

Label alchemy: Target engineering for improved stock selection

Guillaume Coqueret*

Tony Guida[†]

September 20, 2026

Abstract

Machine-learning research on stock returns has focused its attention on features and models, while treating the prediction target (or label) as a given (forward monthly or annual returns most of the time). We show that the choice of labels is crucial. Holding the investable universe (1,960 US stocks), the 122 firm features, and the model fixed, we decompose the label into location, scale and shape operators, combine them in a factorial design, and attribute out-of-sample stock-selection performance to each operator. Reshaping the target alone raises the long-short decile Sharpe ratio from 0.68 (simple return label) to 1.69 for the Gaussianised cross-sectional rank of that same return. These gains are model independent: replacing trees with a linear model or neural networks leads to the same conclusion. The label explains much more of the out-of-sample variation than the model does. We thus conclude that the prediction target deserves the same engineering effort that the field already devotes to models and features.

1 Introduction

A supervised learning problem has three ingredients: the inputs X , the model f , and the target y . In empirical asset pricing, a substantial literature has documented the impact of the first two: hundreds of firm characteristics, cross-sectional standardisation of features, and a litany of models from penalised linear regressions to random forests, gradient boosting, and deep neural networks (see, e.g., [Gu et al., 2020](#)). The third ingredient, the label, is almost always the *raw one-month-ahead excess return*, fed to the model, very often unchanged.

This asymmetry is puzzling, because the label encodes the core of the problem the investor seeks to address. It is the quantity that matters the most, as it is the one that we seek to predict and optimize from. Raw returns are prone to regime shifts (bull versus bear markets) and models may just capture this, whereas money managers are mostly interested in assets that outperform a benchmark. Moreover, the distribution of the label also matters a lot because lottery stocks are likely to account for a larger share of the loss function. Hence, a normalisation, or ranking, can help mitigate this overweighting towards the more risky stocks.

*EMLYON Business School, 144 avenue Jean Jaurès, 69007 Lyon. Email: coqueret@em-lyon.com

[†]AtonRâ Partners, 7 rue de la croix d'or, 1204 Genève. Email: tguida@atonra.ch

This paper isolates label engineering as a treatment variable. We fix the universe and the 122 features, and, for our main experiment, we hold the learner (i.e., model) fixed as a gradient-boosted tree ensemble, so that the only thing that varies is the training target. We subsequently confirm that the conclusions do not depend on that learner by repeating the study under two alternative models, a penalised linear regression (elastic-net) and a feed-forward neural network. We represent the target as a composition of three operators and evaluate fifteen labels, curated to span the three axes: ten constructions that the literature already studied; and five novel ones (Section 3.2). Our contributions can be summarized as follows:

1. **An exhaustive comparison of label constructions.** Our framework allows to attribute performance variation to label configuration. Partial comparisons exist (Bai and Pukthuanthong, 2021; Chen et al., 2026), and Cakici and Zaremba (2026) document across 35 markets that transforming the target matters more than transforming the features; but in each of these studies a target moves several operators at once, so the performance difference cannot be attributed to a specific axis. In contrast, we compound the three axes systematically, in an $L \times S \times T$ factorial of 30 combinations, from which we extract 15 that are most relevant. Our comparisons also run through several correlation-correction procedures: the Model Confidence Set (Hansen et al., 2011), the probability of backtest overfitting (Bailey et al., 2017), and the test of superior predictive ability (Hansen, 2005).
2. **A decomposition that separates location, scale and shape.** Prior work confounds the following three stages: centering with respect to a benchmark, scaling and transforming (e.g., through binarisation). Those are three distinct choices. Writing every construction as a composition of three operations allows to discern the role and impact of each of these operations.
3. **New centered labels.** We introduce sector-demeaned, PCA-residual and Fama-French six-factor (FF6) residual targets, with rank and standardised compositions for the sector and PCA variants. To our knowledge, no prior comparison includes a sector-relative target or a residual from an estimated factor model.

The remainder of the paper proceeds as follows. Section 2 reviews the literature, organizing the related work around the Location, Scale and Shape operators. Section 3 sets out the notation and describes the data. Section 3.2 then develops the taxonomy of label operators: it presents the fifteen constructions we test. Section 4 turns to the experimental design: the walk-forward protocol and the correlation-corrected inference. Section 5 reports the results, including the robustness checks. Section 6 concludes. Two appendices complete the paper: Appendix A gathers the full results for the thirty factorial labels from which the fifteen analysed in details are drawn, and Appendix B documents the feature panel by family.

2 Related work

We structure the literature according to the paper’s prism: an engineered label is obtained from three layers that operate on a starting point (the original target value, e.g. a stock return). We outline them below.

The starting point. Machine-learning prediction of stock returns predates the current wave: neural networks were originally fitted to equity returns in [White \(1988\)](#), and a soft-computing literature followed ([Olson and Mossman, 2003](#); [Atsalakis and Valavanis, 2009](#)). [Gu et al. \(2020\)](#) supplied the modern reference design, and with it the convention that concerns us: the one-month-ahead excess return under squared-error loss, with *features* rank-standardised to $[-1, 1]$ and the *label* untouched. This practice has a rich offspring, from [Freyberger et al. \(2020\)](#), [Kelly et al. \(2019\)](#), to [Chen et al. \(2026\)](#), and across markets by [Leippold et al. \(2022\)](#), [Drobtetz and Otto \(2021\)](#) and [Hanauer and Kalsbach \(2023\)](#). Even the replication critiques of [Cakici et al. \(2023a,b\)](#) and the market-timing study of [Kelly et al. \(2024\)](#) resort to it.

Location, i.e., the benchmark that is subtracted. This is the most common departure from the default choice. The simplest subtraction is the cross-sectional mean: [Tobek and Hronec \(2021\)](#) and [Chen et al. \(2026\)](#) target the market-relative return, arguing that stripping the common component raises the label’s signal-to-noise ratio. In both, the target is one implementation choice inside a broader exercise and is not really studied. Industry-relative signals can set the benchmark to the sector mean (e.g., in [Asness et al., 2000](#); cf. [Moskowitz and Grinblatt, 1999](#)).¹

Learned location. A parallel literature drops the predictive label entirely and lets the model choose the benchmark: instrumented PCA ([Kelly et al., 2019](#)), conditional autoencoders ([Gu et al., 2021](#)), adversarial SDF estimation ([Chen et al., 2024](#)) and asset-pricing trees ([Bryzgalova et al., 2025](#)) estimate latent structure under a pricing-error criterion. These models do not exactly seek to forecast next month’s winners, they are judged by how small their pricing errors are. Whether centering purely relates to the label is not necessarily straightforward: [Hanauer et al. \(2026\)](#) find that much of the sector benefit is learned from the characteristics directly.

Scale: the dispersion that divides the centered target. This axis is the least studied as a label choice and the best documented as a strategy choice, especially in volatility timing. The momentum literature is decisive that dividing by a dispersion matters: [Moskowitz et al. \(2012\)](#) size positions to constant ex-ante volatility (but not return). [Kim et al. \(2016\)](#) attribute much of time-series-momentum ‘alpha’ to the scaling rather than the signal. [Barroso and Santa-Clara \(2015\)](#) find that risk-managing momentum by its trailing volatility “nearly doubles the Sharpe ratio”. [Daniel and Moskowitz \(2016\)](#) and [Harvey et al. \(2018\)](#) concur. All of them scale the *position* but not the training target. The closest in spirit to what is done in the present paper is the volatility-scaled barriers of [López de Prado \(2018\)](#). But isolating *Scale* has, to our knowledge, not been investigated.

Shape: reshaping the marginal. The third step of the process acts on the distribution of the target. [Azevedo et al. \(2023\)](#) find that their best models use a percent-ranked target. [Cakici and Zaremba \(2026\)](#) compare five feature and six target transformations across 35 markets and find that the target side dominates, with rank-based targets ahead on average and cardinal ones in dispersed or skewed cross-sections; each of their targets, however, moves several operators at once, and none neutralises beyond demeaning. Gaussianising has precedent in statistical

¹A larger literature predicts idiosyncratic returns, moving *Location* to a factor-fitted value: [Blitz et al. \(2011\)](#) rank on Fama-French residuals standardised by residual volatility, a *Location* and a *Scale* at once, and roughly double the Sharpe of total-return momentum “mainly due to lower return variability”. The construction extends to residual reversal and idiosyncratic momentum ([Blitz et al., 2013, 2020](#)), to PCA residuals ([Avellaneda and Lee, 2010](#)), and to conditional-latent-factor residuals ([Guijarro-Ordóñez et al., 2026](#)).

genetics (Beasley et al., 2009; McCaw et al., 2020). Finally, perhaps as the strongest counter-example of our critique, López de Prado (2018, 2020) treats labelling itself as the modelling choice: triple-barrier labels are volatility-scaled, path-dependent and categorical.²

Horizon: the window is also a degree of freedom. Every investor plans according to a forward return measured over some window, which is a design choice. Picking one month, or three, or twelve, changes both the target’s statistical character and the strategy it implies. Coqueret (2022) documents the role of memory in characteristic-based models: algorithms perform best when trained on long samples with long-horizon returns, persistent features dominate, and investors are better off predicting annual returns even when rebalancing monthly or quarterly. Coqueret and Guida (2020, §4.5) reach the same conclusion and warn that one-month returns are too weakly autocorrelated to make good labels while six- and twelve-month returns can be better choices. Cakici et al. (2023b) document predictability varying sharply across horizons, firm size and time. Sampling long horizons monthly also overlaps observations, whose inference corrections go back to Hansen and Hodrick (1980) and Britten-Jones et al. (2011). We sidestep the problem instead: every horizon is evaluated on non-overlapping h -month periods (Section 5.3), and the horizon is crossed with the three operators in Sections 5.3 and 5.4.

