

# Problem Set 3 - Submission to JKP Competition

Behruz Ahkmedov (ba554), Nicolas Lopez (nml44), and Marc Utzmeier (mu244)

October 7, 2026

## Executive summary

We forecast monthly US stock returns with an ensemble of ridge regression and XGBoost, trained on all 402 characteristics in the competition data. A Barra-style risk model, mean-variance optimization and a 5% volatility target turn the forecasts into a dollar- and industry-neutral long-short portfolio.

Over the competition test period (Dec 1989–Nov 2023, 408 months, run once with the frozen model), the strategy earned a gross Sharpe ratio of 2.36 (annual return 9.9% at 4.2% volatility; maximum drawdown 8.6%; positive in 76.7% of months and 31 of 34 years), with a market beta of  $-0.03$ . Performance declined over time, from a Sharpe ratio of 4.08 in the 1990s to 1.23 in 2020–2023. The portfolio turns over 1.8 times its capital a month, so costs matter: at 10 basis points per dollar traded, the Sharpe ratio is 1.86 over the full period and about 0.9 since 2010.

## 1 Data and preparation

We use only the competition dataset, built from CRSP and Compustat: 1,369,324 stock-months of relatively liquid US stocks (mega, large and small size groups) from December 1951 to November 2023, with 402 characteristics per stock-month and daily excess returns.

- **Features.** Each characteristic is ranked cross-sectionally within the month and mapped to  $(-0.5, 0.5)$ ; missing values are set to 0 (the median rank). Ranks are then demeaned within each (month, 2-digit SIC industry) cell.
- **Target.** The within-month rank of next-month excess return (`ret_exc_lead1m`), demeaned within industry in the same way.

Our model ranks characteristics to remove the impact of outliers: rather than looking at absolute differences in a given metric, we look only at a stock’s position relative to other stocks.

## 2 Return forecasts

Both forecasters see all 402 ranked traits. With this many noisy and correlated predictors, the main danger is overfitting. We guard against it in three ways. First, ridge regression shrinks every coefficient toward zero, so no single noisy trait gets a large weight; the penalty is chosen on a rolling ten-year validation window the model is not fit on. Second, XGBoost, which adds the nonlinear effects and interactions a linear model misses, is kept deliberately simple: depth-3 trees, a learning rate of 0.05, and early stopping on the same validation window. Third, ranking the traits removes outliers. Each month we convert both forecasts to z-scores and average them 50/50, so neither model dominates because of its scale.

**Forecast quality.** In the development period, the ensemble’s information coefficient (IC), the average monthly rank correlation between forecasts and next-month returns, was 0.115 and was positive in 97.5% of months, against 0.109 for ridge and 0.113 for XGBoost alone. The rank correlation between ridge and XGBoost forecasts was 0.86 in the 1980s and 0.75 in the 1970s. By share of XGBoost’s total gain, the most important traits are short-term price signals, (i) maximum daily returns (`rmax5_21d`, `rmax1_21d`) and (ii) short-term reversal (`ret_1_0`), followed by (iii) momentum (`ret_12_1`) and valuation ratios (`nix_me`, `cop_mev`).

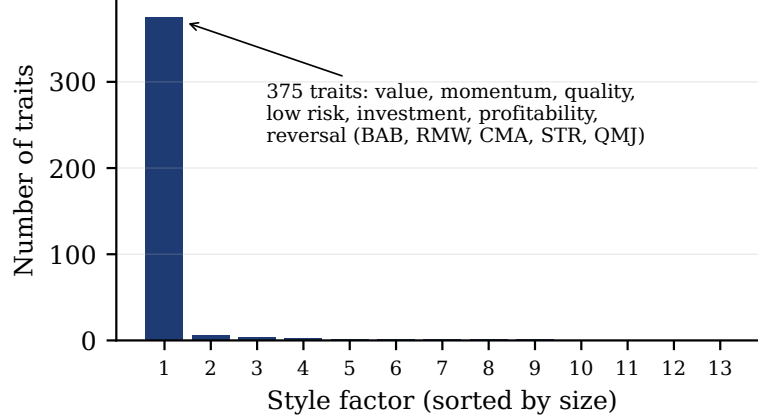
## 3 Portfolio construction and risk model

Every December the forecasting models are refit on all earlier data. Each month, the forecast is turned into weights in four steps, all fixed before the test period.

