The Expected Returns on Machine-Learning Strategies [⋆]

Vitor Azevedo[a,*], Christopher Hoegner[b], Mihail Velikov[c]

[a] *School of Business and Economics, RPTU Kaiserslautern-Landau, Gottlieb-Daimler-Straße 42, 67663 Kaiserslautern, Germany*

[b] *Department of Financial Management and Capital Markets, TUM School of Management, Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany*

[c] *Smeal College of Business, Penn State University, State College, PA 16802, U.S.*

Abstract

We estimate the expected returns of machine learning-based anomaly trading strategies, accounting for three factors often overlooked in the previous literature: transaction costs, post-publication decay, and the post-decimalization era of high liquidity. Despite a cumulative performance reduction averaging about 57% when accounting for these three factors, sophisticated machine learning strategies remain profitable, particularly those employing Long Short-Term Memory (LSTM) models. We estimate that our most effective strategy, the one based on an LSTM model with one hidden layer, has an expected gross (net) Sharpe Ratio of 0.94 (0.84). Our findings contrast with previous literature suggesting that machine learning strategies are unprofitable after accounting for economic constraints and demonstrate persistent return predictability that cannot be explained by common risk factors or limits to arbitrage.

Keywords: Stock market anomalies; machine learning models; return prediction; transaction costs; asset pricing models.

JEL classifications: G11, G12, G14, C45, C58.

[⋆] We thank Christoph Kaserer, Jonathan W. Lewellen, Juhani Linnainmaa, Stefan Nagel, Robert Novy-Marx, Alberto Rossi, along with conference and seminar participants at Banco Santander S.A., the Chicago Quantitative Alliance (CQA) Spring 2024 Conference and the 18th International Joint Conference on Computational and Financial Econometrics and Computational and Methodological Statistics for insightful discussions and helpful comments. None of the authors has conflicts of interest to disclose.

[*] Corresponding author.

Email addresses: vitor.azevedo@rptu.de (Vitor Azevedo), christopher.hoegner@tum.de (Christopher Hoegner), velikov@psu.edu (Mihail Velikov)

1. Introduction

A growing body of literature in finance documents the remarkable ability of machine learning techniques to enhance predictability in the cross-section of stock returns.[1] Studies that use these techniques to extract expected return signals routinely report annualized Sharpe ratios on trading strategies employing these signals in excess of 1.0, with extreme examples exceeding 2.0 (e.g., Freyberger et al., 2020; Chen et al., 2023; Cong et al., 2020), performance that corresponds to about five times the historical market Sharpe ratio of 0.43, estimated over the entire CRSP sample from 1925-2021.

Despite this impressive paper performance, the extent to which these strategies can be implemented in practice remains an ongoing debate. Avramov et al. (2022) argue that trading strategies based on machine learning models extract profitability from difficult-to-arbitrage stocks and during high limits-to-arbitrage market states, and their performance deteriorates in the presence of economic constraints because of high turnover. Blitz et al. (2023) advocate using longer prediction horizons to train the machine learning models and show that those can improve performance even after accounting for 25-basis-points-per-trade transaction costs. Jensen et al. (2022) go further and develop a framework that integrates trading-cost-aware portfolio optimization with machine learning.

Complicating the matters, simply accounting for transaction costs still does not answer the question of what the expected returns on machine learning strategies are and whether we can expect to see similar performance in the future. This question is more subtle because many anomalies were not discovered for significant periods of the samples in which they are typically used for backtesting machine learning strategies. Thus, even if we make the optimistic assumption that the machine learning techniques were available for investors, the anomaly signals were not. Moreover, individual anomaly performance deteriorates post-publication (Mclean and Pontiff, 2016) and has further deteriorated in the more recent sample post-decimalization due to the new era of investment and trading technology (Chordia et al., 2014).

The issue of using pre-publication signals is further exacerbated because combining anomalies known to work well in-sample suffers from severe overfitting bias. Novy-Marx (2016) demonstrates that when anomalies that perform well in-sample are combined, the resulting strategy's backtested performance can be significantly overstated, as the individual signals may have

[1] See, e.g., Gu et al. (2020); Leippold et al. (2022); Hanauer and Kalsbach (2023); Azevedo and Hoegner (2023); Chen et al. (2023); Azevedo et al. (2023); Cakici et al. (2023).

worked solely due to chance. This issue is emphasized by Li et al. (2022) in the context of machine learning. They construct real-time machine learning strategies based on a universe of data-mined fundamental signals and find significantly weaker performance than strategies using anomaly signals from published studies.

In this study, we aim to quantify the effects of each of these issues in backtesting machine-learning-based trading strategies discussed above and to arrive at a more realistic estimate of their expected returns. As a benchmark, we start with backtests that resemble what papers in the literature typically report. That is, we test nine different machine learning strategies that combine up to 320 anomaly signals from the Chen and Zimmermann (2022) dataset using an out-of-sample backtesting period starting in 2000 and ignoring trading costs.[2] To estimate expected returns, we sequentially 1) restrict the sample to more recent periods that more closely resemble current liquidity conditions, 2) restrict the anomaly signals to only published ones, and 3) subtract trading costs.

We find that the performance on the expected returns on machine learning strategies deteriorates significantly compared to typical backtests. On average, the performance across all models declines by approximately 57% relative to that of a typical backtest. However, this reduction varies considerably among different models, ranging from a modest 23% decrease for the best-performing LSTM1 model to a substantial 92% decline for the FFNN2 model. After accounting for post-publication effects, the post-2005 sample period, and trading costs, most machine learning strategies yield Sharpe ratios that fail to surpass the market Sharpe ratio over the same timeframe. The impact on average monthly returns is also stark: what initially appeared as promising returns of 1-2% per month in typical backtests diminishes to a more modest range of 0.2-1.4% per month under more realistic investment conditions.

By far, the biggest impact on performance comes from restricting the anomalies to only published ones. Averaging across nine machine learning strategies, restricting our features to only post-publication anomalies accounts for approximately 26% Sharpe ratio reduction. These results align with Mclean and Pontiff (2016), who find that portfolio returns are 58% lower post-publication. While the transaction costs lower the Sharpe ratio on average by approximately

[2] The machine learning techniques we use include Ordinary Least Square with Huber Loss Function (OLS-H), an Elastic Net (ENET), which combines a Least Absolute Shrinkage and Selection Operator (LASSO) and ridge regression, Feedforward Neural Network (FFNN) with two to five hidden layers (FFNN2, FFNN3, FFNN4, and FFNN5), two variations on Long Short-Term Memory (LSTM) with one and two hidden layers (LSTM1 and LSTM2), and an ensemble model.

15%, restricting the sample to post-2005 reduces the average Sharpe ratio by 11%. In terms of average returns, the typical machine-learning strategy loses about 46 basis points when only using anomalies post-publication. Restricting the sample post-2005 and trading costs cumulatively adds another 57 basis points per month decline in the average machine learning strategy profitability relative to the benchmark case. The breakdown between the two effects is about 34 basis points per month decline due to the post-2005 sample restriction and about 23 basis points per month due to trading costs.

Although more modest under realistic conditions, these machine learning strategies demonstrate value that extends beyond conventional risk factors. Most notably, while their expected returns and Sharpe ratios parallel the broader market, all strategies generate positive Novy-Marx and Velikov (2016) generalized alphas, with five of the nine strategies showing statistical significance. The LSTM2 model stands out as particularly effective, generating a Fama and French (2018) six-factor alpha of 1.26% monthly (t-statistic = 3.24). The practical significance of these results becomes clear when considering portfolio optimization: an investor with access to the net returns of the Fama and French (2018) six factors could substantially improve their portfolio's Sharpe ratio from 1.15 to 1.42 by incorporating the LSTM2 strategy. This material improvement in risk-adjusted returns demonstrates how sophisticated machine learning approaches, especially LSTM models, can meaningfully enhance investment outcomes beyond the capabilities of traditional factor investing.

Our analysis of feature importance reveals that LSTM1's success stems from its selection of fundamental rather than technical signals. The model generates reliable out-of-sample predictions by primarily leveraging event-based and accounting-based indicators while assigning lower weights to market-based signals. Specifically, three categories emerge as the most influential: *Earnings Event*, *Cash Flow Risk*, and *Valuation*. The model's distinctive approach is further evidenced by its remarkably low correlation with other machine learning signals, suggesting it captures unique patterns in the cross-section of returns. To further validate its robustness, we subject the LSTM1 signal to the "Assaying Anomalies" protocol developed by Novy-Marx and Velikov (2023). The signal demonstrates exceptional performance across nearly all dimensions of this comprehensive evaluation framework.

In the final section of our study, we investigate how cost-mitigation techniques and economic constraints affect the expected returns of machine learning strategies. We draw on previous research by Avramov et al. (2022), who examined economic constraints such as excluding the smallest 20% of stocks by market cap or filtering out the 30% of stocks with the highest recent trading costs. We also consider transaction cost mitigation techniques explored by Novy-Marx and Velikov (2016, 2019), including extending holding periods up to four months and implementing a Buy-Hold-Spread (BHS) approach, which involves buying stocks in the top/bottom decile but only selling when they fall out of the top/bottom quintile.

Our results show that these mitigation techniques significantly reduce both turnover and transaction costs. However, in most cases, these benefits are offset by a corresponding decrease in gross average returns. The only technique that consistently improves net performance across all machine learning strategies is extending the holding period to two months, but even this improvement is modest, averaging only seven basis points. These findings suggest that while cost-mitigation strategies can effectively reduce trading costs, they may also diminish the strategies' ability to capture short-term price movements, resulting in a minimal net benefit for most machine-learning approaches in the current market environment.

Our paper contributes to a growing literature on using machine learning in asset pricing settings. Many papers demonstrate the impressive predictive power of machine learning signals in the cross-section of U.S. stock returns. For example, Freyberger et al. (2020) use adaptive group LASSO, Gu et al. (2020) survey and apply many machine learning techniques including elastic net, dimension reduction techniques (PCR and PLS), trees, and neural networks. Cong et al. (2020) apply a deep reinforcement learning model. Simon et al. (2022) use neural networks to optimize portfolio weights as a function of firm characteristics. Chen et al. (2023) apply both feedforward and recurrent neural networks with long short-term memory.

Our main contribution relative to all these studies is our focus on the expected returns of machine learning strategies through careful treatment of trading costs, post-publication decay, and the staleness of historical data. While most of the studies cited above attempt to address the issue of implementability, they do so indirectly through economic constraints in the spirit of Avramov et al. (2022) such as size or turnover cutoffs or using crude trading costs measures. For example, Freyberger et al. (2020) use the Brandt et al. (2009) trading costs imputation based on size and Blitz et al. (2023) use a flat 25 basis points per trade assumption. The Chen and Velikov (2023) effective bid-ask spread measure we employ presents a more realistic estimate of trading costs post-decimalization since it is based on high-frequency TAQ data. To the best of our knowledge, we are also the first to estimate the Novy-Marx and Velikov (2016) generalized alphas for machine learning strategies, which enables a more precise estimation of risk-adjusted net returns. This method directly addresses the frequently overlooked impact of trading costs in asset pricing models, a factor often neglected in studies that use gross return asset pricing models to explain anomalies in net returns.

Furthermore, our study demonstrates that machine learning strategies can be profitable, even in the recent era of high liquidity and when using only discovered anomalies. This is in stark contrast to the conclusions in Avramov et al. (2022). While it is true that machine learning strategies concentrate on historically difficult-to-arbitrage stocks, value-weighting stocks in the portfolios and the sharp decline in trading costs over the past couple of decades combine to result in significant profits for these strategies.

Our study is also related to recent papers that use machine learning techniques to construct improved factor models (Feng et al., 2023), show that technical analysis works, though its profitability decreases through time (Brogaard and Zareei, 2023), explain the post-earnings announcement drift (Hansen and Siggaard, 2023; Meursault et al., 2023), measure firm complexity (Loughran and McDonald, 2023), and uncover sparsity and heterogeneity in firm-level return predictability (Evgeniou et al., 2023).

