Boosted Trees, Clustered Risk: a machine-learning long–short strategy for U.S. equities

Angel Zhao · Yale School of Management · MGT 924 Statistical Foundations, Problem Set 3 · Submission to the JKP Common Task Framework · October 2026

3.74	**[3.19, 4.33]**	**1.38**	**12.44**	**70 bp**
Sharpe ratio, test sample Dec 1989 – Nov 2023	95% bootstrap interval for that Sharpe ratio	Same forecasts, naive rank weights, no risk model	t-stat of the CAPM alpha vs the 1/N market	break-even trading cost per dollar traded

Summary. The strategy forecasts each stock's next-month excess return from all 402 characteristics in the competition data and turns the forecasts into a dollar- and market-beta-neutral portfolio with a mean–variance optimizer. Two ingredients do the work. Gradient-boosted trees, blended with a ridge regression, capture non-linear and interaction effects among the characteristics; and the optimizer prices risk with a covariance matrix of about 2,060 stocks estimated in two complementary ways — a shrunk sample covariance of daily returns, and a Barra-style factor model whose style factors are discovered by hierarchical clustering of the characteristics. Models are refit yearly on trailing data only, and the configuration that trades in the competition's test sample was chosen by the code on the ten years before it. Over the 34-year held-out test sample the final portfolio earns an annualised Sharpe ratio of **3.74**, against 1.38 for the same forecasts with rank weights, 1.08 for a linear (ridge) benchmark and 0.49 for the 1/N market; it is still 3.20 at 10 basis points of trading costs.

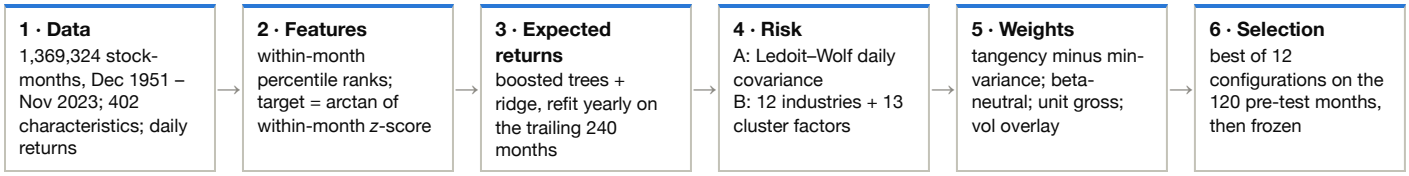


Figure 1. Pipeline. Every step is a deterministic function of the competition data; the portfolio for month *t* uses only information dated *t* or earlier.

1 Data and features

The competition data contain 1,369,324 stock-months of 12,325 relatively liquid U.S. common stocks from Dec 1951 to Nov 2023 with the 402 characteristics of Jensen, Kelly and Pedersen (2023), each stock's excess return over the following month, and daily excess returns, which I use only to estimate risk. The model uses data from Dec 1959 onward, when its first 240-month training window starts. Every characteristic and every stock enters; nothing is selected by hand. Each characteristic is converted to its **percentile rank within the month**, which removes differences in scale across characteristics and over time and neutralises outliers. Missing values stay missing, so the trees can learn from missingness, and the linear model sets them to the median. The training target is arctan(*z*), where *z* is the next-month return standardised within the month: standardising strips out the market-wide component that a dollar-neutral portfolio does not trade, and the arctan bounds the pull of extreme returns.

2 Expected returns

Gradient-boosted trees (GBM). scikit-learn's histogram gradient boosting with 500 trees, learning rate 0.02, up to 255 leaves per tree and at least 2,000 observations per leaf. The large leaf minimum is the main regulariser: trees are deep enough to model interactions (e.g. reversal that depends on volatility or size), but every leaf averages thousands of noisy returns. **Ridge regression** on the centred ranks is the stable linear counterpart; its penalty is chosen inside each training window by time-series cross-validation on the window's last 60 months (criterion: mean monthly rank correlation; chosen values 0.001–10 times the rank variance). Both models are re-estimated every 12 months on the trailing 240 months (on average 442,000 stock-months; one refit of both models takes about 5 minutes on a laptop) and then forecast the next 12 months. Forecasts are standardised within the month and combined as 0.7 GBM + 0.3 ridge (the “blend”).