3 Notation and data

3.1 Notation

This section fixes the notation used henceforth and describes the panel on which our empirical study is built. The notation follows Chapter 1 of Coqueret and Guida (2020). Firms are indexed by $n = 1, \dots, N$ and months by $t = 1, \dots, T$, and we call a single firm-month (t, n) an *instance*, of which the panel holds I . The quantity a strategy ultimately cares about is the forward total return of firm n over the next h months, observed at the one-, three-, six- and twelve-month horizons. Alongside each return, every instance carries a vector of $K = 122$ firm characteristics, $\mathbf{x}_{t,n} = (x_{t,n}^{(1)}, \dots, x_{t,n}^{(K)})$, spanning the momentum, value, quality, growth, profitability, leverage, liquidity and risk families (Appendix B lists all thirteen families and their representative members) and stacked into the feature matrix $\mathbf{X} \in \mathbb{R}^{I \times K}$. Every characteristic is cross-sectionally percent-ranked within its month into $(0, 1]$, the uniformisation of Coqueret and Guida (2020) and the counterpart of the $[-1, 1]$ rank map of Gu et al. (2020), so the features are level-, scale-, and distribution-agnostic. A *label* is any scalar target $y_{t,n}$ derived from the forward return $r_{t,n}^{(h)}$. How that derivation is performed is exactly the subject of this paper.

²Three kinds of contributions enforce other kinds of shaping. The first is classification, which fixes a *Location* (a threshold) and a *Shape* (binarisation) together: Takeuchi and Lee (2013), Krauss et al. (2017) and Fischer and Krauss (2018) predict the probability of beating the monthly median and trade the extremes. Wolff and Echterling (2024) find regularised logistic regression competitive with far more complex learners on that task. Related constructions appear in Rasekhschaffe and Jones (2019), Moritz and Zimmermann (2016) and Huck (2009, 2010). The recurring critics, class imbalance at a zero threshold and the magnitude discarded by binarising, are some trade-offs of this procedure. Another approach is learning to rank, since selection needs only an ordering. Poh et al. (2021) report an approximately threefold Sharpe improvement over regression- or classification-then-sort, the strongest existing evidence that label encoding can dominate model choice. Extensions cover currencies (Poh et al., 2022), data scarcity (Poh et al., 2023), joint time-series and cross-sectional momentum (Tan et al., 2023), list-wise long-short construction (Zhang et al., 2022) and a differentiable rank-IC surrogate (Lin et al., 2026).

3.2 The taxonomy of labels

Each label is a point in a single design space spanned by the three operators:

- **L: Location.** Subtract a benchmark $b_{t,n}$, which can be the following: nothing (raw), the cross-sectional mean or median (market-neutral), the **sector mean** (sector-neutral), or the fitted value from a **statistical-factor (PCA) model** (idiosyncratic). We use *Location* in the broad sense of removing a conditional centre, the cross-sectional mean or median, or, more generally, the residual of a projection onto a systematic subspace. In financial terms, we coin this axis “*neutralization*” (or centering in statistical parlance).
- **S: Scale.** The outcome of the previous stage is divided by a dispersion factor $d_{t,n}$, that can be: nothing (i.e., 1), the firm’s realised **volatility** (Sharpe-like), the cross-sectional standard deviation (completing the z-score), or a robust scale Median Absolute Deviation (henceforth MAD). *Scale* spans two regimes: division by a cross-sectional dispersion (a single scalar per date, order-preserving) and division by a firm’s own volatility (studentisation, which reorders the cross-section). When the per-stock route is taken, the divisor is the trailing realised volatility of the stock’s own monthly returns over the preceding twelve months, estimated strictly before the forward window and floored, date by date, at the second percentile of the cross-section of volatilities, so that near-zero denominators cannot inflate the label.
- **T: Shape.** Finally, the output is reshaped and we allow for the following alternative: identity (no transformation), **percentile rank** (uniform), **Gaussianised rank** (normal), or the **sign of the located value** (binarization for classification, i.e., the beat-the-benchmark indicator $\mathbb{1}\{v > 0\}$). We name this axis *Shape*, echoing the location–scale–shape decomposition of GAMLSS (Rigby and Stasinopoulos, 2005). The operator is written T for the monotone transform it applies. The Gaussianised rank is the rank-based inverse-normal transformation (van der Waerden or Blom scores; Beasley et al., 2009).

Spanning all relevant combinations generates thirty distinct labels. Section 4.1 derives this count from the full cross of the three axes, in which some crossings are redundant because a rank or sign transform absorbs certain scales. Our main study will focus on a subset of them (15) which are listed in Table 1.

Five of the fifteen labels, marked $\star$ in the last column of Table 1, are to our knowledge new to controlled label-comparison studies. They form two families: the sector-relative targets (sector_demean and its standardised composition sector_zscore) and the factor-residual targets, of which the PCA residual (with its percentile form) is the latent member and the FF6 residual the observable one. We outline their construction below:

- **Sector-demean (sector-neutral).** $y_{t,n}^{\text{sec}} = r_{t,n}^{(h)} - \bar{r}_{s(n),t'}^{(h)}$ the return net of the firm’s sector mean. The learner is rewarded only for beating industry peers, so the resulting label is approximately sector-neutral by construction.
- **PCA-residual (idiosyncratic alpha).** Loadings $\mathbf{B}$ on K^* statistical factors are estimated from a **trailing** window of realised returns (standardised, excluding the current month, so there is no look-ahead). The current cross-section is regressed on $\mathbf{B}$ and the **residual**

#	Label	Location L	Scale S	Shape T	New?
1	raw	none	none	identity	
2	demean	market mean	none	identity	
3	median_demean	median	none	identity	
4	zscore	market mean	cs-std	identity	
5	robust_z	median	MAD	identity	
6	sector_demean	sector mean	none	identity	*
7	sector_zscore	sector mean	cs-std	identity	*
8	percentile	none	none	rank	
9	rank_gauss	none	none	rank-Gauss	
10	beat_median	median	none	sign	
11	vol_scaled	none	trailing vol	identity	
12	vol_scaled_rank	none	trailing vol	rank	
13	pca_resid	PCA	none	identity	*
14	pca_resid_pct	PCA	none	rank	*
15	FF6 cross-sectional	FF6 (Fama–MacBeth)	none	identity	*

Table 1: **The fifteen labels, as (L, S, T) compositions.** * = newly tested.

is the label, $y_{t,n}^{\text{pca}} = \hat{\varepsilon}_{t,n}$. This is the component of a stock’s forward return *not* spanned by pervasive factors (statistical alpha).

- **Fama-French residual (idiosyncratic return, observable factors).** The same construction with an *economically identified* factor space: the forward excess return is residualised against the Fama-French–Carhart factors (FF6 = Mkt-RF, SMB, HML, RMW, CMA, and momentum; Fama and French, 2015; Carhart, 1997), $y_{t,n}^{\text{ff}} = (r_{t,n}^{(h)} - r_{t,f}) - \beta_n^\top \mathbf{F}_t$. The residual can be estimated by two routes: a per-stock time-series regression, whose fitted intercept absorbs the stock’s mean return, and with it the cross-sectional signal a selection label needs, or a cross-sectional (Fama–MacBeth; Fama and MacBeth, 1973) regression, which estimates the factor premia date by date and preserves that signal; we use the cross-sectional route throughout.

Between the two members of the factor-residual family, the FF factors are interpretable and quite stable, while the PCA adapts to realised comovement but its factors rotate and are harder to interpret economically. The closest known contribution to this family is the beta-adjusted target of Chen et al. (2026), which deducts a single pre-estimated market beta. Figure 1 illustrates the impact of alternative design choices for labels: it processes a single monthly cross-section of returns through five target constructions.

3.3 Data

The empirical work relies on a single panel of 1,960 US stocks observed monthly from January 2002 to March 2026, a total of 446,679 firm-months over 291 cross-sections, with a median of roughly 1,527 firms trading in any given month. Its characteristics extend the publicly available factor investing dataset of Coqueret and Guida (2020). For the centered labels of Section 3.2 we merge in each firm’s sector from the North American Industry Classification System (NAICS).