Finally, we also add to the debate on the virtue of complexity. Kelly et al. (2023) establish the rationale for using machine learning to model expected returns and theoretically show that "complex" models should outperform "simple" models. Consistent with their findings, our strongest results are obtained using the most sophisticated machine learning models - the LSTM. All of our machine learning strategies are stronger compared to the ones in Chen and Velikov (2023), who find that using simpler combination techniques results in measly expected returns for strategies based on sorts of individual stock return predictors.

2. Data and Methodology

This section describes our data and methodology.

2.1. Data Sources, Samples and Pre-Processing

Our anomaly data come from Chen and Zimmermann (2022), who provide the most comprehensive dataset of replicated anomalies.[3] We use the March 2022 version of their signals. Motivated by Kelly et al. (2023), we download all 320 anomalies to ensure we provide our machine learning signals with the largest set of characteristics possible.

[3] For a detailed description of the methodology and anomaly composition, as well as the corresponding code and dataset, see their website: https://www.openassetpricing.com

We follow common practice and include only common equity stocks (CRSP share code 10 or 11). Our sample is from March 1957 to December 2021, totaling ~3.4mn stock-month observations over nearly 65 years. The anomaly signals have varying ranges of values over which they are defined, making it more difficult for neural networks to estimate suitable parameters during training (Singh and Singh, 2020). Consequently, we follow the current literature by percent-ranking all anomaly features into the same range [-1;1] (Kelly et al., 2019; Freyberger et al., 2020; Gu et al., 2020). Missing values are replaced with 0.

We enrich the anomaly dataset with further relevant data points following Gu et al. (2020) and include eight key macroeconomic predictors of Welch and Goyal (2008), namely dividend-price ratio, earnings-price ratio, book-to-market ratio, net equity expansion, treasury-bill rate, term-spread, default spread, and stock variance. The objective is to inform our models about the current macroeconomic situation and make them capable of setting it into context for anomaly returns. Furthermore, we incorporate the 49 Fama and French (1997) industry classification indicators publicly available on the Kenneth R. French data library[4] into the feature set. We use one-hot encoding to ensure that our studied models do not suffer multi-collinearity issues, leading to 48 additional features. Next-month returns, market capitalization, and further metadata are obtained from the Center for Research in Security Prices (CRSP). This leads to $320 + 8 + 48 = 376$ features per observation for our models.

[Figure 1 about here.]

Figure 1 reports the number of input features we use over time. For the first year of our asset pricing tests, 2005, our models only use the 137 anomalies from Chen and Zimmermann (2022) with publication dates up to 2004, resulting in $137 + 48 + 8 = 193$ features. The Chen and Zimmermann (2022) database ends in 2016, at which point our list of features reaches its maximum of 376.

To train our machine learning models and tune hyperparameters without any data snooping or look-ahead bias, we follow standard machine learning practice and split the overall dataset into three subsets: a training-, validation-, and testing-set. To allow our model to learn from new information and adapt to the non-stationarity characteristics of stock return time series, we follow the latest literature using an expanding window approach for the training data (Gu et al.,

[4] For more information, see https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

2020; Azevedo and Hoegner, 2023). We re-train the models annually to include new data while keeping a fixed-length moving validation set of six years and a one-year out-of-sample test set. For example, for the out-of-sample year 2005, we use all available stock-month observations from Mar 1957 to Dec 1998 to train the model. We then tune hyperparameters based on the six-year validation set from Jan 1999 to Dec 2004 to ensure the temporal ordering of the observations in the training process. The out-of-sample test set, which we use to evaluate our models regarding machine learning metrics and long-short portfolio performance, contains predictions for January 2005 to December 2021 (i.e., each new year, we move this approach one year ahead, extending the training set).

The portfolio construction process follows common practices in anomaly research. We construct decile portfolios based on the models' next-month stock return predictions, calculating the long-short gross excess return on a monthly rebalancing frequency. Transaction costs are estimated using the composite high-frequency Chen and Velikov (2023) effective bid-ask spread estimator and applied in the calculation of the net excess return following Detzel et al. (2023). To ensure an openly accessible and thus replicable construction and testing protocol, we use the methodology and library of Novy-Marx and Velikov (2023). Further construction details can be found in their paper and on their website.[5]

2.2. Machine learning algorithms and evaluation methodology

Our choice of machine learning models is motivated by prior literature, which shows that neural networks outperform traditional linear regressions as well as penalized ones such as elastic nets (Gu et al., 2020; Chen et al., 2023; Azevedo and Hoegner, 2023; Azevedo et al., 2023; Avramov et al., 2022). We follow Gu et al. (2020), and include the non-constrained OLS-H, a regularized linear model using ENET, and four feedforward neural networks (FFNNs) with hidden layers ranging from two to five.[6] We extend this core model set with two (one- and two-hidden layers, respectively) recurrent neural networks with long short-term memory (LSTMs), which are designed to capture long-term dependencies. Finally, we create an (ENSEMBLE) model by taking the average of all deep-learning models (FFNNs and LSTMs).[7]

Following Chen et al. (2023), we use hyperparameter tuning to find the best set of parameters in each model. To optimize the tuning parameters for each model, we initially implemented a

[5] Documentation available at `http://assayinganomalies.com/`.

[6] A more detailed explanation of the models can be found in the internet appendix of Gu et al. (2020).

[7] We utilize the HPC SLURM cluster from RPTU Kaiserslautern-Landau and run our linear and deep learning models on an AMD EPYC 7262 node with eight cores and 1TB of RAM.

grid search strategy. For efficiency and computational practicality, we selected a representative subset of the data, comprising 20% of the total dataset, to fine-tune these parameters. Once established, these parameters were consistently applied throughout the expanding window estimation process.[8]

However, applying hyperparameter tuning to find the best set of parameters in each model was not feasible for neural network models due to their extensive computational demands and the wide variability in their parameter ranges. For the neural network approaches, we apply a fixed set of parameters. Like Gu et al. (2020), we use the geometric rule to derive the specific neuron configuration for our 2-, 3-, 4-, and 5-hidden-layer FFNNs, and similarly for our 1- and 2-hidden layer LSTMs. All neural networks use an ADAM optimizer with a learning rate of 0.01, the mean squared error (MSE) validation metric, and 200 epochs with a batch size of 10,000 observations. As an activation function, we employ Leaky Rectified Linear Unit (ReLU) because it helps to address the issue of "dead neurons," where neurons stop learning when the gradient becomes zero.[9] We apply dynamic learning rate shrinkage by factor 5 when our validation metrics have not improved for ten epochs of training. We also regularize the models through an early stopping callback, which stops training when validation metrics have not improved for 20 epochs. Furthermore, we minimize the effect of random parameter initialization within the training process by training each model five times and taking the average of the prediction results.

We evaluate the actual out-of-sample trading strategy performance using common portfolio metrics, mainly gross and net excess return of the long-short portfolios, their statistical significance, Sharpe ratio, information ratio, turnover ratio, and transaction costs, as well as the R2. Also, we test the model against the most comprehensive factor model to date, the Fama and French (2018) six-factor model (FF6). To the best of our knowledge, we are the first to estimate the Novy-Marx and Velikov (2016) generalized alphas to machine-learning-based strategies in order to evaluate the ability of these strategies to expand the net-of-costs mean-variance efficient frontier based on the six factors alone.

[8] Our linear models that used hyperparameters were trained with a (training and validation) sample up to 1999 to predict returns from 2000 on.

[9] Leaky ReLU allows a small, non-zero gradient when the input is negative, which can lead to better learning in deep networks compared to (traditional) ReLU. We further discuss the differences between ReLU and Leaky ReLU in the Appendix.

2.3. Turnover and cost mitigation techniques

Novy-Marx and Velikov (2016) and Novy-Marx and Velikov (2019) study the impact of cost-mitigation techniques on the profitability of anomaly trading strategies after accounting for trading costs. In addition to their proposed three techniques, we add further variations and filters to test the effect of cost mitigation on the net performance of machine-learning-based strategies.

One major driver to reduce transaction costs is to reduce turnover rate. Most straightforwardly, this can be achieved using an increased holding period/reduced rebalancing frequency for the portfolio construction. Since we train our models based on monthly predictions, we extend the holding period mitigation to 2-, 3-, or 4-months (H2, H3, H4). Additionally, we create quintile- instead of decile portfolios (QUINTILE), hypothesizing that this could reduce turnover while keeping a significant signal-to-noise ratio. As a more complex variation of adapted overall holding period and number of portfolios, we apply the Buy-Hold-Spread (BHS) technique outlined by Novy-Marx and Velikov (2016) to enter a position for the top-/bottom-decile but exit a position only if they fall out of the top-/bottom-quintile.

Furthermore, we test multiple stock universe filters and weightings. We follow Fama-French in creating a high-market cap filter that excludes the bottom 20% of stocks by market capitalization for our tradable stock universe (HMCAP20). We hypothesize that those high market capitalization stocks are more liquid, i.e., they should face lower transaction costs. In addition, we follow the approach of Novy-Marx and Velikov (2016) in directly filtering out stocks with high previous transaction costs, i.e., using only stocks in the bottom 70%-percentile transaction costs (LTC 70). As an alternative way to incorporate transaction costs into the portfolio construction process, we weight predictions to 25% with the transaction cost percentile (TCWEIGHTED75).[10] Finally, we create two combination strategies that use (a) both H2 and BHS (COMBO1) and (b) H2, BHS, and TCWEIGHTED75 (COMBO2).

3. Machine learning strategy performance decomposition

[Figure 2 about here.]

Figure 2 illustrates the out-of-sample Sharpe ratios for various machine learning strategies

[10] Example: a prediction of 1% excess return but with a transaction cost at the 80% percentile will result in $75\% * 1.0\% + 25\% * (1 - 80\%) = 0.8\%$, while the same return prediction with lower transaction cost at the 20% percentile will result in a $75\% * 1.0\% + 25\% * (1 - 20\%) = 0.95\%$ signal for the models.

applied to stock return prediction. These strategies span a range of complexity, from simpler models like Ordinary Least Squares (OLS) with Huber loss and elastic net to more sophisticated approaches such as feedforward neural networks with varying depths (2-5 layers), long short-term memory (LSTM) models with 1-2 layers, and an ensemble model that combines the neural networks and LSTMs.

The figure presents the Sharpe ratios under different scenarios, with each successive scenario providing a more refined and realistic estimate of the strategies' expected performance. The baseline scenario uses the full set of anomalies and data starting in 2000. The next scenario restricts the anomalies to only those that have been published. The third scenario further restricts the sample to the post-2005 period. Finally, the last scenario incorporates trading costs. For comparison, the market's Sharpe ratio over this period is also shown, assuming no trading costs.

The key observations from Figure 2 reveal several important insights about the performance of machine learning strategies for stock return prediction. First, simpler strategies like OLS-H and elastic net and even FFNN significantly underperform the market when realistic constraints are considered. In contrast, the more sophisticated LSTM-based strategies demonstrate the most impressive performance, consistent with the idea that complexity can be beneficial for return prediction. The ensemble model combining deep learning approaches also performs well, though not as strongly as the LSTM models. Importantly, the figure illustrates how accounting for post-publication effects, focusing on the post-2005 period, and incorporating trading costs successively reduces the Sharpe ratios of all strategies. However, even under these more realistic conditions, the LSTM and ensemble strategies still manage to outperform the market, highlighting their potential value for investors.

[Table 1 about here.]

Table 1 provides a more detailed look at the average monthly returns and associated t-statistics (in brackets) of the machine learning strategies across the different scenarios introduced in Figure 2. The key findings from Table 1 provide a more nuanced understanding of the performance of machine learning strategies for stock return prediction across different scenarios. In the baseline scenario, most strategies generate average returns that are 2-3 times higher than the market. However, restricting the anomalies to only those that have been published reduces returns by an average of 46 basis points per month, while further restricting the sample to the post-2005 period subtracts another 34 basis points per month on average. After accounting for all three effects – post-publication, post-2005, and trading costs – the cumulative reduction in returns amounts to approximately 104 basis points per month. Despite these reductions, the LSTM models stand out by continuing to generate significant positive returns and alphas, even net of trading costs. These findings complement the insights from Figure 2, highlighting that while machine learning strategies' expected returns are more moderate when accounting for real-world considerations, the LSTM models, in particular, still demonstrate potential for meaningful outperformance.