3 Risk models

With about 2,060 stocks but only 756 days of returns, the sample covariance matrix is singular, and a mean–variance optimizer would exploit its errors. I estimate risk in two complementary ways. **Model A** shrinks the sample covariance of the last 756 daily excess returns toward a scaled identity (Ledoit and Wolf, 2004). **Model B** is a Barra-style factor model, $\Sigma = X\Omega X' + D$, following Jensen, Kelly, Malamud and Pedersen (2024). Its 25 factors are the 12 Fama–French industries plus 13 *style* factors that are discovered rather than chosen. Every 12 months the 402 characteristics are grouped by average-linkage hierarchical clustering of their rank-correlation matrix over the last 60 months (distance $1 - |\rho|$), as Jensen, Kelly and Pedersen (2023) do for factor returns (I cluster the characteristics themselves, because exposures are what enter a risk model). A cluster's exposure is the average rank of its members after aligning their signs with the cluster's first principal component, standardised within the month. In practice the clusters are unbalanced: at the last refit one broad cluster holds 341 of the 402 characteristics and contrasts large, liquid, low-beta, cheap firms with small, illiquid, risky ones; a second holds 23 momentum, reversal and earnings-news signals; a

third holds 6 skewness measures; and ten small clusters hold narrow accounting signals. Model B is therefore close to an industry model with three broad style factors. Daily cross-sectional ridge regressions of returns on the exposures give factor returns and residuals; Ω is an exponentially weighted covariance of factor returns (half-life 504 days for correlations, 84 days for volatilities) and D an exponentially weighted residual variance (half-life 84 days). Model A sees every source of co-movement but noisily; model B is stable but only sees the risks it is given. Averaging their portfolios diversifies the two kinds of estimation error.

4 From forecasts to weights

Given forecasts μ (within-month *z*-scores) and a covariance matrix Σ, the book is the tangency portfolio minus the global minimum-variance portfolio under $\Sigma_\lambda = \Sigma + \lambda I$, with λ equal to the average stock variance: $w = \Sigma_\lambda^{-1}\mu - (1'\Sigma_\lambda^{-1}\mu \,/\, 1'\Sigma_\lambda^{-1}1)\,\Sigma_\lambda^{-1}1$, up to scale. This is the mean–variance-efficient zero-cost portfolio, and the ridge term pulls it toward the forecasts themselves, which guards against error maximisation. For model B the inverse uses the Woodbury identity. Each book is then projected to zero market beta (trailing 252-day beta on the equal-weighted universe) and scaled to unit gross exposure, $\Sigma|w| = 1$. An optional overlay scales month *t* by the ratio of the median past ex-ante volatility to the current one (from model B), clipped to [0.5, 2]. This gives twelve configurations: {GBM, blend} × {A, B, A+B} × {overlay, none}.

5 Validation protocol and safeguards

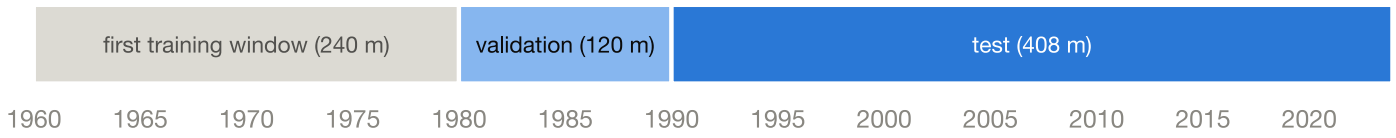


Figure 2. Estimation scheme. Forecasting models and clusters are refit yearly on trailing data and risk models monthly. The 12 configurations are compared on the 120 pre-test months; the best (GBM + ridge blend, risk models A+B, vol overlay) is frozen for the test sample.

