ell=0}^{4}$, where $\mathrm{utb}_j = \min\left(w_j^+ / \max(\gamma_j, \varepsilon), 5\right)$, that is, the clipped ratio of upper wick to absolute body.
Lagged lower-to-body ratio	5	$\{\mathrm{ltb}_{19-\ell}\}_{\ell=0}^{4}$, where $\mathrm{ltb}_j = \min\left(w_j^- / \max(\gamma_j, \varepsilon), 5\right)$, that is, the clipped ratio of lower wick to absolute body.
Mean upper wick fraction	1	$\bar{\mathrm{uf}}_5 = \frac{1}{5}\sum_{j\in\mathcal{I}_5} \mathrm{uf}_j$, that is, the mean upper-wick fraction over the last five bars.
Mean lower wick fraction	1	$\bar{\mathrm{lf}}_5 = \frac{1}{5}\sum_{j\in\mathcal{I}_5} \mathrm{lf}_j$, that is, the mean lower-wick fraction over the last five bars.
Mean total wick fraction	1	$\bar{\mathrm{wf}}_5 = \frac{1}{5}\sum_{j\in\mathcal{I}_5} \mathrm{wf}_j$, that is, the mean total-wick fraction over the last five bars.
Mean wick asymmetry	1	$\bar{\mathrm{wa}}_5 = \frac{1}{5}\sum_{j\in\mathcal{I}_5} \mathrm{wa}_j$, that is, the mean wick asymmetry over the last five bars.
Long upper-wick count	1	$\sum_{j\in\mathcal{I}_5} \mathbf{1}\{\mathrm{uf}_j > 0.50\}$, that is, the number of bars in the last five for which the upper wick exceeds half of the total daily range.
Long lower-wick count	1	$\sum_{j\in\mathcal{I}_5} \mathbf{1}\{\mathrm{lf}_j > 0.50\}$, that is, the number of bars in the last five for which the lower wick exceeds half of the total daily range.
Close-near-high count	1	$\sum_{j\in\mathcal{I}_5} \mathbf{1}\{\mathrm{cl}_j > 0.80\}$, that is, the number of bars in the last five for which the close lies in the upper fifth of the daily range.
Close-near-low count	1	$\sum_{j\in\mathcal{I}_5} \mathbf{1}\{\mathrm{cl}_j < 0.20\}$, that is, the number of bars in the last five for which the close lies in the lower fifth of the daily range.
Current upper wick fraction	1	uf_{19}, that is, the upper-wick share of the most recent bar.
Current lower wick fraction	1	lf_{19}, that is, the lower-wick share of the most recent bar.
Total	40	Wick translated feature block.

Note: This table defines the Wick translated feature block $\mathcal{W}_{i,t}$. All features are deterministic functions of the renderer-aligned baseline coordinates defined above, which are constructed from the common dated input record in Equation (A3.1). The block is added as the second translated feature block in the cumulative translation and captures wick length, wick asymmetry, wick-to-body ratios, and close location within the daily range, with an emphasis on the current bar and the four immediately preceding bars. The resulting block contains 40 features.

Direction block $D_{i,t}$

Table A3.3: Direction Features in the Translated Feature Block $D_{i,t}$

Feature group	Count	Definition
Lagged body sign	5	$\{\mathrm{sgn}(\beta_{19-\ell})\}_{\ell=0}^{4}$, that is, the sign of the candle body for the current bar and the four immediately preceding bars.
Lagged close-change sign	5	$\{\mathrm{sgn}(\Delta c_{19-\ell})\}_{\ell=0}^{4}$, that is, the sign of the close-to-close change for the same five bars.
Lagged second difference	4	$\{\Delta^2 c_{19-\ell}\}_{\ell=0}^{3}$, that is, discrete second differences of the close series for the four most recent admissible lags.
Adjacent body-fraction product	4	$\{\mathrm{bf}_{19-\ell}\mathrm{bf}_{18-\ell}\}_{\ell=0}^{3}$, that is, products of adjacent signed body fractions that preserve local directional sequencing.
Adjacent close-change product	4	$\{\Delta c_{19-\ell}\Delta c_{18-\ell}\}_{\ell=0}^{3}$, that is, products of adjacent close-to-close changes.
Reversal count (5)	1	$\mathrm{switch}_5(\Delta c)$, that is, the number of sign changes in close-to-close moves over $\mathcal{I}_5$.
Reversal count (10)	1	$\mathrm{switch}_{10}(\Delta c)$, that is, the number of sign changes in close-to-close moves over $\mathcal{I}_{10}$.
Position of max high	1	$\mathrm{posmax}_5(h)$, that is, the location of the highest high within the last five bars, rescaled to $[0,1]$.
Position of min low	1	$\mathrm{posmin}_5(l)$, that is, the location of the lowest low within the last five bars, rescaled to $[0,1]$.
Position of max body	1	$\mathrm{posmax}_5(\gamma)$, that is, the location of the largest absolute candle body within the last five bars, rescaled to $[0,1]$.
Last bar is max high	1	$\mathbf{1}\{\mathrm{posmax}_5(h)=1\}$, that is, an indicator that the final bar attains the highest high within the last five bars.
Last bar is min low	1	$\mathbf{1}\{\mathrm{posmin}_5(l)=1\}$, that is, an indicator that the final bar attains the lowest low within the last five bars.
Trend then reversal	1	$\mathbf{1}\{s_{16}=s_{17}=s_{18}\neq 0,\ s_{19}=-s_{18}\}$, where $s_j=\mathrm{sgn}(\Delta c_j)$, that is, the first three close changes in the last-four-change sequence share the same nonzero sign and the final one reverses.
Reversal then trend	1	$\mathbf{1}\{s_{17}=s_{18}=s_{19}\neq 0,\ s_{16}=-s_{17}\}$, where $s_j=\mathrm{sgn}(\Delta c_j)$, that is, the first change reverses and the final three changes share the same nonzero sign.
Positive run length	1	$\mathrm{run}_5^{+}(\Delta c)$, that is, the number of consecutive strictly positive close changes when scanning backward from the final bar.
Negative run length	1	$\mathrm{run}_5^{-}(\Delta c)$, that is, the number of consecutive strictly negative close changes when scanning backward from the final bar.
Signed move (2 bars)	1	$c_{19}-c_{18}$, that is, the final one-step close change.
Signed move (3 bars)	1	$c_{19}-c_{17}$, that is, the final two-step net close change.
Signed move (5 bars)	1	$c_{19}-c_{15}$, that is, the net close change across the last five bars.
Net body	1	$\sum_{j\in\mathcal{I}_5}\beta_j$, that is, the sum of signed candle bodies over $\mathcal{I}_5$.
High-low span	1	$\max_{j\in\mathcal{I}_5}h_j-\min_{j\in\mathcal{I}_5}l_j$, that is, the total price span across the last five bars.
Peak after trough	1	$\mathbf{1}\{\mathrm{posmax}_5(h)>\mathrm{posmin}_5(l)\}$, that is, an indicator that the local peak occurs after the local trough within the last five bars.
Endpoint close rank	1	$\mathrm{ranklast}_5(c)$, that is, the rank of the final close within the last five closes, scaled to $[0,1]$.
Total	40	Direction translated feature block.

Note: This table defines the Direction translated feature block $D_{i,t}$. All features are deterministic functions of the renderer-aligned baseline coordinates defined above, which are constructed from the common dated input record in Equation (A3.1). The block is added as the third translated feature block in the cumulative translation and is designed to preserve local sequencing, reversals, run length, endpoint position, and short-horizon directional structure visible in the ordered price path. The resulting block contains 40 features.

Volume block $\mathcal{V}_{i,t}$

Table A3.4: Volume Features in the Translated Feature Block $\mathcal{V}_{i,t}$

Feature group	Count	Definition
Lagged body × volume	5	$\{\mathrm{bf}_{19-\ell} v_{19-\ell}\}_{\ell=0}^{4}$, that is, the signed body fraction multiplied by mapped volume for the current bar and the four immediately preceding bars.
Lagged abs. body × volume	5	$\{\mathrm{abf}_{19-\ell} v_{19-\ell}\}_{\ell=0}^{4}$, that is, the absolute body fraction multiplied by mapped volume over the same five bars.
Lagged range × volume	5	$\{\mathrm{rr}_{19-\ell} v_{19-\ell}\}_{\ell=0}^{4}$, that is, the relative bar range multiplied by mapped volume over the most recent five bars.
Lagged upper wick × volume	5	$\{\mathrm{uf}_{19-\ell} v_{19-\ell}\}_{\ell=0}^{4}$, that is, the upper-wick fraction multiplied by mapped volume over the most recent five bars.
Lagged lower wick × volume	5	$\{\mathrm{lf}_{19-\ell} v_{19-\ell}\}_{\ell=0}^{4}$, that is, the lower-wick fraction multiplied by mapped volume over the most recent five bars.
Lagged gap × volume	5	$\{g_{19-\ell} v_{19-\ell}\}_{\ell=0}^{4}$, that is, the gap variable multiplied by mapped volume over the most recent five bars.
Body–volume correlation	1	$\mathrm{corr}_5(\mathrm{bf}, v)$, that is, the Pearson correlation over $\mathcal{I}_5$ between signed body fraction and mapped volume.
Abs. body–volume correlation	1	$\mathrm{corr}_5(\mathrm{abf}, v)$, that is, the Pearson correlation over $\mathcal{I}_5$ between absolute body fraction and mapped volume.
Range–volume correlation	1	$\mathrm{corr}_5(\mathrm{rr}, v)$, that is, the Pearson correlation over $\mathcal{I}_5$ between relative range and mapped volume.
Volume slope	1	$\mathrm{slope}_5(v)$, that is, the OLS slope of mapped volume over $\mathcal{I}_5$.
Volume surprise	1	$v_{19} - \bar{v}_4^-$, where $\bar{v}_4^- = \frac{1}{4}\sum_{j=15}^{18} v_j$, that is, current mapped volume relative to the mean of the previous four bars.
Current abs. body × volume	1	$\mathrm{abf}_{19} v_{19}$, that is, the current-bar absolute body fraction multiplied by current mapped volume.
Current upside breakout × volume	1	$(c_{19} - H_5^-) v_{19}$, that is, the current close relative to the previous-five high, multiplied by current mapped volume.
Current downside breakout × volume	1	$(L_5^- - c_{19}) v_{19}$, that is, the previous-five low relative to the current close, multiplied by current mapped volume.
Current close position × volume	1	$\dfrac{c_{19} - L_5^-}{\max(H_5^- - L_5^-, \varepsilon)} v_{19}$, that is, the position of the current close within the previous-five high–low range, interacted with current mapped volume.
Total	39	Volume translated feature block.

Note: This table defines the Volume translated feature block $\mathcal{V}_{i,t}$. All features are deterministic functions of the renderer-aligned baseline coordinates defined above, which are constructed from the common dated input record in Equation (A3.1). The block is added as the fourth translated feature block in the cumulative translation and combines recent price-shape objects with mapped volume through lagged interactions, short-horizon price–volume correlations, and current-bar breakout and position measures scaled by volume. The resulting block contains 39 features.

Path block $P_{i,t}$

Table A3.5: Path Features in the Translated Feature Block $P_{i,t}$

Feature group	Count	Definition
Close efficiency	2	$\mathrm{eff}_5(c)$ and $\mathrm{eff}_{10}(c)$, that is, the efficiency of the close path over the last five and last ten bars, defined as net displacement divided by total path length.
Close path length	2	$\mathrm{pathlen}_5(c)$ and $\mathrm{pathlen}_{10}(c)$, that is, the total variation of the close path over the last five and last ten bars.
Absolute close curvature	2	$\frac{1}{5}\sum_{j\in\mathcal{I}_5}\lvert\Delta^2 c_j\rvert$ and $\frac{1}{10}\sum_{j\in\mathcal{I}_{10}}\lvert\Delta^2 c_j\rvert$, that is, the mean absolute second difference of the close series over $\mathcal{I}_5$ and $\mathcal{I}_{10}$.
Signed close curvature	2	$\frac{1}{5}\sum_{j\in\mathcal{I}_5}\Delta^2 c_j$ and $\frac{1}{10}\sum_{j\in\mathcal{I}_{10}}\Delta^2 c_j$, that is, the mean signed second difference of the close series over $\mathcal{I}_5$ and $\mathcal{I}_{10}$.
Drawup and drawdown (5)	2	$c_{19}-\min_{j\in\mathcal{I}_5} c_j$ and $\max_{j\in\mathcal{I}_5} c_j - c_{19}$, that is, the last-five drawup and drawdown relative to the final close.
Drawup minus drawdown (5)	1	$\left(c_{19}-\min_{j\in\mathcal{I}_5} c_j\right)-\left(\max_{j\in\mathcal{I}_5} c_j - c_{19}\right)$, that is, the asymmetry between drawup and drawdown over the last five bars.
Drawup and drawdown (10)	2	$c_{19}-\min_{j\in\mathcal{I}_{10}} c_j$ and $\max_{j\in\mathcal{I}_{10}} c_j - c_{19}$, that is, the last-ten drawup and drawdown relative to the final close.
Line deviation	2	$\mathrm{areadev}_5(c)$ and $\mathrm{areadev}_{10}(c)$, that is, the mean absolute deviation of the close path from the straight line joining the first and last points over the last five and last ten bars.
Turning count	2	$\mathrm{turn}_5(c)$ and $\mathrm{turn}_{10}(c)$, that is, the number of sign changes in the first differences of the close path over the last five and last ten bars.
Endpoint close rank	1	$\mathrm{ranklast}_{10}(c)$, that is, the rank of the final close within the last ten closes, scaled to $[0,1]$.
Peak-trough ordering	1	$\mathrm{posmax}_{10}(c)-\mathrm{posmin}_{10}(c)$, that is, the difference between the normalized first-occurrence positions of the local peak and trough within the last ten closes.
Endpoint minus mean	1	$c_{19}-\bar{c}_{10}$, where $\bar{c}_{10}=\frac{1}{10}\sum_{j\in\mathcal{I}_{10}} c_j$, that is, the final close minus the average close over $\mathcal{I}_{10}$.
Typical-price efficiency	1	$\mathrm{eff}_5(\mathrm{tp})$, where $\mathrm{tp}_j=(h_j+l_j+c_j)/3$, that is, the efficiency of the typical-price path over the last five bars.
Absolute typical-price curvature	1	$\frac{1}{5}\sum_{j\in\mathcal{I}_5}\lvert\Delta^2\mathrm{tp}_j\rvert$, that is, the mean absolute second difference of the typical-price path over the last five bars.
Total	22	Path translated feature block.

Note: This table defines the Path translated feature block $P_{i,t}$. All features are deterministic functions of the renderer-aligned baseline coordinates defined above, which are constructed from the common dated input record in Equation (A3.1). The block is added as the fifth translated feature block in the cumulative translation and summarizes geometric properties of the recent price path, including efficiency, curvature, drawup and drawdown asymmetry, deviation from a straight-line path, turning behavior, and endpoint location within the recent window. The resulting block contains 22 features.

A4 Formal Definition of Perturbation Operators

This subsection records the exact pixel-space definitions of the six perturbation operators used in the main-text analysis. Each operator corresponds one-to-one with a perturbation label reported in Table 3.4, and is a deterministic map from the rendered chart image to a modified image; the fitted model is held fixed throughout. The definitions below allow the perturbed evaluation in the analysis to be reproduced exactly from the renderer's fixed geometry.

Because the renderer is deterministic and uses fixed geometry, each perturbation can be written as a deterministic pixel-space map applied to the rendered image

$$X_{i,t} = \mathcal{R}(W_{i,t}) \in \{0, \ldots, 255\}^{224 \times 224 \times 3}. \tag{A4.1}$$

For expositional convenience, pixel coordinates are indexed from 1 rather than from 0; the definitions below are therefore the exact one-indexed counterparts of the implementation.

For row index $u \in \{1, \ldots, 224\}$, column index $v \in \{1, \ldots, 224\}$, and channel index $c \in \{1, 2, 3\}$, let $[X_{i,t}]_{u,v,c}$ denote the corresponding RGB intensity. Let $\mathbf{0} \in \mathbb{R}^3$ denote the black pixel. Under the implemented renderer, the price panel occupies rows

$$\mathcal{R}^P = \{(u, v) : 11 \leq u \leq 166,\ 1 \leq v \leq 224\}, \tag{A4.2}$$

and the volume panel occupies rows

$$\mathcal{R}^V = \{(u, v) : 171 \leq u \leq 214,\ 1 \leq v \leq 224\}. \tag{A4.3}$$

The 20 rendered bars are centered horizontally, with bar width 6 pixels and inter-bar gap 4 pixels. Writing $m \in \{1, \ldots, 20\}$ for the bar index from earliest to latest, the mth bar occupies columns

$$\mathcal{C}_m = \{v : 15 + 10(m-1) \leq v \leq 20 + 10(m-1)\}, \tag{A4.4}$$

where $m = 1$ corresponds to date $\tau = t - 19$ and $m = 20$ corresponds to date $\tau = t$.