1. **Expected returns.** Each stock’s forecast is multiplied by its trailing 12-month volatility, so the same forecast implies a larger expected return for a more volatile stock (#3 in Table 1, paired  $t = 3.79$ ).
2. **Risk model.** A Barra-style factor model [3, 12] with industry factors (2-digit SIC) and 13 style factors. Factor covariances and stock-specific variances are exponentially weighted averages of daily returns (half-lives of 84 days

for volatilities and 504 days for correlations). The style factors come from hierarchical clustering of the traits (average linkage on  $1 - |\rho|$ , pre-1990 data); each is the signed average of the traits in its cluster (Figure 1).

3. **Optimization.** Mean-variance weights, constrained to zero net exposure in every industry, so the portfolio is also dollar-neutral.
4. **Volatility targeting.** The portfolio is scaled to 5% annualized volatility, based on the realized volatility of the current weights over the past 63 trading days. This raised the development Sharpe ratio from 6.00 to 6.55 (#5b,  $t = 3.76$ ).



**Figure 1: What the 13 style factors contain.** Number of traits in each style factor of the risk model (average linkage, pre-1990 data). One factor holds 375 of the 402 traits.

**Strengths and limits.** Because most traits are only weakly correlated, clustering yields one broad style factor (375 traits) and twelve narrow ones. The risk model therefore mainly controls industry exposure and one blended style, and volatility targeting does most of the risk scaling: test-period volatility was 4.2% against the 5% target, and an investor can lever the portfolio to any desired level. After finding this imbalance, we tested two richer risk models to address it: Ward-linkage clustering, which forms balanced groups of 13–75 traits, and five IPCA factors [9]. Both lowered the development Sharpe ratio, to 6.57 and 5.10 (#16–17), most likely because they penalize the style tilts that drive the forecast. The trade-off is that the portfolio carries style exposure its risk model only partly measures: in the development period the model predicted 1.63% volatility against 3.41% realized. Volatility targeting limits this risk but reacts with a lag, so a sudden style reversal could cost more than predicted. A next version should select risk models on how well they forecast realized volatility, not on Sharpe ratio.

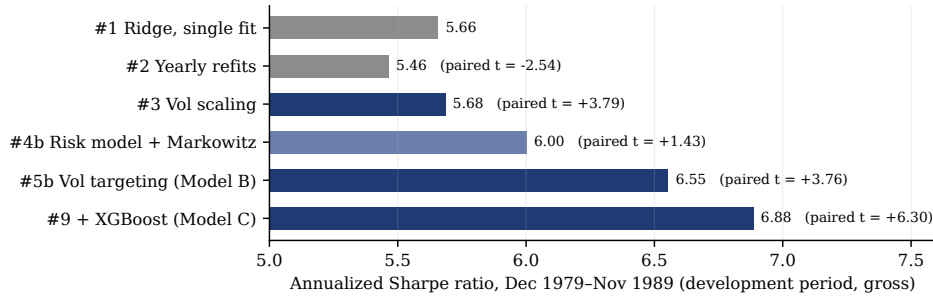
## 4 Validation protocol and research log

**Protocol.** With hundreds of characteristics, it is easy to find a model that fits one sample by luck [6, 7]. We tested every design change on a development period (Dec 1979–Nov 1989, 120 months), treated as a test period: the model was trained only on earlier data. A change was adopted only if its paired  $t$ -statistic against the incumbent exceeded 2, with

$$d_t = \frac{r_t^{\text{new}}}{\widehat{\text{sd}}(r^{\text{new}})} - \frac{r_t^{\text{base}}}{\widehat{\text{sd}}(r^{\text{base}})}, \quad t = \frac{\bar{d}}{\widehat{\text{sd}}(d)/\sqrt{T}}.$$

Changes that passed were confirmed on a second period (Dec 1969–Nov 1979) under rules fixed before looking ( $t > 0$  for single changes,  $t > 2$  for combinations). The pass bar and all settings of each experiment were written down before the run, and Model C was run on the competition period (Dec 1989–Nov 2023) once, at the end, after it was frozen.

**Results.** We tested 16 changes and adopted five (Table 1, Figure 2). Apart from the yearly refit, which the 34-year test period requires, every adopted step had a development  $t$ -statistic above 3.7 (#3, #5b, #9), above the  $t > 3$  hurdle Harvey et al. [6] recommend when many hypotheses are tested [see also 1]. The rejections follow a pattern: richer risk models (#11, #16, #17) lowered the Sharpe ratio (Section 3), and possible refinements such as weighting recent years more heavily did not help. Because the development period was reused for every test, its Sharpe ratios are optimistic; the competition period is the only clean out-of-sample evaluation.