- **No look-ahead.** The forecast for month *t* uses models trained on months up to *t*–1, whose targets are realised by *t*; risk models use returns up to the last trading day of month *t*; the configuration choice uses only pre-test months. I verified this the way the competition does: re-running `main()` on truncated data reproduces every earlier weight, on synthetic data, on a 60-stock extract of the real data, and on the full real data cut at Dec 1992 (all 402 characteristics, 37 test months: largest difference $\approx 10^{-16}$, i.e. floating-point rounding).
- **Determinism.** Inputs are sorted canonically and nothing is random (fixed seed, no subsampling): repeated runs and runs on shuffled inputs give identical weights.
- **Assumptions.** The objective is the Sharpe ratio of the monthly return $\Sigma_i\, w_{i,t}\, r_{i,t+1}$ over the months flagged `ctff_test`, before trading costs. As it is scale-free, weights are normalised to unit gross exposure each month (the overlay varies this scale).
- **No cherry-picking, honest selection.** All stocks and characteristics enter; the only performance-based choice (1 of 12 configurations) is made by the code on pre-test data. Selection inflates the winner's validation Sharpe ratio (6.62), but its deflated Sharpe ratio (Bailey and López de Prado, 2014) — the probability of a positive true Sharpe ratio given 12 trials — exceeds 0.99.

6 Results

I built the strategy up from the assignment's suggested baseline (ranked characteristics, ridge forecast, long–short portfolio), adding one ingredient at a time; Figure 4 decomposes the Sharpe ratio along these steps. All results are out of sample: each month's portfolio uses only models estimated on earlier data.

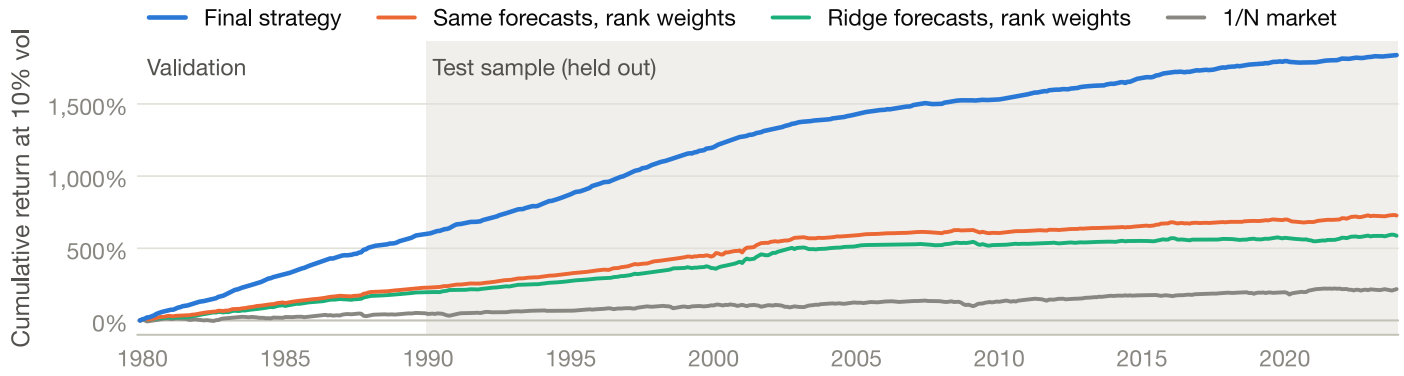


Figure 3. Cumulative performance. Cumulative sum of monthly returns, each series scaled ex post to 10% annualised volatility so that slopes are proportional to Sharpe ratios. Shaded: held-out test sample.

Table 1. Performance. Monthly returns of unit-gross portfolios before trading costs; "rank weights" are dollar-neutral weights linear in the forecast's cross-sectional rank. Intervals: moving-block bootstrap (6-month blocks, 10,000 draws); alpha t-statistics use Newey–West (6 lags) standard errors; drawdowns are computed after scaling each series to 10% annualised volatility.