The same cross-section (Nov 2015, 1,462 firms) under five target constructions

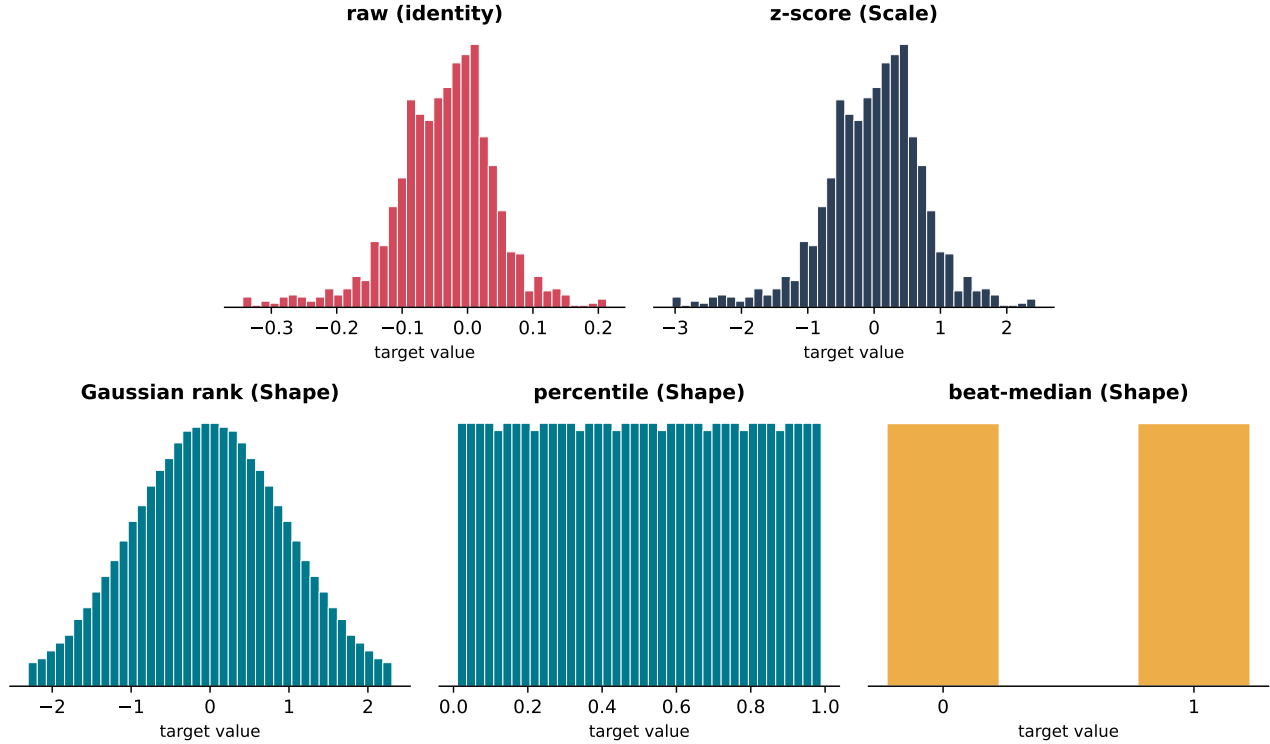


Figure 1: **Example of a label’s marginal distribution under five target constructions.** A representative monthly cross-section (Nov 2015 with 1,462 firms with a realised one-month return), showing the same returns mapped through five constructions. Scale (z-score) rescales the raw distribution but preserves its shape; the Shape operators reshape the marginal outright, into a Gaussian (rank-Gauss), a uniform (percentile), or a two-point law (beat-median). Location operators re-centre rather than reshape and are omitted here.

Because a training target must be realised (known) before it can be used, restricting the sample to instances with a valid one-month return leaves 386,583 rows with non-missing information. We leave undefined characteristics unchanged in our baseline study because gradient-boosted trees handle missing values natively. For the extension on linear and neural learners in Section 5.7, we resort to median imputation.

4 Experimental design

This section specifies the experiment: the fixed model and training protocol, the combinatorial grid that combines the three operators, and the inference stack that corrects for selection bias. Every label is trained with the same fixed gradient-boosted-tree configuration (LightGBM: 300 trees, learning rate 0.05, 63 leaves, 70% feature and row subsampling, a minimum of 100 samples per leaf, and an L2 penalty of 1.0), and Section 5.7 repeats the horse-race under linear and neural network models. All targets, including the binary label, are fit under squared-error loss, hence

the binary case is treated as a regression on its $\{0, 1\}$ value. Holding both the hyper-parameters *and* the objective constant across labels ensures that any performance difference is attributable to the target alone. Had we switched to a log-loss classifier for the binary label, the change of loss function would have perturbed the comparison.³

The evaluation is a strict expanding-window walk-forward with annual refitting. Each January, the model is retrained on every instance whose label is fully realised, with an embargo of h months to rule out look-ahead, and is then used to score the next twelve monthly cross-sections. Out-of-sample coverage runs from January 2010 to March 2026, i.e., roughly 200 months. Each month, we rank stocks by their model score, sort them into decile portfolios, and hold an equal-weighted portfolio of long top-decile versus short bottom-decile stocks. Equal weights are convenient because they allow to sidestep a portfolio-optimisation stage. Moreover, equal-weighting is known to be hard to outperform out of sample (DeMiguel et al., 2009) and requires no estimation of expected returns or covariances.

For each label, we report three groups of statistics: (i) its rank-IC, the monthly Spearman correlation of score with realised return, together with the information ratio $ICIR = \overline{IC} / \sigma_{IC}$ and its t -statistic, computed as ICIR scaled by the square root of the number of evaluation months (the Newey–West-corrected counterpart is reported in Table A1); (ii) the annualised return, volatility, Sharpe ratio, maximum drawdown and monthly hit rate of the long-short book; and (iii) the turnover of the long book. Each label is therefore judged on its objective, i.e., the ability to discriminate according to relative performance.

4.1 The grid and inference correction

A label in our study is determined by a *combination* (sometimes referred to as a ‘cell’), i.e., a triple (L, S, T) with

$$\begin{aligned} L &\in \{\text{none, market, sector, PCA, FF6}\} \\ S &\in \{\text{none, cs-std, vol}\}, \\ T &\in \{\text{identity, rank-Gauss, binary}\}, \end{aligned} \tag{1}$$

where the FF6 Location uses the cross-sectional (Fama–MacBeth) residual and the vol Scale divides by the floored trailing realised volatility, the constructions established in Section 3.2. Not every crossing yields a distinct label: cross-sectional standardisation is absorbed by both the rank-Gauss and the sign transforms (a positive scalar per date changes no within-month ordering and no sign), and the per-stock volatility scale is absorbed by the sign transform, which removes 15 degenerate crossings and leaves the 30 distinct cells listed in Table A1.

Thirty labels built from the same returns are highly correlated, so naive per-label t -tests and a ‘best-Sharpe’ claim would be selection-biased. To make cells comparable, every label is scored on the same two metrics, rank-IC and decile-spread return, over identical out-of-sample periods. On these series, we apply a five-layer inference procedure:

³Nevertheless, as a robustness check, we also fit the binary label with a log-loss classifier: the two produce highly correlated cross-sectional rankings (mean rank correlation 0.78 / 0.84 / 0.88 / 0.91 at the 1/3/6/12-month horizons) and essentially the same decile Sharpe (1.35 vs 1.35 for the monthly forward return), so no conclusion depends on the choice.

1. **Per-label:** Newey–West (HAC) t -statistic on the rank-IC series; Probabilistic Sharpe Ratio on the spread.
2. **Multiplicity:** Romano–Wolf StepM (Romano and Wolf, 2005; family-wise error) and Benjamini–Yekutieli (false discovery rate), both exploiting the cross-label correlation via a joint bootstrap.
3. **Overfitting probability:** the Probability of Backtest Overfitting (PBO; Bailey et al., 2017) via combinatorially-symmetric cross-validation (252 in/out partitions).
4. **Set inference:** the Model Confidence Set (MCS), the labels statistically indistinguishable from the best, and Hansen’s (2005) Superior Predictive Ability (SPA) test of the best label against the naive raw-return benchmark.
5. **Attribution:** a factorial ANOVA / η^2 variance decomposition of performance onto L, S, T and their interactions.

We resort to a multivariate stationary block bootstrap (Politis and Romano, 1994) that resamples **common time blocks across all labels**, preserving the cross-label correlation that makes StepM and the MCS more powerful than a Bonferroni correction. Again, the reason is the correlation: Bonferroni divides the significance level by the number of comparisons as if the tests were independent, but thirty labels built from the same returns are far from independent.

5 Results

This section reports the empirical results. We start with the last five years of the out-of-sample period, 2021-2026, and then present the full-sample comparison of all fifteen labels along its robustness across horizons.

5.1 Late out-of-sample results

Before displaying results for the full sample, we report outcomes for the final five years of the out-of-sample period, January 2021 to March 2026, where the contrast between the raw and the engineered targets is widest (Table 2).

Label	rank-IC	ICIR	LS Sharpe	Max DD
raw (control)	−0.023	−0.22	− 0.26	−31%
z-score	−0.002	−0.02	+1.16	−9%
percentile	+0.014	+0.17	+0.91	−23%

Table 2: **Label comparison over the final five years of the out-of-sample period**, January 2021 to March 2026. z-score and percentile are the constructions numbered 4 and 8 in Table 1, the cross-sectionally standardised return and the cross-sectional percentile rank. Portfolios are equal-weighted long-short decile books built from the model scores.

Over 2021-2026 the naive raw-return target produces an *outright negative* Sharpe, while re-shaping the *same* signals through a cross-sectional standardisation (z-score) or a percentile rank yields a strongly positive long-short strategy. Features and model are identical, and only the label differs. The two engineered rows are the simplest representative of each distributional axis, Scale and Shape, and they are not the strongest constructions over this window: the rank-Gauss and sector z-score compositions reach Sharpe ratios of 1.46 and 1.32 in the same period, and the full fifteen-label comparison follows in Section 5.2.