4. The expected returns on machine learning-based strategies

Figure 3 shows the cumulative *net* performance of the machine learning strategies. It reveals that the outperformance of the LSTM and ENSEMBLE strategies is largely due to their impressive performance during the Great Recession. All other machine learning strategies exhibited a severe drawdown in 2009, which the LSTM strategies completely and the ENSEMBLE strategy to some extent are able to avoid. The LSTM1 strategy also performed better in the aftermath of the COVID-19 pandemic at the end of 2020, while all other strategies suffered losses.

[Figure 3 about here.]

Figure 4 presents a correlation heatmap of the returns generated by our nine machine-learning strategies. This visualization allows us to assess the degree of similarity in the return patterns across different approaches. The color scale ranges from dark blue (indicating a strong positive correlation) to white (indicating weak or no correlation).

[Figure 4 about here.]

Several interesting patterns emerge from this heatmap. First, we can observe that the LSTM-based strategies (LSTM1 and LSTM2) exhibit relatively low correlations with the other approaches. This suggests that the LSTM models capture unique patterns in the data not identified by the other techniques. The ENSEMBLE strategy, as expected, shows moderate correlations with most other strategies, reflecting its nature as a combination of multiple approaches. Interestingly, the feedforward neural networks (FFNN2 through FFNN5) show strong correlations with each other, indicating that increasing the number of layers beyond two may not substantially alter the patterns being captured. The linear models (OLS-H and ENET) also show strong correlations with each other but lower correlations with the more complex models, highlighting the potential benefits of non-linear approaches in capturing complex market dynamics.

[Table 2 about here.]

Table 2 reports the performance metrics for our nine machine-learning-based strategies after accounting for trading costs. We can observe that costs significantly impact performance, reducing average returns across the board. The trading costs for the strategies, reported in the last column, range between 19 and 27 basis points per month, rendering the average returns to FFNN2 and FFNN5 insignificant. The two-sided portfolio turnover, reported in the second-to-last column, varies between 121.7% and 139.76% per month, classifying the machine-learning strategies as high-turnover anomalies based on the Novy-Marx and Velikov (2016) taxonomy.

Nevertheless, we still observe economically and statistically significant average returns and generalized six-factor model alphas for the majority of the models. Among the FFNN, the strategy with three hidden layers, as proposed by Gu et al. (2020), reports a generalized FF6 alpha of 0.69% (t-stat of 2.32). The clear winners are the LSTM models; the LSTM1 (LSTM2) earns an impressive 1.39% with t-stat of 3.47 (1.28% with t-stat of 3.35) per month and a generalized FF6 alpha of 1.18% with t-stat of 3.02 (1.26% with t-stat of 3.24).

[Table 3 about here.]

Table 3 presents the ex-post mean-variance efficient (MVE) weights and Sharpe ratios for portfolios combining the Fama-French 6-factor model (FF6) with each of the machine learning strategies. The MVE weights represent the optimal allocation to each factor and the machine learning strategy that would have maximized the portfolio's Sharpe ratio over the sample period based on the realized returns and covariances.

The table shows that the FF6 factors alone yield an in-sample Sharpe ratio of 1.15. The market factor (MKT) receives a consistent weight of around 30% across all portfolios, while the portfolios all short the value factor (HML). The MVE portfolios generally substitute the machine learning strategy for the profitability factor (RMW).

Consistent with the findings from Figure 2 and Table 1, the LSTM models demonstrate the most impressive improvement in the MVE portfolio's Sharpe ratio. The LSTM2 model increases the Sharpe ratio from 1.15 to 1.42, with a 19% allocation to the strategy. The ensemble model, which combines the deep learning approaches, also performs well, increasing the Sharpe ratio to 1.29 with a 19% allocation.

The feedforward neural network strategies receive allocations ranging from 1% to 19%, with the 3-layer network (FFNN3) showing the best improvement in the Sharpe ratio among the FFNN models. The elastic net strategy receives a 20% allocation and increases the Sharpe ratio to 1.28.

These results suggest that the machine learning strategies, particularly the LSTM and ensemble models, have the potential to significantly enhance the mean-variance efficiency of a factor-based portfolio. By combining these strategies with the FF6 factors, investors could have achieved higher risk-adjusted returns over the sample period compared to using the factors alone. However, it is important to note that these are ex-post results based on realized returns, and future performance may vary.

5. Zooming in on the LSTM1 strategy performance

To understand the superior ability of LSTM1 to predict stock returns, we analyze the relative importance of the stock-level predictors and their respective categories for the model's performance. We employ a nullification method similar to Gu et al. (2020), where each predictor is systematically set to zero in the test sample. Then, we observe the annual reduction in out-of-sample R^2 caused by nullifying each predictor. Finally, we rank the feature importance annually on a [0,1] scale, where 1 indicates the most influential predictor.

[Figure 5 about here.]

Figure 5 presents a heatmap of the top 50 stock-level characteristics over time, sorted by their average importance rankings. From 2005 to 2021, the earnings announcement return (*AnnouncementReturn*) emerges as the most important feature. The average percent rank from this feature is 0.9063, which indicates that this predictor introduced by Chan et al. (1996) ranked as the most influential category in 90% of the years.

Following *AnnouncementReturn*, the next most influential features are the change in analyst coverage (*ChNAnalyst*), cash-flow to price variance ($VarCF$), Return seasonality years 2 to 5 (*MomSeason*), and operating cash flows to price quarterly (*cfpq*). Their average ranks are 0.8729, 0.7398, 0.7375, and 0.7335, respectively. Our findings contrast to Gu et al. (2020), which report a dominance of recent price trends (e.g., short-term reversal, stock momentum,

momentum change, industry momentum, recent maximum return, and long-term reversal). A possible reason for this divergence is the difference in model architectures. While Gu et al. (2020) apply linear models, tree-based models, and feedforward neural networks, our LSTM approach uses feedback loops to capture time dependencies, learns long-term patterns, and selectively retains or discards information to enhance prediction accuracy. This different architecture with lower importance to price trends can help explain why LSTM1 performed well even during the momentum crash of 2009.

[Figure 6 about here.]

We then aggregate the individual signals into economic categories based on the classification from Chen and Zimmermann (2022). Figure 6 presents the heatmap of these categories over time. The *Earnings Event* category is the most significant, which is expected since it includes the two highest-ranking anomalies (*AnnouncementReturn* and *ChNAnalyst*). The next most important categories are *Cash Flow Risk*, which includes *VarCF* and return on asset volatility (*roavol*), and *Valuation*, which includes up to 30 features (depending on the publication year). In contrast, the *Momentum* category ranks 25th out of 35 categories, and *Short-Term Reversal* ranks 23rd. Overall, our results indicate that LSTM1 provides stable out-of-sample predictions by leveraging mostly event-based and accounting-based signals and not giving so much importance to market-based signals.

Among the macroeconomic variables, the dividend-price ratio (DP) has an average percent rank of 1 (i.e., the variable with the highest influence in all years from our test sample). Book-to-market ratio (BM) emerges as the second most influential variable with an average percent rank of 0.5966, while treasure-bill rate (TBL) is the third with an average rank (0.4958). These results indicate that valuation metrics greatly influence the accuracy of the LSTM1.

Next, we assess the robustness of LSTM1 as a predictor for the cross-section of stock returns. In the Online Appendix, we evaluate the performance of the LSTM1 strategy using the "Assaying Anomalies" protocol proposed by Novy-Marx and Velikov (2023). This protocol consists of five key steps: (1) providing basic signal diagnostics, (2) checking return predictability, (3) comparing performance to the universe of known anomalies, (4) finding closely related anomalies, and (5) controlling for the entire factor zoo.

In the first step, the LSTM1 signal demonstrates good coverage, with a consistent 25th to 75th percentile range and a stable mean. The second step reveals strong gross profitability for the LSTM1 strategy, with significant excess returns and alphas relative to various factor models.

The strategy's performance is robust to alternative portfolio construction methods and trading costs, and a double sort on size and the LSTM1 signal shows that the strategy's returns are not concentrated among microcaps.

Comparing the LSTM1 strategy to the anomaly universe in the third step, the strategy exhibits an impressive Sharpe ratio and alpha relative to other anomalies, both gross and net of trading costs. In the fourth step, the LSTM1 strategy shows low correlations with other anomalies and does not cluster tightly with any specific group. Furthermore, the strategy's predictive power remains significant when controlling for the most closely related anomalies using Fama-MacBeth regressions and spanning tests. Finally, in the fifth step, the LSTM1 strategy's performance is evaluated in the context of combination strategies that pool information from the entire anomaly universe. While the results are mixed, the LSTM1 signal does not consistently enhance the combination strategies' performance, suggesting that its predictive power may already be captured by the collective information in the anomaly universe.

Overall, the LSTM1 strategy performs well across the majority of the "Assaying Anomalies" protocol tests. It demonstrates robust predictive power, generates significant returns and alphas, and appears to capture unique information not fully represented by existing anomalies. However, the strategy's incremental contribution to combination strategies that already exploit the full anomaly universe is less clear. These findings suggest that the LSTM1 strategy is a promising approach for stock return prediction, but its value may be most pronounced when used in isolation or in combination with a more limited set of anomalies.

6. Applying turnover and transaction cost mitigation strategies

The machine-learning-based strategies examined thus far were designed without regard for trading costs. Recently, Blitz et al. (2023) show that training the models in a longer time horizon can lead to higher returns. Furthermore, Avramov et al. (2022) show that excluding firms with economic constraints can reduce the significance of machine learning models. In this section, we investigate different mitigation techniques and their impact on the net performance of machine learning-based strategies.

[Figure 7 about here.]

Table 4 and Figure 7 show the impact of the previously outlined mitigation approaches on our four classes of model architectures, namely linear models, FFNNs, LSTMs, and the ensemble model. We show the impact of absolute differences in the net excess return portfolio metrics and generalized FF6 alpha and relative changes in turnover and transaction costs. As the results show, most of the cost-mitigation techniques significantly reduce turnover and, as a result, transaction costs.

[Table 4 about here.]

This decrease in transaction costs, however, is only beneficial if it is not accompanied by a larger reduction in gross returns. As we can observe in Table 4, the average change net excess returns across the nine machine learning models is negative for all but one mitigation technique. This implies that the drop in the gross average returns due to the mitigation techniques more than compensates for the reduced trading costs. This is likely because our testing sample period, which consists of the last two decades, is marked by higher liquidity and significantly lower trading costs after the introduction of the decimal trading system (Chordia et al., 2014; Chen and Velikov, 2023). The only technique that seems to marginally improve the net average returns across the nine machine learning strategies is the two-month holding period. Not surprisingly, the stock universe filters have a smaller impact on turnover but a similar impact on transaction costs, as they aim to reduce the weight of high-cost stocks. However, the change in net excess returns for these methods is similarly negative.

7. Conclusion

This study provides a comprehensive assessment of machine learning-based anomaly trading strategies, focusing on their expected returns in real-world conditions. By accounting for transaction costs, post-publication decay, and the post-decimalization era of high liquidity, we offer a more realistic estimate of these strategies' performance than typically reported in the literature.

Our findings challenge some prevailing notions in the field while confirming others. Contrary to claims in prior literature, we find that sophisticated machine learning strategies, particularly those employing Long Short-Term Memory (LSTM) models, remain profitable even after accounting for real-world constraints. The best-performing LSTM strategy earns net out-of-sample monthly returns of 1.39% and a six-factor generalized net alpha of 1.18%, with statistically significant t-statistics. This performance persists despite high turnover rates of up to 129% and the inclusion of some difficult-to-arbitrage stocks.

A key contribution of our study is the decomposition of backtested returns to arrive at expected returns. We find that restricting the sample to the post-2005 period has the largest impact on performance, followed by post-publication effects and trading costs. Cumulatively, these factors reduce strategy performance by 23-92% compared to typical backtests. However, the fact that some strategies, particularly LSTM models, continue to outperform even under these constraints is evidence of the potential of advanced machine learning techniques in asset pricing.