	Final strategy	Same forecasts, rank weights	Ridge forecasts, rank weights	1/N market
Test sample, Dec 1989 – Nov 2023 (408 months)				
Sharpe ratio (annualised)	**3.74**	**1.38**	**1.08**	**0.49**
95% bootstrap interval	[3.19, 4.33]	[1.02, 1.87]	[0.73, 1.52]	[0.15, 0.88]
Mean return, % p.a. (unit gross)	8.0%	8.0%	6.0%	9.3%
Volatility, % p.a.	2.1%	5.8%	5.6%	19.0%
CAPM alpha t-stat (vs 1/N)	12.44	7.43	6.26	–2.95
Beta to the 1/N market	–0.00	–0.07	–0.15	1.00
Max drawdown at 10% vol	–12%	–19%	–23%	–34%
Positive months	87%	71%	66%	61%
Validation sample, Dec 1979 – Nov 1989 (120 months)				
Sharpe ratio (annualised)	**6.62**	**3.28**	**2.77**	**0.50**
95% bootstrap interval	[5.55, 8.70]	[2.63, 4.28]	[2.07, 3.59]	[–0.09, 1.33]
Mean return, % p.a. (unit gross)	13.1%	12.2%	10.3%	9.4%
Volatility, % p.a.	2.0%	3.7%	3.7%	18.7%
CAPM alpha t-stat (vs 1/N)	15.88	10.76	10.49	2.52
Beta to the 1/N market	–0.00	–0.05	–0.09	1.00
Max drawdown at 10% vol	–1%	–6%	–7%	–18%
Positive months	97%	85%	81%	58%

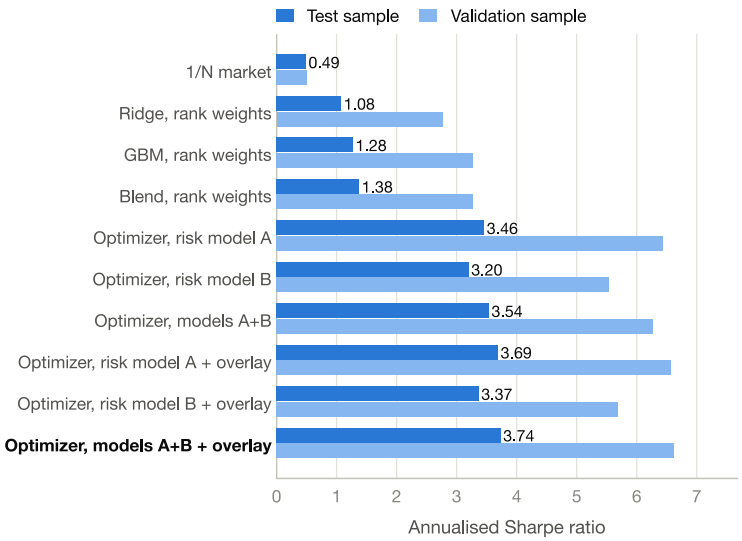


Figure 4. Where the Sharpe ratio comes from. The first four bars map forecasts to weights by rank; the others use the GBM + ridge blend forecasts with the mean–variance optimizer (all beta-neutral). Bold: the selected configuration.

Table 2. All twelve configurations. Annualised Sharpe ratios in the validation (selection) and test samples; the selected configuration is shaded.

Forecast	Covariance	No overlay		Vol overlay	
		Val.	Test	Val.	Test
GBM	model A	6.21	3.42	6.44	3.70
GBM	model B	5.59	3.29	5.74	3.45
GBM	models A+B	6.17	3.56	6.51	3.76
GBM + ridge blend	model A	6.44	3.46	6.56	3.69
GBM + ridge blend	model B	5.54	3.20	5.70	3.37
GBM + ridge blend	models A+B	6.26	3.54	**6.62**	**3.74**

The test sample ranks the configurations much as the validation sample did (rank correlation 0.87). Every configuration earns a test Sharpe ratio between 3.20 and 3.76, so the result does not hinge on the selection step.