5.2 Full results (OOS 2010–2026)

Label	rank-IC	ICIR	IC t	LS ret	LS vol	Sharpe	Max DD	Hit	Turn. L	Turn. S
rank_gauss	0.032	0.41	5.8	19.8%	11.7%	1.69	−10%	74%	0.48	0.52
sector_zscore *	0.025	0.29	4.0	19.0%	12.2%	1.56	−14%	70%	0.43	0.55
zscore	0.024	0.29	4.1	19.0%	12.5%	1.53	−10%	70%	0.40	0.54
robust_z	0.024	0.25	3.5	18.9%	13.0%	1.46	−14%	72%	0.41	0.55
beat_median	0.030	0.39	5.4	14.9%	11.0%	1.35	−15%	63%	0.58	0.57
percentile	0.032	0.38	5.3	16.5%	12.4%	1.34	−23%	70%	0.54	0.54
pca_resid_pct *	0.023	0.32	4.5	14.4%	11.3%	1.28	−23%	69%	0.51	0.53
FF6 cross-sectional *	0.018	0.29	4.0	13.8%	10.8%	1.28	−13%	63%	0.42	0.52
pca_resid *	0.015	0.19	2.6	15.0%	13.0%	1.16	−11%	66%	0.39	0.52
demean	0.019	0.19	2.7	15.8%	13.7%	1.15	−15%	66%	0.40	0.58
sector_demean *	0.016	0.16	2.3	15.6%	14.7%	1.06	−16%	65%	0.41	0.58
median_demean	0.018	0.17	2.4	14.9%	14.2%	1.05	−19%	66%	0.41	0.57
raw (control)	0.006	0.06	0.9	9.7%	14.3%	0.68	−31%	57%	0.57	0.67
vol_scaled_rank	0.027	0.30	4.2	8.3%	12.6%	0.66	−29%	59%	0.56	0.52
vol_scaled	0.008	0.10	1.4	4.0%	11.1%	0.36	−40%	54%	0.71	0.69

Table 3: **Label comparison, OOS Jan 2010 – Mar 2026 (194 months).** * = new. IC t is the plain t -statistic of the monthly rank-IC series (ICIR scaled by $\sqrt{194}$); the Newey–West-corrected t -statistics for the factorial cells are reported in Table A1. Turn. L and Turn. S are the average monthly fraction of names replaced in the long (top-decile) and short (bottom-decile) books.

Table 3 reports the fifteen labels over the full out-of-sample period, January 2010 to March 2026. The raw forward return ranks 13th of 15: a Sharpe of 0.68 against the leader’s 1.69, the deepest drawdown of any unscaled label (−31%), and a rank-IC indistinguishable from zero ($t = 0.9$). With the same model and features, most engineered labels reach IC t -statistics between 2.3 (sector_demean) and 5.8 (rank_gauss).

We find that scaling is the dominant lever. Moving from raw (0.68) or demean (1.15) to zscore (1.53) lifts the Sharpe by 0.4-0.85 and cuts the maximum drawdown from −31% to −10%. The three z-score compositions hold three of the top five rows. The mechanism comes likely from the distribution: the raw target’s fat right tail lets a few extreme firm-months dominate the squared-error loss, which standardisation (and, even more so, ranking) removes.

With regards to Shape, the Gaussianised rank combines the strengths of the z-score and the percentile. When location is fixed, the zscore earns the higher Sharpe (1.53 against 1.34) and percentile the higher rank-IC (0.032 against 0.024): ranking predicts the *ordering* better,

but z-scoring builds the better *portfolio*. The rank_gauss label, the percentile mapped through the inverse normal, captures both: the best Sharpe (1.69), IC t (5.8) and hit rate (74%), and a maximum drawdown (−10%) matching the smallest in Table 3. The rank-to-Gaussian map is the ‘RankGauss’ input normalisation of applied machine learning (Jahrer, 2017). Applying it to the *label*, which turns return regression into a learning-to-rank problem (Poh et al., 2021), is, to our knowledge, novel.

With Location, neutralisation helps modestly, and works best when combined with scale. demean, sector_demean and median_demean cluster at Sharpe $\approx$ 1.05-1.15, above raw but below the standardised labels. Composed with scale, sector_zscore reaches 1.56, second overall, at a long-book turnover of 0.43 (leader 0.48, other z-score compositions 0.40-0.41) and a short book of 0.55: the sector neutralisation earns its rank without additional trading.

The novel centered labels work well. All five $\star$ labels beat the raw control: sector_zscore (1.56, #2), pca_resid_pct (1.28, #7), FF6 cross-sectional (1.28, #8), pca_resid (1.16, #9) and sector_demean (1.06, #11). pca_resid pairs low volatility with a shallow drawdown (−11%), as expected of a target stripped of factor return; its percentile form adds roughly 0.12 of Sharpe by replacing magnitude with order.

Finally, per-stock risk-scaling is inferior to cross-sectional standardisation. The volatility-scaled labels of Section 3.2 earn a positive, significant IC but low Sharpes (0.36 for vol_scaled, 0.66 for vol_scaled_rank): the decile returns are flat at the top, so the label separates losers but cannot rank winners. Dividing by each firm’s own volatility changes the problem, from expected return to return per unit of risk, and raises turnover (long book 0.71 against roughly 0.40 for the standardised labels, short book 0.69 against 0.55). The damage comes from the drift between trailing and future volatility, which reorders the cross-section.

5.3 Horizon robustness

We repeat the full sweep at the 1-, 3-, 6- and 12-month horizons, each evaluated on non-overlapping h -month periods (annualised with $12/h$) to avoid overlap-inflated significance. The number of independent periods falls with horizon ($194 \rightarrow 64 \rightarrow 32 \rightarrow 16$), so longer-horizon t -statistics are mechanically weaker. Table 4 reports the long-short Sharpe at every horizon, ordered by the R1M ranking, with each label’s within-horizon rank in parentheses: a column shows a horizon’s ranking, a row reports how far a label changes as the horizon increases.

Three main findings:

1. **The ranking is stable.** The four constructions rank_gauss, sector_zscore, zscore and robust_z hold four of the five top ranks at every horizon, with rank_gauss never below third. Raw returns never climb above ninth of fifteen and the volatility-scaled label finishes last at every horizon except R6M, where it ties ninth. (Rows use the Section 3.2 constructions.)
2. **The label-engineering effect is largest at short horizons.** The best-minus-raw Sharpe gap shrinks with horizon ($\approx$ 1.01 at R1M, 0.60 at R3M, 0.25 at R6M, 0.30 at R12M), and both the absolute Sharpe ratios and the spread between labels compress: the payoff to label design is greatest where the signal is the most noisy.
3. **The factor-neutral labels are competitive at every horizon** (FF6 and PCA residuals reach Sharpe $\approx$ 1.1-1.3), *provided they are constructed cross-sectionally*.

Label	R1M	R3M	R6M	R12M
rank_gauss	1.69 (1)	1.27 (1)	0.67 (3)	0.62 (2)
sector_zscore *	1.56 (2)	1.15 (3)	0.68 (2)	0.68 (1)
zscore	1.53 (3)	1.18 (2)	0.66 (4)	0.56 (3)
robust_z	1.46 (4)	0.93 (5)	0.69 (1)	0.53 (5)
beat_median	1.35 (5)	0.82 (10)	0.57 (7)	0.29 (11)
percentile	1.34 (6)	1.00 (4)	0.55 (8)	0.54 (4)
pca_resid_pct *	1.28 (=7)	0.89 (=6)	0.33 (14)	−0.02 (13)
FF6 cross-sectional *	1.28 (=7)	0.74 (12)	0.44 (=9)	0.25 (12)
pca_resid *	1.16 (9)	0.86 (9)	0.38 (13)	0.34 (10)
demean	1.15 (10)	0.87 (8)	0.58 (6)	0.45 (7)
sector_demean *	1.06 (11)	0.89 (=6)	0.61 (5)	0.51 (6)
median_demean	1.05 (12)	0.78 (11)	0.39 (12)	0.41 (8)
raw (control)	0.68 (13)	0.67 (13)	0.44 (=9)	0.38 (9)
vol_scaled_rank	0.66 (14)	0.23 (15)	0.24 (15)	−0.14 (15)
vol_scaled	0.36 (15)	0.54 (14)	0.44 (=9)	−0.03 (14)

Table 4: **Long-short Sharpe by horizon.** Each label’s within-horizon rank is reported in parentheses (non-overlapping evaluation). Labels with equal Sharpe at the reported precision share a rank, marked “=”. * = new compared to the literature.

5.4 Variance attribution and correlation-corrected inference

Sections 5.2 and 5.3 compared the fifteen curated labels one at a time. This section runs the fully-crossed grid of Section 4.1, whose thirty combinations depend on Location, Scale and Shape jointly, through the inference stack of Section 4.1: the variance decomposition first (Table 5), the correlation-corrected comparisons after it. Table A1 (Appendix A) lists the thirty options ranked by out-of-sample Sharpe, with their Newey–West t -statistics and Probabilistic Sharpe Ratios.

Variance attribution (ANOVA). The decomposition determines how much of the performance variation across the thirty combinations each design dimension explains. The response is each label’s mean out-of-sample rank-IC over the 194 evaluation months. η^2 is the fraction of the variance across combinations for which one of the three dimension accounts for (a Location η^2 of 0.83 means that the Location operator alone explains 83% of the variation).

Table 5 therefore answers the question on the two sub-grids that are complete: Location×Scale with Shape fixed at identity, and Location×Shape with Scale fixed at none, each a full 5×3 cross of fifteen combinations (the cells themselves are in Table A1). The table shows not the cells but the decomposition, one panel per sub-grid: within a panel the axes are orthogonal, the shares unambiguous, and the three rows sum to one per column. The marginal three-way decomposition below then treats the full 30-cell grid with Type-II sums of squares.

Sampling uncertainty comes from the same stationary block bootstrap as Section 4.1, re-sampling common time blocks across all cells and recomputing the shares on each draw. At the monthly horizon the 90% intervals are wide but informative: in the Location×Scale panel, Location’s 0.51-0.91 and Scale’s 0.03-0.28 do not overlap, so Location’s lead is solid. In the Location×Shape panel, Shape’s 0.30-0.75 overlaps Location’s 0.10-0.45, so the ordering is less

obvious.