The implementation does not mask only the body pixels of the first or last five bars. Instead, it masks the full enclosing vertical strip spanning those bars, including the gaps between them. Accordingly, define

$$\mathcal{R}^F = \{(u, v) : 1 \le u \le 224,\ 15 \le v \le 60\}\,, \tag{A4.5}$$

for the first-five-bars strip, and

$$\mathcal{R}^L = \{(u, v) : 1 \le u \le 224,\ 165 \le v \le 210\}\,, \tag{A4.6}$$

for the last-five-bars strip.

With these regions in hand, the perturbation operators are defined as follows.

The volume-panel mask zeros out all pixels in the volume panel:

$$[\Pi_{\text{vol}}(X_{i,t})]_{u,v,c} = \begin{cases} 0, & (u, v) \in \mathcal{R}^V, \\ [X_{i,t}]_{u,v,c}\,, & \text{otherwise.} \end{cases} \tag{A4.7}$$

The first-five-bars mask zeros out the enclosing strip corresponding to the earliest five rendered bars:

$$[\Pi_{\text{first5}}(X_{i,t})]_{u,v,c} = \begin{cases} 0, & (u, v) \in \mathcal{R}^F, \\ [X_{i,t}]_{u,v,c}\,, & \text{otherwise.} \end{cases} \tag{A4.8}$$

The last-five-bars mask zeros out the enclosing strip corresponding to the most recent five rendered bars:

$$[\Pi_{\text{last5}}(X_{i,t})]_{u,v,c} = \begin{cases} 0, & (u, v) \in \mathcal{R}^L, \\ [X_{i,t}]_{u,v,c}\,, & \text{otherwise.} \end{cases} \tag{A4.9}$$

The grayscale perturbation replaces each RGB pixel by a common luminance value

computed from the implemented channel weights. Define

$$g_{i,t}(u,v) = \mathrm{round}\Big(0.2989\,[X_{i,t}]_{u,v,1} + 0.5870\,[X_{i,t}]_{u,v,2} + 0.1140\,[X_{i,t}]_{u,v,3}\Big)\,, \tag{A4.10}$$

with the result clipped to the integer range $\{0,\ldots,255\}$. Then

$$[\Pi_{\mathrm{gray}}(X_{i,t})]_{u,v,c} = g_{i,t}(u,v), \qquad c \in \{1,2,3\}. \tag{A4.11}$$

Thus, the perturbation preserves spatial structure but removes color differences between up and down days.

The horizontal-flip perturbation reverses the image along the horizontal axis:

$$[\Pi_{\mathrm{flip}}(X_{i,t})]_{u,v,c} = [X_{i,t}]_{u,\,225-v,\,c}\,. \tag{A4.12}$$

This preserves local shapes and colors but reverses left-to-right temporal order.

The wick-removal perturbation is implementation-specific. In the rendered chart, the removable set is defined as the set of pure-white pixels in the price panel. Formally, let

$$\mathcal{R}^W(X_{i,t}) = \Big\{(u,v) \in \mathcal{R}^P : [X_{i,t}]_{u,v,1} = [X_{i,t}]_{u,v,2} = [X_{i,t}]_{u,v,3} = 255\Big\}\,. \tag{A4.13}$$

The perturbation then zeros out exactly those pixels:

$$[\Pi_{\mathrm{wick}}(X_{i,t})]_{u,v,c} = \begin{cases} 0, & (u,v) \in \mathcal{R}^W(X_{i,t})\,, \\ [X_{i,t}]_{u,v,c}\,, & \text{otherwise.} \end{cases} \tag{A4.14}$$

Under the implemented renderer, these pure-white price-panel pixels correspond to wick strokes, so Equation A4.14 removes wick information without altering candlestick bodies, moving-average overlays, or the volume panel.

For each perturbation $\pi \in \{\text{vol}, \text{first5}, \text{last5}, \text{gray}, \text{flip}, \text{wick}\}$, the perturbed score is then

$$s_{i,t}^{H,\pi} = \sigma(f_\theta(\Pi_\pi(X_{i,t}))) \,, \tag{A4.15}$$

which is the object used in the perturbation analysis of Subsection 3.3. Because the fitted Hiera parameters are held fixed throughout, all variation induced by Π_π is attributable to the information removed or altered by the corresponding perturbation operator.

B Proofs of the Decomposition Results

This appendix collects the formal results behind Section 2.2. The first result is a population identity: under nested representations, the oracle image–tabular gap separates into encoding loss and tabular capacity loss. The second result formalizes the distillation diagnostic: under a teacher-consistency interpretation of the image score, the residual $1 - R^2$ after score distillation converges to the share of latent signal not recoverable from the tabular representation.

The results are stated under squared-error loss because both projection arguments and the empirical R^2 diagnostic are squared-error objects. The final subsection relates this formulation to the cross-entropy training used in estimation.

B1 Population representation decomposition

Proof of Proposition 1. The nesting condition implies two identities. First, by iterated expectations,

$$p_Z^\star(Z) = \mathbb{E}[Y \mid Z] = \mathbb{E}[\mathbb{E}[Y \mid X] \mid Z] = \mathbb{E}[p_X^\star(X) \mid Z].$$

Second, for any $g \in \mathcal{F}$, $g(Z)$ is X-measurable, and hence

$$\mathbb{E}[(Y - g(Z))^2] = \mathbb{E}[(Y - p_X^\star(X))^2] + \mathbb{E}[(p_X^\star(X) - g(Z))^2].$$

It remains to decompose the second term. For any $g \in \mathcal{F}$,

$$p_X^\star(X) - g(Z) = \{p_X^\star(X) - p_Z^\star(Z)\} + \{p_Z^\star(Z) - g(Z)\}.$$

The two terms are orthogonal because the first has conditional mean zero given Z, while the second is Z-measurable. Therefore,

$$\mathbb{E}[(p_X^\star(X) - g(Z))^2] = \epsilon_{\mathcal{G}} + \mathbb{E}[(p_Z^\star(Z) - g(Z))^2].$$

Minimizing over $g \in \mathcal{F}$ gives

$$\inf_{g \in \mathcal{F}} \mathbb{E}[(p_X^\star(X) - g(Z))^2] = \epsilon_{\mathcal{G}} + \kappa_{\mathcal{F}}.$$

Combining this identity with the risk decomposition above proves the result. □

Remark 1. *The population identity contains no separate target-noise term. The irreducible variance of Y around the conditional mean is common to the image and tabular oracle risks and cancels in the difference. Target noise enters the feasible empirical problem because fitted models are trained on noisy realized labels; this channel is studied through the distillation exercise formalized in Proposition 2.*

B2 Distillation residual

This subsection formalizes the score-distillation diagnostic. The result does not introduce a separate population target-noise term. Instead, it characterizes the residual score-space variation left after the tabular student is trained on the image-model score.

Assumption 2 (Teacher approximation)**.** *The approximation error of the teacher image-model score $T(X) = \hat{p}^H(X)$ is bounded for some finite constant δ_T:*

$$\mathbb{E}[(T(X) - p_W^\star(W))^2] \leq \delta_T^2$$

Assumption 3 (Student approximation)**.** *The tabular student class satisfies*

$$\inf_{g \in \mathcal{F}} \mathbb{E}[(m_T(Z) - g(Z))^2] \xrightarrow{p} 0.$$

Assumption 4 (Nondegenerate teacher score)**.**

$$\mathrm{Var}(T(X)) > 0.$$

These assumptions clarify the interpretation of the diagnostic. Teacher approximation bounds the distance between the fitted Hiera score and the latent full-window conditional signal; the bound need not vanish, so the distillation result is primarily an accounting of fitted-score recoverability. Student approximation requires the tabular class to approximate the teacher's conditional projection on Z when trained on the teacher score. Nondegeneracy only rules out a zero-variance teacher score, so that the score-alignment R^2 is well defined.

Proof of Proposition 2. By the definition of score-alignment R^2,

$$1 - R^2(g_T^\star, T) = \frac{\mathbb{E}[(T(X) - g_T^\star(Z))^2]}{\mathrm{Var}(T(X))}.$$

It is therefore enough to characterize the numerator and denominator.

Let $m_T(Z) = \mathbb{E}[T(X) \mid Z]$. For any $g \in \mathcal{F}$,

$$T(X) - g(Z) = \{T(X) - m_T(Z)\} + \{m_T(Z) - g(Z)\}.$$

The two terms are orthogonal because the first has conditional mean zero given Z, whereas the second is Z-measurable. Hence

$$\mathbb{E}[(T(X) - g(Z))^2] = \mathbb{E}[(T(X) - m_T(Z))^2] + \mathbb{E}[(m_T(Z) - g(Z))^2].$$

Minimizing over $g \in \mathcal{F}$ gives

$$\mathbb{E}[(T(X) - g_T^\star(Z))^2] = \mathbb{E}[(T(X) - m_T(Z))^2] + \inf_{g \in \mathcal{F}} \mathbb{E}[(m_T(Z) - g(Z))^2].$$

By Assumption 3, the second term is $o_p(1)$. Therefore,

$$\mathbb{E}[(T(X) - g_T^\star(Z))^2] = \mathbb{E}[(T(X) - m_T(Z))^2] + o_p(1).$$

Dividing by $\mathrm{Var}(T(X))$, which is nonzero by assumption, gives

$$1 - R^2(g_T^\star, T) = \frac{\mathbb{E}[(T(X) - m_T(Z))^2]}{\mathrm{Var}(T(X))} + o_p(1).$$

For the second statement, let $a = T(X) - m_T(Z)$ and $b = p_W^\star(W) - h_Z^\star(Z)$. Then

$$\left|\mathbb{E}[a^2] - \mathbb{E}[b^2]\right| \leq (\|a\|_2 + \|b\|_2)\, \|a - b\|_2.$$

Moreover,

$$a - b = \{T(X) - p_W^\star(W)\} - \{m_T(Z) - h_Z^\star(Z)\}.$$

Since conditional expectation is an L^2-contraction,

$$\|m_T(Z) - h_Z^\star(Z)\|_2 = \|\mathbb{E}[T(X) - p_W^\star(W) \mid Z]\|_2 \leq \|T(X) - p_W^\star(W)\|_2.$$

Hence, by Assumption 2,

$$\|a - b\|_2 \leq 2\|T(X) - p_W^\star(W)\|_2 \leq 2\delta_T.$$

Since $T(X)$, $m_T(Z)$, $p_W^\star(W)$, and $h_Z^\star(Z)$ are scores in $[0, 1]$, we have $\|a\|_2 + \|b\|_2 \leq 2$. Therefore,

$$\left|\mathbb{E}[(T(X) - m_T(Z))^2] - \mathbb{E}[(p_W^\star(W) - h_Z^\star(Z))^2]\right| \leq 4\delta_T.$$

Combining this bound with the first statement yields the asserted bound for $1 - R^2(g_T^\star, T)$.

□

B3 Relation to cross-entropy training

The image and tabular classifiers are trained by binary cross-entropy, whereas Propositions 1 and 2 are stated under squared-error loss. This is deliberate. For binary outcomes, squared error is the Brier score, a strictly proper scoring rule whose population minimizer is the conditional probability, the same Bayes predictor targeted by cross-

entropy.

The squared-error formulation also matches the empirical distillation statistic: R^2 is a projection measure based on squared deviations between the student score and the image-model score. Thus, the training loss determines how the scores are estimated, while the appendix results characterize the population projection objects used to interpret their alignment.

The residual H–L spread reported in the empirical section gives the portfolio-space magnitude of the residual score component. It is not a separate identification result for the squared-error encoding object.

C Deep Tabular Benchmarks

This appendix documents two deep tabular benchmark families used as additional complexity checks in the translation exercise: a multilayer perceptron (MLP) and a continuous-feature FT-Transformer. Both models are estimated on the same stage-specific translated tabular specifications $Z_{i,t}^{(k)}$, using the same Hiera-aligned class-balanced training subset, the same validation period, and the same out-of-sample evaluation grid as the other tabular benchmarks. The MLP is included as a flexible feed-forward deep tabular benchmark, while the FT-Transformer follows the continuous-feature tokenization approach of Gorishniy et al. (2021). The purpose of these models is not to replace the main benchmark families, but to assess whether the central conclusions depend on restricting tabular comparisons to linear and moderately flexible nonlinear learners.

For each translation stage $k \in \{0, \ldots, 5\}$, let

$$Z_{i,t}^{(k)} \in \mathbb{R}^{d_k} \tag{C0.1}$$

denote the corresponding stage-specific tabular predictor vector, where d_k varies with the included translated feature blocks. Before estimation, both deep tabular benchmarks use a common standardization constructed from the exact selected training rows only. Let

$\mu \in \mathbb{R}^{d_5}$ and $\sigma \in \mathbb{R}^{d_5}$ denote the elementwise mean and standard deviation of the union feature space formed by the fully translated vector $Z_{i,t}^{(5)}$. For stage k, the model input is the corresponding standardized subvector

$$\widetilde{Z}_{i,t}^{(k)} = \left[\frac{Z_{i,t}^{(k)} - \mu^{(k)}}{\sigma^{(k)}} \right]_{-8}^{8}, \tag{C0.2}$$

where $\mu^{(k)}$ and $\sigma^{(k)}$ collect the coordinates relevant for stage k, and $[\cdot]_{-8}^{8}$ denotes elementwise clipping to the interval $[-8, 8]$.

Let $b \in \{M, F\}$ index the MLP and FT-Transformer benchmarks. For each stage k, model b maps $\widetilde{Z}_{i,t}^{(k)}$ into a scalar logit

$$\ell_{i,t}^{b,k} = f_{\theta_k}^{b}\left(\widetilde{Z}_{i,t}^{(k)}\right), \tag{C0.3}$$

and the corresponding probability forecast is

$$s_{i,t}^{b,k} = \sigma\left(\ell_{i,t}^{b,k}\right) = \frac{1}{1 + \exp\left(-\ell_{i,t}^{b,k}\right)}. \tag{C0.4}$$

Thus, $s_{i,t}^{b,k}$ is interpreted as the estimated probability that the five-day-ahead cumulative return is positive.

For a training sample of size N, parameters are estimated by minimizing the binary cross-entropy objective

$$\mathcal{L}^{b,k}(\theta_k) = -\frac{1}{N} \sum_{(i,t)} \left[y_{i,t}^{(5)} \log s_{i,t}^{b,k} + \left(1 - y_{i,t}^{(5)}\right) \log\left(1 - s_{i,t}^{b,k}\right) \right]. \tag{C0.5}$$

Optimization uses AdamW, gradient clipping, and early stopping based on validation log loss. For each model family and translation stage, the reported specification is the checkpoint attaining the lowest validation log loss.

The MLP benchmark is a four-hidden-layer feed-forward network with batch normalization and dropout after each hidden affine map. Writing

$$h_{i,t}^{M,k,(0)} = \widetilde{Z}_{i,t}^{(k)}, \tag{C0.6}$$

the hidden layers satisfy

$$h_{i,t}^{M,k,(\ell)} = \mathcal{D}^{(\ell)}\Big(\phi\Big(\mathrm{BN}^{(\ell)}\Big(W^{(\ell)} h_{i,t}^{M,k,(\ell-1)} + b^{(\ell)}\Big)\Big)\Big), \qquad \ell = 1, \ldots, 4, \tag{C0.7}$$

where $\phi(\cdot)$ is the ReLU activation and $\mathcal{D}^{(\ell)}(\cdot)$ denotes dropout. The output logit is

$$\ell_{i,t}^{M,k} = w^{\mathrm{out}\prime} h_{i,t}^{M,k,(4)} + b^{\mathrm{out}}. \tag{C0.8}$$

This benchmark allows for nonlinear interactions among the translated tabular predictors while remaining substantially more structured than the image model.