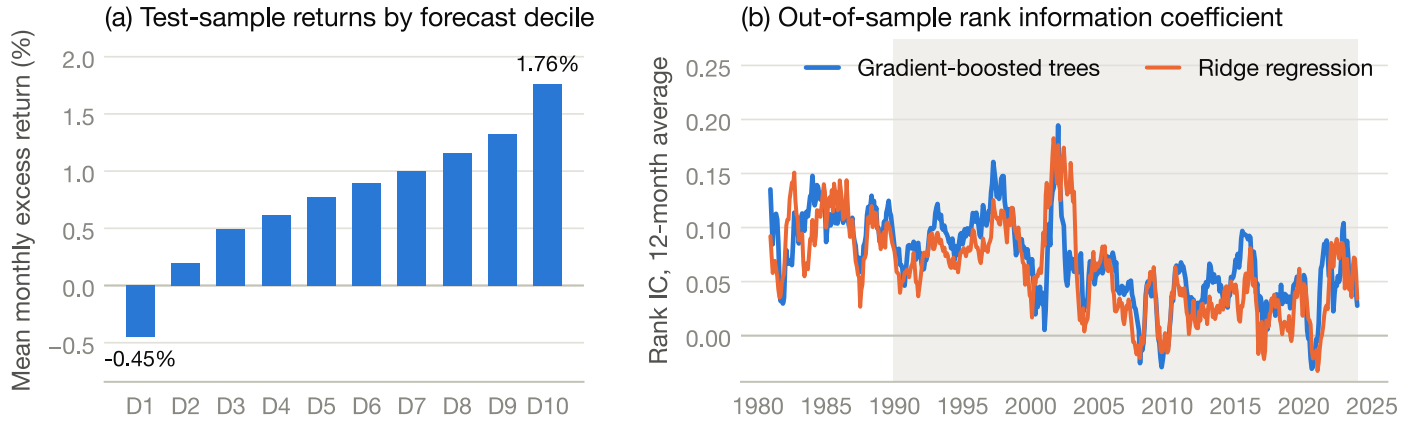


Figure 5. Forecast quality. (a) Mean excess return of stocks sorted each month into deciles of the final forecast, test sample (D10 – D1: 2.22% per month, t = 7.0). (b) 12-month moving average of the monthly Spearman correlation between forecasts and realised returns; shaded: test sample.

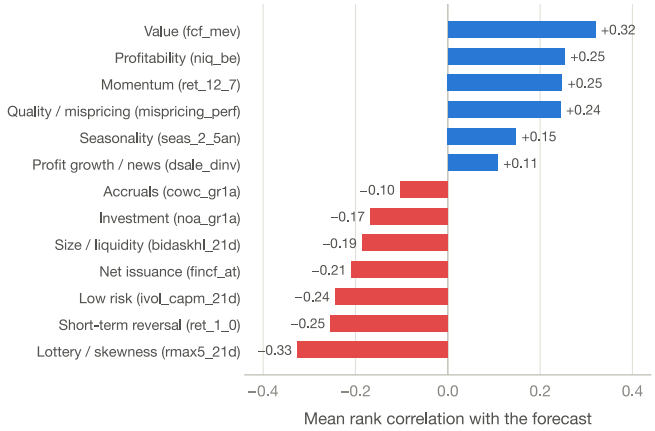


Figure 6. What the forecast loads on. Mean within-month rank correlation between the final forecast and characteristics, test sample. For each of 13 economic themes, the member characteristic with the largest absolute correlation is shown (code in brackets).

Table 3. Robustness. Turnover is $\Sigma|w_t - w_{t-1}|$ per month; net returns subtract cost × turnover.