Effect (η^2)	R1M	R3M	R6M	R12M
<i>Location $\times$ Scale design (Shape fixed at identity)</i>				
Location	0.83	0.27	0.26	0.41
Scale	0.11	0.64	0.68	0.44
Location $\times$ Scale	0.07	0.09	0.06	0.15
<i>Location $\times$ Shape design (Scale fixed at none)</i>				
Location	0.25	0.20	0.48	0.47
Shape	0.52	0.54	0.37	0.40
Location $\times$ Shape	0.23	0.26	0.14	0.13

Table 5: η^2 **variance decomposition by horizon**, on the two fully-crossed 15 sub-designs (constructions of Section 3.2: cross-sectional FF6 in Location, floored trailing realised volatility in Scale). Within each panel the two main effects and their interaction sum to one per column, up to rounding. The largest share per panel and horizon is in bold.

These are *pairwise* shares, each to be read with its held-out axis in mind. In the Location $\times$ Scale design (Shape fixed at identity), Location leads at the monthly horizon (0.83): with the Shape axis removed, the benchmark is what remains to separate the identity-transformed labels. Cross-sectional Scale then takes over at 3-6 months (0.64, 0.68). In the Location $\times$ Shape design (Scale fixed at none) the distributional Shape leads at 1-3 months (0.52, 0.54) and Location at 6-12 months (0.48, 0.47). Because each pairwise share is inflated by the axis held out, the clean answer is the *marginal* 3-way decomposition below, which isolates each axis net of the other two.

Marginal contribution of each dimension (3-way ANOVA). The two 15-combination sub-designs in Table 5 give *pairwise* decompositions. In order to isolate each dimension's contribution *net of the other two*, we complete the cross to its full 30 *non-degenerate cells* (the 25-cell grid plus the 5 rank $\times$ vol cells, the only place a per-stock scale is not absorbed by the rank transform) and fit a 3-way Type-II ANOVA:

Marginal η^2	R1M	R3M	R6M	R12M
Location	0.25	0.14	0.35	0.48
Scale	0.05	0.15	0.33	0.16
Shape	0.38	0.23	0.09	0.09
residual (interactions + noise)	0.36	0.35	0.17	0.20

Table 6: **Marginal contribution of each step**: Type-II three-way ANOVA η^2 over the 30 non-degenerate factorial cells.

The picture is horizon-dependent: Shape is the largest marginal lever at short horizons (0.38 at 1M, 0.23 at 3M). Location takes over at long horizons (0.35 at 6M, 0.48 at 12M). Scale is the smallest axis overall. Interactions are sizeable at one month (residual 0.36). Unlike the panels of

Table 5, the columns need not sum to one: under Type-II sums of squares each axis is credited only with the variance it explains net of the other two, and in an incomplete grid these unique contributions and the residual do not partition the total exactly. The Scale effect is identified by the combinations where a scale is not absorbed: the three scale levels are distinct under the identity Shape, and the per-stock scale survives under the rank transform, which is what the five vol×rank-Gauss combinations contribute.

Multiplicity across cells (StepM and BY). At the monthly horizon, 27 of the 30 cells have a significant rank-IC at the 5% level on unadjusted two-sided Newey–West p -values. The Romano–Wolf StepM procedure (family-wise error rate 5%, one-sided) retains 25 of the 30; its statistic is the plain t -ratio of each combination’s rank-IC series, the serial and cross-cell dependence being carried by the common-block bootstrap that sets the critical values, not the Newey–West t of Table A1. The Benjamini–Yekutieli control of the false discovery rate at 5%, applied to the same p -values, retains the same 25. The five cells that fail both are the raw control, its standardised, sign and risk-scaled variants, and the sector-demeaned identity cell: only the compositions closest to the unengineered target drop out, and every rank-Gauss cell survives.

The confidence set (MCS). At R1M the 90% Model Confidence Set retains 16 of 30 cells: rank-Gauss or cross-sectionally standardised transforms on light neutralisation, including risk-scaled combinations under the floored construction of Section 3.2. The best is sector-neutral rank-Gauss (Sharpe ≈ 1.74), narrowly ahead of the unneutralised and market rank-Gauss combinations (≈ 1.69). At 3-12 months the MCS retains 20-28 of 30: the viable combinations become largely indistinguishable, reflecting both the smaller non-overlapping samples and the genuine result that the specific choice matters little at longer horizons. The set is tight where the data are informative and wide where they are not, as the procedure is designed to behave (Hansen et al., 2011).

Backtest-overfitting probability (PBO). PBO stays far below the 0.5 threshold at every horizon: 0.06 (R1M), 0.29 (R3M), 0.32 (R6M), 0.25 (R12M). The in-sample-best cell reliably lands in the upper half of the out-of-sample ranking.

Superiority over the raw benchmark (SPA). Hansen’s SPA test against the raw-return benchmark rejects the null at R1M ($p = 0.002$) and R3M ($p = 0.004$). At 6 and 12 months, the point estimates still favour the engineered labels but the test is insignificant ($p = 0.11$ and 0.21), where the raw label is less bad (Section 5.3) and the samples are small. The data-snooping-robust evidence is thus stronger for the shorter horizons. We recall that Table A1 lists the results over the whole grid.

5.5 Sensitivity to seed choice

The only stochastic component of the learner is the tree ensemble’s row and feature subsampling, whose seed is fixed, so the walk-forward is exactly reproducible. To confirm that the ranking does not depend on that seed, we re-estimate each of the 30 combinations 100 times with distinct random states on a single expanding split (train ≤ 2017 , test 2018-2026). The dispersion is small and uniform: the seed-induced standard deviation of the decile Sharpe is 0.06-0.14 across all 30 options, against a spread of roughly 1.3 between the best and worst combinations, and the ordering remains the same under every seed (Figure 2). Hence our results hold, irrespective of the random seed.

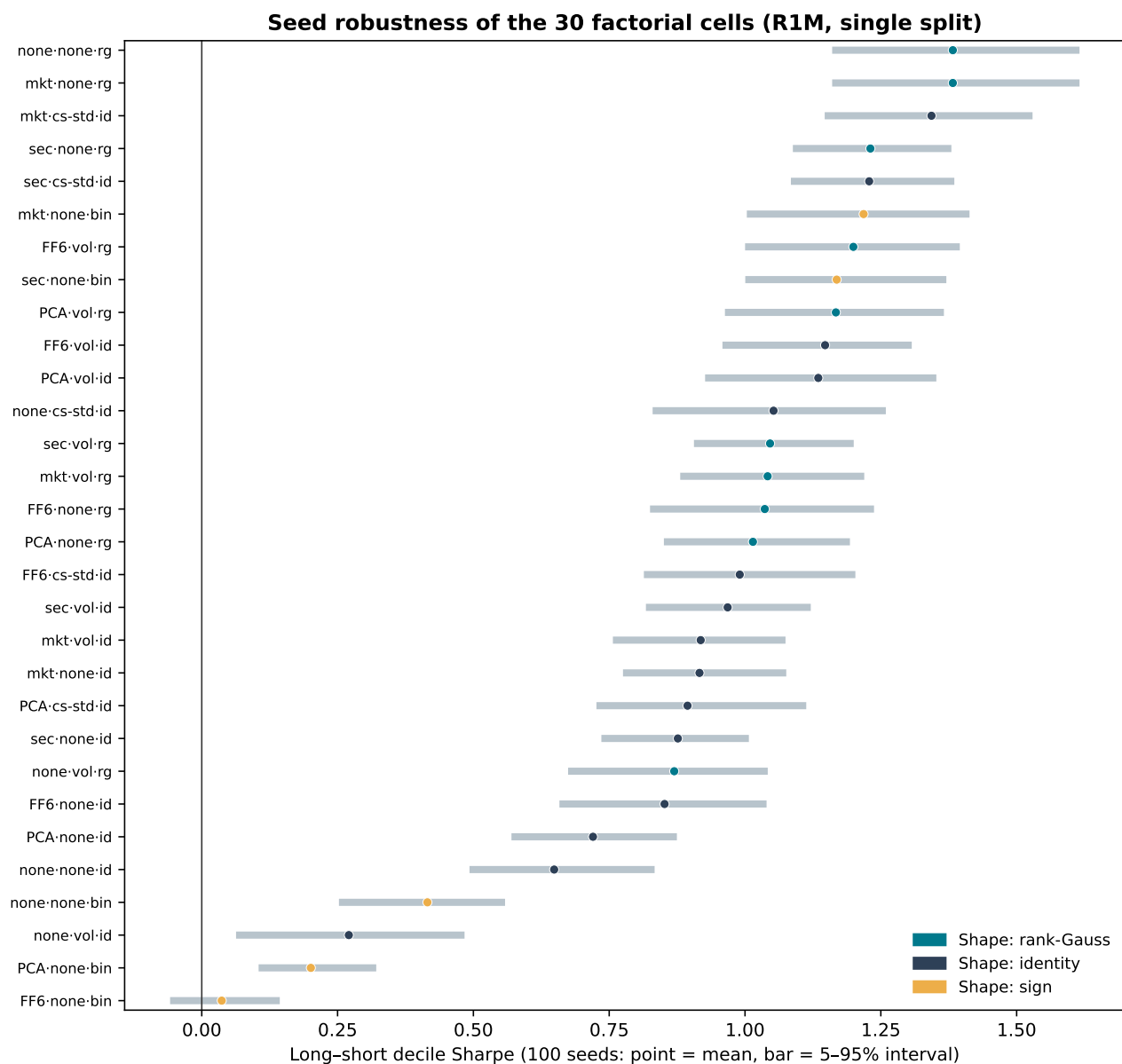


Figure 2: Robustness to sampling. Each of the 30 combinations refit under 100 random seeds on a single expanding split (train ≤ 2017 , test 2018-2026; levels therefore differ from Table 3). Points are mean long-short Sharpe ratios, bars the 5-95% seed intervals, colours the Shape. The intervals are narrow relative to the gaps between cells: the ranking is not impacted much by random seeds.