The FT-Transformer benchmark follows the continuous-feature formulation of Gorishniy et al. (2021). For stage k, write

$$\widetilde{Z}_{i,t}^{(k)} = \Big(\tilde{z}_{i,t,1}^{(k)}, \ldots, \tilde{z}_{i,t,d_k}^{(k)}\Big)'. \tag{C0.9}$$

Each scalar feature is converted into a token embedding through a feature-specific affine map,

$$e_{i,t,j}^{k} = \tilde{z}_{i,t,j}^{(k)}\, a_j^k + c_j^k, \qquad e_{i,t,j}^{k} \in \mathbb{R}^{d_{\mathrm{tok}}}, \qquad j = 1, \ldots, d_k, \tag{C0.10}$$

where $a_j^k, c_j^k \in \mathbb{R}^{d_{\mathrm{tok}}}$ are learned feature-specific parameters. These tokens are stacked together with a learned classification token,

$$H_{i,t}^{F,k,(0)} = \Big[e_{\mathrm{cls}}; e_{i,t,1}^{k}; \ldots; e_{i,t,d_k}^{k}\Big] \in \mathbb{R}^{(d_k+1)\times d_{\mathrm{tok}}}. \tag{C0.11}$$

For transformer block m, the representation is updated by

$$\widetilde{H}_{i,t}^{F,k,(m)} = H_{i,t}^{F,k,(m-1)} + \mathrm{MHA}\Big(\mathrm{LN}\Big(H_{i,t}^{F,k,(m-1)}\Big)\Big), \tag{C0.12}$$

$$H_{i,t}^{F,k,(m)} = \widetilde{H}_{i,t}^{F,k,(m)} + \mathrm{FFN}\Big(\mathrm{LN}\Big(\widetilde{H}_{i,t}^{F,k,(m)}\Big)\Big), \qquad m = 1, \ldots, M. \tag{C0.13}$$

After the final block, the model forms its logit from the transformed classification token,

$$\ell_{i,t}^{F,k} = w^{\text{out}\prime}\,\text{LN}\Big(H_{i,t,0}^{F,k,(M)}\Big) + b^{\text{out}}, \tag{C0.14}$$

where $H_{i,t,0}^{F,k,(M)}$ denotes the final representation of the classification token. In the implementation used here, the FT-Transformer is kept deliberately shallow, so that differences across translated tabular specifications are driven primarily by the included translated feature blocks rather than by aggressive architecture scaling.

Table C0.1 reports the exact architectures and training hyperparameters for the two deep tabular benchmarks.

Table C0.2 then reports out-of-sample portfolio performance for the deep tabular benchmarks across the same three translation views used elsewhere in the appendix: the cumulative ladder, one-block additions to the baseline tabular specification, and one-block omissions from the fully translated specification.

While Table C0.2 summarizes the high-minus-low spread performance of the deep tabular benchmarks, it does not show how those results map into the full decile structure. Table C0.3 therefore reports the decile portfolio profiles for the Base and Full specifications for each deep tabular model.

Table C0.1: Deep Tabular Benchmark Architectures and Training Hyperparameters

Setting	MLP	FT-Transformer
Reference architecture	Feed-forward deep tabular benchmark	Continuous-feature FT-Transformer
Methodological reference	Gorishniy et al. (2021)	Gorishniy et al. (2021)
Input dimension	Stage-specific d_s	Stage-specific d_s
Input scaling	Selected-train standardization, clipped to $[-8, 8]$	Selected-train standardization, clipped to $[-8, 8]$
Hidden / token dimension	(256, 192, 128, 64)	$d_{\text{token}} = 96$
Depth	4 hidden layers	3 transformer blocks
Attention heads	Not applicable	8
FFN expansion	Not applicable	$2.0 \times d_{\text{token}}$
Activation	ReLU	ReLU in block FFN
Normalization	BatchNorm1d after each hidden layer	LayerNorm in each block and final LayerNorm
Dropout	0.15 after each hidden layer	Attention = 0.10, FFN = 0.10, residual = 0.00
Output layer	Linear scalar logit head	Linear scalar logit head on the final [CLS] token
Optimizer	AdamW	AdamW
Learning rate	10^{-3}	8×10^{-4}
Weight decay	10^{-4}	10^{-5}
Batch size	8192	4096
Maximum epochs	50	50
Early stopping patience	6	6
Minimum improvement threshold	10^{-4}	10^{-4}
Gradient clipping	Max norm = 1.0	Max norm = 1.0
Selection criterion	Lowest validation log loss	Lowest validation log loss
Loss	Binary cross-entropy with logits	Binary cross-entropy with logits

Note: This table reports the architectures and training hyperparameters for the deep tabular benchmarks. The models are estimated separately on each tabular specification, so the input dimension d_s is stage-specific. In the cumulative translation ladder, the predictor count rises from 220 in Base to 421 in Full, and the one-block addition and one-block omission exercises use the corresponding intermediate dimensions implied by the included translated feature blocks. Both models are trained on the same selected training rows and full validation split as the other tabular benchmarks. Input scaling uses selected-train standardization with clipping to $[-8, 8]$ for both models. CUDA runs use automatic mixed precision when available.

Table C0.2: Deep Tabular Benchmarks on the Translated Specification Ladder, 2006–2025

Model	Specification	Features	H–L Return	H–L Sharpe	ΔSharpe vs. Base
			Panel A: Cumulative Translation		
FT-Transformer	Base	220	0.1387	0.722	0.000
	+ Geometry	280	0.1506	1.167	0.445
	+ Wick	320	0.1605	1.192	0.470
	+ Direction	360	0.1449	0.998	0.276
	+ Volume	399	0.0905	0.608	-0.114
	+ Path	421	0.1535	1.180	0.458
MLP	Base	220	0.1299	0.922	0.000
	+ Geometry	280	0.1311	0.930	0.008
	+ Wick	320	0.1486	1.064	0.142
	+ Direction	360	0.2023	1.597	0.675
	+ Volume	399	0.1680	1.332	0.410
	+ Path	421	0.1303	1.207	0.285
			Panel B: One Block Additions		
FT-Transformer	Base	220	0.1387	0.722	0.000
	Base + Geometry	280	0.1506	1.167	0.445
	Base + Wick	260	0.1301	1.124	0.402
	Base + Direction	260	0.1776	0.921	0.199
	Base + Volume	259	0.1318	1.006	0.284
	Base + Path	242	0.2324	1.398	0.676
MLP	Base	220	0.1299	0.922	0.000
	Base + Geometry	280	0.1311	0.930	0.008
	Base + Wick	260	0.1301	0.929	0.007
	Base + Direction	260	0.1765	1.324	0.402
	Base + Volume	259	0.1243	0.958	0.036
	Base + Path	242	0.1734	1.304	0.382

(Continued on next page)

Table C0.2: Deep Tabular Benchmarks on the Translated Specification Ladder (Continued)

Model	Specification	Features	H–L Return	H–L Sharpe	ΔSharpe vs. Base
			Panel C: One Block Omissions		
FT-Transformer	Full	421	0.1535	1.180	0.458
	Full − Geometry	361	0.2039	1.057	0.335
	Full − Wick	381	0.1944	1.339	0.617
	Full − Direction	381	0.1472	1.148	0.426
	Full − Volume	382	0.1560	1.201	0.479
	Full − Path	399	0.0905	0.608	-0.114
MLP	Full	421	0.1303	1.207	0.285
	Full − Geometry	361	0.1823	1.414	0.492
	Full − Wick	381	0.1378	1.125	0.203
	Full − Direction	381	0.1671	1.404	0.482
	Full − Volume	382	0.2064	1.667	0.745
	Full − Path	399	0.1680	1.332	0.410

Note: This table reports out-of-sample portfolio diagnostics for the deep tabular benchmarks on the translated specification ladder over 2006–2025 on the common evaluation sample. Panel A reports cumulative translation results, Panel B reports one-block additions to the baseline tabular specification, and Panel C reports one-block omissions from the fully translated specification. Base denotes the baseline tabular specification, and Full denotes the translated specification that includes all translated feature blocks. Features reports the number of predictors in the corresponding specification. H–L Return is the annualized return of the equal-weighted high-minus-low decile spread formed from the model scores and rebalanced every five trading days. H–L Sharpe is the corresponding Sharpe ratio computed from the rebalance-return series. ΔSharpe vs. Base reports the change in H–L Sharpe ratio relative to Base within the corresponding model family.

Table C0.3: Decile Portfolio Profiles for the Deep Tabular Base and Full Specifications, 2006–2025

Portfolio	Base Return	Base Sharpe	Base TO	Full Return	Full Sharpe	Full TO
			Panel A: FT-Transformer			
Low	0.0611	0.402	0.665	0.0350	0.214	0.834
2	0.0821	0.492	0.820	0.0798	0.454	0.882
3	0.1238	0.678	0.854	0.1029	0.540	0.896
4	0.1108	0.614	0.873	0.1279	0.685	0.899
5	0.1227	0.656	0.879	0.1051	0.585	0.897
6	0.1354	0.691	0.879	0.1275	0.682	0.899
7	0.1624	0.774	0.867	0.1690	0.715	0.894
8	0.1214	0.618	0.849	0.1406	0.733	0.894
9	0.1380	0.693	0.838	0.1813	0.904	0.889
High	0.1998	0.815	0.841	0.1885	0.948	0.870
H–L	0.1387	0.722	0.753	0.1535	1.180	0.852
			Panel B: MLP			
Low	0.0665	0.402	0.850	0.0605	0.354	0.861
2	0.1009	0.541	0.897	0.1074	0.551	0.896
3	0.0963	0.539	0.904	0.0842	0.456	0.903
4	0.1121	0.585	0.902	0.0972	0.511	0.897
5	0.1050	0.565	0.900	0.1186	0.644	0.899
6	0.1141	0.604	0.897	0.1408	0.717	0.897
7	0.1763	0.742	0.895	0.1002	0.538	0.899
8	0.1291	0.680	0.897	0.1957	0.861	0.900
9	0.1604	0.832	0.901	0.1614	0.867	0.901
High	0.1965	0.992	0.909	0.1908	1.082	0.890
H–L	0.1299	0.922	0.880	0.1303	1.207	0.876

Note: This table reports decile portfolio profiles for the deep tabular benchmarks on the common evaluation sample over 2006–2025. Panel A reports FT-Transformer results and Panel B reports MLP results. Base denotes the baseline tabular specification with 220 predictors, and Full denotes the translated specification that includes all translated feature blocks with 421 predictors. At each rebalance date, stocks are sorted into equal-weighted deciles by model-implied probability of a positive 5-day return. High denotes the highest-score decile, Low denotes the lowest-score decile, and H–L denotes the equal-weighted high-minus-low decile spread. Return denotes annualized close-to-close arithmetic mean return, Sharpe denotes Sharpe ratio computed from the corresponding rebalance-return series, and TO denotes average one-way turnover per rebalance.

D Supplementary Results - Benchmarks

This section collects supplementary benchmark evidence that supports the main common-information comparison between Hiera and the baseline tabular benchmarks. The tables below add forecast-space diagnostics, portfolio decompositions, and residual-content tests that complement the headline benchmark results in the main text while leaving the underlying evaluation design unchanged. The first item is a calibration check on Hiera's predicted probability against the natural out-of-sample class proportion, motivated by the class-balanced training design described in Subsection 2.3.

D1 Additional Forecast Diagnostics

This subsection reports supplementary forecast-space diagnostics for Hiera and the baseline tabular benchmarks. The tables below provide split-level classification summaries, tail-spread decompositions under alternative tail cuts, and the full information-coefficient diagnostics, thereby expanding the benchmark comparison beyond the headline measures reported in the main text.

Table D2.1 focuses on probability fit and classification performance at the split level. The next table turns to forecast discrimination in the tails of the score distribution, which is closer to the decile-based forecast-space comparisons emphasized in the main text.

Table D1.1 shows that the same model ordering remains visible across alternative tail cuts, not only at the 10% threshold used in the headline specification. Table D1.2 then reports the corresponding full-cross-section ranking diagnostics through Pearson and Spearman information coefficients.

D2 Portfolio Decomposition

This subsection decomposes the benchmark portfolio results into their long-side and short-side components relative to the equal-weighted market portfolio. The purpose is to

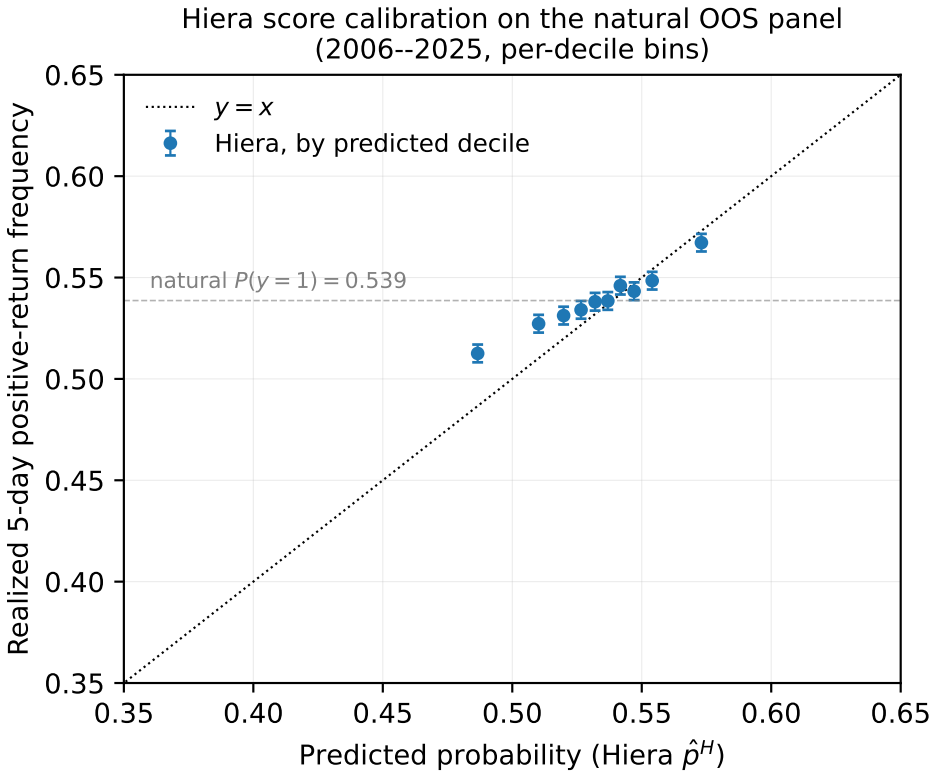


Figure D0.1: Hiera score calibration on the natural OOS panel, 2006–2025. Each point reports, for one predicted-probability decile of the saved Hiera score $\hat{p}^H$, the mean predicted probability on the horizontal axis and the realized 5-day positive-return frequency on the vertical axis (vertical bars are Wald 95% confidence intervals). The grey dashed line is the natural panel prior $P(y{=}1) = 0.539$ and the dotted line is the 45° reference. The score is monotone in the realized hit rate and the mean predicted probability is within 0.6 percentage points of the natural prior; the small lift in the lowest predicted decile reflects mild under-prediction at low scores, and the highest decile is slightly over-predicted. Hiera was trained on a class-balanced 600,000-row subset (300,000 positive and 300,000 negative) so a perfect 45° alignment on the natural panel is not the design objective; the figure verifies that the score nonetheless tracks the realized hit rate closely on the panel used for distillation and downstream decomposition.

Alt text: Calibration plot for Hiera predictions on the 2006–2025 out-of-sample panel. Ten points represent predicted-probability deciles, with mean predicted probability on the horizontal axis and realized 5-day positive-return frequency on the vertical axis; vertical error bars show 95% confidence intervals. The points rise monotonically from approximately 0.51 realized frequency at a predicted probability of 0.49 to approximately 0.57 at 0.58. Most points lie close to the dotted 45-degree calibration line and cluster around the dashed natural positive-return frequency of 0.539, with the lowest decile above the calibration line and the highest decile slightly below it.

Table D1.1: Tail-Spread Decomposition for Hiera and the Baseline Tabular Benchmarks

Model	Tail Cut	Top Hit Rate	Bottom Hit Rate	Tail Spread	Newey–West t-stat
Hiera	5%	0.5512	0.5197	0.0315	5.030
	10%	0.5506	0.5237	0.0269	5.487
	20%	0.5489	0.5288	0.0201	5.479
HistGB	5%	0.5509	0.5287	0.0222	3.990
	10%	0.5524	0.5314	0.0211	4.475
	20%	0.5467	0.5312	0.0155	4.161
Logit	5%	0.5494	0.5330	0.0164	2.629
	10%	0.5465	0.5319	0.0147	2.804
	20%	0.5432	0.5335	0.0097	2.302

Note: This table reports tail-spread decompositions for Hiera and the baseline tabular benchmarks on the common evaluation sample over 2006–2025. At each rebalance date, stocks are sorted by model score, and the top and bottom tails are then defined using the indicated tail cut. Top Hit Rate is the frequency of positive 5-day-ahead outcomes in the top tail, Bottom Hit Rate is the corresponding frequency in the bottom tail, and Tail Spread is the difference between the two. The reported Newey–West t-statistics are based on the time series of date-level tail spreads and use five lags. The underlying panel contains 1,006 rebalance dates, and the unconditional frequency of positive 5-day-ahead outcomes on the common evaluation sample is 0.5386.