Sharpe ratio by sub-period, test sample				
	Final	Rank wts	Ridge	1/N
1989–1999 (121 m)	6.77	3.18	2.25	0.63
2000–2009 (120 m)	3.87	1.03	1.08	0.23
2010–2019 (120 m)	3.08	1.39	0.67	0.78
2020–2023 (47 m)	1.39	0.60	0.32	0.41
Final strategy after trading costs, test sample				
Cost per $ traded	0 bp	5 bp	10 bp	20 bp
Net Sharpe ratio	3.74	3.47	3.20	2.66
Break-even cost per $ traded				70 bp
Size groups, test sample				
	Mega	Large	Small	
Share of gross exposure	18.2%	31.7%	50.0%	
Share of stocks held	18.1%	32.3%	49.6%	

Reading the evidence. The forecasts are informative out of sample: in the test sample the mean monthly rank information coefficient of the boosted trees is 0.061 (t = 10.4), higher than ridge's 0.052 (t = 8.4), and the blend reaches 0.064. Returns rise monotonically across forecast deciles (Figure 5a). Economically, the forecast combines well-known themes (Figure 6): it buys cheap stocks with high cash-flow yields (0.32), profitable firms (0.25), intermediate-horizon winners (0.25) and high-quality firms (0.24), and sells lottery-like stocks with extreme recent daily gains (–0.33), last month's winners (–0.25), volatile stocks (–0.24) and firms raising external financing (–0.21). The trees improve on the linear model because they can capture non-linear effects and interactions among these signals, as Gu, Kelly and Xiu (2020) also find. Holding the forecasts fixed, replacing rank weights by the risk-model optimizer raises the test-sample Sharpe ratio from 1.38 to 3.74; a paired block bootstrap puts the difference at 2.36 with a 95% interval of [1.90, 2.82], which excludes zero. The optimizer keeps the mean return (8.0% against 8.0% a year at unit gross) but cuts volatility from 5.8% to 2.1%, by hedging the common-factor risk that rank weights leave in the portfolio. The portfolio is not a disguised market bet (beta –0.00; CAPM-alpha t-statistic 12.44). Measured against the equal-weighted market, the long side adds 3.7% and the short side 4.3% per year at unit gross, so both sides earn their keep.

7 What an investor should expect

- **Gross Sharpe ratio.** 3.74 over 34 years (95% interval [3.19, 4.33]), but it has fallen decade by decade: 6.77 in 1989–1999, 3.87 in 2000–2009, 3.08 in 2010–2019, 1.39 in 2020–2023 (Table 3). The same forecasts with rank weights fall in step (3.18, 1.03, 1.39, 0.60), so the decline reflects weaker signals rather than weaker portfolio construction; it is consistent with the post-publication decay of anomaly returns (McLean and Pontiff, 2016), and several characteristics in the data were only discovered during the test period. The recent sub-periods are the better guide: I would plan for a gross Sharpe ratio closer to their 1.39–3.08 than to the full-sample 3.74.
- **Costs.** The portfolio holds about 2,060 names and trades about 100% of its gross exposure each month. At 10 bp per dollar traded the Sharpe ratio falls to 3.20 (to 2.66 at 20 bp); it breaks even at 70 bp. A production version should add a turnover penalty, as in Jensen, Kelly, Malamud and Pedersen (2024).
- **Risk and leverage.** Weights are dollar- and beta-neutral with unit gross exposure (0.50–1.42 with the overlay); at that leverage the strategy earned 8.0% a year with 2.1% volatility, and returns scale linearly with leverage. Drawdowns of –12.2% at 10% volatility occurred in the test sample.
- **Known limitations.** The clusters behind model B are unbalanced (Section 3), and a minimum cluster size or Ward linkage would likely improve it; the optimizer ignores trading costs; and while no choice in the code uses test-sample data, the design borrows ideas from published research and public entries whose evidence overlaps the test period, so the test sample is not perfectly pristine.