5.6 Feature importance across labels and horizons

The comparisons so far establish that the label impacts performance. This section tries to figure out *how*. If different labels led the model to select materially different features, varying the label would alter the information set. For every (label, horizon) pair, we fit the fixed tree regressor on the full labeled sample, then record the ensemble’s feature importance values for the 122 signals, and finally ask whether labels change the feature set.

Across labels (fixed horizon). For monthly returns, the ranking of important features is broadly shared across the standardised and rank labels: momentum (rel_str_252d, mom_252; Jegadeesh and Titman, 1993), short-horizon volatility (vol_22, idio_vol_252d), analyst breadth (EPS_FY1_Breadth) and dividend yield lead for most, echoing the dominant signals in Gu et al. (2020). It seems that labels reweight the *loss*, but not the feature set. The neutralised labels (sector/PCA/FF6 residual) mainly shed forward-earnings breadth and dividend yield, and lean instead on cash-flow quality and trading-activity signals. The per-stock volatility scale is the one operator that visibly changes the feature set rather than reweighting it, pulling risk measures such as total volatility and beta into the top ranks.

Across horizons (label-averaged). The important features rotate with the horizon (Figure 3): technical and analyst-flow signals (EPS_FY1_Breadth, vol_22, rel_str_252d) dominate at one month, while slow fundamentals (Div_yld, size_log, EPS) rise toward twelve months. Longer labels are better matched by more persistent predictors (Coqueret and Guida, 2020), and this explains the compression of Section 5.3. The distributional operators are effective because they shrink the noise of the monthly cross-section, hence as the model shifts onto slower, less noise-dominated signals, there is less for a Scale or Shape operator to improve on.

5.7 Model-agnosticism: the impact of the learner

Everything above uses gradient-boosted trees. To rule out a tree artefact, we re-run the horse-race under a penalised linear model (elastic-net) and a small feed-forward network (a 64-32 ReLU MLP, hereafter MLP-1), and add a second network of the same architecture implemented independently end to end (MLP-2: different framework, optimiser and initialisation). Results are gathered in Table 7. Features, labels, walk-forward, embargo and evaluation are identical to Section 4. Because linear and neural learners need finite, comparably scaled inputs, each refit fits a median imputer and a standardiser on the training window only, and standardises the target within that window. Such an affine map leaves every cross-sectional ranking, hence every decile portfolio, unchanged. All labels use the constructions of Section 3.2.

Table 7 makes two points. First, the same labels lead under all three models: the z-score / rank-Gauss family and the cross-sectionally constructed residual targets occupy the top of every column, while the raw forward return sits in the bottom quartile of each (13th of 15 under the trees and the elastic-net, 14th under MLP-1). The order within the leading group shifts, but its membership does not. Second, a two-way variance decomposition of the forty-five label $\times$ learner Sharpes attributes $\eta^2 = 0.87$ to the label and 0.04 to the learner (residual 0.09), roughly twenty times as much: a fixed label’s Sharpe changes by a median of 0.27 across learners, against a 1.0-1.3 spread across labels within any learner. The Spearman correlation with the tree ordering is 0.92 for the elastic-net and 0.74 for MLP-1, with the disagreements concentrated within

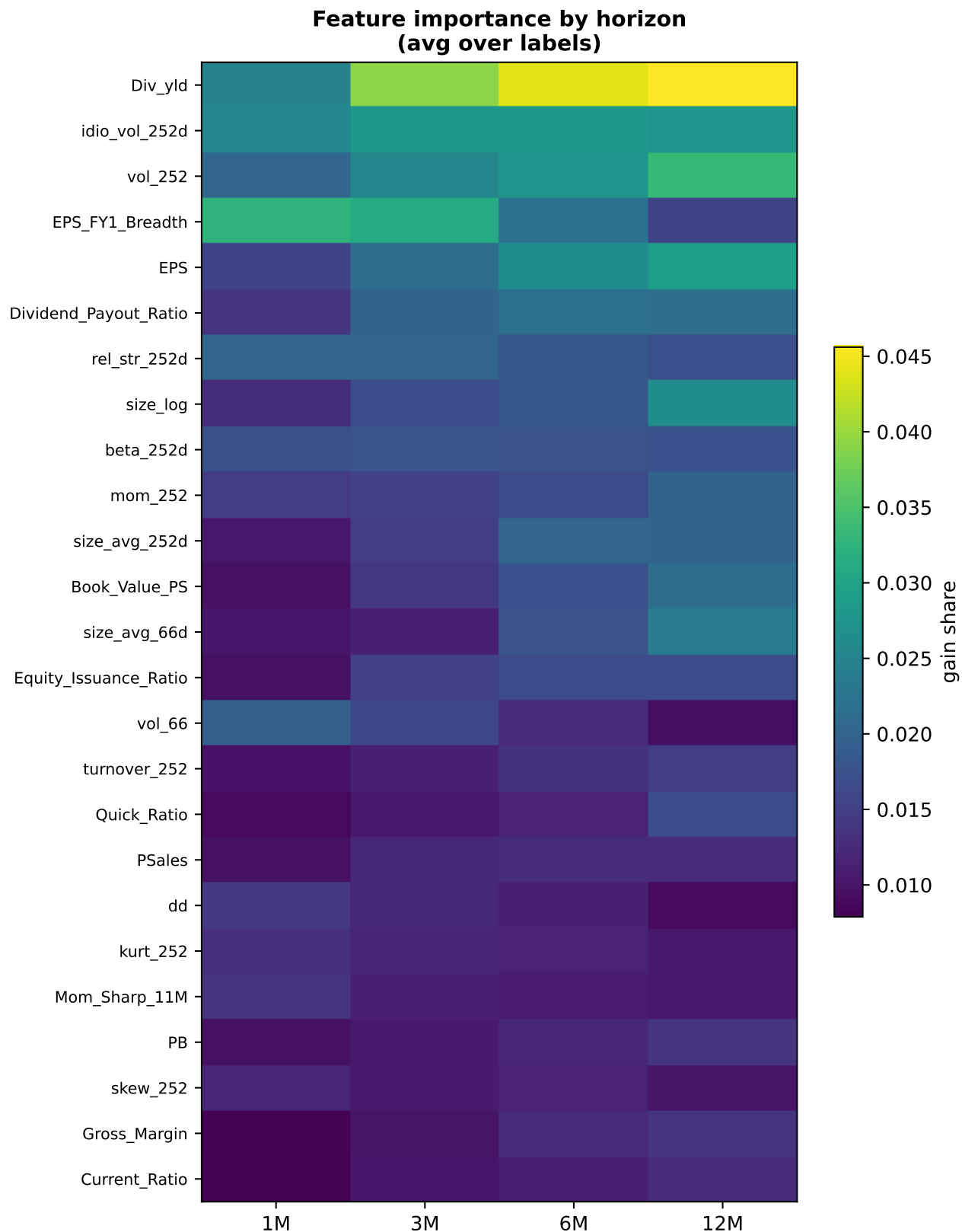


Figure 3: **Tree-ensemble gain share per feature**, averaged over labels, by horizon. Dividend yield and idiosyncratic and total volatility dominate at longer horizons while forward-earnings breadth leads at one month.

Label	GBT	elastic-net	MLP-1
rank_gauss	1.69 (1)	1.52 (1)	1.31 (7)
sector_zscore	1.56 (2)	1.32 (4)	1.43 (5)
zscore	1.53 (3)	1.41 (2)	1.45 (=3)
robust_z	1.46 (4)	1.19 (=6)	1.45 (=3)
beat_median	1.35 (5)	1.27 (5)	1.22 (9)
percentile	1.34 (6)	1.38 (3)	1.37 (6)
ff6_cs †	1.28 (=7)	1.10 (8)	1.48 (2)
pca_resid_pct	1.28 (=7)	1.19 (=6)	1.51 (1)
pca_resid	1.16 (9)	0.87 (=9)	1.23 (8)
demean	1.15 (10)	0.86 (11)	1.00 (11)
sector_demean	1.06 (11)	0.82 (12)	1.04 (10)
median_demean	1.05 (12)	0.87 (=9)	0.92 (12)
raw (control)	0.68 (13)	0.76 (13)	0.72 (14)
vol_scaled_rank †	0.66 (14)	0.22 (15)	0.51 (15)
vol_scaled †	0.36 (15)	0.57 (14)	0.83 (13)

Table 7: **Label ranking across learners** (R1M, out-of-sample long-short Sharpe), with each label’s within-learner rank in parentheses; rows ordered by the tree result. Labels with equal Sharpe at the reported precision share a rank, marked “=”. † marks a Section 3.2 construction.

the weakest, volatility-scaled labels.

Across horizons. Table 8 repeats the decomposition at each horizon over the twelve core labels. The label dominates everywhere: η^2_{label} runs from 0.84 at one month to 0.59 at twelve, against η^2_{learner} of 0.04-0.27, and the margin narrows as the horizon increases, consistent with the findings of Section 5.3. The orderings agree as well: ρ with the trees is 0.74-0.87 for the elastic-net, and MLP-2, implemented independently of MLP-1, tracks the trees at 0.89, 0.82 and 0.83 from one to six months before loosening at the sample-thin annual horizon (0.49). At three months it is steadier than MLP-1 (0.82 against 0.41). The fact that an independent neural implementation reproduces the ordering corroborates our baseline findings.