Table D1.2: Out-of-Sample Information Coefficients for Hiera and the Baseline Tabular Benchmarks, 2006–2025

Model	Pearson IC	Newey–West t-stat	Spearman IC	Newey–West t-stat	Dates
Hiera	0.02394	8.129	0.02634	7.925	1,006
HistGB	0.02057	7.406	0.02064	6.281	1,006
Logit	0.01442	4.882	0.01591	4.227	1,006

Note: This table reports out-of-sample information coefficients for Hiera and the baseline tabular benchmarks on the common evaluation sample over 2006–2025. Pearson IC and Spearman IC are the mean date-level cross-sectional Pearson and rank correlations, respectively, between model scores and subsequent 5-day cumulative close-to-close returns. The reported Newey–West t-statistics are based on the corresponding time series of date-level information coefficients and use three lags. The common evaluation sample contains 1,006 rebalance dates.

Table D2.1: Full Classification Diagnostics for Hiera and the Baseline Tabular Benchmarks

Model	Split	Obs.	Positive Rate	Mean Pred. Prob.	Accuracy	Balanced Acc.	Log Loss	AUC	Brier
Hiera	Validation	250,556	0.5370	0.5323	0.5347	—	0.6897	0.5231	0.2483
	Test	499,409	0.5386	0.5328	0.5369	0.5047	0.6900	0.5151	0.2484
HistGB	Training	600,000	0.5000	0.5000	0.5791	0.5791	0.6810	0.6129	0.2440
	Validation	250,556	0.5370	0.4994	0.5160	0.5142	0.6930	0.5213	0.2499
	Test	499,409	0.5386	0.4998	0.5102	0.5085	0.6941	0.5128	0.2505
Logit	Training	600,000	0.5000	0.5005	0.5299	0.5299	0.6961	0.5412	0.2500
	Validation	250,556	0.5370	0.5028	0.5140	0.5128	0.6962	0.5172	0.2509
	Test	499,409	0.5386	0.5018	0.5090	0.5087	0.6977	0.5103	0.2516

Note: This table reports split-level classification diagnostics for Hiera and the baseline tabular benchmarks. For the tabular benchmarks, the Training split is the balanced 600,000-observation estimation sample used for model fitting. The Validation split contains 250,556 observations from 2004–2005, and the Test split contains 499,409 observations from the common evaluation sample over 2006–2025. Positive Rate is the share of observations with a positive 5-day cumulative close-to-close return. Mean Pred. Prob. is the average predicted probability of a positive outcome. Balanced Acc. is the average of sensitivity and specificity. Hiera training diagnostics are not reported because full training-set prediction summaries were not archived during training. Hiera validation diagnostics are recomputed from the best-validation predictions retained during model selection; balanced accuracy for that validation summary was not archived separately and is therefore omitted.

clarify where the benchmark differences arise within the cross section, rather than only through the aggregate High–Low spread reported in the main text.

Table D2.2: Long-Side and Short-Side Decomposition of Out-of-Sample Portfolio Performance, 2006–2025

Model	High Ret	EW Ret	Low Ret	High−EW Ret	High−EW SR	EW−Low Ret	EW−Low SR	H–L Ret	H–L SR
Hiera	0.2326	0.1257	0.0259	0.1069	1.173	0.0998	1.353	0.2067	1.496
HistGB	0.2298	0.1257	0.0541	0.1041	1.191	0.0716	1.031	0.1757	1.343
Logit	0.2066	0.1257	0.0663	0.0809	1.001	0.0594	0.715	0.1403	1.037

Note: This table reports a long-side and short-side decomposition of out-of-sample equal-weighted decile portfolio performance for Hiera and the baseline tabular benchmarks on the common evaluation sample over 2006–2025. At each rebalance date, stocks are sorted into deciles by model score. High denotes the highest-score decile, Low denotes the lowest-score decile, and EW denotes the equal-weighted market portfolio that holds all stocks in the evaluation universe. Ret denotes annualized close-to-close arithmetic mean return, and SR denotes Sharpe ratio computed from the corresponding rebalance-return series. High−EW Ret and EW−Low Ret decompose the overall H–L spread into its long-side and short-side components relative to EW. H–L Ret and H–L SR report the equal-weighted high-minus-low decile spread formed from the same score sorts and rebalanced every five trading days.

D3 Residual Content Diagnostics

This subsection reports supplementary residual-content diagnostics for Hiera relative to the baseline tabular benchmarks. The tables below document score overlap, conditional-sort asymmetry, and the broader spanning structure behind the main-text conclusion that Hiera retains distinct predictive content under common information.

Table D3.1: Score Overlap and Conditional-Sort Asymmetry, 2006–2025

Comparison	Overall Pearson	Overall Spearman	Mean Date-Level Pearson	Mean Date-Level Spearman
Panel A: Pairwise score correlations				
Hiera vs. HistGB	0.6706	0.6374	0.5936	0.5632
Hiera vs. Logit	0.4835	0.5710	0.3957	0.4646
Logit vs. HistGB	0.5923	0.6216	0.5260	0.5420
Panel B: Conditional-sort spreads				
Conditional sort	Annualized Return	Sharpe Ratio	Avg. Bins Used	Periods
Hiera within Logit deciles	0.1299	1.315	10.00	1,006
Logit within Hiera deciles	0.0524	0.464	10.00	1,006
Hiera within HistGB deciles	0.1308	1.227	10.00	1,006
HistGB within Hiera deciles	0.0736	0.754	10.00	1,006

Note: This table reports score overlap and conditional-sort asymmetry for Hiera and the baseline tabular benchmarks on the common evaluation sample over 2006–2025. Panel A reports pairwise score correlations. Overall correlations are computed by pooling all stock-date observations in the common evaluation sample. Mean date-level correlations are computed by first measuring the cross-sectional correlation within each rebalance date and then averaging across rebalance dates. Panel B reports equal-weighted conditional high-minus-low spreads. For example, "Hiera within Logit deciles" first assigns stocks to deciles on the Logit score and then forms the Hiera high-minus-low spread within each Logit decile; the reported annualized return and Sharpe ratio are then averaged across the ten conditioning deciles. Avg. Bins Used reports the average number of non-empty conditioning deciles per rebalance date. Periods reports the number of rebalance dates used in the conditional-sort evaluation.

Table D3.1 provides a descriptive view of how closely the benchmark scores move together and how asymmetric the conditional-sort evidence is across model orderings. Table D3.2 then reports the corresponding regression-based residual-content diagnostics in both cross-sectional and portfolio space.

Table D3.2: Nested Cross-Sectional and Time-Series Spanning Diagnostics, 2006–2025

Specification	Hiera Coef.	HistGB Coef.	Logit Coef.	Avg. R^2	Median R^2
Panel A: Fama–MacBeth-style cross-sectional spanning					
Hiera only	0.001163***	—	—	0.0102	0.003744
HistGB only	—	0.001048***	—	0.0097	0.003506
Logit only	—	—	0.000625***	0.0096	0.003549
Hiera + HistGB	0.000804***	0.000572***	—	0.0150	0.008076
Hiera + Logit	0.001093***	—	0.000205	0.0167	0.009181
All three	0.000791***	0.000582***	0.000021	0.0208	0.013271
Panel B: Time-series portfolio spanning					
Dependent spread / controls	Alpha	HistGB Beta	Logit Beta	R^2	Periods
EW−Low / HistGB only	0.000923***	0.744***	—	0.4911	1,006
EW−Low / Logit only	0.001454***	—	0.447***	0.2531	1,006
EW−Low / Logit + HistGB	0.000873***	0.639***	0.169**	0.5176	1,006
H–L / HistGB only	0.001625***	0.710***	—	0.4524	1,006
H–L / Logit only	0.002381***	—	0.618***	0.3666	1,006
H–L / Logit + HistGB	0.001496***	0.510***	0.297***	0.5010	1,006

Note: This table reports nested cross-sectional and time-series spanning diagnostics for Hiera relative to the baseline tabular benchmarks on the common evaluation sample over 2006–2025. Panel A reports Fama–MacBeth-style cross-sectional spanning regressions using standardized model scores. The reported coefficients are average date-level slope estimates, and inference is based on Newey–West-adjusted standard errors with three lags. Avg. R^2 and Median R^2 summarize the distribution of date-level cross-sectional fit across rebalance dates. Panel B reports time-series spanning regressions for Hiera-based spread returns on benchmark spread returns. Alpha is the intercept per rebalance period and is not annualized. EW denotes the equal-weighted market portfolio that holds all stocks in the evaluation universe, so EW−Low is the equal-weighted market-minus-low spread and H–L is the high-minus-low spread. Time-series spanning inference is based on Newey–West-adjusted standard errors with three lags. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

E Supplementary Results - Perturbations

This section collects supplementary evidence from the inference-time perturbation analysis. The tables below extend the main-text perturbation results by reporting the full decile structure under each perturbation, additional score-stability diagnostics, and the corresponding residual-content fragility tests relative to the baseline tabular benchmarks.

E1 Performance and Stability

This subsection reports the level-performance and score-stability consequences of the inference-time perturbations. The first table gives the full decile structure under each perturbed Hiera ranking, while the second summarizes how strongly each perturbation changes stock-level scores, ranks, and portfolio outcomes relative to the unperturbed image.

Table E2.1 shows how each perturbation changes the full return, Sharpe-ratio, and turnover structure across the score-sorted deciles. Table E1.1 then summarizes the same exercise more directly through score-stability measures and return differences relative to the unperturbed image.

E2 Residual Content and Fragility

This subsection turns from level performance to the fragility of Hiera's residual content under the same inference-time perturbations. The tables below ask whether the chart components disrupted by each perturbation matter not only for the raw Hiera ranking, but also for the part of Hiera's signal that remains distinct from the baseline tabular benchmarks.

Table E2.2 reports the full conditional-sort decomposition under the perturbed Hiera rankings, separating the conditional spread into its long-side and short-side components within Logit and HistGB deciles. Table E2.3 then gives the corresponding regression-based

Table E1.1: Score Stability and Portfolio-Difference Diagnostics under Inference-Time Image Perturbations, 2006–2025

Panel A: Score stability relative to the unperturbed image

Perturbation	Mean $\lvert\Delta p\rvert$	Mean Abs. Rank Shift	Score Pearson	Score Spearman	High-Decile Overlap	Low-Decile Overlap
Mask last 5 bars	0.0193	0.2588	0.3391	0.3364	0.2169	0.2909
Horizontal flip	0.0184	0.2683	0.3339	0.2981	0.1886	0.3172
Remove wicks	0.0198	0.1655	0.7376	0.7076	0.4532	0.5700
Mask volume panel	0.0073	0.0694	0.9375	0.9357	0.7722	0.7873
Grayscale image	0.0113	0.1367	0.8104	0.7929	0.5591	0.6123
Mask first 5 bars	0.0062	0.0760	0.9455	0.9316	0.7768	0.7855

Panel B: Return differences relative to the unperturbed image

Perturbation	Unpert. − Pert. High Ret	NW t-stat	Unpert. − Pert. Low Ret	NW t-stat	Unpert. − Pert. H–L Ret	NW t-stat
Mask last 5 bars	0.0460	1.663	-0.0624	-3.803	0.1085	3.066
Horizontal flip	0.0352	1.399	-0.0483	-2.722	0.0835	2.386
Remove wicks	0.0417	2.285	-0.0206	-1.454	0.0624	2.694
Mask volume panel	0.0226	2.154	-0.0035	-0.469	0.0261	1.966
Grayscale image	-0.0174	-0.969	-0.0151	-1.625	-0.0023	-0.110
Mask first 5 bars	0.0052	0.482	0.0052	0.690	-0.0001	-0.005

Note: This table reports score-stability and portfolio-difference diagnostics for Hiera under inference-time image perturbations on the common evaluation sample over 2006–2025. The model is held fixed throughout; perturbations are applied to the rendered input image at evaluation only, with no retraining. Panel A compares each perturbed Hiera score with the corresponding score from the unperturbed image. Mean $\lvert\Delta p\rvert$ is the average absolute change in predicted positive-outcome probability. Mean Abs. Rank Shift is the average absolute change in within-date rank. Score Pearson and Score Spearman are the mean date-level cross-sectional correlations between perturbed and unperturbed scores. High-Decile Overlap and Low-Decile Overlap report the share of names that remain in the unperturbed High and Low deciles after the perturbation. Panel B reports differences in annualized equal-weighted portfolio returns, computed as close-to-close arithmetic mean returns as in Table E2.1, between the unperturbed-image sort and the perturbed sort. Positive values in the High Ret and H–L Ret columns indicate deterioration under perturbation. Negative values in the Low Ret column indicate that the perturbed Low decile earns a higher return than the corresponding Low decile under the unperturbed image, which also compresses the spread. The reported Newey–West t-statistics are based on the time series of periodic return differences using three lags.

fragility diagnostics through nested cross-sectional and time-series spanning tests.

Table E2.1: Full Decile Structure under Inference-Time Image Perturbations, 2006–2025

Portfolio	Unperturbed image	Mask last 5 bars	Horizontal flip	Remove wicks	Mask volume panel	Grayscale image	Mask first 5 bars
Panel A: Annualized returns							
Low	0.0259	0.0883	0.0742	0.0465	0.0293	0.0410	0.0206
2	0.0790	0.1055	0.1082	0.0669	0.0904	0.0859	0.0762
3	0.0706	0.0938	0.1255	0.0964	0.0908	0.0713	0.0981
4	0.1064	0.1088	0.1209	0.1072	0.1055	0.0782	0.0961
5	0.1046	0.1252	0.1239	0.0910	0.1316	0.0989	0.1094
6	0.1305	0.1007	0.1190	0.1255	0.1113	0.1351	0.1066
7	0.1457	0.1378	0.1141	0.1500	0.1313	0.1383	0.1672
8	0.1746	0.1746	0.1229	0.1647	0.1358	0.1664	0.1590
9	0.1876	0.1357	0.1507	0.2182	0.2214	0.1921	0.1964
High	0.2326	0.1865	0.1974	0.1908	0.2100	0.2500	0.2274
H–L	0.2067	0.0982	0.1232	0.1443	0.1806	0.2090	0.2068
Panel B: Sharpe ratios							
Low	0.154	0.501	0.434	0.282	0.171	0.240	0.123
2	0.456	0.589	0.500	0.381	0.505	0.489	0.437
3	0.400	0.527	0.679	0.509	0.499	0.398	0.500
4	0.542	0.605	0.625	0.599	0.534	0.427	0.529
5	0.546	0.673	0.670	0.479	0.692	0.535	0.591
6	0.712	0.560	0.626	0.649	0.615	0.713	0.577
7	0.764	0.712	0.621	0.765	0.695	0.734	0.734
8	0.764	0.760	0.680	0.843	0.745	0.736	0.833
9	0.967	0.728	0.788	0.924	0.951	0.960	1.006
High	1.143	0.913	0.989	0.997	1.061	1.219	1.132
H–L	1.496	0.722	0.980	1.061	1.354	1.505	1.527
Panel C: Average one-way turnover							
Low	0.777	0.771	0.696	0.757	0.807	0.789	0.775
2	0.876	0.869	0.849	0.869	0.878	0.877	0.881
3	0.890	0.887	0.877	0.888	0.892	0.889	0.894
4	0.896	0.892	0.888	0.894	0.894	0.895	0.897
5	0.895	0.895	0.892	0.896	0.897	0.894	0.895
6	0.895	0.895	0.889	0.895	0.895	0.896	0.894
7	0.890	0.892	0.887	0.893	0.890	0.893	0.889
8	0.884	0.882	0.877	0.885	0.883	0.885	0.886
9	0.877	0.868	0.859	0.878	0.880	0.875	0.885
High	0.860	0.837	0.788	0.862	0.876	0.838	0.873
H–L	0.818	0.804	0.742	0.810	0.841	0.813	0.824

Note: This table reports the full decile structure for Hiera under the unperturbed image and each inference-time image perturbation on the common evaluation sample over 2006–2025. The model is held fixed throughout; perturbations are applied to the input image at evaluation only, with no retraining. At each rebalance date, stocks are sorted into deciles by the corresponding perturbed Hiera score. Panel A reports annualized close-to-close arithmetic mean returns, Panel B reports Sharpe ratios computed from the corresponding rebalance-return series, and Panel C reports average one-way turnover per rebalance. High denotes the highest-score decile, Low denotes the lowest-score decile, and H–L denotes the equal-weighted high-minus-low decile spread formed from the same score sorts and rebalanced every five trading days.