**Figure 2: How Model C was built.** Development-period Sharpe ratio after each adopted step (gross of costs). Paired  $t$ -statistics are against the previous step.

**Table 1: Research log: every model change tested (development period unless noted). Sharpe ratios are gross of costs.**

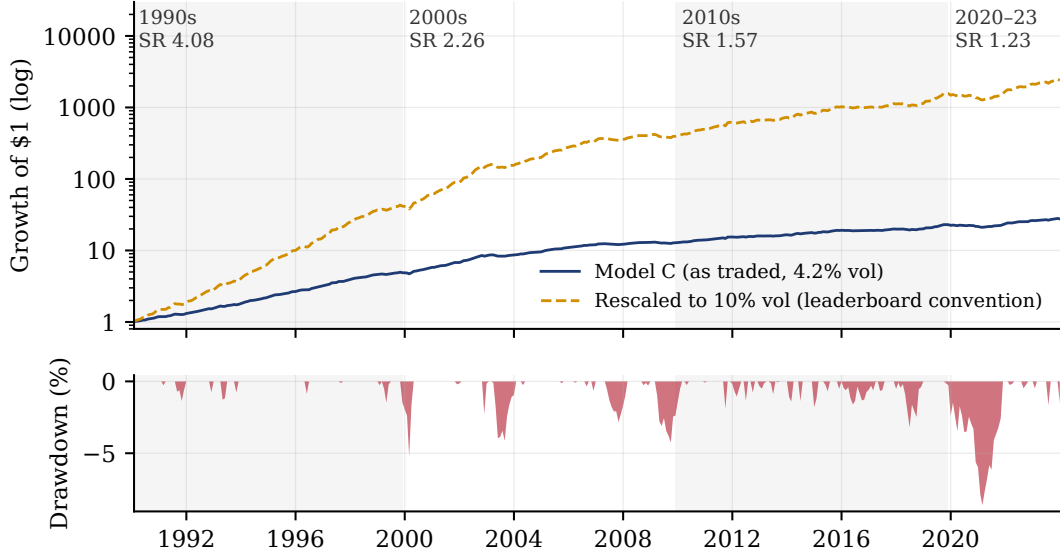
| #  | Change                                                    | Sharpe: base $\rightarrow$ new                   | Paired $t$ | Decision                   |
|----|-----------------------------------------------------------|--------------------------------------------------|------------|----------------------------|
| 1  | Ridge, single fit, linear weights                         | n/a $\rightarrow$ 5.66                           | n/a        | Starting point             |
| 2  | Yearly refits (expanding window)                          | 5.66 $\rightarrow$ 5.46                          | -2.54      | Kept (see note)            |
| 3  | Stock volatility scaling ( $w \propto \hat{\mu}/\sigma$ ) | 5.46 $\rightarrow$ 5.68                          | 3.79       | Adopted                    |
| 4a | Risk model + Markowitz, dollar-neutral                    | 5.68 $\rightarrow$ 5.93                          | 1.05       | Rejected                   |
| 4b | Risk model + Markowitz, dollar + industry neutral         | 5.68 $\rightarrow$ 6.00                          | 1.43       | Carried forward (see note) |
| 5a | Volatility targeting on #3                                | 5.68 $\rightarrow$ 5.84                          | 0.99       | Rejected                   |
| 5b | Volatility targeting on #4b (= Model B) [13]              | 6.00 $\rightarrow$ 6.55                          | 3.76       | Passed                     |
| 6  | Confirmation, 1970s: #3 vs. naive; B vs. #3               | 4.21 $\rightarrow$ 4.31; 4.31 $\rightarrow$ 4.90 | 1.66; 3.50 | Confirmed; Model B adopted |
| 7  | Stronger ridge penalty (tie-break rule)                   | 6.55 $\rightarrow$ 6.30                          | -4.22      | Rejected                   |
| 8  | Recency weighting of training years                       | 6.55 $\rightarrow$ 6.48                          | -6.62      | Rejected                   |
| 9  | Ridge + XGBoost ensemble [2, 4]                           | 6.55 $\rightarrow$ 6.89                          | 6.30       | Passed                     |
| 10 | Confirmation, 1970s: ensemble vs. B                       | 4.90 $\rightarrow$ 5.71                          | 9.70       | Confirmed; Model C adopted |
| 11 | Ridge forecast added as a risk factor                     | 6.89 $\rightarrow$ 5.34                          | -7.98      | Rejected                   |
| 12 | Random-feature ridge as third member [8]                  | 6.89 $\rightarrow$ 6.94                          | 1.66       | Rejected                   |
| 13 | Risk-adjusted target (return / trailing vol)              | 6.89 $\rightarrow$ 6.76                          | -2.18      | Rejected                   |
| 14 | Factor momentum as third member [5]                       | 6.89 $\rightarrow$ 6.94                          | 0.33       | Rejected                   |
| 15 | XGBoost seed averaging (5 seeds)                          | 6.89 $\rightarrow$ 6.89                          | 0.50       | Rejected                   |
| 16 | IPCA risk model, $K = 5$ [9, 10]                          | 6.89 $\rightarrow$ 5.10                          | -7.25      | Rejected                   |
| 17 | Ward-linkage style clusters                               | 6.89 $\rightarrow$ 6.57                          | -2.37      | Rejected                   |