Our analysis of cost-mitigation techniques reveals an interesting paradox. While these techniques successfully reduce turnover and trading costs, they generally do not improve net anomaly performance in the recent high-liquidity era. This finding suggests that the benefits of reduced costs are often offset by a decrease in the strategy's ability to capture short-term price movements.

The superior performance of LSTM models, especially during the Great Recession when many other strategies faltered, highlights the potential for machine learning to capture complex, non-linear relationships in financial data. The low correlation between LSTM models and other machine-learning approaches further underscores their unique contribution to the field of quantitative finance.

Our rigorous evaluation of the LSTM1 signal using the "Assaying Anomalies" protocol demonstrates its robustness across multiple dimensions. This comprehensive testing framework provides strong evidence for the signal's validity and predictive power, even when controlling for a wide array of known anomalies and risk factors.

These findings have significant implications for both academic research and practical investment strategy design. For academics, our results underscore the importance of considering real-world constraints when evaluating new predictive signals or models. They also highlight the potential for advanced machine learning techniques to uncover persistent sources of alpha in financial markets.

For practitioners, our study offers valuable insights into the realistic performance expectations for machine learning-based strategies. While the raw performance of these strategies may be less impressive than some backtests, the persistent outperformance of certain models, particularly LSTM, indicates their potential value in investment portfolios.

Future research could extend this work in several directions. First, investigating the specific features or architectures that contribute to the superior performance of LSTM models could yield valuable insights. Second, exploring the integration of transaction cost models directly into the machine learning framework might lead to strategies that are inherently more robust to trading frictions. Finally, examining the performance of these strategies across different asset classes or international markets could provide a more comprehensive understanding of their efficacy.

In conclusion, while our study tempers some of the more optimistic claims about machine learning in asset pricing, it also provides compelling evidence for the continued relevance and potential of these techniques. As we venture deeper into the age of artificial intelligence and big data, research at the intersection of machine learning and finance will undoubtedly continue to yield valuable insights and innovative investment approaches.

References

Avramov, D., Cheng, S., and Metzker, L. (2022). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*.

Azevedo, V. and Hoegner, C. (2023). Enhancing stock market anomalies with machine learning. *Review of Quantitative Finance and Accounting*, 60(1):195–230.

Azevedo, V., Kaiser, S., and Müller, S. (2023). Stock market anomalies and machine learning across the globe. *Journal of Asset Management*, Forthcoming:1–23.

Blitz, D., Hanauer, M. X., Hoogteijling, T., and Howard, C. (2023). The term structure of machine learning alpha. *SSRN Electronic Journal*, pages 1–40.

Brandt, M. W., Santa-Clara, P., and Valkanov, R. (2009). Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *The Review of Financial Studies*, 22(9):3411–3447.

Brogaard, J. and Zareei, A. (2023). Machine learning and the stock market. *Journal of Financial and Quantitative Analysis*, 58(4):1431–1472.

Cakici, N., Fieberg, C., Metko, D., and Zaremba, A. (2023). Machine learning goes global: Cross-sectional return predictability in international stock markets. *Journal of Economic Dynamics and Control*, 155:104725.

Chan, L. K. C., Jegadeesh, N., and Lakonishok, J. (1996). Momentum strategies. *The Journal of Finance*, 51(5):1681–1713.

Chen, A. Y. and Velikov, M. (2023). Zeroing in on the expected returns of anomalies. *Journal of Financial and Quantitative Analysis*, 58(3):968–1004.

Chen, A. Y. and Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Review of Finance*, 27(2):207–264.

Chen, L., Pelger, M., and Zhu, J. (2023). Deep learning in asset pricing. *Management Science*, (Forthcoming).

Chordia, T., Subrahmanyam, A., and Tong, Q. (2014). Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? *Journal of Accounting and Economics*, 58(1):41–58.

Cong, L., Tang, K., Wang, J., and Zhang, Y. (2020). AlphaPortfolio for Investment and Economically Interpretable AI. *SSRN Electronic Journal*.

Detzel, A., Novy-Marx, R., and Velikov, M. (2023). Model comparison with transaction costs. *The Journal of Finance*, 78(3):1743–1775.

Evgeniou, T., Guecioueur, A., and Prieto, R. (2023). Uncovering sparsity and heterogeneity in firm-level return predictability using machine learning. *Journal of Financial and Quantitative Analysis*, 58(8):3384–3419.

Fama, E. F. and French, K. R. (1997). Industry costs of equity. *Journal of Financial Economics*, 43(2):153–193.

Fama, E. F. and French, K. R. (2018). Choosing factors. *Journal of Financial Economics*, 128(2):234–252.

Feng, G., He, J., Polson, N. G., and Xu, J. (2023). Deep learning in characteristics-sorted factor models. In *Journal of Financial and Quantitative Analysis*. Cambridge University Press.

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

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

Hanauer, M. X. and Kalsbach, T. (2023). Machine learning and the cross-section of emerging market stock returns. *Emerging Markets Review*, 55:101022.

Hansen, J. H. and Siggaard, M. V. (2023). Double machine learning: Explaining the post-earnings announcement drift. *Journal of Financial and Quantitative Analysis*, Forthcoming.

Jensen, T. I., Kelly, B. T., Malamud, S., and Pedersen, L. H. (2022). Machine learning and the implementable efficient frontier. *SSRN Electronic Journal*, pages 1–67.

Kelly, B., Malamud, S., and Zhou, K. (2023). The virtue of complexity in return prediction. *The Journal of Finance*.

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.

Leippold, M., Wang, Q., and Zhou, W. (2022). Machine learning in the Chinese stock market. *Journal of Financial Economics*, 145(2):64–82.

Li, B., Rossi, A., Yan, X., and Zheng, L. (2022). Real-time machine learning in the cross-section of stock returns. *Working paper*.

Loughran, T. and McDonald, B. (2023). Measuring firm complexity. *Journal of Financial and Quantitative Analysis*.

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

Meursault, V., Liang, P. J., Routledge, B. R., and Scanlon, M. M. (2023). PEAD.txt: Post-earnings-announcement drift using text. In *Journal of Financial and Quantitative Analysis*, volume 58, pages 2299–2326. Cambridge University Press.

Novy-Marx, R. (2016). Testing strategies based on multiple signals. *Working paper*.

Novy-Marx, R. and Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *Review of Financial Studies*, 29(1):104–147.

Novy-Marx, R. and Velikov, M. (2019). Comparing cost-mitigation techniques. *Financial Analysts Journal*, 75(1):85–102.

Novy-Marx, R. and Velikov, M. (2023). Assaying anomalies. *SSRN Electronic Journal*, pages 1–36.

Simon, F., Weibels, S., and Zimmermann, T. (2022). Deep parametric portfolio policies. *SSRN Electronic Journal*.

Singh, D. and Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97:105524.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium Prediction. *Review of Financial Studies*, 21(4):1455–1508.

Table 1: Decomposition of Monthly Returns (%)

Model	Baseline	Post-Publication	Post-2005 Sample	Net of Costs	Total Change
Panel A: Market					
MKT	0.61	0.61	0.86	0.86	+0.25
	[2.18]	[2.18]	[2.82]	[2.82]	
Panel B: Linear Models					
OLS-H	1.30	0.80	0.42	0.22	-1.08
	[3.62]	[2.02]	[1.18]	[0.62]	
ENET	1.48	1.15	0.83	0.64	-0.84
	[5.42]	[4.25]	[3.16]	[2.43]	
Panel C: Neural Networks					
FFNN2	1.90	0.69	0.41	0.16	-1.74
	[5.37]	[2.12]	[1.20]	[0.48]	
FFNN3	1.82	1.23	1.01	0.76	-1.06
	[5.51]	[3.76]	[3.25]	[2.43]	
FFNN4	2.04	1.37	0.98	0.71	-1.33
	[5.54]	[3.85]	[2.76]	[2.02]	
FFNN5	2.14	1.49	1.17	0.90	-1.24
	[5.54]	[3.86]	[3.09]	[2.39]	
Panel D: LSTM Models					
LSTM1	1.80	1.94	1.60	1.39	-0.41
	[4.86]	[4.99]	[3.97]	[3.47]	
LSTM2	1.85	1.74	1.49	1.28	-0.57
	[5.21]	[4.60]	[3.86]	[3.35]	
Panel E: Ensemble					
ENSEMBLE	1.89	1.63	1.09	0.85	-1.04
	[5.29]	[4.74]	[3.27]	[2.56]	
Panel F: ML Average					
Average (ML)	1.80	1.34	1.00	0.77	-1.03

This table presents the decomposition of monthly returns (in percentages) for various machine-learning strategies across different scenarios. The 'Baseline' scenario uses the full set of anomalies and data starting in 2000. 'Post-Publication' restricts anomalies to only those that have been published. 'Post-2005 Sample' further restricts the sample to the post-2005 period. 'Net of Costs' incorporates trading costs. 'Total Change' shows the difference between 'Net of Costs' and 'Baseline' returns. T-statistics are reported in square brackets below the corresponding returns. ML = Machine Learning, OLS = Ordinary Least Squares, ENET = Elastic Net, FFNN = Feedforward Neural Network, LSTM = Long Short-Term Memory.

Table 2: Out-of-Sample Net Performance of Machine-Learning Anomaly Strategies

Model architecture	Net monthly excess return (%)	Net monthly Sharpe Ratio	Generalized alpha FF6 (%)	Two-sided turnover (%)	Transaction costs (%)
OLS-H	0.22	0.15	0.39	121.7	0.20
	[0.62]		[1.50]		
ENET	0.64	0.59	0.55	139.76	0.19
	[2.43]		[2.20]		
FFNN2	0.16	0.12	0.06	130.25	0.25
	[0.48]		[0.20]		
FFNN3	0.76	0.59	0.69	127.05	0.26
	[2.43]		[2.32]		
FFNN4	0.71	0.49	0.59	126.31	0.26
	[2.02]		[1.78]		
FFNN5	0.90	0.58	0.54	126.80	0.27
	[2.39]		[1.56]		
LSTM1	1.39	0.84	1.18	128.04	0.21
	[3.47]		[3.02]		
LSTM2	1.28	0.81	1.26	128.97	0.21
	[3.35]		[3.24]		
ENSEMBLE	0.85	0.62	0.73	130.95	0.24
	[2.56]		[2.31]		

This table presents the out-of-sample net performance metrics for various machine learning anomaly strategies from January 2005 to December 2021. All portfolios are constructed as value-weighted long-short decile portfolios with NYSE breakpoints and a one-month holding period/rebalancing frequency. 'Net monthly excess return' and 'Generalized alpha FF6' (generalized alpha relative to the Fama-French 6-factor model) are reported in percentages, with t-statistics in square brackets. 'Net monthly Sharpe Ratio' is the annualized Sharpe Ratio estimated as the strategy's mean net return scaled by its standard deviation. 'Two-sided turnover' is the average monthly two-sided portfolio turnover in percentages. 'Transaction costs' are estimated using the Chen and Velikov (2023) high-frequency combination effective bid-ask spread estimator and reported in percentages. OLS = Ordinary Least Squares, ENET = Elastic Net, FFNN = Feedforward Neural Network, LSTM = Long Short-Term Memory.