Horizon	$\eta^2(\text{label})$	$\eta^2(\text{learner})$	$\rho(\text{enet})$	$\rho(\text{MLP-1})$	$\rho(\text{MLP-2})$
R1M	0.84	0.06	0.87	0.71	0.89
R3M	0.70	0.04	0.76	0.41	0.82
R6M	0.67	0.18	0.78	0.89	0.83
R12M	0.59	0.27	0.74	0.61	0.49

Table 8: **Model-agnosticism by horizon** (twelve core labels). η^2 from a two-way (label $\times$ learner) decomposition of the out-of-sample Sharpes over {GBT, elastic-net, MLP-1}. ρ is the Spearman rank correlation of each learner’s label ordering with GBT. MLP-2 is the independently implemented network of the same architecture, evaluated at all four horizons.

6 Conclusion

In this paper, we treat the prediction target as a crucial design variable. On a 24-year panel of 1,960 US stocks, we hold the 122 characteristics and the learner fixed, decompose the label into Location, Scale and Shape operators, and vary only that composition. The choice is essential: the long-short decile Sharpe moves from 0.68 for the raw forward return (with rank information coefficient indistinguishable from zero), to 1.69 for the Gaussianised cross-sectional rank and to 1.74 for the best labeling option.

The coding of the target explains the largest share of the variance in performance at the monthly horizon. The label ordering is robust to a change of model, with the label explaining at least twice the out-of-sample variance the learner does at every horizon, and many times more at the shorter ones. Results hold even upon re-estimation under one hundred random seeds, under correlation-aware inference against multiplicity, backtest overfitting and data snooping.

These conclusions are useful to anyone building supervised models for the cross-section of stock returns: the target deserves the same engineering scrutiny the field already applies to the features. We see three directions for future research. The first is a net-of-cost re-ranking built from the saved turnover. The second is replication in other markets and at other frequencies, where the cross-market evidence of [Cakici and Zaremba \(2026\)](#) suggests that our conclusions replicate externally. The third is learning objectives, such as learning-to-rank or quantile losses, which target the leading Shape directly.

A The full factorial grid

The 30 non-degenerate cells of the $L \times S \times T$ factorial (Section 4.1), each trained and scored under the identical protocol of Section 4, are listed below in full and ranked by out-of-sample long-short decile Sharpe (R1M, 194 months). Location, Scale and Shape give the cell’s coordinates; rank-IC and its Newey–West t are on the non-overlapping series, and PSR is the Probabilistic Sharpe Ratio of the long-short spread (the probability that the true Sharpe ratio exceeds zero given the sample length, skewness and kurtosis), the per-cell layer 1 of the inference stack of Section 4.1. The compositions that the curated Table 3 does not name individually, for instance sector-neutral rank-Gauss (the top cell) and the residual-over-idiosyncratic-volatility cells (PCA or FF6 location with a trailing-vol scale), are all present here.

Location	Scale	Shape	rank-IC	IC t	Sharpe	PSR
sector	none	rank-Gauss	0.029	5.5	1.74	1.000
raw	none	rank-Gauss	0.032	6.4	1.69	1.000
market	none	rank-Gauss	0.032	6.4	1.69	1.000
sector	cs-std	identity	0.024	4.1	1.65	1.000
market	cs-std	identity	0.024	4.2	1.53	1.000
FF6	vol	rank-Gauss	0.023	4.8	1.46	1.000
FF6	none	rank-Gauss	0.026	5.8	1.46	1.000
FF6	cs-std	identity	0.022	4.7	1.41	1.000
PCA	vol	rank-Gauss	0.018	3.4	1.41	1.000
sector	none	sign	0.029	5.3	1.40	1.000
PCA	vol	identity	0.018	3.3	1.38	1.000
FF6	vol	identity	0.019	4.1	1.33	1.000
PCA	none	rank-Gauss	0.022	4.6	1.32	1.000
market	none	sign	0.028	4.9	1.28	1.000
FF6	none	identity	0.018	3.9	1.28	1.000
market	vol	rank-Gauss	0.024	3.3	1.25	1.000
PCA	cs-std	identity	0.015	2.9	1.25	1.000
sector	vol	rank-Gauss	0.022	3.2	1.24	1.000
raw	vol	rank-Gauss	0.028	4.7	1.19	1.000
sector	vol	identity	0.018	2.7	1.18	1.000
PCA	none	identity	0.015	2.7	1.16	1.000
market	none	identity	0.019	2.8	1.15	1.000
market	vol	identity	0.020	3.0	1.12	1.000
raw	cs-std	identity	0.009	1.5	1.09	1.000
sector	none	identity	0.016	2.5	1.06	1.000
raw	none	identity	0.006	0.9	0.68	0.998
raw	none	sign	0.014	2.1	0.44	0.959
raw	vol	identity	0.008	1.3	0.36	0.921
PCA	none	sign	0.013	2.6	0.07	0.607
FF6	none	sign	0.015	2.9	-0.09	0.360

Table A1: All 30 factorial combinations (R1M, out-of-sample), ranked by long-short decile Sharpe. PSR is the Probabilistic Sharpe Ratio of the long-short spread.

B The feature panel by family

The 122 firm characteristics are held fixed throughout the paper, and Section 3 describes their common preprocessing; this appendix documents what they are without enumerating all of them. Table B1 groups the panel into thirteen economic families, with the count and representative members of each.

Family	Count	Representative members
Momentum and trend	11	mom_252, rel_str_252d, trend_composite
Risk and return distribution	9	vol_252, idio_vol_252d, beta_252d
Size	3	size_log, size_avg_252d
Liquidity and trading activity	9	dollar_turnover, adv_252, abn_volume_21
Valuation	13	PB, PE, EV_EBITDA
Payout and issuance	5	Div_yld, Buyback_Yield, Equity_Issuance_Ratio
Profitability: margins	11	Gross_Margin, EBIT_Margin, Net_Profit_Margin
Profitability: returns on capital	14	ROE, ROIC, GP_to_Assets_Novy_Marx
Cash-flow generation	10	OCF_to_Assets, FCF_to_Assets, FCF_to_NOA
Earnings quality and operating efficiency	7	Sloan_Accruals, Net_Income_to_CFO, Asset_Turn
Growth and investment	12	Asset_Growth_1Y, Sales_Growth_1Y, Capex_Intensity
Leverage and solvency	11	Debt_Assets, Interest_Coverage, Current_Ratio
Analyst estimates and surprises	7	EPS_FY1_Rev30d, EPS_Quarterly_Surprise, Est_Revenue_Growth
Total	122	

Table B1: The 122 characteristics, grouped into thirteen families.

The families deliberately over-represent the fundamental dimensions (margins, returns on capital, cash-flow generation) relative to the price-based ones. Section 5.6 nonetheless shows that the models' attention to features changes with the horizon, from the technical and analyst-flow families at one month toward the slower fundamental families as the horizon increases to twelve months.

References

- Clifford S. Asness, R. Burt Porter, and Ross L. Stevens. Predicting stock returns using industry-relative firm characteristics. *SSRN Working Paper*, 213872, 2000. doi: 10.2139/ssrn.213872.
- George S. Atsalakis and Kimon P. Valavanis. Surveying stock market forecasting techniques – part II: Soft computing methods. *Expert Systems with Applications*, 36(3):5932–5941, 2009. doi: 10.1016/j.eswa.2008.07.006.
- Marco Avellaneda and Jeong-Hyun Lee. Statistical arbitrage in the US equities market. *Quantitative Finance*, 10(7):761–782, 2010. doi: 10.1080/14697680903124632.
- Vitor Azevedo, Georg Sebastian Kaiser, and Sebastian Müller. Stock market anomalies and machine learning across the globe. *Journal of Asset Management*, 24(5):419–441, 2023. doi: 10.1057/s41260-023-00318-z.
- Yang Bai and Kuntara Pukthuanthong. Machine learning classification and portfolio construction: Does the loss function matter? *Working paper*, 2108.02283, 2021. doi: 10.48550/arXiv.2108.02283. arXiv v7, revised September 2026; earlier versions carried different titles.
- David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. The probability of backtest overfitting. *Journal of Computational Finance*, 20(4):39–69, 2017. doi: 10.21314/jcf.2016.322. published online September 2016; print issue April 2017.
- Pedro Barroso and Pedro Santa-Clara. Momentum has its moments. *Journal of Financial Economics*, 116(1):111–120, 2015. doi: 10.1016/j.jfineco.2014.11.010.
- T. Mark Beasley, Stephen Erickson, and David B. Allison. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics*, 39(5):580–595, 2009. doi: 10.1007/s10519-009-9281-0.
- David Blitz, Joop Huij, and Martin Martens. Residual momentum. *Journal of Empirical Finance*, 18(3):506–521, 2011. doi: 10.1016/j.jempfin.2011.01.003.
- David Blitz, Joop Huij, Simon Lansdorp, and Marno Verbeek. Short-term residual reversal. *Journal of Financial Markets*, 16(3):477–504, 2013. doi: 10.1016/j.finmar.2012.10.005.
- David Blitz, Matthias X. Hanauer, and Milan Vidojevic. The idiosyncratic momentum anomaly. *International Review of Economics & Finance*, 69:932–957, 2020. doi: 10.1016/j.iref.2020.05.008.
- Mark Britten-Jones, Anthony Neuberger, and Ingmar Nolte. Improved inference in regression with overlapping observations. *Journal of Business Finance & Accounting*, 38(5–6):657–683, 2011. doi: 10.1111/j.1468-5957.2011.02244.x.
- Svetlana Bryzgalova, Markus Pelger, and Jason Zhu. Forest through the trees: Building cross-sections of stock returns. *Journal of Finance*, 80(5):2447–2506, 2025. doi: 10.1111/jofi.13477.
- Nusret Cakici and Adam Zaremba. Getting the target right in return prediction. *SSRN Working Paper*, 6615698, 2026. doi: 10.2139/ssrn.6615698.