#2 lowers the Sharpe ratio through higher risk, with forecast accuracy unchanged, but is required to forecast a 34-year test period. #4b missed the bar on its own and was adopted only as part of Model B (#4b + #5b), which beat the incumbent #3 in the development period (paired  $t = 3.43$ ) and again in the 1970s ( $t = 3.50$ , #6). Rejected changes were not run on the 1970s.

## 5 Out-of-sample results (competition test period)

**Headline.** Model C earned a Sharpe ratio of 2.36 over 408 months ( $t = 13.8$ ), with an annual return of 9.9% at 4.2% volatility (Figure 3, Table 2). Returns were positive in 76.7% of months and in 31 of 34 years; the three negative years were 2009, 2016 and 2020 (Figure 4). Rescaled to the leaderboard’s 10% volatility, the annual return is about 24%. Monthly returns are close to normal (skewness 0.15, excess kurtosis 0.35), so the Sharpe ratio is not driven by a few extreme months.

**Market exposure.** The portfolio is dollar-neutral by construction and also close to beta-neutral. Regressing its monthly returns on the value-weighted market excess return, which we compute from the competition data, gives a beta of  $-0.03$  (Newey-West  $t = -1.72$ ) and an annual alpha of 10.1% ( $t = 11.4$ ). The strategy held up when the market fell: in the 41 worst market months (bottom decile), it averaged 1.26% a month, above its full-sample average of 0.82%, and was positive in 85% of them. In October 2008 the market fell 18.4% and the strategy gained 0.2%.



**Figure 3: Cumulative performance, Dec 1989–Nov 2023 (gross of costs).** *Top:* growth of \$1 (log scale), as traded (4.2% realized volatility) and rescaled to 10% annual volatility as on the leaderboard. Shaded bands mark decades, with the Sharpe ratio of each. *Bottom:* drawdown of cumulative (non-compounded) returns as traded; the maximum is 8.6%.

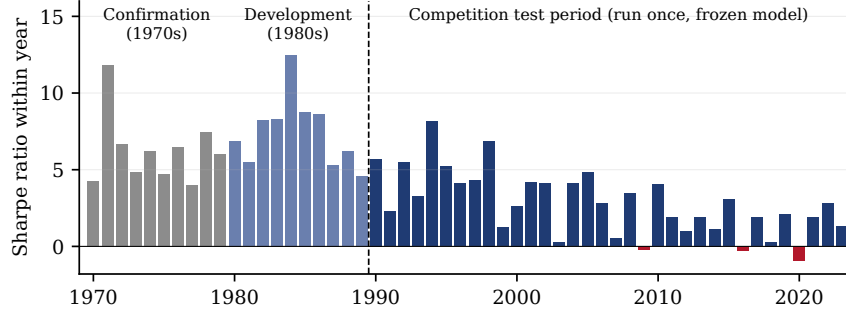
**Drawdowns.** The largest drawdown, 8.6%, ran from late 2019 to early 2021, through the COVID-19 crash and rebound, and was recovered by late 2021. Before that, the largest was 5.2%, from late 1999 to February 2000 at the peak of the technology bubble, which also holds the worst single month (−2.8%). The rolling 36-month Sharpe ratio fell below 1 in 10% of windows, with a low of 0.43 in early 2021, and was 1.99 at the end of the sample.