Table 3: Net Ex-Post Mean-Variance Efficient Weights and Sharpe Ratios

2lModel	Mean-Variance Efficient Weights (%)							Sharpe	ΔSharpe
	MKT	SMB	HML	RMW	CMA	UMD	ML	Ratio	
FF6	28	9	-22	59	22	3		1.15	
FF6+OLS-H	30	17	-25	47	19	0	12	1.21	+0.06
FF6+ENET	29	11	-27	49	17	0	20	1.28	+0.13
FF6+FFNN2	28	9	-23	59	22	2	1	1.15	+0.00
FF6+FFNN3	30	9	-33	52	23	0	19	1.29	+0.14
FF6+FFNN4	32	8	-30	54	23	0	13	1.24	+0.09
FF6+FFNN5	31	7	-28	55	24	0	11	1.22	+0.07
FF6+LSTM1	33	8	-28	44	19	4	20	1.38	+0.24
FF6+LSTM2	31	4	-25	46	22	3	19	1.42	+0.27
FF6+ENSEMBLE	33	6	-35	52	24	0	19	1.29	+0.14

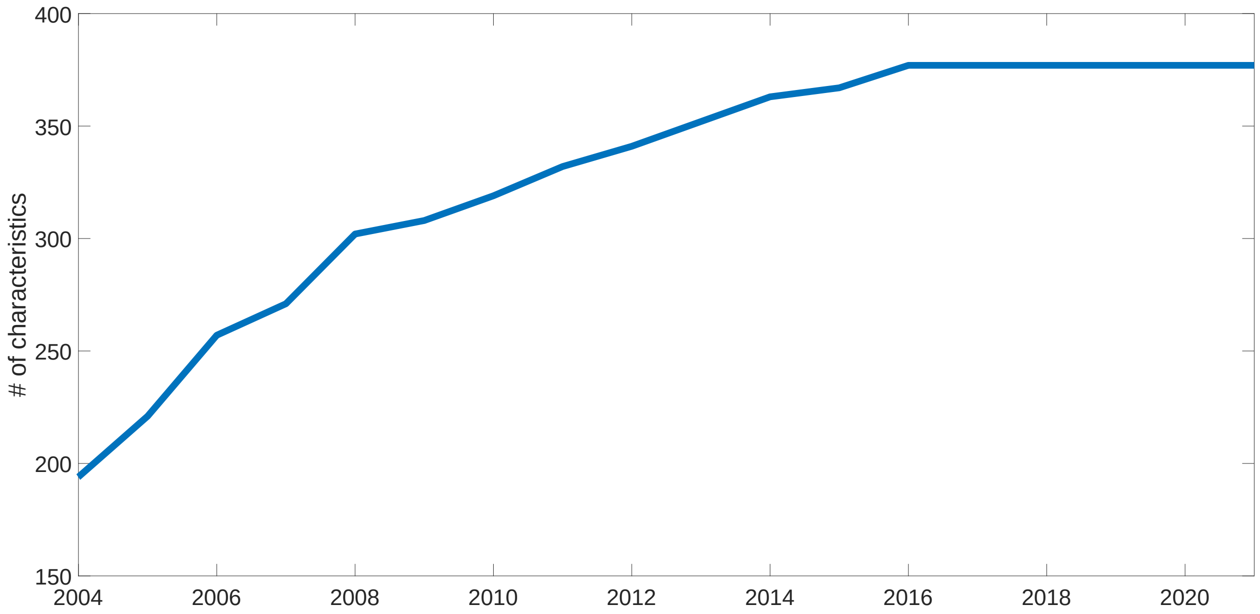
This table presents the net ex-post mean-variance efficient (MVE) weights and Sharpe ratios for portfolios combining the Fama-French 6-factor model (FF6) with each machine-learning strategy. MKT = Market, SMB = Size, HML = Value, RMW = Profitability, CMA = Investment, UMD = Momentum, ML = Machine Learning strategy. ΔSharpe shows the improvement in Sharpe ratio compared to the FF6 baseline. OLS-H = Ordinary Least Squares with Huber loss, ENET = Elastic Net, FFNN = Feedforward Neural Network, LSTM = Long Short-Term Memory.

Table 4: Average mitigation technique effect across different model architectures

	relative change in %		absolute average Δ to baseline	
Mitigation technique	Two-sided turnover	Transaction costs	Net excess return in p.p.	[t-stat]
H2	-45.89	-46.39	0.07	0.23
H3	-62.31	-62.89	-0.20	-0.53
H4	-71.16	-72.20	-0.24	-0.46
BHS	-53.14	-56.22	-0.13	0.19
QUINTILE	-12.23	-34.47	-0.31	-0.17
HMCAP20	-1.72	-29.71	-0.14	-0.31
LTC70	3.12	-53.42	-0.09	-0.42
TCWEIGHTED75	-1.49	-16.61	-0.04	0.00
COMBO1	-72.16	-73.14	-0.36	-0.85
COMBO2	-72.18	-76.64	-0.38	-0.87

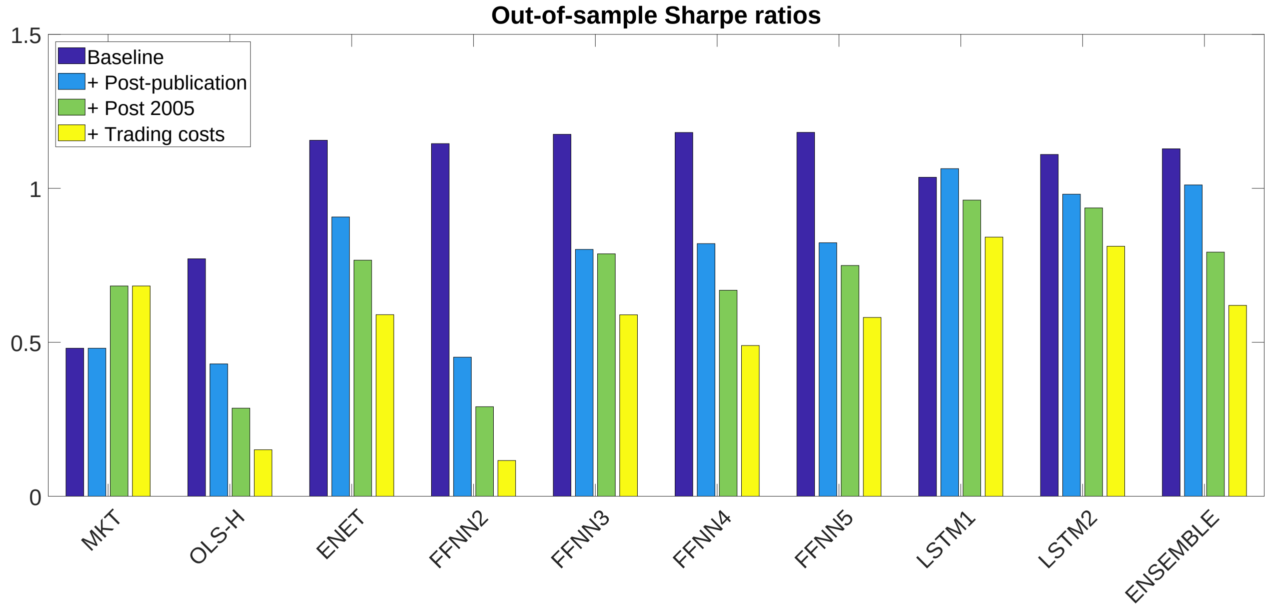
This table shows the average effect on the baseline model of applying each mitigation technique in our out-of-sample period from January 2005 to December 2021. We apply different techniques to our sample: Extended holding period/reduced rebalancing frequency (H2, H3, and H4 for 2-, 3-, and 4-month periods), a BHS, quintile instead of decile portfolios (QUINTILE), a high market cap filter (HMCAP20), a low transaction cost filter (LTC70), and a weighting of next month's predicted returns by estimated transaction costs (TCWEIGHTED75). We report the impact in % change of the respective variable compared to the baseline model without mitigations.

Figure 1: Number of characteristics used over time



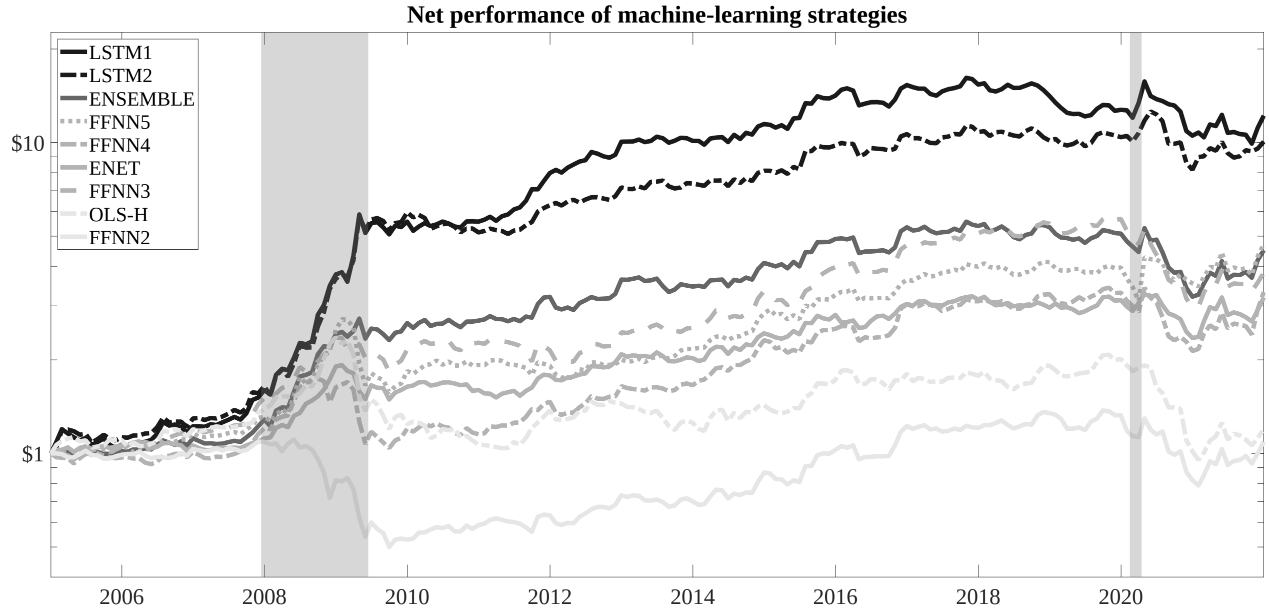
The figure plots the number of characteristics used in the construction of our machine-learning strategies over time.

Figure 2: Return decomposition



The figure describes the monthly performance of our baseline models gross and net of trading costs in the out-of-sample period from January 2005 to December 2021. The transaction costs are estimated using the Chen and Velikov (2023) high-frequency combination effective bid-ask spread estimator. The bars show the average monthly gross and net excess returns of the value-weighted long-short decile portfolios based on NYSE breakpoints. The *t*-statistic is in brackets. The labels below the columns show the respective relative drop in return in % due to the introduction of transaction costs.