Table E2.2: Conditional-Sort Fragility Decomposition under Inference-Time Image Perturbations, 2006–2025

	Unperturbed image	Mask last 5 bars	Horizontal flip	Remove wicks	Mask volume panel	Grayscale image	Mask first 5 bars
Panel A: Within Logit deciles — annualized returns							
High−EW	0.0584	0.0524	0.0539	0.0447	0.0507	0.0914	0.0742
EW−Low	0.0714	0.0281	0.0426	0.0555	0.0446	0.0587	0.0694
H–L	0.1299	0.0805	0.0965	0.1001	0.0954	0.1501	0.1436
Panel B: Within Logit deciles — Sharpe ratios							
High−EW	0.944	0.547	0.727	0.709	0.778	1.030	1.097
EW−Low	1.112	0.438	0.671	0.819	0.674	0.863	1.073
H–L	1.315	0.617	0.879	0.958	0.913	1.189	1.398
Panel C: Within HistGB deciles — annualized returns							
High−EW	0.0669	0.0293	0.0625	0.0239	0.0527	0.0710	0.0721
EW−Low	0.0639	0.0180	0.0475	0.0465	0.0557	0.0383	0.0587
H–L	0.1308	0.0473	0.1100	0.0704	0.1084	0.1092	0.1308
Panel D: Within HistGB deciles — Sharpe ratios							
High−EW	0.912	0.328	0.754	0.391	0.705	0.980	0.992
EW−Low	1.055	0.278	0.737	0.608	0.846	0.603	0.929
H–L	1.227	0.378	0.940	0.633	0.943	1.000	1.205

Note: This table reports the full conditional-sort decomposition under inference-time image perturbations on the common evaluation sample over 2006–2025. The model is held fixed throughout; perturbations are applied to the rendered input image at evaluation only, with no retraining. In Panels A and B, stocks are first assigned to deciles on the Logit score and then to deciles on the corresponding perturbed Hiera score within each Logit decile. In Panels C and D, the same procedure is applied using HistGB as the conditioning score. High denotes the highest-score decile, Low denotes the lowest-score decile, and EW denotes the equal-weighted market portfolio within the corresponding conditioning decile. High−EW and EW−Low report the long-side and short-side components of the conditional spread, and H–L reports the equal-weighted conditional high-minus-low spread. Panels A and C report annualized close-to-close arithmetic mean returns, and Panels B and D report Sharpe ratios computed from the corresponding rebalance-return series.

Table E2.3: Nested Spanning Fragility under Inference-Time Image Perturbations, 2006–2025

	Unperturbed image	Mask last 5 bars	Horizontal flip	Remove wicks	Mask volume panel	Grayscale image	Mask first 5 bars
Panel A: Fama–MacBeth-style slopes on the perturbed Hiera score							
With Logit + HistGB	0.000791***	0.000371**	0.000385**	0.000613***	0.000634***	0.000787***	0.000749***
With HistGB only	0.000804***	0.000369**	0.000380**	0.000614***	0.000648***	0.000788***	0.000768***
With Logit only	0.001093***	0.000501***	0.000545***	0.000920***	0.000944***	0.001076***	0.001071***
Panel B: Time-series spanning alphas for the perturbed Hiera H–L spread							
On Logit + HistGB	0.001496***	0.000813	0.001487***	0.000266	0.001133***	0.001709***	0.001501***
On HistGB only	0.001625***	0.000856	0.001444***	0.000386	0.001278***	0.001790***	0.001630***
On Logit only	0.002381***	0.001238**	0.002098***	0.001177**	0.001890***	0.002661***	0.002383***

Note: This table reports nested spanning fragility diagnostics for Hiera under inference-time image perturbations on the common evaluation sample over 2006–2025. The model is held fixed throughout; perturbations are applied to the rendered input image at evaluation only, with no retraining. Panel A reports average date-level slopes on the perturbed Hiera score from Fama–MacBeth-style cross-sectional regressions under the indicated benchmark controls. Panel B reports time-series spanning intercepts for the perturbed Hiera H–L spread under the corresponding benchmark controls. In Panel A, inference is based on Newey–West-adjusted standard errors with three lags. In Panel B, the reported intercepts are per rebalance period, are not annualized, and use Newey–West-adjusted standard errors with three lags. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

F Supplementary Results - Translation

This section collects supplementary evidence for the translation exercise. The tables below expand the main-text comparison between Hiera and the translated tabular benchmarks by reporting the full decile structures, tail-hit decompositions, and the residual-content diagnostics after translated feature blocks have been introduced.

F1 Decile Structures

This subsection reports the full decile structures for the translated tabular benchmarks. The tables below separate the cumulative translation ladder, one-block additions to the baseline tabular specification, and one-block omissions from the fully translated specification so that the portfolio consequences of the translated feature blocks can be read at the decile level rather than only through headline spread statistics.

Table F2.2 reports the full decile structure along the cumulative translation ladder, showing how the score-sorted portfolios evolve as translated feature blocks are added sequentially. Table F2.3 then isolates the same objects for one-block additions to the baseline tabular specification.

Where Table F2.3 isolates the stand-alone contribution of each translated feature block relative to Base, Table F2.4 turns to the complementary omission exercise by removing one translated feature block at a time from the fully translated specification.

F2 Tail-Spread Decompositions

This subsection reports the corresponding forecast-space tail-hit evidence for the translated tabular benchmarks. The tables below mirror the same three translation views as above—the cumulative translation, one-block additions, and one-block omissions—but now focus on high-tail and low-tail hit rates and the resulting tail spreads.

Table F2.1 reports the tail-hit decomposition along the cumulative translation ladder,

Table F2.1: Tail-Spread Decomposition for the Cumulative Translation Ladder, 2006–2025

Metric	Base	+ Geometry	+ Wick	+ Direction	+ Volume	+ Path
Panel A: HistGB						
High Hit @ 10%	0.5524	0.5523	0.5523	0.5496	0.5510	0.5492
Low Hit @ 10%	0.5314	0.5291	0.5292	0.5279	0.5278	0.5273
Tail Spread @ 5%	0.0222	0.0323	0.0303	0.0324	0.0310	0.0318
Tail Spread @ 10%	0.0211	0.0231	0.0231	0.0217	0.0232	0.0219
Δ Spread vs. Base @ 10%	0.0000	+0.0021	+0.0020	+0.0007	+0.0021	+0.0009
Tail Spread @ 20%	0.0155	0.0174	0.0183	0.0166	0.0171	0.0181
Newey–West t-stat @ 10%	4.475	5.009	4.896	4.557	4.767	4.662
Panel B: Logit						
High Hit @ 10%	0.5465	0.5486	0.5489	0.5496	0.5489	0.5475
Low Hit @ 10%	0.5319	0.5336	0.5339	0.5316	0.5333	0.5335
Tail Spread @ 5%	0.0164	0.0211	0.0216	0.0223	0.0224	0.0220
Tail Spread @ 10%	0.0147	0.0150	0.0150	0.0181	0.0156	0.0140
Δ Spread vs. Base @ 10%	0.0000	+0.0004	+0.0003	+0.0034	+0.0009	-0.0007
Tail Spread @ 20%	0.0097	0.0103	0.0107	0.0134	0.0132	0.0129
Newey–West t-stat @ 10%	2.804	2.978	2.921	3.457	3.145	2.833

Note: This table reports tail-spread decompositions for the cumulative translation ladder on the common evaluation sample over 2006–2025. Panel A reports HistGB results and Panel B reports Logit results. Base denotes the baseline tabular specification. The columns labeled + Geometry, + Wick, + Direction, + Volume, and + Path add the translated feature blocks sequentially, so each column contains the preceding specification plus the indicated translated feature block. At each rebalance date, stocks are sorted by model score. High Hit @ 10% and Low Hit @ 10% denote the frequencies of positive 5-day-ahead outcomes in the highest-score and lowest-score 10% tails, respectively. Tail Spread @ q% is the mean date-level difference in positive-outcome frequency between the high and low q% tails. Δ Spread vs. Base @ 10% is measured relative to Base within the corresponding model family. The reported Newey–West t-statistics are based on the time series of date-level tail spreads at the 10% cut and use five lags.

Table F2.2: Full Decile Returns (Sharpe Ratios) for the Cumulative Translation Ladder, 2006–2025

Portfolio	Base	+ Geometry	+ Wick	+ Direction	+ Volume	+ Path
			Panel A: HistGB			
Low	0.0541 (0.321)	0.0444 (0.259)	0.0386 (0.231)	0.0443 (0.261)	0.0441 (0.262)	0.0444 (0.258)
2	0.0575 (0.333)	0.0716 (0.405)	0.0688 (0.390)	0.0710 (0.414)	0.0654 (0.375)	0.0552 (0.319)
3	0.0798 (0.449)	0.0728 (0.412)	0.1028 (0.556)	0.0822 (0.448)	0.1006 (0.531)	0.0911 (0.509)
4	0.1068 (0.565)	0.0994 (0.537)	0.1084 (0.566)	0.0913 (0.493)	0.0878 (0.476)	0.0923 (0.494)
5	0.1257 (0.660)	0.1400 (0.718)	0.1137 (0.598)	0.1063 (0.549)	0.1104 (0.575)	0.1434 (0.696)
6	0.1221 (0.666)	0.1006 (0.537)	0.1118 (0.613)	0.1216 (0.647)	0.1242 (0.668)	0.1202 (0.643)
7	0.1478 (0.742)	0.1338 (0.721)	0.1329 (0.705)	0.1448 (0.769)	0.1256 (0.666)	0.1191 (0.650)
8	0.1517 (0.803)	0.1545 (0.817)	0.1330 (0.716)	0.1560 (0.823)	0.1635 (0.863)	0.1426 (0.788)
9	0.1811 (0.782)	0.1978 (0.850)	0.2203 (0.921)	0.2135 (0.901)	0.2034 (0.873)	0.1907 (0.979)
High	0.2298 (1.147)	0.2418 (1.232)	0.2269 (1.169)	0.2261 (1.162)	0.2321 (1.214)	0.2575 (1.095)
High-EW	0.1041 (1.191)	0.1161 (1.425)	0.1011 (1.246)	0.1003 (1.219)	0.1063 (1.341)	0.1318 (0.960)
EW-Low	0.0715 (1.029)	0.0813 (1.213)	0.0871 (1.209)	0.0814 (1.151)	0.0816 (1.145)	0.0812 (1.164)
H–L	0.1757 (1.343)	0.1974 (1.595)	0.1882 (1.514)	0.1817 (1.448)	0.1879 (1.509)	0.2131 (1.247)
			Panel B: Logit			
Low	0.0663 (0.399)	0.0711 (0.412)	0.0724 (0.418)	0.0556 (0.320)	0.0509 (0.306)	0.0464 (0.277)
2	0.0683 (0.390)	0.0702 (0.411)	0.0684 (0.402)	0.0698 (0.400)	0.0734 (0.403)	0.0813 (0.451)
3	0.0955 (0.522)	0.0830 (0.454)	0.0773 (0.423)	0.0619 (0.349)	0.0768 (0.429)	0.0634 (0.354)
4	0.1019 (0.565)	0.1160 (0.592)	0.1042 (0.579)	0.1029 (0.571)	0.0867 (0.479)	0.0991 (0.548)
5	0.1155 (0.591)	0.1050 (0.562)	0.1166 (0.585)	0.1036 (0.561)	0.1064 (0.587)	0.1028 (0.570)
6	0.1366 (0.693)	0.1117 (0.597)	0.1178 (0.625)	0.1381 (0.715)	0.1322 (0.690)	0.1235 (0.637)
7	0.1618 (0.684)	0.1429 (0.752)	0.1493 (0.782)	0.1451 (0.757)	0.1575 (0.749)	0.1670 (0.816)
8	0.1455 (0.764)	0.1752 (0.755)	0.1833 (0.783)	0.1766 (0.732)	0.1540 (0.824)	0.1539 (0.798)
9	0.1587 (0.822)	0.1636 (0.849)	0.1512 (0.791)	0.1827 (0.937)	0.1969 (0.845)	0.2066 (0.885)
High	0.2066 (1.061)	0.2180 (1.121)	0.2161 (1.100)	0.2203 (1.153)	0.2220 (1.168)	0.2128 (1.137)
High-EW	0.0809 (1.001)	0.0923 (1.147)	0.0904 (1.109)	0.0946 (1.269)	0.0963 (1.268)	0.0871 (1.205)
EW-Low	0.0594 (0.714)	0.0546 (0.674)	0.0532 (0.665)	0.0701 (0.872)	0.0748 (1.063)	0.0792 (1.143)
H–L	0.1403 (1.037)	0.1469 (1.112)	0.1436 (1.079)	0.1647 (1.281)	0.1711 (1.401)	0.1663 (1.424)

Note: This table reports the full decile structure for the cumulative translation ladder on the common evaluation sample over 2006–2025. Panel A reports HistGB results and Panel B reports Logit results. Base denotes the baseline tabular specification. The columns labeled + Geometry, + Wick, + Direction, + Volume, and + Path add the translated feature blocks sequentially, so each column contains the preceding specification plus the indicated translated feature block. Entries report annualized close-to-close arithmetic mean returns, with Sharpe ratios shown in parentheses and computed from the corresponding rebalance-return series. At each rebalance date, stocks are sorted into equal-weighted deciles by model score. High denotes the highest-score decile, Low denotes the lowest-score decile, EW denotes the equal-weighted market portfolio that holds all stocks in the evaluation universe, High-EW is the long-side component relative to EW, EW-Low is the short-side component relative to EW, and H–L is the equal-weighted high-minus-low decile spread formed from the same score sorts and rebalanced every five trading days.

Table F2.3: Full Decile Returns (Sharpe Ratios) for One-Block Additions to the Baseline Specification, 2006–2025

Portfolio	Base	Base + Geometry	Base + Wick	Base + Direction	Base + Volume	Base + Path
			Panel A: HistGB			
Low	0.0541 (0.321)	0.0444 (0.259)	0.0395 (0.234)	0.0466 (0.282)	0.0547 (0.313)	0.0348 (0.210)
2	0.0575 (0.333)	0.0716 (0.405)	0.0956 (0.482)	0.0560 (0.319)	0.0753 (0.412)	0.0760 (0.419)
3	0.0798 (0.449)	0.0728 (0.412)	0.0741 (0.420)	0.0879 (0.478)	0.0810 (0.460)	0.0659 (0.369)
4	0.1068 (0.565)	0.0994 (0.537)	0.1053 (0.589)	0.0740 (0.419)	0.0991 (0.542)	0.1048 (0.586)
5	0.1257 (0.660)	0.1400 (0.718)	0.1181 (0.639)	0.0928 (0.511)	0.1047 (0.568)	0.1110 (0.606)
6	0.1221 (0.666)	0.1006 (0.537)	0.1393 (0.615)	0.1345 (0.719)	0.1144 (0.624)	0.1158 (0.616)
7	0.1478 (0.742)	0.1338 (0.721)	0.1206 (0.654)	0.1474 (0.745)	0.1466 (0.761)	0.1321 (0.694)
8	0.1517 (0.803)	0.1545 (0.817)	0.1362 (0.741)	0.1714 (0.836)	0.1695 (0.885)	0.1518 (0.787)
9	0.1811 (0.782)	0.1978 (0.850)	0.1916 (0.981)	0.1941 (0.852)	0.1802 (0.790)	0.1924 (0.948)
High	0.2298 (1.147)	0.2418 (1.232)	0.2364 (1.168)	0.2518 (1.262)	0.2315 (1.142)	0.2721 (1.135)
High-EW	0.1041 (1.191)	0.1161 (1.425)	0.1108 (1.205)	0.1262 (1.443)	0.1058 (1.205)	0.1464 (1.033)
EW-Low	0.0715 (1.029)	0.0813 (1.213)	0.0862 (1.253)	0.0790 (1.119)	0.0710 (0.911)	0.0908 (1.344)
H–L	0.1757 (1.343)	0.1974 (1.595)	0.1970 (1.497)	0.2052 (1.548)	0.1768 (1.270)	0.2373 (1.325)
			Panel B: Logit			
Low	0.0663 (0.399)	0.0711 (0.412)	0.0660 (0.386)	0.0543 (0.325)	0.0659 (0.400)	0.0627 (0.376)
2	0.0683 (0.390)	0.0702 (0.411)	0.0768 (0.438)	0.0608 (0.346)	0.0842 (0.467)	0.0701 (0.397)
3	0.0955 (0.522)	0.0830 (0.454)	0.0755 (0.422)	0.0762 (0.422)	0.0938 (0.511)	0.0839 (0.455)
4	0.1019 (0.565)	0.1160 (0.592)	0.1039 (0.585)	0.1062 (0.586)	0.0976 (0.508)	0.0865 (0.474)
5	0.1155 (0.591)	0.1050 (0.562)	0.1365 (0.667)	0.1100 (0.567)	0.1110 (0.602)	0.1352 (0.679)
6	0.1366 (0.693)	0.1117 (0.597)	0.1274 (0.669)	0.1189 (0.636)	0.1195 (0.642)	0.1307 (0.693)
7	0.1618 (0.684)	0.1429 (0.752)	0.1625 (0.705)	0.1417 (0.737)	0.1196 (0.632)	0.1410 (0.742)
8	0.1455 (0.764)	0.1752 (0.755)	0.1438 (0.761)	0.1671 (0.732)	0.1617 (0.815)	0.1510 (0.781)
9	0.1587 (0.822)	0.1636 (0.849)	0.1625 (0.838)	0.2017 (0.968)	0.1972 (0.849)	0.1806 (0.769)
High	0.2066 (1.061)	0.2180 (1.121)	0.2020 (1.046)	0.2196 (1.133)	0.2064 (1.084)	0.2152 (1.141)
High-EW	0.0809 (1.001)	0.0923 (1.147)	0.0763 (0.957)	0.0940 (1.152)	0.0807 (1.048)	0.0895 (1.135)
EW-Low	0.0594 (0.714)	0.0546 (0.674)	0.0597 (0.726)	0.0713 (0.846)	0.0598 (0.778)	0.0630 (0.757)
H–L	0.1403 (1.037)	0.1469 (1.112)	0.1361 (1.010)	0.1653 (1.197)	0.1405 (1.120)	0.1525 (1.173)

Note: This table reports the full decile structure for specifications that add one translated feature block directly to the baseline tabular specification on the common evaluation sample over 2006–2025. Panel A reports HistGB results and Panel B reports Logit results. Base denotes the baseline tabular specification. The columns labeled Base + Geometry, Base + Wick, Base + Direction, Base + Volume, and Base + Path add the indicated translated feature block directly to Base, one block at a time. Entries report annualized close-to-close arithmetic mean returns, with Sharpe ratios shown in parentheses and computed from the corresponding rebalance-return series. At each rebalance date, stocks are sorted into equal-weighted deciles by model score. High denotes the highest-score decile, Low denotes the lowest-score decile, EW denotes the equal-weighted market portfolio that holds all stocks in the evaluation universe, High-EW is the long-side component relative to EW, EW-Low is the short-side component relative to EW, and H–L is the equal-weighted high-minus-low decile spread formed from the same score sorts and rebalanced every five trading days.