**Table 2: Performance by period (Model C, gross of trading costs)**

| Period                                        | Months     | Sharpe      | $t(\text{Sharpe})$ | Ann. return | Ann. vol    | Max DD      | Hit rate     |
|-----------------------------------------------|------------|-------------|--------------------|-------------|-------------|-------------|--------------|
| <i>Pre-competition (pseudo out-of-sample)</i> |            |             |                    |             |             |             |              |
| Confirmation, Dec 1969–Nov 1979               | 120        | 5.71        | 18.06              | 21.8%       | 3.8%        | 1.2%        | 95.0%        |
| Development, Dec 1979–Nov 1989                | 120        | 6.89        | 21.77              | 23.5%       | 3.4%        | 0.7%        | 96.7%        |
| <i>Competition test period (run once)</i>     |            |             |                    |             |             |             |              |
| <b>Full, Dec 1989–Nov 2023</b>                | <b>408</b> | <b>2.36</b> | <b>13.76</b>       | <b>9.9%</b> | <b>4.2%</b> | <b>8.6%</b> | <b>76.7%</b> |
| 1990–1999                                     | 120        | 4.08        | 12.90              | 16.0%       | 3.9%        | 1.9%        | 87.5%        |
| 2000–2009                                     | 120        | 2.26        | 7.14               | 9.9%        | 4.4%        | 4.3%        | 78.3%        |
| 2010–2019                                     | 120        | 1.57        | 4.96               | 5.6%        | 3.6%        | 3.2%        | 68.3%        |
| 2020–2023                                     | 48         | 1.23        | 2.46               | 5.3%        | 4.3%        | 7.1%        | 66.7%        |

Years run December–November.  $t(\text{Sharpe}) = \text{annualized Sharpe} \times \sqrt{\text{years}}$ . Max DD is computed on cumulative (non-compounded) returns. Monthly turnover ( $\sum_i |w_{i,t} - w_{i,t-1}|$ ) averages 2.15 (1970s), 2.19 (1980s) and 1.77 (test period).

**Decay.** Performance fell steadily. A regression of yearly Sharpe ratios on time gives a decline of 1.3 per decade, and the test-period Sharpe ratio is about a third of the development-period 6.89, although every decade remains significantly positive ( $t \geq 2.46$ ). We see three likely causes, which we cannot separate statistically. (i) *The traits were chosen by the literature.* The 402 characteristics are ones academics found to predict returns, often using pre-1990 data, so the pre-1990 results overstate what a model built in 1990 could have achieved. This is a property of the dataset, not look-ahead in our pipeline, which passes the truncated-data test (Table 3). (ii) *Signals decay after publication.* McLean and Pontiff [11] find that anomaly returns are about 58% lower after publication, and many of our traits were published during the test period. (iii) *XGBoost became much simpler.* Early stopping kept 192–379 trees in every refit from 1990 to 2008 but only 12–64 from 2009 on, when the ten-year validation window moved to post-1999 data, so XGBoost’s half of the forecast captured little nonlinear structure.



**Figure 4: Sharpe ratio by year, 1970–2023 (gross of costs).** Colors mark the confirmation (1970s), development (1980s) and competition test (1990–2023) periods; negative years are in red. Test-period Sharpe ratios are lower and more variable.

## 6 What an investor should expect

**Trading costs.** All results above are gross of costs. With monthly turnover of 1.77, the break-even cost is about 47 basis points per dollar traded, and at 10 and 25 basis points the Sharpe ratio falls to 1.86 and 1.09 (Table 3). Costs matter more in recent years because the gross edge is smaller: at 10 basis points the net Sharpe ratio is 3.53 in the 1990s but 0.88 in 2010–2019 and 0.95 in 2020–2023.