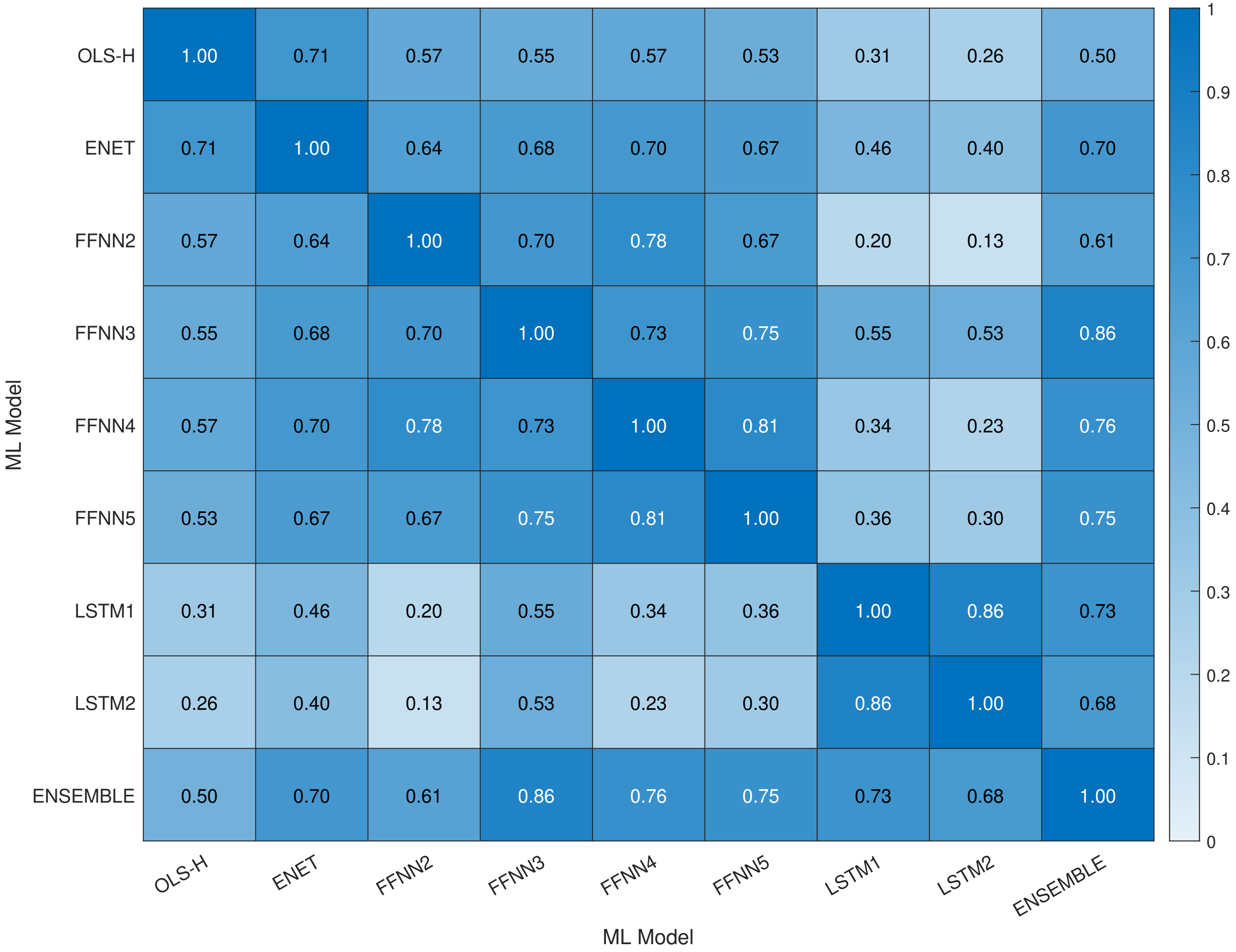
Figure 3: Net performance of machine-learning strategies



The figure describes the monthly performance of machine-learning anomaly strategies net of trading costs in the out-of-sample period from January 2005 to December 2021. The transaction costs are estimated using the Chen and Velikov (2023) high-frequency combination effective bid-ask spread estimator. The bars show the average monthly gross and net excess returns of the value-weighted long-short decile portfolios based on NYSE breakpoints.

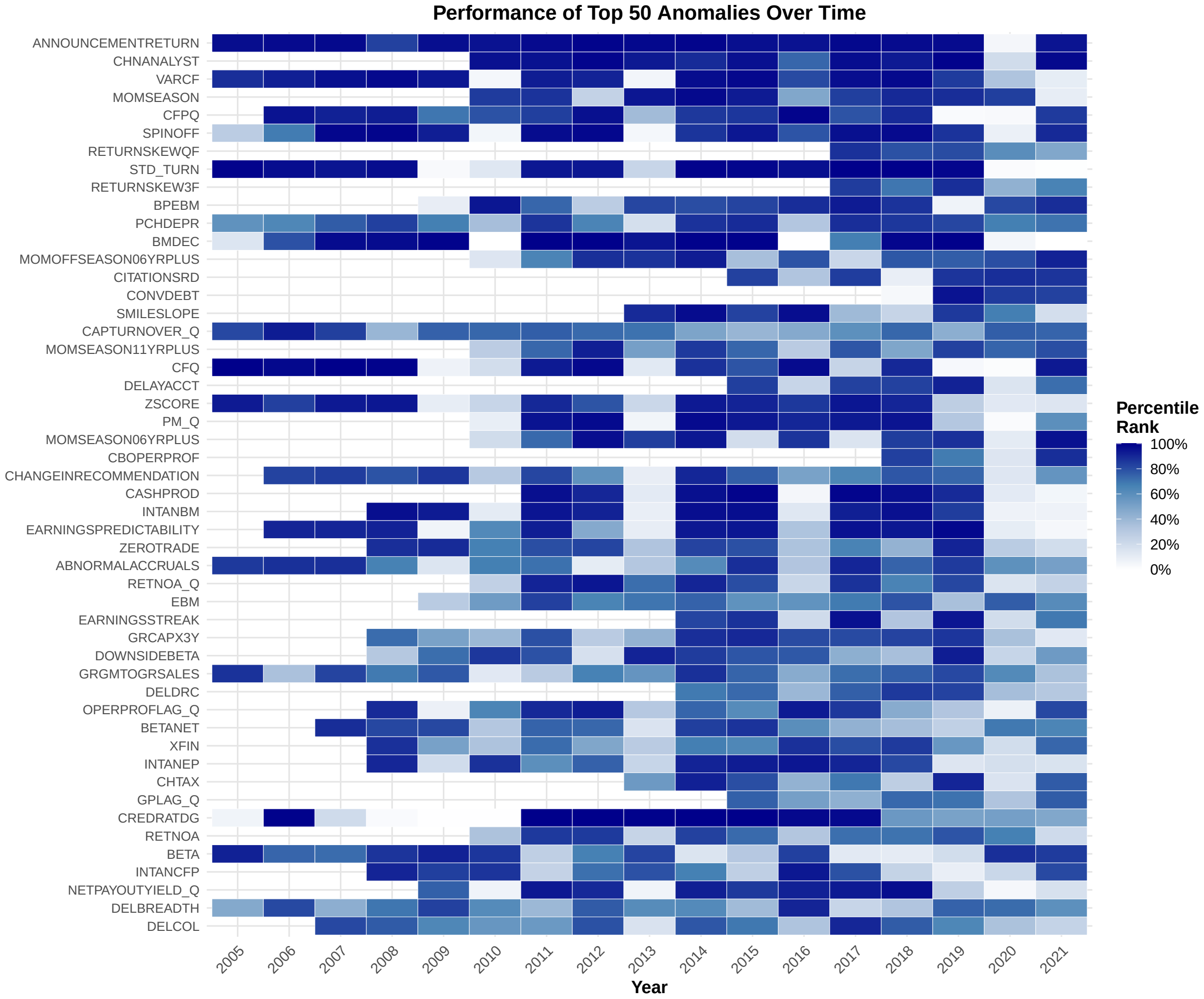
.

Figure 4: Correlation heatmap



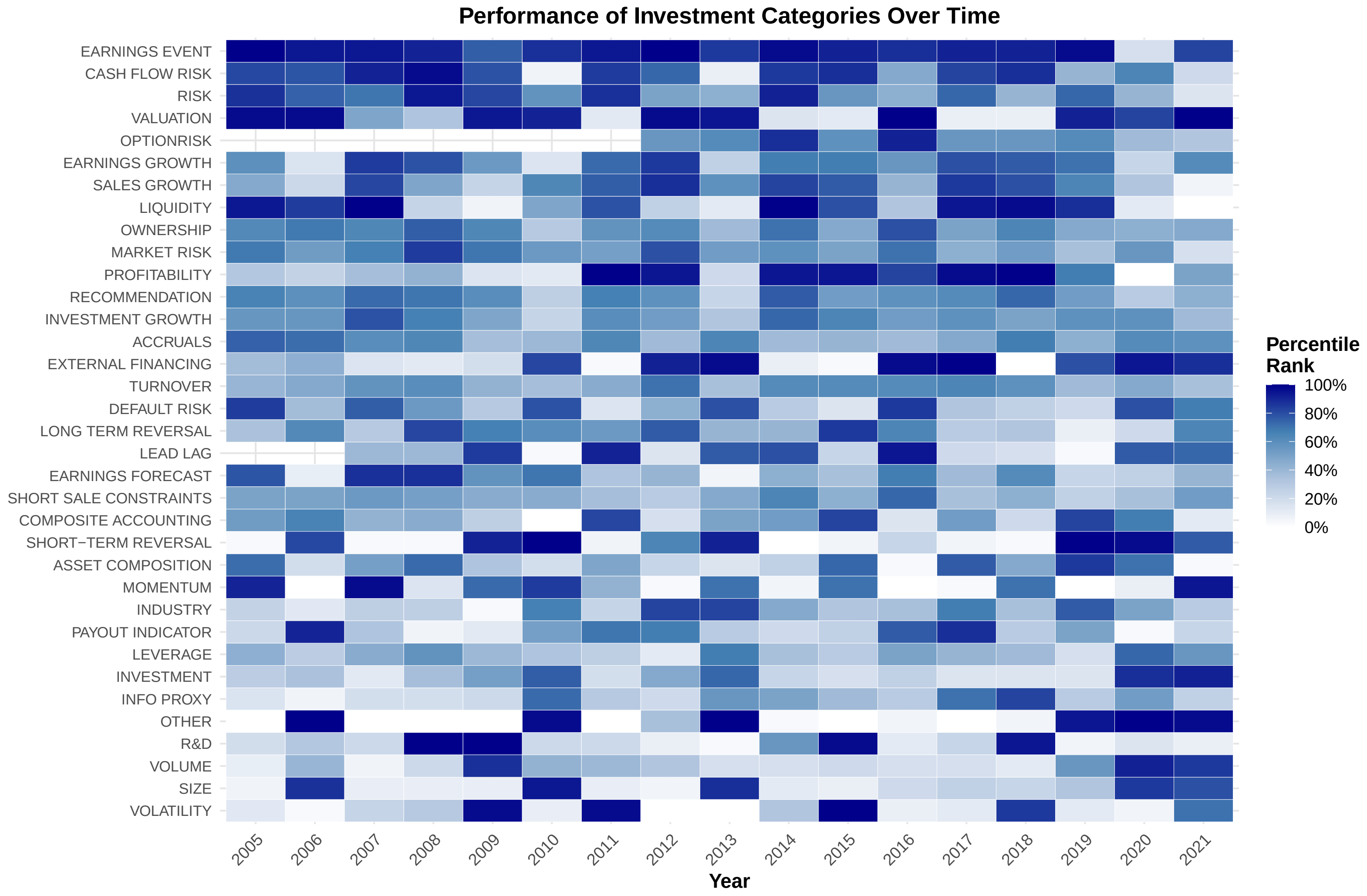
This figure plots the pairwise correlations between the returns of nine different machine-learning strategies. The color scale ranges from dark blue (strong positive correlation) to white (weak or no correlation). Strategies include Ordinary Least Squares with Huber Loss (OLS-H), Elastic Net (ENET), Feedforward Neural Networks with 2-5 layers (FFNN2-5), Long Short-Term Memory networks with 1-2 layers (LSTM1-2), and an ensemble model (ENSEMBLE). The heatmap reveals varying degrees of correlation between different strategies, with LSTM-based approaches showing the lowest correlations with other methods.

Figure 5: Top 50 Firm-level characteristics importance heatmap



This figure plots the top 50 firm-level characteristics' importance over time for the LSTM1 signal. We calculate the reduction in R^2 every year by nullifying each predictor from our test sample. Then, characteristics are (percent) ranked yearly on a [0,1] scale, where 1 denotes the most influential. Columns represent years, with color gradients from dark blue (most influential) to white (least influential), indicating relative feature importance.