Table F2.4: Full Decile Returns (Sharpe Ratios) for One-Block Omissions from the Fully Translated Specification, 2006–2025

Portfolio	Full	Full − Geometry	Full − Wick	Full − Direction	Full − Volume	Full − Path
			Panel A: HistGB			
Low	0.0444 (0.258)	0.0247 (0.146)	0.0431 (0.252)	0.0580 (0.337)	0.0345 (0.205)	0.0441 (0.262)
2	0.0552 (0.319)	0.0670 (0.377)	0.0755 (0.425)	0.0585 (0.338)	0.0750 (0.420)	0.0654 (0.375)
3	0.0911 (0.509)	0.0704 (0.394)	0.0730 (0.408)	0.0785 (0.434)	0.0718 (0.398)	0.1006 (0.531)
4	0.0923 (0.494)	0.1188 (0.629)	0.1044 (0.562)	0.1029 (0.542)	0.0949 (0.529)	0.0878 (0.476)
5	0.1434 (0.696)	0.1251 (0.659)	0.1069 (0.593)	0.1262 (0.645)	0.1168 (0.608)	0.1104 (0.575)
6	0.1202 (0.643)	0.1225 (0.657)	0.1312 (0.674)	0.1310 (0.701)	0.1495 (0.754)	0.1242 (0.668)
7	0.1191 (0.650)	0.1303 (0.688)	0.1396 (0.722)	0.1204 (0.642)	0.1209 (0.665)	0.1256 (0.666)
8	0.1426 (0.788)	0.1540 (0.803)	0.1415 (0.748)	0.1538 (0.807)	0.1390 (0.739)	0.1635 (0.863)
9	0.1907 (0.979)	0.2054 (0.897)	0.1936 (0.997)	0.1646 (0.864)	0.2117 (0.889)	0.2034 (0.873)
High	0.2575 (1.095)	0.2384 (1.224)	0.2484 (1.063)	0.2627 (1.141)	0.2427 (1.275)	0.2321 (1.214)
High-EW	0.1318 (0.960)	0.1128 (1.377)	0.1226 (0.915)	0.1371 (1.030)	0.1171 (1.509)	0.1063 (1.341)
EW-Low	0.0812 (1.164)	0.1009 (1.481)	0.0826 (1.194)	0.0677 (0.983)	0.0911 (1.303)	0.0816 (1.145)
H–L	0.2131 (1.247)	0.2137 (1.744)	0.2052 (1.229)	0.2047 (1.236)	0.2082 (1.729)	0.1879 (1.509)
			Panel B: Logit			
Low	0.0464 (0.277)	0.0426 (0.254)	0.0461 (0.276)	0.0559 (0.337)	0.0479 (0.276)	0.0509 (0.306)
2	0.0813 (0.451)	0.0788 (0.438)	0.0833 (0.459)	0.0756 (0.420)	0.0653 (0.370)	0.0734 (0.403)
3	0.0634 (0.354)	0.0744 (0.419)	0.0601 (0.336)	0.0836 (0.441)	0.0663 (0.375)	0.0768 (0.429)
4	0.0991 (0.548)	0.1069 (0.596)	0.0886 (0.492)	0.0976 (0.530)	0.1010 (0.567)	0.0867 (0.479)
5	0.1028 (0.570)	0.0998 (0.533)	0.1214 (0.632)	0.1042 (0.561)	0.1361 (0.708)	0.1064 (0.587)
6	0.1235 (0.637)	0.1209 (0.644)	0.1453 (0.756)	0.1221 (0.655)	0.1227 (0.621)	0.1322 (0.690)
7	0.1670 (0.816)	0.1677 (0.804)	0.1490 (0.751)	0.1417 (0.735)	0.1343 (0.709)	0.1575 (0.749)
8	0.1539 (0.798)	0.1489 (0.770)	0.1439 (0.749)	0.1584 (0.820)	0.1771 (0.895)	0.1540 (0.824)
9	0.2066 (0.885)	0.2069 (0.893)	0.2065 (0.893)	0.1628 (0.846)	0.1863 (0.805)	0.1969 (0.845)
High	0.2128 (1.137)	0.2102 (1.122)	0.2132 (1.137)	0.2551 (1.095)	0.2194 (1.157)	0.2220 (1.168)
High-EW	0.0871 (1.205)	0.0845 (1.166)	0.0875 (1.202)	0.1294 (0.957)	0.0938 (1.299)	0.0963 (1.268)
EW-Low	0.0792 (1.143)	0.0832 (1.197)	0.0796 (1.131)	0.0697 (1.002)	0.0778 (0.965)	0.0748 (1.063)
H–L	0.1663 (1.424)	0.1676 (1.413)	0.1671 (1.419)	0.1991 (1.181)	0.1715 (1.360)	0.1711 (1.401)

Note: This table reports the full decile structure for one-block omissions from the fully translated specification on the common evaluation sample over 2006–2025. Panel A reports HistGB results and Panel B reports Logit results. Full denotes the translated specification that includes all translated feature blocks. The columns labeled Full − Geometry, Full − Wick, Full − Direction, Full − Volume, and Full − Path remove the indicated translated feature block from Full, one block at a time. Full − Path coincides with the specification that includes Geometry, Wick, Direction, and Volume but excludes Path. Entries report annualized close-to-close arithmetic mean returns, with Sharpe ratios shown in parentheses and computed from the corresponding rebalance-return series. At each rebalance date, stocks are sorted into equal-weighted deciles by model score. High denotes the highest-score decile, Low denotes the lowest-score decile, EW denotes the equal-weighted market portfolio that holds all stocks in the evaluation universe, High-EW is the long-side component relative to EW, EW-Low is the short-side component relative to EW, and H–L is the equal-weighted high-minus-low decile spread formed from the same score sorts and rebalanced every five trading days.

which shows how forecast-space discrimination changes as translated feature blocks are added sequentially. Table F3.1 then isolates the same tail-spread objects for one-block additions relative to Base.

Table F3.1 is informative about stand-alone block additions, but it does not show whether a given translated feature block remains important once the others are already present. Table F2.5 addresses that omission margin by removing one translated feature block at a time from the fully translated specification.

Table F2.5: Tail-Spread Decomposition for One-Block Omissions from the Fully Translated Specification, 2006–2025

Metric	Full	Full − Geometry	Full − Wick	Full − Direction	Full − Volume	Full − Path
			Panel A: HistGB			
High Hit @ 10%	0.5492	0.5520	0.5495	0.5556	0.5524	0.5510
Low Hit @ 10%	0.5273	0.5274	0.5281	0.5281	0.5268	0.5278
Tail Spread @ 5%	0.0318	0.0330	0.0319	0.0320	0.0311	0.0310
Tail Spread @ 10%	0.0219	0.0246	0.0214	0.0275	0.0255	0.0232
Δ Spread vs. Base @ 10%	+0.0009	+0.0035	+0.0004	+0.0064	+0.0045	+0.0021
Tail Spread @ 20%	0.0181	0.0173	0.0175	0.0178	0.0172	0.0171
Newey–West t-stat @ 10%	4.662	5.227	4.543	5.858	5.330	4.767
			Panel B: Logit			
High Hit @ 10%	0.5475	0.5478	0.5474	0.5484	0.5505	0.5489
Low Hit @ 10%	0.5335	0.5315	0.5328	0.5333	0.5308	0.5333
Tail Spread @ 5%	0.0220	0.0232	0.0222	0.0222	0.0236	0.0224
Tail Spread @ 10%	0.0140	0.0163	0.0146	0.0151	0.0197	0.0156
Δ Spread vs. Base @ 10%	-0.0007	+0.0016	-0.0001	+0.0005	+0.0050	+0.0009
Tail Spread @ 20%	0.0129	0.0127	0.0126	0.0120	0.0146	0.0132
Newey–West t-stat @ 10%	2.833	3.281	2.947	3.026	3.844	3.145

Note: This table reports tail-spread decompositions for one-block omissions from the fully translated specification on the common evaluation sample over 2006–2025. Panel A reports HistGB results and Panel B reports Logit results. Full denotes the translated specification that includes all translated feature blocks. The columns labeled Full − Geometry, Full − Wick, Full − Direction, Full − Volume, and Full − Path remove the indicated translated feature block from Full, one block at a time. Full − Path coincides with the specification that includes Geometry, Wick, Direction, and Volume but excludes Path. At each rebalance date, stocks are sorted by model score. High Hit @ 10% and Low Hit @ 10% denote the frequencies of positive 5-day-ahead outcomes in the highest-score and lowest-score 10% tails, respectively. Tail Spread @ q% is the mean date-level difference in positive-outcome frequency between the high and low q% tails. Δ Spread vs. Base @ 10% is measured relative to the baseline tabular specification within the corresponding model family, not relative to Full. The reported Newey–West t-statistics are based on the time series of date-level tail spreads at the 10% cut and use five lags.

F3 Residual Content after Translation

This subsection reports the reverse-direction residual-content diagnostics after translation, asking whether Hiera continues to retain distinct information once the translated tabular benchmarks are taken into account. The resulting evidence complements the main-text translation analysis by showing whether narrowing the unconditional gap to Hiera is

Table F3.1: Tail-Spread Decomposition for One-Block Additions to the Baseline Specification, 2006–2025

Metric	Base	Base + Geometry	Base + Wick	Base + Direction	Base + Volume	Base + Path
			Panel A: HistGB			
High Hit @ 10%	0.5524	0.5523	0.5509	0.5549	0.5503	0.5511
Low Hit @ 10%	0.5314	0.5291	0.5265	0.5289	0.5290	0.5285
Tail Spread @ 5%	0.0222	0.0323	0.0338	0.0309	0.0271	0.0254
Tail Spread @ 10%	0.0211	0.0231	0.0244	0.0260	0.0213	0.0226
Δ Spread vs. Base @ 10%	0.0000	+0.0021	+0.0033	+0.0049	+0.0002	+0.0015
Tail Spread @ 20%	0.0155	0.0174	0.0184	0.0194	0.0153	0.0175
Newey–West t-stat @ 10%	4.475	5.009	4.875	5.506	4.566	4.594
			Panel B: Logit			
High Hit @ 10%	0.5465	0.5486	0.5465	0.5465	0.5468	0.5492
Low Hit @ 10%	0.5319	0.5336	0.5333	0.5303	0.5322	0.5328
Tail Spread @ 5%	0.0164	0.0211	0.0145	0.0201	0.0198	0.0245
Tail Spread @ 10%	0.0147	0.0150	0.0132	0.0162	0.0147	0.0164
Δ Spread vs. Base @ 10%	0.0000	+0.0004	-0.0015	+0.0015	-0.0000	+0.0017
Tail Spread @ 20%	0.0097	0.0103	0.0080	0.0127	0.0083	0.0124
Newey–West t-stat @ 10%	2.804	2.978	2.513	3.120	3.050	3.139

Note: This table reports tail-spread decompositions for specifications that add one translated feature block directly to the baseline tabular specification on the common evaluation sample over 2006–2025. Panel A reports HistGB results and Panel B reports Logit results. Base denotes the baseline tabular specification. The columns labeled Base + Geometry, Base + Wick, Base + Direction, Base + Volume, and Base + Path add the indicated translated feature block directly to Base, one block at a time. At each rebalance date, stocks are sorted by model score. High Hit @ 10% and Low Hit @ 10% denote the frequencies of positive 5-day-ahead outcomes in the highest-score and lowest-score 10% tails, respectively. Tail Spread @ q% is the mean date-level difference in positive-outcome frequency between the high and low q% tails. Δ Spread vs. Base @ 10% is measured relative to Base within the corresponding model family. The reported Newey–West t-statistics are based on the time series of date-level tail spreads at the 10% cut and use five lags.

accompanied by subsumption or instead by continued partial overlap between the image and translated tabular representations.