- Nusret Cakici, Christian Fieberg, Daniel Metko, and Adam Zaremba. Machine learning goes global: Cross-sectional return predictability in international stock markets. *Journal of Economic Dynamics and Control*, 155:104725, 2023a. doi: 10.1016/j.jedc.2023.104725.
- Nusret Cakici, Christian Fieberg, Daniel Metko, and Adam Zaremba. Predicting returns with machine learning across horizons, firm size, and time. *Journal of Financial Data Science*, 5(4): 119–144, 2023b. doi: 10.3905/jfds.2023.1.139.
- Mark M. Carhart. On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82, 1997. doi: 10.1111/j.1540-6261.1997.tb03808.x.
- Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Management Science*, 70(2):714–750, 2024. doi: 10.1287/mnsc.2023.4695.
- Minghui Chen, Matthias X. Hanauer, and Tobias Kalsbach. Design choices, machine learning, and the cross-section of stock returns. *SSRN Working Paper*, 5031755, 2026. doi: 10.2139/ssrn.5031755. first posted November 2024; revised June 2026.
- Guillaume Coqueret. Persistence in factor-based supervised learning models. *Journal of Finance and Data Science*, 8:12–34, 2022. doi: 10.1016/j.jfds.2021.10.002.
- Guillaume Coqueret and Tony Guida. *Machine learning for factor investing*. Chapman & Hall/CRC, 2020. doi: 10.1201/9781003034858.
- Kent Daniel and Tobias J. Moskowitz. Momentum crashes. *Journal of Financial Economics*, 122(2):221–247, 2016. doi: 10.1016/j.jfineco.2015.12.002.
- Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies*, 22(5):1915–1953, 2009. doi: 10.1093/rfs/hhm075.
- Wolfgang Drobetz and Tizian Otto. Empirical asset pricing via machine learning: Evidence from the European stock market. *Journal of Asset Management*, 22(7):507–538, 2021. doi: 10.1057/s41260-021-00237-x.
- Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015. doi: 10.1016/j.jfineco.2014.10.010.
- Eugene F. Fama and James D. MacBeth. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3):607–636, 1973. doi: 10.1086/260061.
- Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669, 2018. doi: 10.1016/j.ejor.2017.11.054.
- Joachim Freyberger, Andreas Neuhierl, and Michael Weber. Dissecting characteristics nonparametrically. *Review of Financial Studies*, 33(5):2326–2377, 2020. doi: 10.1093/rfs/hhz123.
- Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273, 2020. doi: 10.1093/rfs/hhaa009.

- Shihao Gu, Bryan Kelly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450, 2021. doi: 10.1016/j.jeconom.2020.07.009.
- Jorge Guijarro-Ordóñez, Markus Pelger, and Greg Zanolli. Deep learning statistical arbitrage. *Management Science*, 72(9):7502–7549, 2026. doi: 10.1287/mnsc.2022.03132. published online December 2025; print issue September 2026.
- Matthias X. Hanauer and Tobias Kalsbach. Machine learning and the cross-section of emerging market stock returns. *Emerging Markets Review*, 55:101022, 2023. doi: 10.1016/j.ememar.2023.101022.
- Matthias X. Hanauer, Amar Soebhag, Marc Stam, and Tobias Hoogteijling. Do machine learning models need to be sector experts? *SSRN Working Paper*, 5224253, 2026. doi: 10.2139/ssrn.5224253. first posted May 2025; revised February 2026.
- Lars Peter Hansen and Robert J. Hodrick. Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. *Journal of Political Economy*, 88(5):829–853, 1980. doi: 10.1086/260910.
- Peter R. Hansen, Asger Lunde, and James M. Nason. The model confidence set. *Econometrica*, 79(2):453–497, 2011. doi: 10.3982/ecta5771.
- Peter Reinhard Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005. doi: 10.1198/073500105000000063.
- Campbell R. Harvey, Edward Hoyle, Russell Korgaonkar, Sandy Rattray, Matthew Sargaison, and Otto van Hemert. The impact of volatility targeting. *Journal of Portfolio Management*, 45(1):14–33, 2018. doi: 10.3905/jpm.2018.45.1.014.
- Nicolas Huck. Pairs selection and outranking: An application to the S&P 100 index. *European Journal of Operational Research*, 196(2):819–825, 2009. doi: 10.1016/j.ejor.2008.03.025.
- Nicolas Huck. Pairs trading and outranking: The multi-step-ahead forecasting case. *European Journal of Operational Research*, 207(3):1702–1716, 2010. doi: 10.1016/j.ejor.2010.06.043.
- Michael Jahrer. 1st place solution: Porto Seguro’s safe driver prediction (RankGauss normalization). Kaggle competition discussion forum, 2017. URL <https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/discussion/44629>.
- Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48(1):65–91, 1993. doi: 10.1111/j.1540-6261.1993.tb04702.x.
- Bryan Kelly, Semyon Malamud, and Kangying Zhou. The virtue of complexity in return prediction. *Journal of Finance*, 79(1):459–503, 2024. doi: 10.1111/jofi.13298.
- Bryan T. Kelly, Seth Pruitt, and Yinan Su. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524, 2019. doi: 10.1016/j.jfineco.2019.05.001.

- Abby Y. Kim, Yiuman Tse, and John K. Wald. Time series momentum and volatility scaling. *Journal of Financial Markets*, 30:103–124, 2016. doi: 10.1016/j.finmar.2016.05.003.
- Christopher Krauss, Xuan Anh Do, and Nicolas Huck. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2):689–702, 2017. doi: 10.1016/j.ejor.2016.10.031.
- Markus Leippold, Qian Wang, and Wenyu Zhou. Machine learning in the Chinese stock market. *Journal of Financial Economics*, 145(2):64–82, 2022. doi: 10.1016/j.jfineco.2021.08.017.
- Yan Lin, Yihong Su, and Yi Yang. LambdaRankIC: Directly optimizing rank IC for financial prediction. *Working paper*, 2605.00501, 2026. doi: 10.48550/arXiv.2605.00501.
- Marcos López de Prado. *Advances in financial machine learning*. Wiley, 2018.
- Marcos López de Prado. *Machine learning for asset managers*. Cambridge University Press, 2020. doi: 10.1017/9781108883658.
- Zachary R. McCaw, Jacqueline M. Lane, Richa Saxena, Susan Redline, and Xihong Lin. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*, 76(4):1262–1272, 2020. doi: 10.1111/biom.13214.
- Benjamin Moritz and Tom Zimmermann. Tree-based conditional portfolio sorts: The relation between past and future stock returns. *SSRN Working Paper*, 2740751, 2016. doi: 10.2139/ssrn.2740751.
- Tobias J. Moskowitz and Mark Grinblatt. Do industries explain momentum? *Journal of Finance*, 54(4):1249–1290, 1999. doi: 10.1111/0022-1082.00146.
- Tobias J. Moskowitz, Yao Hua Ooi, and Lasse Heje Pedersen. Time series momentum. *Journal of Financial Economics*, 104(2):228–250, 2012. doi: 10.1016/j.jfineco.2011.11.003.
- Dennis Olson and Charles Mossman. Neural network forecasts of Canadian stock returns using accounting ratios. *International Journal of Forecasting*, 19(3):453–465, 2003. doi: 10.1016/S0169-2070(02)00058-4.
- Daniel Poh, Bryan Lim, Stefan Zohren, and Stephen Roberts. Building cross-sectional systematic strategies by learning to rank. *Journal of Financial Data Science*, 3(2):70–86, 2021. doi: 10.3905/jfds.2021.1.060.
- Daniel Poh, Bryan Lim, Stefan Zohren, and Stephen Roberts. Enhancing cross-sectional currency strategies by context-aware learning to rank with self-attention. *Journal of Financial Data Science*, 4(3):89–107, 2022. doi: 10.3905/jfds.2022.1.099.
- Daniel Poh, Stephen Roberts, and Stefan Zohren. Transfer ranking in finance: applications to cross-sectional momentum with data scarcity. *Journal of Systematic Investing*, 3(1), 2023. doi: 10.52354/jsi.3.1.ii.

- Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.
- Keywan Christian Rasekhschaffe and Robert C. Jones. Machine learning for stock selection. *Financial Analysts Journal*, 75(3):70–88, 2019. doi: 10.1080/0015198x.2019.1596678.
- Robert A. Rigby and D. Mikis Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005. doi: 10.1111/j.1467-9876.2005.00510.x.
- Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005. doi: 10.1111/j.1468-0262.2005.00615.x.
- Lawrence Takeuchi and Yu-Ying Lee. *Applying deep learning to enhance momentum trading strategies in stocks*. Technical report, CS229, Stanford University, 2013.
- Wee Ling Tan, Stephen Roberts, and Stefan Zohren. Spatio-temporal momentum: Jointly learning time-series and cross-sectional strategies. *Journal of Financial Data Science*, 5(3):107–129, 2023. doi: 10.3905/jfds.2023.1.130.
- Ondrej Tobek and Martin Hronec. Does it pay to follow anomalies research? machine learning approach with international evidence. *Journal of Financial Markets*, 56:100588, 2021. doi: 10.1016/j.finmar.2020.100588.
- Halbert White. Economic prediction using neural networks: The case of IBM daily stock returns. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 2, pages 451–458. IEEE, 1988. doi: 10.1109/icnn.1988.23959.
- Dominik Wolff and Fabian Echterling. Stock picking with machine learning. *Journal of Forecasting*, 43(1):81–102, 2024. doi: 10.1002/for.3021.
- Xin Zhang, Lan Wu, and Zhixue Chen. Constructing long-short stock portfolio with a new listwise learn-to-rank algorithm. *Quantitative Finance*, 22(2):321–331, 2022. doi: 10.1080/14697688.2021.1939117.