Figure 6: Category importance heatmap

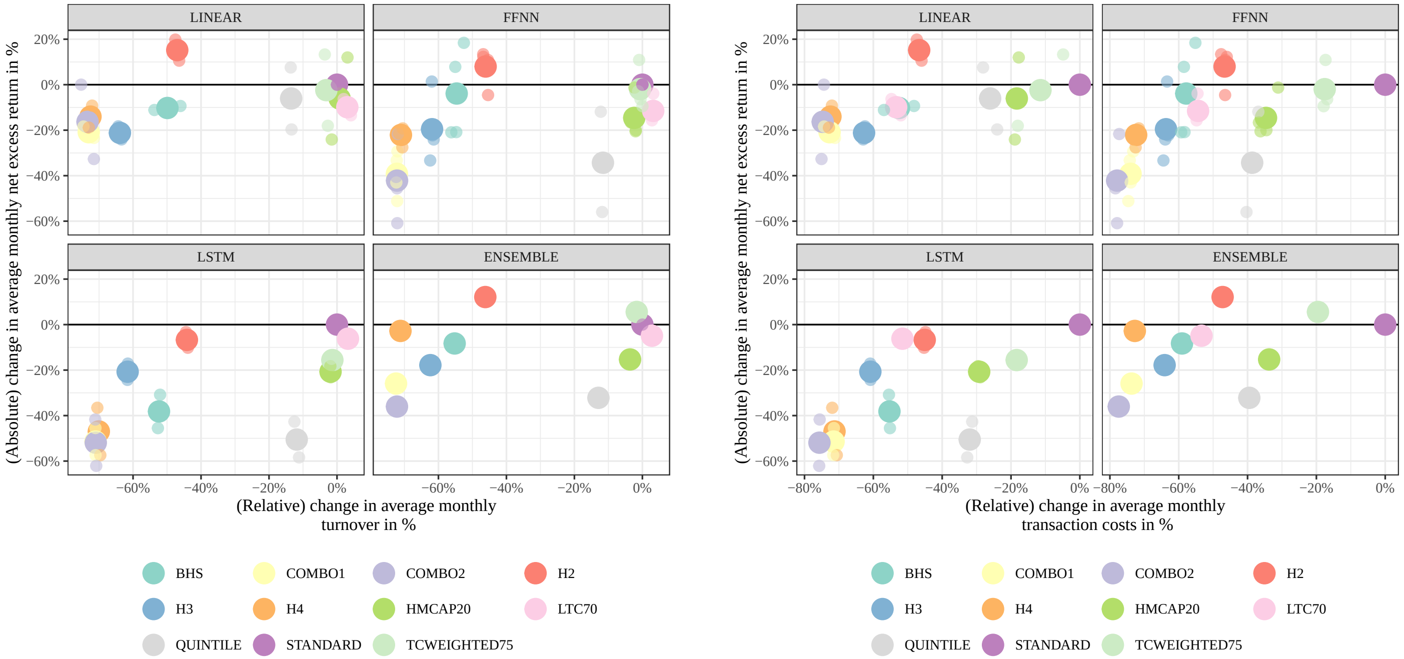


This figure plots the category importance over time for the LSTM1 signal. We calculate the reduction in R^2 every year by nullifying all predictors within each category from our test sample. Then, categories are ranked yearly on a [0,1] scale, where 1 denotes the most influential. The category classification is based on Chen and Zimmermann (2022). The alternative categories (investment alternative and profitability alternative) are combined with their respective categories (investment and profitability) to save space. Columns represent years, with color gradients from dark blue (most influential) to white (least influential), indicating relative category importance.

Figure 7: Net return impact of different turnover and cost mitigation techniques on portfolio performance

(a) Turnover-Performance effect

(b) Transaction costs-Performance effect



The table illustrates the turnover and transaction cost relations to the respective change in net excess return in the out-of-sample period from Jan 2005 to Dec 2021. All changes are in % compared to the baseline model without mitigation techniques in a transaction cost environment.

Appendix A.

Elastic Net Regression

The Elastic Net is a linear regression model that combines the L1 and L2 regularization of the Lasso and Ridge regression methods. This approach is beneficial when dealing with highly correlated independent variables. The objective function of the Elastic net is as follows:

$$\text{Minimize}\left(\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + \frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2\right)\right) \tag{A.1}$$

where y_i represents the response variable, x_{ij} are the predictors, β_j are the coefficients, λ is the regularization parameter, and α is the mixing parameter between Lasso and Ridge penalties.

Feedforward neural networks (FFNNs)

Feedforward Neural Networks (FFNNs) are one of the most used forms of artificial neural networks. In these networks, information moves in a single direction—from the input layer through one or more hidden layers to the output layer. Each layer consists of interconnected neurons that apply activation functions to process the input. In the following, we demonstrate the mathematical formulation of a feedforward neural network (FFNN) with two hidden layers with a Leaky ReLU activation function. In the first hidden layer, the transformation of the input vector x is given by:

$$H_1 = LeakyReLU(W_1 \cdot x + b_1) \tag{A.2}$$

here, the input x is multiplied by the weight matrix W_1 and added to the bias vector b_1 using the Leaky ReLU activation function. In the second hidden layer, the output of the first hidden layer, H_1, is further processed as follows:

$$H_2 = LeakyReLU(W_2 \cdot H_1 + b_2). \tag{A.3}$$

Similarly, H_1 is multiplied by the weight matrix W_2 and added to the bias vector b_2, with the Leaky ReLU activation applied again. In the output layer, H_2 is transformed using the weight matrix W_3 and bias vector b_3, followed by the Leaky ReLU activation function to generate the final prediction $\hat{y}$:

$$\hat{y} = LeakyReLU(W_3 \cdot H_2 + b_3). \tag{A.4}$$

The weight matrices W_1, W_2, and W_3 correspond to the transformations in each layer, while the bias vectors b_1, b_2, and b_3 adjust the outputs before the activation functions are applied. The Leaky ReLU activation function is defined as:

$$LeakyReLU(z) = \max(\alpha z, z)$$

where α is a constant, which allows small negative values for z instead of setting them to zero. In our paper, we use $\alpha = 0.3$.

The main difference between ReLU and Leaky ReLU lies in how they handle negative input values. For any input $z < 0$, the output of ReLU is set to 0. This can lead to the issue of "dead neurons," where neurons stop learning when the gradient becomes zero. Leaky ReLU introduces a small slope (αz) for negative input values rather than setting them to zero, mitigating the dead neuron problem. This allows the network to propagate negative values through the layers, preserving information and gradients even for negative inputs.

Backpropagation and Initialization of Weights and Biases

Backpropagation is a fundamental algorithm in supervised learning used for training FFNNs. The core idea of backpropagation is to adjust the network's weights and biases to minimize the error between the predicted and actual outputs. Mathematically, this involves computing the gradient of the loss function with respect to each weight and bias in the network. As an example, consider a network with L layers, each with weights $W^{(l)}$ and biases $b^{(l)}$ for layer l. In the process of backpropagation, initially the forward pass computes the output of the network for a given input. For any layer l, the output $H^{(l)}$ is obtained as:

$$H^{(l)} = \sigma(W^{(l)} \cdot H^{(l-1)} + b^{(l)}) \tag{A.5}$$

where σ represents the activation function, and $H^{(0)}$ is the input to the network. Then, the loss (error) $\mathcal{L}$ is computed by comparing the predicted output $\hat{y}$ with the actual output y. In our paper, we apply Mean Squared Error (MSE) as the loss function, which is expressed as:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2. \tag{A.6}$$

After calculating the loss, in the backward pass, the gradient of the loss with respect to each weight and bias is computed. For a weight $W_{ij}^{(l)}$ in layer l, the gradient is calculated as:

$$\frac{\partial \mathcal{L}}{\partial W_{ij}^{(l)}} = \frac{\partial \mathcal{L}}{\partial H_i^{(l)}} \cdot \frac{\partial H_i^{(l)}}{\partial W_{ij}^{(l)}}. \tag{A.7}$$

The final step is to update the weights and biases, adjusting them in the direction that reduces the loss. This is typically performed using gradient descent. For the weights $W_{ij}^{(l)}$ and biases $b_i^{(l)}$, the updates are applied as follows:

$$W_{ij}^{(l)} = W_{ij}^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial W_{ij}^{(l)}} \tag{A.8}$$

$$b_i^{(l)} = b_i^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial b_i^{(l)}} \tag{A.9}$$

where η represents the learning rate, which controls the step size for the updates. In our paper, we use a learning rate of 0.01.

Long Short-Term Memory (LSTM) Networks

LSTMs are a type of Recurrent Neural Network (RNN) that is capable of learning long-term dependencies in sequential data, which is particularly useful in financial time series forecasting. An LSTM unit has several key mechanisms that manage the flow of information. The forget gate controls which information should be discarded from the cell state, while the input gate updates the cell state with new relevant data. Finally, the output gate determines what the next hidden state should be, allowing the network to carry important information through time.

The operations inside an LSTM cell can be formulated as follows:

$$\begin{aligned}
\text{Forget Gate:} \quad & f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) && \text{(A.10)}\\
\text{Input Gate:} \quad & i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) && \text{(A.11)}\\
\text{Cell State Update:} \quad & \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) && \text{(A.12)}\\
\text{Final Cell State:} \quad & C_t = f_t * C_{t-1} + i_t * \tilde{C}_t && \text{(A.13)}\\
\text{Output Gate:} \quad & o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) && \text{(A.14)}\\
\text{Output:} \quad & h_t = o_t * \tanh(C_t) && \text{(A.15)}
\end{aligned}$$

Where f_t, i_t, and o_t are the activations of the forget, input, and output gates, respectively; C_t is the cell state; h_t is the hidden state; W and b are the weights and biases for each gate, and x_t is the input at time t.

Online Appendix for The Expected Returns on Machine-Learning Strategies: Long-Short Term Memory Model signal and the Cross Section of Stock Returns

Vitor Azevedo Christopher Hoegner Mihail Velikov

December 16, 2024

Abstract

This report studies the asset pricing implications of Long-Short Term Memory Model signal (LSTM1), and its robustness in predicting returns in the cross-section of equities using the protocol proposed by Novy-Marx and Velikov (2023). A value-weighted long/short trading strategy based on LSTM1 achieves an annualized gross (net) Sharpe ratio of 1.04 (0.89), and monthly average abnormal gross (net) return relative to the Fama and French (2015) five-factor model plus a momentum factor of 86 (68) bps/month with a t-statistic of 3.86 (3.13), respectively. Its gross monthly alpha relative to these six factors plus the six most closely related strategies from the factor zoo (Momentum based on FF3 residuals, Net external financing, Share issuance (1 year), Equity Duration, Days with zero trades, Earnings surprise streak) is 77 bps/month with a t-statistic of 3.63.

1 Introduction

The following automatically generated report tests the asset pricing implications of Long-Short Term Memory Model signal (LSTM1), and its robustness in predicting returns in the cross-section of equities. It is produced using the methodology of Novy-Marx and Velikov (2023), from input data consisting of firm-month observations for the proposed predictor.[1]

2 Signal diagnostics

Figure 1 plots descriptive statistics for the LSTM1 signal. Panel A plots the time-series of the mean, median, and interquartile range for LSTM1. On average, the cross-sectional mean (median) LSTM1 is 0.01 (0.01) over the 2004 to 2021 sample, where the starting date is determined by the availability of the input LSTM1 data. The signal's interquartile range spans -0.01 to 0.15. Panel B of Figure 1 plots the time-series of the coverage of the LSTM1 signal for the CRSP universe. On average, the LSTM1 signal is available for 92.64% of CRSP names, which on average make up 98.04% of total market capitalization.

3 Does LSTM1 predict returns?

Table 1 reports the performance of portfolios constructed using a value-weighted, quintile sort on LSTM1 using NYSE breaks. The first two lines of Panel A report monthly average excess returns for each of the five portfolios and for the long/short portfolio that buys the high LSTM1 portfolio and sells the low LSTM1 portfolio. The rest of Panel A reports the portfolios' monthly abnormal returns relative to the five most common factor models: the CAPM, the Fama and French (1993) three-factor

[1] It used version v0.4.1 of the publicly available code repository at https://github.com/velikov-mihail/AssayingAnomalies. See more details at http://AssayingAnomalies.com.

model (FF3) and its variation that adds momentum (FF4), the Fama and French (2015) five-factor model (FF5), and its variation that adds momentum factor used in Fama and French (2018) (FF6). The table shows that the long/short LSTM1 strategy earns an average return of 0.99% per month with a t-statistic of 4.29. The annualized Sharpe ratio of the strategy is 1.04. The alphas range from 0.86% to 1.09% per month and have t-statistics exceeding 3.86 everywhere. The lowest alpha is with respect to the FF6 factor model.

Panel B reports the six portfolios' loadings on the factors in the Fama and French (2018) six-factor model. The long/short strategy's most significant loading is 0.58, with a t-statistic of 4.44 on the RMW factor. Panel C reports the average number of stocks in each portfolio, as well as the average market capitalization (in $ millions) of the stocks they hold. In an average month, the five portfolios have at least 603 stocks and an average market capitalization of at least $2,916 million.