8 Implementation, reproducibility and disclosures

The submission is one Python file, `angel_zhao_ctf.py`, whose `main(chars, features, daily_ret)` returns `id, eom, w` for every test-sample stock-month. It reads no files and uses no external data; dependencies are pinned in `requirements.txt` (Python 3.13, also tested on 3.12 with identical weights; numpy 2.3.3, pandas 2.3.2, scipy 1.16.2, scikit-learn 1.7.2) and a full run takes about 3.5 hours on an 8-core laptop (Apple M2, 8 GB). **Prior work.** The arctan target, the ridge-stabilised tangency-minus-minimum-variance portfolio and the Ledoit–Wolf daily covariance follow Paul Geertsema's public CTF entry; beta neutralisation follows Abhinav Keshri's public entry; the factor risk model follows Jensen, Kelly, Malamud and Pedersen (2024) and the clustering Jensen, Kelly and Pedersen (2023). AI assistance (Anthropic's Claude) was used to write and test code and to draft this document; every number in it is computed from the output of the submitted code.

A Appendix: complete specification

Every value below is a fixed literal in `angel_zhao_ctf.py`; nothing is tuned on the test sample.

Inputs	`chars` (id, eom, ret_exc_lead1m, ctff_test, sic and the characteristics listed in `features`) and `daily_ret` (id, date, ret_exc); rows sorted by (eom, id), duplicate (eom, id) pairs dropped, so the output does not depend on input order
Features and target	within-month percentile ranks (average ties); missing values kept for the trees and set to the median rank for ridge; target $\arctan(z)$ of the within-month z-score of ret_exc_lead1m
Training scheme	refit every 12 months, starting 120 months before the first `ctff_test` month; trailing 240-month window, at most 600,000 most recent stock-months
Gradient-boosted trees	scikit-learn HistGradientBoostingRegressor: squared error loss, 500 iterations, learning rate 0.02, at most 255 leaves, at least 2,000 observations per leaf, 255 bins, no L2 penalty, no early stopping, random_state 42
Ridge regression	centred ranks; penalty $\lambda/12$ with $\lambda \in \{0.001, 0.01, 0.1, 1, 10\}$ chosen by mean monthly rank IC on the window's last 60 months, then refit on the whole window
Forecast	blend = $z(0.7 \cdot z_{GBM} + 0.3 \cdot z_{ridge})$, z = within-month standardisation; GBM forecast = z_{GBM}
Risk model A	Ledoit–Wolf shrinkage toward a scaled identity of the last 756 daily excess returns (stocks with at least 250 observations), scaled by 21 to a monthly horizon
Risk model B	12 Fama–French industries (from SIC) + 13 clusters (average linkage on $1 - \lvert\rho\rvert$ of rank correlations over 60 months, characteristics with at least 50% coverage, refit yearly); daily cross-sectional ridge regressions (penalty $0.0001 \cdot n$); Ω: EWMA over 1,260 days, half-lives 504 (correlations) and 84 (volatilities); D: EWMA with half-life 84 over 252 days (at least 200 observations), floored at its 5% quantile
Portfolio	tangency minus minimum-variance under $\Sigma + \lambda I$, $\lambda = 1 \times$ mean variance; books A and B each scaled to unit gross; A+B is their average; projected onto zero beta (252-day beta on the equal-weighted universe, at least 60 observations) and zero net exposure; unit gross exposure
Volatility overlay	multiply by median(past ex-ante volatility) / current ex-ante volatility from model B, clipped to [0.5, 2], once 12 months of history exist
Selection	highest Sharpe ratio over the 120 months before the first `ctff_test` month among {GBM, blend} × {A, B, A+B} × {overlay, none}; frozen for all test months

References

Bailey, D. H. and M. López de Prado (2014). The deflated Sharpe ratio. *Journal of Portfolio Management* 40(5), 94–107.

Gu, S., B. Kelly and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies* 33(5), 2223–2273.

Jensen, T. I., B. Kelly and L. H. Pedersen (2023). Is there a replication crisis in finance? *Journal of Finance* 78(5), 2465–2518.

Jensen, T. I., B. Kelly, S. Malamud and L. H. Pedersen (2024). Machine learning and the implementable efficient frontier. Working paper.

McLean, R. D. and J. Pontiff (2016). Does academic research destroy stock return predictability? *Journal of Finance* 71(1), 5–32.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88(2), 365–411.