Table F3.2: Residual Content of Hiera after Controlling for the Translated Tabular Benchmarks: Full Specification Ladder, 2006–2025

		Cross-sectional regression	Portfolio spanning regression	Conditional sorts: annualized H–L return	
Family	Specification	FM $\beta_{\text{Hiera}\|\text{Model}}$	TS $\alpha_{\text{Hiera}\|\text{Model}}$	Hiera within model	Model within Hiera
		Panel A: Cumulative Translation			
HistGB	Base	0.000803***	0.001625***	0.1909	0.0768
	+ Geometry	0.000745***	0.001217***	0.1361	0.1570
	+ Wick	0.000800***	0.001199***	0.1271	0.1094
	+ Direction	0.000736***	0.001396***	0.1219	0.0404
	+ Volume	0.000771***	0.001427***	0.1348	0.1350
	+ Path	0.000748***	0.002672***	0.1625	0.1343
Logit	Base	0.001092***	0.002381***	0.1287	0.0790
	+ Geometry	0.001071***	0.002194***	0.2570	0.0372
	+ Wick	0.001066***	0.002258***	0.1577	0.0488
	+ Direction	0.000979***	0.001986***	0.1722	0.0873
	+ Volume	0.000962***	0.001624***	0.1636	0.0957
	+ Path	0.000947***	0.001761***	0.1233	0.1033
		Panel B: One Block Additions			
HistGB	Base	0.000803***	0.001625***	0.1909	0.0768
	Base + Geometry	0.000745***	0.001217***	0.1361	0.1570
	Base + Wick	0.000877***	0.001369***	0.1442	0.0493
	Base + Direction	0.000651***	0.001146**	0.1017	0.1546
	Base + Volume	0.000898***	0.001596***	0.1252	-0.0368
	Base + Path	0.000637***	0.002405***	0.2187	0.1570
Logit	Base	0.001092***	0.002381***	0.1287	0.0790
	Base + Geometry	0.001071***	0.002194***	0.2570	0.0372
	Base + Wick	0.001092***	0.002421***	0.1540	0.0086
	Base + Direction	0.000994***	0.002071***	0.1509	0.0870
	Base + Volume	0.001057***	0.002222***	0.1133	0.0483
	Base + Path	0.001052***	0.002196***	0.1394	0.0821

(Continued on next page)

Table F3.2: Residual Content of Hiera after Controlling for the Translated Tabular Benchmarks: Full Specification Ladder (Continued)

		Cross-sectional regression	Portfolio spanning regression	Conditional sorts: annualized H–L return	
Family	Specification	FM $\beta_{\text{Hiera}\|\text{Model}}$	TS $\alpha_{\text{Hiera}\|\text{Model}}$	Hiera within model	Model within Hiera
		Panel C: One Block Omissions			
HistGB	Full	0.000748^{***}	0.002672^{***}	0.1625	0.1343
	Full − Geometry	0.000717^{***}	0.001125^{**}	0.1332	0.1548
	Full − Wick	0.000761^{***}	0.002674^{***}	0.1951	0.1777
	Full − Direction	0.000790^{***}	0.002671^{***}	0.1609	0.1327
	Full − Volume	0.000710^{***}	0.000897^{**}	0.1581	0.1429
	Full − Path	0.000771^{***}	0.001427^{***}	0.1348	0.1350
Logit	Full	0.000947^{***}	0.001761^{***}	0.1233	0.1033
	Full − Geometry	0.000966^{***}	0.001736^{***}	0.1661	0.1215
	Full − Wick	0.000956^{***}	0.001736^{***}	0.1354	0.0806
	Full − Direction	0.000984^{***}	0.002666^{***}	0.1304	0.1367
	Full − Volume	0.000966^{***}	0.002018^{***}	0.1491	0.0970
	Full − Path	0.000962^{***}	0.001624^{***}	0.1636	0.0957

Note: This table reports residual-content tests for Hiera after controlling for the translated tabular benchmarks on the common evaluation sample over 2006–2025. Panel A reports cumulative translation results, Panel B reports one-block additions to the baseline tabular specification, and Panel C reports one-block omissions from the fully translated specification. FM $\beta_{\text{Hiera}|\text{Model}}$ is the average date-level slope on the Hiera score from a Fama–MacBeth-style cross-sectional regression that includes both the Hiera score and the translated tabular benchmark score. TS $\alpha_{\text{Hiera}|\text{Model}}$ is the intercept from a time-series regression of the Hiera H–L spread return on the translated tabular benchmark H–L spread return. This intercept is reported per rebalance period and is not annualized. The conditional-sort columns report annualized equal-weighted H–L returns only. In the Hiera within model column, stocks are first assigned to deciles on the translated tabular benchmark score and then to deciles on the Hiera score within each translated-benchmark decile. In the model within Hiera column, the sorting order is reversed. Regression-based significance tests use Newey–West-adjusted inference with three lags. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

F4 Murray-Style Cumulative-Path Benchmark

This subsection reports a focused robustness check based on the cumulative-path methodology in Murray et al. (2024). The purpose is not to expand the benchmark horse race with another full benchmark family. Instead, the exercise asks whether the gains from the translated tabular specifications can be reduced to a generic path-based summary drawn from adjacent chart-based work. To keep the comparison within the common-information design, the Murray-20 (M20) block is constructed from the same 20-trading-day history used to define the image and baseline tabular representations. For each stock-date observation, let day 1 denote the oldest day in the 20-bar window and day $k \in \{1, \ldots, 20\}$ the kth day in that same window. The M20 block is then

$$M_{i,t}^{\mathrm{M20}} = \left(\frac{C_{i,t-19}}{O_{i,t-19}} - 1,\ \frac{C_{i,t-18}}{O_{i,t-19}} - 1,\ \ldots,\ \frac{C_{i,t}}{O_{i,t-19}} - 1 \right),$$

so each element records the cumulative return from the first opening price in the window to the closing price on a later day within that same window. The resulting M20 features are estimated on the same selected training rows as the main tabular benchmarks and evaluated on the same out-of-sample panel used in the post-translation comparison, so the exercise changes only the path representation rather than the underlying information set or evaluation sample.

In addition to the estimated M20-based models, we also construct a model-agnostic M20 signal that does not rely on fitted model weights. For each rebalance date, each of the twenty M20 coordinates is standardized cross-sectionally across stocks, and the resulting standardized coordinates are then averaged with equal weights within stock-date. Formally, if $M_{i,t,k}^{\mathrm{M20}}$ denotes coordinate k of the M20 block and

$$Z_{i,t,k}^{\mathrm{M20}} = \frac{M_{i,t,k}^{\mathrm{M20}} - \bar{M}_{t,k}^{\mathrm{M20}}}{\mathrm{sd}\left(M_{\cdot,t,k}^{\mathrm{M20}}\right)},$$

then the model-agnostic M20 signal is

$$S_{i,t}^{\mathrm{M20}} = \frac{1}{20} \sum_{k=1}^{20} Z_{i,t,k}^{\mathrm{M20}}.$$

This construction yields a simple path composite that summarizes the cross-sectional position of each stock's cumulative return path within the same 20-day window, while imposing no estimated weighting scheme on the underlying coordinates.

Table F4.1 reports two sets of results. Panel A compares the incremental addition of the M20 block to the baseline tabular representation with the corresponding fully translated specification in each model family. Panel B reports residual-content tests for Hiera and the translated tabular specifications after conditioning on the model-agnostic M20 signal.

Table F4.1: External Murray-Style Cumulative-Path Robustness Check, 2006–2025

Family	Specification	Pearson IC	H–L Return	H–L Sharpe
Panel A: Incremental comparison within the common-information design				
HistGB	Base	0.0206	0.1757	1.343
	Base + M20	0.0239	0.1768	1.209
	Full	0.0225	0.2131	1.247
Logit	Base	0.0144	0.1403	1.037
	Base + M20	0.0177	0.1699	0.977
	Full	0.0204	0.1663	1.424
Hiera		0.0239	0.2067	1.496
Panel B: Residual content after conditioning on the model-agnostic M20 signal				
Model		FM $\beta_{\mathrm{Model\|M20}}$	TS $\alpha_{\mathrm{Model\|M20}}$	TS R^2
Hiera		0.0010***	0.0035***	0.1169
HistGB Base		0.0009***	0.0030***	0.1014
HistGB Full		0.0010***	0.0037***	0.0507
Logit Full		0.0009***	0.0029***	0.0576

Note: This table reports a focused robustness check against a Murray-style cumulative-path representation over 2006–2025 on the common evaluation sample. The M20 block consists of twenty cumulative-return coordinates from the same 20-day input window used elsewhere in the analysis, where coordinate k equals the cumulative return from the opening price on the first day of the window to the closing price on day k of that same window. To preserve alignment with the image dataset, the path block is constructed from the same history used to define the image sample, and the M20-based models are trained on the same selected training rows and evaluated on the same out-of-sample panel as the main post-translation comparison. Panel A compares the incremental addition of the M20 block to the baseline tabular representation with the corresponding fully translated specification in each model family. Panel B reports residual-content tests after conditioning on a model-agnostic M20 signal, defined as the equal-weighted average of the date-standardized M20 coordinates. Pearson IC is the mean date-level cross-sectional Pearson correlation between model scores and realized five-day-ahead returns. H–L Return and H–L Sharpe are the annualized return and Sharpe ratio of the equal-weighted high-minus-low decile spread rebalanced every five trading days. FM $\beta_{\mathrm{Model|M20}}$ is the average date-level slope on the indicated score from a Fama–MacBeth-style cross-sectional regression that includes both the indicated score and the model-agnostic M20 signal. TS $\alpha_{\mathrm{Model|M20}}$ is the intercept from a time-series regression of the indicated model's H–L spread return on the corresponding H–L spread formed from that same model-agnostic M20 signal. This intercept is reported per rebalance period and is not annualized. TS R^2 is the corresponding regression R^2. Regression-based significance tests use Newey–West-adjusted inference with three lags. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

F5 What does the encoding-loss residual covary with?

The variance-share decomposition in Table 3.12 attributes about 22% of Hiera's cross-sectional score variance to an encoding-loss share $\epsilon_{\mathcal{G}}$ that the structured tabular feature space does not reproduce, and the perturbation analysis localizes this share to recent-bar geometry, intrabar extrema, and temporal order. A natural further question is whether the residual that drives $\epsilon_{\mathcal{G}}$ at the per-observation level covaries with standard cross-sectional liquidity or volatility characteristics. If it does, the residual would be a noisier proxy for an existing characteristic, sharpening the interpretation. If it does not, the residual is identified at the cross-sectional level as a different kind of information.

We compute the per-observation residual of Hiera's score after projecting on the distilled-HistGB student score within each rebalance date, and regress this residual cross-sectionally on three observation-level characteristics constructed from the same CRSP daily OHLCV data the renderer ingests: Amihud illiquidity (the 20-day mean of |daily return|/dollar volume, scaled by 10^6), Parkinson volatility (the 20-day Parkinson high-low range estimator), and idiosyncratic volatility (the 60-day rolling standard deviation of close-to-close returns net of the equal-weighted cross-section). Each characteristic is standardized within each rebalance date so the reported coefficients are comparable in size. The residual has a cross-sectional standard deviation of about 0.011 on the Hiera probability scale. Table F5.1 reports the Fama–MacBeth average date-level slopes with Newey–West t-statistics at three lags.

Table F5.1: Cross-sectional characteristics of the Hiera-on-distilled residual, 2006–2025

Characteristic	FM coefficient	NW(3) t-stat	p
Amihud illiquidity (z)	−0.000030	−1.49	0.137
Parkinson 20-day vol (z)	−0.000002	−0.05	0.957
Idiosyncratic vol, 60-day (z)	−0.000029	−0.84	0.401

Note: Fama–MacBeth panel regression of the per-observation residual of Hiera's score after a within-date OLS projection on the distilled HistGB student score, against three standardized stock-level characteristics computed from CRSP daily OHLCV on the closure panel ($n_{\text{complete}} = 499{,}152$ stock-date observations across 1,006 rebalance dates over 2006–2025). The residual is on Hiera's probability scale (standard deviation across observations ≈ 0.011). Coefficients are the time-series mean of date-level cross-sectional OLS slopes; t-statistics use Newey–West-adjusted inference at three lags. Amihud illiquidity and Parkinson volatility are measured on the renderer's 20-trading-day input window; idiosyncratic volatility uses a longer 60-day estimation horizon. The test isolates whether the encoding-loss residual is a noisier counterpart of standard liquidity or volatility proxies.

The three coefficients are economically and statistically indistinguishable from zero. The

residual is not a noisier version of Amihud illiquidity, Parkinson volatility, or idiosyncratic volatility on this panel. We read this as a clean null finding rather than a structural mechanism statement: it disciplines the interpretation by ruling out the most obvious microstructure proxies, which would otherwise be a leading candidate to subsume the chart-encoding component. The residual is identified by the perturbation analysis to recent-bar geometry, intrabar extrema, and temporal order, and is not collinear with broad liquidity or volatility characteristics at the cross-section.

G Supplementary Results - Ensemble

This section reports additional diagnostics for the 50/50 probability-average ensemble of Hiera and the fully translated HistGB model. The purpose is to summarize how the ensemble compares with its two underlying components in forecast space, in portfolio space, and in residual-content tests based on score overlap and spanning. Table G0.1 reports the full headline comparison across the three models. Table G0.2 reports score overlap and two-score cross-sectional spanning diagnostics for the ensemble relative to each component model. Table G0.3 reports the corresponding leg-level time-series spanning results for the long side, the short side, and the overall spread.

Table G0.1: Headline Diagnostics for the Hiera–Translated HistGB Ensemble, 2006–2025

Model	Log Loss	Brier	AUC	Pearson IC	Spearman IC	Tail Spread @ 10%	H–L Return	H–L Sharpe
Ensemble (50/50)	0.691	0.249	0.515	0.026	0.027	0.0268	0.2540	1.398
Hiera	0.690	0.248	0.515	0.024	0.026	0.0263	0.2067	1.496
HistGB (Full)	0.694	0.251	0.513	0.023	0.023	0.0220	0.2131	1.247

Note: This table reports headline forecast-space and portfolio-space diagnostics for the 50/50 probability-average ensemble of Hiera and the fully translated HistGB model on the common aligned out-of-sample panel over 2006–2025. The ensemble score is the equal-weighted average of the two model probabilities. Pearson IC and Spearman IC are the mean date-level cross-sectional Pearson and rank correlations, respectively, between model scores and realized 5-day-ahead returns. Tail Spread @ 10% is the mean date-level difference in the frequency of positive 5-day-ahead outcomes between the top and bottom score deciles. H–L Return and H–L Sharpe are the annualized return and Sharpe ratio of the equal-weighted high-minus-low decile spread rebalanced every five trading days.

Table G0.2: Score Overlap and Cross-Sectional Spanning for the Hiera–Translated HistGB Ensemble, 2006–2025

Comparison	Pool. ρ_P	Pool. ρ_S	Score R^2	Date-avg. ρ_P	Date-avg. ρ_S	Date-avg. R^2	β_{Ens} (t)	β_{Comp} (t)
Ensemble vs. Hiera	0.867	0.838	0.751	0.830	0.801	0.692	0.001 (4.44)	0.000 (1.55)
Ensemble vs. HistGB (Full)	0.953	0.953	0.907	0.941	0.939	0.887	0.002 (4.90)	-0.000 (-1.56)

Note: This table reports score-overlap and two-score cross-sectional spanning diagnostics for the ensemble relative to Hiera and the fully translated HistGB model on the common aligned out-of-sample panel over 2006–2025. Pool. ρ_P and Pool. ρ_S are pooled Pearson and Spearman score correlations computed across all stock-date observations in the aligned panel. Score R^2 is the corresponding pooled score R^2. Date-avg. ρ_P, Date-avg. ρ_S, and Date-avg. R^2 are computed within rebalance date and then averaged across rebalance dates. β_{Ens} and β_{Comp} are the average date-level coefficients from Fama–MacBeth-style cross-sectional regressions that include both the ensemble score and the comparator score; reported t-statistics use Newey–West-adjusted inference with three lags. Comp denotes the non-ensemble model named in the comparison label.

Table G0.3: Leg-Level Portfolio Spanning for the Hiera–Translated HistGB Ensemble, 2006–2025

Comparison	Portfolio	$\alpha_{\text{Ens}\|\text{Comp}}$	t-stat	p-value	$\alpha_{\text{Comp}\|\text{Ens}}$	t-stat	p-value
Ensemble vs. Hiera	High–EW	0.0021	2.18	0.029	0.0015	3.01	0.003
	EW–Low	0.0004	2.12	0.034	0.0004	2.07	0.039
	H–L	0.0019	2.43	0.015	0.0019	2.08	0.037
Ensemble vs. HistGB (Full)	High–EW	0.0005	2.67	0.007	-0.0002	-0.92	0.360
	EW–Low	0.0005	3.10	0.002	-0.0000	-0.25	0.806
	H–L	0.0008	3.46	0.001	-0.0002	-0.64	0.523

Note: This table reports leg-level time-series spanning diagnostics for the ensemble relative to Hiera and the fully translated HistGB model on the common aligned out-of-sample panel over 2006–2025. High–EW denotes the annualized return component of the highest-score decile relative to the equal-weighted market portfolio, EW–Low denotes the corresponding market-relative short-side component, and H–L denotes the overall equal-weighted high-minus-low decile spread. $\alpha_{\text{Ens}|\text{Comp}}$ is the intercept from a time-series regression of the ensemble leg return on the corresponding comparator leg return, and $\alpha_{\text{Comp}|\text{Ens}}$ is defined analogously in the reverse direction. All reported alphas are per rebalance period, are not annualized, and use Newey–West-adjusted inference with three lags. Comp denotes the non-ensemble model named in the comparison label.

H Robustness Exercises

This section collects three robustness exercises that complement the main-text results: a transaction-cost grid applied to the headline spread returns, a pre/post-2020 subperiod split with a structural-break test, and a multi-horizon re-evaluation of the existing models at one-day and twenty-day holding periods. Each exercise is computed on the common out-of-sample panel and is presented as a robustness check on the main-text results.

H1 Transaction Costs

The headline portfolio results are gross of trading costs. This subsection assesses whether the cross-model ordering is preserved once a reasonable cost wedge is imposed. For each model m and rebalance date t, let $r^{HL}_{m,t}$ denote the gross H–L spread return and $\tau_{m,t}$ the one-way turnover of the spread portfolio. For a one-way cost c in basis points, the net spread return is

$$r^{HL,\text{net}}_{m,t}(c) \;=\; r^{HL}_{m,t} \;-\; 2\,(c/10000)\,\tau_{m,t},$$

where the factor of two reflects that the H–L spread incurs costs on both legs. Annualized net returns and Sharpe ratios are computed from $\{r^{HL,\text{net}}_{m,t}(c)\}_t$ over the 1,006 rebalance dates in 2006–2025. We report a standard one-way cost grid covering $c \in \{5, 10, 20, 40\}$ basis points. The 20-basis-point regime exceeds the typical effective spread for top-500 names in standard microstructure references, and the 40-basis-point regime is included as a stress test rather than a realistic central case. A stock-level Corwin–Schultz adjustment would require per-stock leg-entry decompositions of turnover, which lie outside the scope of the portfolio decomposition used here; the standard grid is conservative relative to that adjustment and is used throughout.