**Table 3: Robustness and implementation checks**

| Check                                                                                                     | Period / model   | Result                                |
|-----------------------------------------------------------------------------------------------------------|------------------|---------------------------------------|
| <i>Trading costs (cost per dollar traded <math>\times</math> monthly turnover)</i>                        |                  |                                       |
| 0 / 10 / 25 / 50 bp                                                                                       | Test, Model C    | Sharpe 2.36 / 1.86 / 1.09 / $-0.17$   |
| 0 / 10 / 25 / 50 bp                                                                                       | 1980s, Model B   | Sharpe 6.55 / 5.77 / 4.59 / 2.67      |
| <i>Null models (full pipeline)</i>                                                                        |                  |                                       |
| Shuffled target                                                                                           | 1980s, Model B   | Sharpe $-0.79$ , IC 0.005             |
| Pure-noise forecasts                                                                                      | 1980s, Model B   | Sharpe $-0.70$ , IC $-0.004$          |
| <i>Data and timing</i>                                                                                    |                  |                                       |
| Returns rebuilt from daily data                                                                           | 1980s, Model B   | Sharpe 6.55 (vs. 6.55)                |
| Skip first trading day of each month                                                                      | 1980s, Model B   | Sharpe 5.90 (vs. 6.55)                |
| $\text{corr}(\text{ret\_1\_0}_t, \text{target}_{t-1}) / \text{corr}(\text{ret\_1\_0}_t, \text{target}_t)$ | 1980s            | 0.9997 / 0.0024                       |
| <i>Reproducibility and look-ahead</i>                                                                     |                  |                                       |
| Same run twice                                                                                            | Competition code | Max abs. weight diff. 0               |
| Truncated data (cut 1992-12-31 vs. 1995-12-31)                                                            | Competition code | Max abs. weight diff. 0               |
| XGBoost seeds 42–46 averaged vs. seed 42                                                                  | 1980s, Model C   | Return corr. 0.999, paired $t = 0.50$ |

Test-period costs: monthly cost =  $c \times \sum_i |w_{i,t} - w_{i,t-1}|$  computed from the submitted weights (first month: full build-up), ignoring price drift between rebalances. Model B = Model C without XGBoost.

**Capacity and implementation.** The portfolio holds about 2,170 stocks a month, with the largest position averaging 0.5% of gross exposure, but our net figures leave out real frictions. A flat cost per dollar understates trading costs for small stocks, which make up 44% of gross exposure, and price impact grows with fund size. We also ignore shorting costs, and the volatile stocks we short tend to be the most expensive or hardest to borrow. The edge also depends on trading promptly: a one-day delay lowered the 1980s Sharpe ratio from 6.55 to 5.90. Running at the leaderboard’s 10% volatility roughly doubles gross exposure (from 1.55) and costs.

**Risks.** Three risks matter most. First, the signal has weakened every decade, and nothing in our results says the decline has stopped. Second, the risk model underestimates volatility and only partly sees the portfolio’s style tilts (Section 3), so a sudden style reversal could cost more than predicted. Third, the strategy loses when the volatile stocks it shorts rally sharply, as in November 2020 and January 2021; these months, not the COVID-19 crash, drove its largest drawdown.

**Bottom line.** Based on the period since 2010, we expect a net Sharpe ratio of about 1 at 10 basis points before shorting costs and price impact. Returns have a correlation of  $-0.11$  with the stock market (beta  $-0.03$ ), so the strategy can add return to an equity portfolio without adding market risk. The full-period net figure of 1.86 is likely optimistic, since the strong 1990s drive much of it.

## References

- [1] D. Bailey and M. López de Prado. The deflated Sharpe ratio. *Journal of Portfolio Management*, 40(5), 2014.
- [2] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
- [3] R. Grinold and R. Kahn. *Active Portfolio Management*, 2nd ed. McGraw-Hill, 2000.
- [4] S. Gu, B. Kelly, and D. Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2020.
- [5] T. Gupta and B. Kelly. Factor momentum everywhere. *Journal of Portfolio Management*, 45(3), 2019.
- [6] C. R. Harvey, Y. Liu, and H. Zhu. . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 2016.
- [7] T. I. Jensen, B. Kelly, and L. H. Pedersen. Is there a replication crisis in finance? *Journal of Finance*, 78(5), 2023.
- [8] B. Kelly, S. Malamud, and K. Zhou. The virtue of complexity in return prediction. *Journal of Finance*, 79(1), 2024.
- [9] B. Kelly, S. Pruitt, and Y. Su. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3), 2019.
- [10] B. Kelly, S. A. Korsaye, S. Pruitt, and Y. Su. Instrumented principal component analysis. Working paper, April 21, 2026.
- [11] R. D. McLean and J. Pontiff. Does academic research destroy stock return predictability? *Journal of Finance*, 71(1), 2016.
- [12] J. Menchero, D. Orr, and J. Wang. *The Barra US Equity Model (USE4)*. MSCI, 2011.
- [13] A. Moreira and T. Muir. Volatility-managed portfolios. *Journal of Finance*, 72(4), 2017.