The baseline comparison is robust to transaction costs: Hiera's Sharpe ratio exceeds that of HistGB Base at every point in the cost grid. Relative to HistGB + Geometry, however, the ordering changes with the cost assumption. HistGB + Geometry leads gross of costs, with a Sharpe ratio of 1.595 versus 1.496 for Hiera, and retains a smaller lead at 5 basis

Table H1.1: H–L Spread Returns and Sharpe Ratios Net of Transaction Costs, 2006–2025

	Gross	5 bps	10 bps	20 bps	40 bps	Turnover
Panel A: Annualized H–L Return						
Hiera	0.2068	0.1656	0.1243	0.0418	−0.1231	0.818
HistGB Base	0.1757	0.1326	0.0895	0.0033	−0.1692	0.855
HistGB + Geometry	0.1974	0.1536	0.1099	0.0224	−0.1526	0.868
HistGB Full	0.2132	0.1695	0.1259	0.0386	−0.1360	0.866
Logit Base	0.1408	0.0970	0.0531	−0.0345	−0.2099	0.870
Logit Full	0.1663	0.1224	0.0784	−0.0095	−0.1853	0.872
HistGB distilled	0.2102	0.1665	0.1229	0.0356	−0.1390	0.866
Panel B: H–L Sharpe Ratio						
Hiera	1.496	1.198	0.900	0.303	−0.891	
HistGB Base	1.343	1.013	0.684	0.025	−1.293	
HistGB + Geometry	1.595	1.242	0.888	0.182	−1.232	
HistGB Full	1.247	0.992	0.736	0.226	−0.795	
Logit Base	1.040	0.716	0.392	−0.255	−1.549	
Logit Full	1.424	1.048	0.672	−0.080	−1.584	
HistGB distilled	1.445	1.145	0.845	0.244	−0.955	

Note: Annualized H–L spread returns (Panel A) and Sharpe ratios (Panel B) for the image model, the baseline and translated tabular benchmarks, and the HistGB-distilled model over 2006–2025 on the common evaluation sample. One-way transaction costs of 5, 10, 20, and 40 basis points are applied to spread turnover with a factor of two to reflect costs on both legs. Turnover reports the average one-way turnover per rebalance for the H–L spread. HistGB + Geometry is the strongest translated HistGB specification by Sharpe (Sharpe 1.595 versus 1.247 for HistGB Full and 1.343 for HistGB Base, see Table 3.6). HistGB Full is the cumulative all-block specification and is included for direct comparison with the distilled model, which is trained on the full translated predictor set. All values are computed from the per-rebalance H–L return series and aggregate turnover series underlying the main-text portfolio diagnostics; the gross-of-cost figures reconcile to within rounding of the headline numbers in Table 3.1 and Table 3.6.

points, with Sharpe ratios of 1.242 and 1.198, respectively. At 10 basis points, Hiera moves ahead, with a Sharpe ratio of 0.900 versus 0.888, and its relative position improves further at higher costs. Hiera also retains its gross Sharpe advantage over the distilled model, and that advantage increases slightly as the cost wedge rises. These changes reflect Hiera's lower one-way turnover, 0.818, compared with 0.868 for HistGB + Geometry and 0.866 for the distilled model. At 20 basis points, HistGB Base is close to zero and Logit Base is negative; at 40 basis points, all reported models have negative annualized returns and Sharpe ratios.

H2 Subperiod Stability

To assess whether the representation comparison is driven by particular subperiods or market regimes, this subsection splits the out-of-sample window into a pre-2020 segment (2006–2019) and a post-2020 segment (2020–2025). The split corresponds to the natural pre/post-COVID break in market-microstructure conditions and is held in common with

the subperiod partitions used elsewhere in the analysis. The same equal-weighted decile sort and five-day rebalance protocol is applied within each subperiod.

Table H2.1: Subperiod Stability: H–L Returns and Sharpe Ratios

	2006–2019		2020–2025		2006–2025	
Model	Return	Sharpe	Return	Sharpe	Return	Sharpe
Hiera	0.2196	1.950	0.1761	0.951	0.2067	1.496
HistGB Base	0.2031	1.796	0.1135	0.687	0.1757	1.343
HistGB + Geometry	0.2106	1.921	0.1680	1.106	0.1974	1.595
HistGB Full	0.2430	1.340	0.1434	0.999	0.2131	1.247
Logit Base	0.1372	1.115	0.1512	0.940	0.1403	1.037
Logit Full	0.1684	1.592	0.1615	1.158	0.1663	1.424
HistGB distilled	0.2259	1.913	0.1733	0.888	0.2102	1.445

Note: Annualized H–L spread returns and Sharpe ratios within each subperiod for the image model, baseline tabular models, the strongest translated HistGB specification (HistGB + Geometry, Sharpe 1.595), HistGB Full, Logit Full, and the HistGB-distilled model. All values are computed from the per-rebalance H–L return series underlying the main-text portfolio diagnostics. The 2006–2019 and 2020–2025 partitions correspond to the natural pre/post-COVID split used throughout the analysis.

Three patterns are visible. First, the chart advantage over the baseline tabular benchmark widens, not narrows, in the post-2020 segment. The Hiera versus HistGB Base Sharpe gap is 0.154 in 2006–2019 (1.950 vs 1.796) and 0.264 in 2020–2025 (0.951 vs 0.687): the absolute gap widens by a factor of about 1.7, and measured relative to the baseline's own Sharpe the advantage rises from 8.6% to 38.4%. The same pattern appears in raw H–L returns. Second, the translated tabular models and the distilled model are markedly more stable across the regime change than the baseline HistGB. HistGB + Geometry has a Sharpe of 1.921 in the early window and 1.106 in the late window, retaining roughly 60% of its early-period Sharpe; HistGB Base retains only 38%. The distilled model degrades in line with Hiera (0.888 versus Hiera's 0.951 in the late window). Third, Logit Base improves marginally in raw H–L return in the post-2020 window (0.1372 to 0.1512), though its Sharpe ratio edges down (1.115 to 0.940); even on this risk-adjusted basis it deteriorates far less than HistGB Base, whose Sharpe falls from 1.796 to 0.687. The tree-based benchmark is therefore more sensitive to the regime change than the linear benchmark on the same predictor set. The representation comparison documented in the main text is therefore not driven by a particular subperiod; if anything, the post-2020 evidence sharpens the chart-tabular distinction rather than dissolving it. Figure H2.1 shows the same pattern at a finer time scale: Hiera and the strongest translated tabular

benchmarks remain near the top of the cross-model ranking throughout the out-of-sample window, while the baseline tabular Sharpe drifts down through the late 2010s and again in 2020–2025.

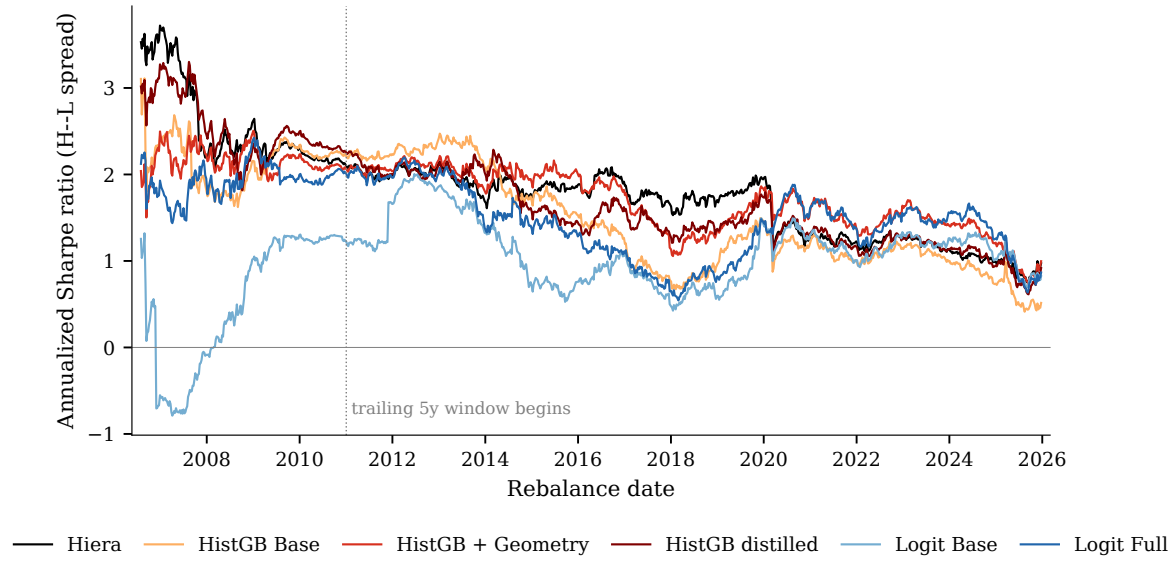


Figure H2.1: Rolling Sharpe Ratio of the H–L Spread for Selected Models, 2006–2025. Each series reports the annualized Sharpe ratio of the equal-weighted high-minus-low decile spread over the common evaluation sample for Hiera, the baseline tabular benchmarks HistGB Base and Logit Base, the strongest cumulative-translation specification within each tabular family (HistGB + Geometry and Logit Full), and the distilled HistGB specification. Sharpe ratios are annualized using the periods-per-year scaling $252/5 \approx 50.4$. Before five years of out-of-sample data have accumulated, each series is computed over an expanding window from the start of the panel to the current rebalance date; thereafter, it is computed over a trailing five-year window of 252 rebalances. A minimum of 30 rebalance observations is required before a series is shown, and the dotted vertical line marks the date at which the trailing window first contains a full five years of data. Hiera and HistGB + Geometry remain near the top of the cross-model ranking throughout the out-of-sample period. The baseline HistGB Sharpe declines through the late 2010s and again from 2020 onward, while the linear benchmarks trail the tabular tree models for most of the sample. The figure visualizes at a finer time scale the subperiod pattern reported in Table H2.1: the chart-based and translated tabular spreads show no detectable structural break over the out-of-sample window, whereas the baseline tabular benchmark's spread does, with the located shift dated to 2012.

Alt text: Line chart of annualized rolling Sharpe ratios for six high-minus-low portfolio spreads from 2006 to 2025. Hiera, HistGB Base, HistGB plus Geometry, HistGB distilled, Logit Base, and Logit Full are shown as separate lines. Most series begin between about 1.5 and 3.5, converge near 2 around 2010–2013, decline through the late 2010s, recover partly around 2020, and finish between roughly 0.4 and 1.0 by 2025. Hiera and HistGB plus Geometry are generally among the highest-performing series, whereas Logit Base is notably weak early in the sample and remains below most other models for much of the period. A vertical dotted line around 2011 marks the point at which the trailing five-year estimation window begins.

H2.0.1 Structural-break test.

The pre-2020 / post-2020 split is held in common with the subperiod partition used elsewhere in the analysis. To assess whether the data themselves locate a structural break in the H–L spread series at or near the partition, we run an Andrews (1993) sup-Wald test for a one-time mean shift at an unknown date with 15% trimming. For Hiera's H–L spread the sup-Wald statistic is 5.50, attained at January 2009; for HistGB augmented with the Recent Geometry block it is 5.46 at January 2010. Both lie below the 5% asymptotic critical value of 8.85, so in neither case can we reject a constant-mean specification. For HistGB Base the statistic is 11.09, attained at February 2012 and significant at the 5% level, making the baseline tabular benchmark the only series in which the test locates a break — and that break predates the COVID partition by eight years rather than coinciding with it. Two caveats temper this reading. First, a failure to reject is evidence of limited power, not proof of structural constancy, so the Hiera and translated-benchmark results are best read as the absence of a *detectable* break rather than as positive stability. Second, because the HistGB Base break is dated to 2012, it does not by itself identify the post-2020 widening. We therefore run two tests aimed directly at the relevant quantity. A sup-Wald test on the chart-minus-baseline spread *difference* (Hiera − HistGB Base) yields only 2.49, and 1.71 for Hiera − (HistGB + Geometry), both far below 8.85: the gap itself exhibits no detectable break at any date. A fixed-date Chow test at the January 2020 partition likewise fails to reject for every series (HistGB Base 1.52, Hiera 0.28, HistGB + Geometry 0.38, and the Hiera-minus-baseline difference 0.95, all below the χ^2_1 critical value of 3.84), even though the post-2020 spread mean is lower for HistGB Base than before (0.0022 versus 0.0040 per rebalance). The structural-break evidence therefore supports a measured reading: the only formally identified instability is a gradual pre-2020 level drift specific to the baseline tabular benchmark, whereas the post-2020 widening documented in Table H2.1 is economically visible in the point estimates but is not sharp enough to register as a discrete regime change at conventional levels.

H3 Multi-Horizon Robustness

The headline comparison uses a five-day-ahead directional target. This subsection reports the same H–L spread comparison at one-day and twenty-day holding horizons using the existing $h = 5$-trained models. The objective is to test whether the chart-tabular ordering documented in the main text is specific to the training horizon or robust across return-prediction horizons.

For each model $m \in$ {Hiera, HistGB Base, Logit Base, HistGB Full, Logit Full, HistGB distilled} and each rebalance date, stocks are sorted into equal-weighted deciles on the model's score and held for $h \in \{1, 5, 20\}$ trading days. Annualized returns and Sharpe ratios are computed using the appropriate periods-per-year scaling (252 for $h = 1$, 50.4 for $h = 5$, 12.6 for $h = 20$).

Table H3.1: Multi-Horizon Robustness (Strict Test): H–L Spread Returns and Sharpe Ratios, 2006–2025

	$h = 1$		$h = 5$		$h = 20$	
Model	Return	Sharpe	Return	Sharpe	Return	Sharpe
Hiera	0.3382	2.011	0.2067	1.496	0.1067	0.754
HistGB Base	0.2549	1.527	0.1757	1.343	0.0915	0.735
Logit Base	0.2453	1.437	0.1403	1.037	0.0694	0.597
HistGB Full	0.3283	2.134	0.2131	1.247	0.1007	0.674
Logit Full	0.2920	2.293	0.1663	1.424	0.0751	0.672
HistGB distilled	0.2865	1.963	0.2102	1.445	0.0959	0.653

Note: Annualized H–L spread returns and Sharpe ratios for six model scores at three holding horizons ($h \in \{1, 5, 20\}$ trading days) on the common out-of-sample panel over 2006–2025. All scores are produced by models trained on the five-day directional target ($h = 5$). For each rebalance date and model, stocks are sorted into equal-weighted deciles on the model's score and held for h trading days. Returns are annualized by the periods-per-year factor $252/h$ and Sharpe ratios by its square root, $\sqrt{252/h}$. The number of rebalance dates is 1,006 at $h = 1$ and $h = 5$ and 1,003 at $h = 20$ (the last three windows lack a full 20-day forward return).

Three patterns are visible. First, all six models earn substantially higher annualized returns and Sharpe ratios at $h = 1$ than at $h = 5$, consistent with short-horizon reversal being a meaningful component of the cross-sectional signal at the one-day holding period. Second, at $h = 20$ the spreads and Sharpe ratios are roughly half their $h = 5$ values for every model, indicating that the chart-based signal decays over longer horizons but does not vanish. Third, the cross-model ordering against the baseline tabular benchmarks is largely preserved across horizons. Hiera leads HistGB Base and Logit Base on annualized H–L return at $h = 1$ (0.3382 vs 0.2549 and 0.2453), at $h = 5$ (0.2067 vs 0.1757 and 0.1403),

and at $h = 20$ (0.1067 vs 0.0915 and 0.0694). Against the translated tabular models the ordering compresses, particularly at $h = 1$ where Logit Full translated attains a higher Sharpe than Hiera but a lower annualized return. The compression at $h = 1$ is consistent with the shortest-horizon predictability being dominated by short-term reversal patterns that simpler models capture well. Across all three horizons, the headline conclusion that Hiera's signal carries non-trivial predictive content over the non-translated tabular benchmarks is robust.

We note one limitation: the tabular benchmarks are trained on the five-day directional target. A fully symmetric horse race that retrained Logit and HistGB on $h = 1$ and $h = 20$ directional targets would give the tabular models the home-field advantage at each horizon and is left as a planned extension. The strict version reported here is the cleaner test of horizon-specificity: if a horizon-specific tabular retraining were necessary to recover the chart-tabular gap at non-five-day horizons, the $h = 5$-trained signals would not predict at $h = 1$ and $h = 20$ in the first place, which the table contradicts.