Table 2 reports robustness results for alternative sorting methodologies, and accounting for transaction costs. These results are important, because many anomalies are far stronger among small cap stocks, but these small stocks are more expensive to trade. Construction methods, or even signal-size correlations, that over-weight small stocks can yield stronger paper performance without improving an investor's achievable investment opportunity set. Panel A reports gross returns and alphas for the long/short strategies made using various different protfolio constructions. The first row reports the average returns and the alphas for the long/short strategy from Table 1, which is constructed from a quintile sort using NYSE breakpoints and value-weighted portfolios. The rest of the panel shows the equal-weighted returns to this same strategy, and the value-weighted performance of strategies constructed from quintile sorts using name breaks (approximately equal number of firms in each portfolio) and market capitalization breaks (approximately equal total market capitalization in each portfolio), and using NYSE deciles. The average return is lowest for the quintile sort using cap breakpoints and value-weighted portfolios, and equals 74 bps/month with a t-statistics of 3.87. Out of the twenty-five alphas reported in Panel A, the t-statistics for twenty-five exceed two, and for twenty-five exceed three.

Panel B reports for these same strategies the average monthly net returns and the generalized net alphas of Novy-Marx and Velikov (2016). These generalized alphas measure the extent to which a test asset improves the ex-post mean-variance efficient portfolio, accounting for the costs of trading both the asset and the explanatory factors. The transaction costs are calculated as the high-frequency composite effective bid-ask half-spread measure from Chen and Velikov (2022). The net average returns reported in the first column range between 64-145bps/month. The lowest return, (64 bps/month), is achieved from the quintile sort using cap breakpoints and value-weighted portfolios, and has an associated t-statistic of 3.32. Out of the twenty-five construction-methodology-factor-model pairs reported in Panel B, the LSTM1 trading strategy improves the achievable mean-variance efficient frontier spanned by the factor models in twenty-five cases, and significantly expands the achievable frontier in twenty-five cases.

Table 3 provides direct tests for the role size plays in the LSTM1 strategy performance. Panel A reports the average returns for the twenty-five portfolios constructed from a conditional double sort on size and LSTM1, as well as average returns and alphas for long/short trading LSTM1 strategies within each size quintile. Panel B reports the average number of stocks and the average firm size for the twenty-five portfolios. Among the largest stocks (those with market capitalization greater than the 80th NYSE percentile), the LSTM1 strategy achieves an average return of 66 bps/month with a t-statistic of 3.13. Among these large cap stocks, the alphas for the LSTM1 strategy relative to the five most common factor models range from 59 to 69 bps/month with t-statistics between 2.84 and 3.34.

4 How does LSTM1 perform relative to the zoo?

Figure 2 puts the performance of LSTM1 in context, showing the long/short strategy performance relative to other strategies in the "factor zoo." It shows Sharpe ratio histograms, both for gross and net returns (Panel A and B, respectively), for 212 documented anomalies in the zoo.[2] The vertical red line shows where the Sharpe ratio for the LSTM1 strategy falls in the distribution. The LSTM1 strategy's gross (net) Sharpe ratio of 1.04 (0.89) is greater than 100% (100%) of anomaly Sharpe ratios, respectively.

Figure 3 plots the growth of a $1 invested in these same 212 anomaly trading strategies (gray lines), and compares those with the growth of a $1 invested in the LSTM1 strategy (red line).[3] Ignoring trading costs, a $1 invested in the LSTM1 strategy would have yielded $4.93 which ranks the LSTM1 strategy in the top 1% across the 212 anomalies. Accounting for trading costs, a $1 invested in the LSTM1 strategy would have yielded $3.45 which ranks the LSTM1 strategy in the top 1% across the 212 anomalies.

Figure 4 plots percentile ranks for the 212 anomaly trading strategies in terms of gross and Novy-Marx and Velikov (2016) net generalized alphas with respect to the CAPM, and the Fama-French three-, four-, five-, and six-factor models from Table 1, and indicates the ranking of the LSTM1 relative to those. Panel A shows that the LSTM1 strategy gross alphas fall between the 97 and 99 percentiles across the five factor models. Panel B shows that, accounting for trading costs, a large fraction of anomalies have not improved the investment opportunity set of an investor with access to the factor models over the 200412 to 202111 sample. For example, 47%

[2] The anomalies come from March, 2022 release of the Chen and Zimmermann (2022) open source asset pricing dataset.

[3] The figure assumes an initial investment of $1 in T-bills and $1 long/short in the two sides of the strategy. Returns are compounded each month, assuming, as in Detzel et al. (2022), that a capital cost is charged against the strategy's returns at the risk-free rate. This excess return corresponds more closely to the strategy's economic profitability.

(53%) of the 212 anomalies would not have improved the investment opportunity set for an investor having access to the Fama-French three-factor (six-factor) model. The LSTM1 strategy has a positive net generalized alpha for five out of the five factor models. In these cases LSTM1 ranks between the 99 and 99 percentiles in terms of how much it could have expanded the achievable investment frontier.

5 Does LSTM1 add relative to related anomalies?

With so many anomalies, it is possible that any proposed, new cross-sectional predictor is just capturing some combination of known predictors. It is consequently natural to investigate to what extent the proposed predictor adds additional predictive power beyond the most closely related anomalies. Closely related anomalies are more likely to be formed on the basis of signals with higher absolute correlations. Figure 5 plots a name histogram of the correlations of LSTM1 with 211 filtered anomaly signals.[4] Figure 6 also shows an agglomerative hierarchical cluster plot using Ward's minimum method and a maximum of 10 clusters.

A closely related anomaly is also more likely to price LSTM1 or at least to weaken the power LSTM1 has predicting the cross-section of returns. Figure 7 plots histograms of t-statistics for predictability tests of LSTM1 conditioning on each of the 211 filtered anomaly signals one at a time. Panel A reports t-statistics on β_{LSTM1} from Fama-MacBeth regressions of the form $r_{i,t} = \alpha + \beta_{LSTM1} LSTM1_{i,t} + \beta_X X_{i,t} + \epsilon_{i,t}$, where X stands for one of the 211 filtered anomaly signals at a time. Panel B plots t-statistics on α from spanning tests of the form: $r_{LSTM1,t} = \alpha + \beta r_{X,t} + \epsilon_t$, where $r_{X,t}$ stands for the returns to one of the 211 filtered anomaly trading strategies at a time. The strategies employed in the spanning tests are constructed using

[4] When performing tests at the underlying signal level (e.g., the correlations plotted in Figure 5), we filter the 212 anomalies to avoid small sample issues. For each anomaly, we calculate the common stock observations in an average month for which both the anomaly and the test signal are available. In the filtered anomaly set, we drop anomalies with fewer than 100 common stock observations in an average month.

quintile sorts, value-weighting, and NYSE breakpoints. Panel C plots t-statistics on the average returns to strategies constructed by conditional double sorts. In each month, we sort stocks into quintiles based one of the 211 filtered anomaly signals. Then, within each quintile, we sort stocks into quintiles based on LSTM1. Stocks are finally grouped into five LSTM1 portfolios by combining stocks within each anomaly sorting portfolio. The panel plots the t-statistics on the average returns of these conditional double-sorted LSTM1 trading strategies conditioned on each of the 211 filtered anomalies.

Table 4 reports Fama-MacBeth cross-sectional regressions of returns on LSTM1 and the six anomalies most closely-related to it. The six most-closely related anomalies are picked as those with the highest combined rank where the ranks are based on the absolute value of the Spearman correlations in Panel B of Figure 5 and the R^2 from the spanning tests in Figure 7, Panel B. Controlling for each of these signals at a time, the t-statistics on the LSTM1 signal in these Fama-MacBeth regressions exceed 8.45, with the minimum t-statistic occurring when controlling for Earnings surprise streak. Controlling for all six closely related anomalies, the t-statistic on LSTM1 is 6.96.

Similarly, Table 5 reports results from spanning tests that regress returns to the LSTM1 strategy onto the returns of the six most closely-related anomalies and the six Fama-French factors. Controlling for the six most-closely related anomalies individually, the LSTM1 strategy earns alphas that range from 70-84bps/month. The minimum t-statistic on these alphas controlling for one anomaly at a time is 3.15, which is achieved when controlling for Earnings surprise streak. Controlling for all six closely-related anomalies and the six Fama-French factors simultaneously, the LSTM1 trading strategy achieves an alpha of 77bps/month with a t-statistic of 3.63.

6 Does LSTM1 add relative to the whole zoo?

Finally, we can ask how much adding LSTM1 to the entire factor zoo could improve investment performance. Figure 8 plots the growth of $1 invested in trading strategies that combine multiple anomalies following Chen and Velikov (2022). The combinations use either the 162 anomalies from the zoo that satisfy our inclusion criteria (blue lines) or these 162 anomalies augmented with the LSTM1 signal.[5] We consider six different methods for combining signals.

Panel A shows results using "Average rank" as the combination method. This method sorts stocks on the basis of forecast excess returns, where these are calculated on the basis of their average cross-sectional percentile rank across return predictors, and the predictors are all signed so that higher ranks are associated with higher average returns. For this method, $1 investment in the 162-anomaly combination strategy grows to $1.78, while $1 investment in the combination strategy that includes LSTM1 grows to $1.77.

Panel B shows results using "Weighted-Average rank" as the combination method. This method sorts stocks on the basis of forecast excess returns, where these are calculated as weighted-average cross-sectional percentile rank across return predictors, and the predictors are all signed so that higher ranks are associated with higher average returns and the weights are determined by the average returns over the past ten years to the long/short strategies based on the individual signals. For this method, $1 investment in the 162-anomaly combination strategy grows to $0.94, while $1 investment in the combination strategy that includes LSTM1 grows to $1.08.

Panel C shows results using "Fama-MacBeth" as the combination method. This method sorts stocks on the basis of forecast excess returns, where these are calculated from Fama and MacBeth (1973) regressions following Haugen and Baker (1996) and

[5] We filter the 207 Chen and Zimmermann (2022) anomalies and require for each anomaly the average month to have at least 40% of the cross-sectional observations available for market capitalization on CRSP in the period for which LSTM1 is available.

Lewellen (2015) using only data in the investor's information set at the time of portfolio formation. The estimation uses rolling ten years of data, so the actual strategies begin ten years later for this combination method. For this method, $1 investment in the 162-anomaly combination strategy grows to $1.39, while $1 investment in the combination strategy that includes LSTM1 grows to $1.32.

Panel D shows results using "Partial Least Squares" as the combination method. This method sorts stocks on the basis of forecast excess returns, where these are calculated from partial least squares (PLS) filtering procedure following Light et al. (2017) using only data in the investor's information set at the time of portfolio formation. The estimation uses rolling ten years of data, so the actual strategies begin ten years later for this combination method. For this method, $1 investment in the 162-anomaly combination strategy grows to $2.68, while $1 investment in the combination strategy that includes LSTM1 grows to $2.26.

Panel E shows results using "IPCA" as the combination method. This method sorts stocks on the basis of forecast excess returns, where these are calculated from the instrumented principal component analysis (IPCA) procedure of Kelly et al. (2019) using only data in the investor's information set at the time of portfolio formation. The estimation uses rolling ten years of data, so the actual strategies begin ten years later for this combination method. For this method, $1 investment in the 162-anomaly combination strategy grows to $1.02, while $1 investment in the combination strategy that includes LSTM1 grows to $0.80.

Panel F shows results using "LASSO" as the combination method. This method sorts stocks on the basis of forecast excess returns, where these are estimated by least absolute shrinkage and selection operator (LASSO) using only data in the investor's information set at the time of portfolio formation. Following Chen and Velikov (2022), LASSO penalty (λ) is selected by minimizing the mean squared error (MSE) estimated by 5-fold cross validation. The estimation uses rolling ten years of data, so the actual strategies begin ten years later for this combination method. For this method, \$1 investment in the 162-anomaly combination strategy grows to \$1.26, while \$1 investment in the combination strategy that includes LSTM1 grows to \$0.96.

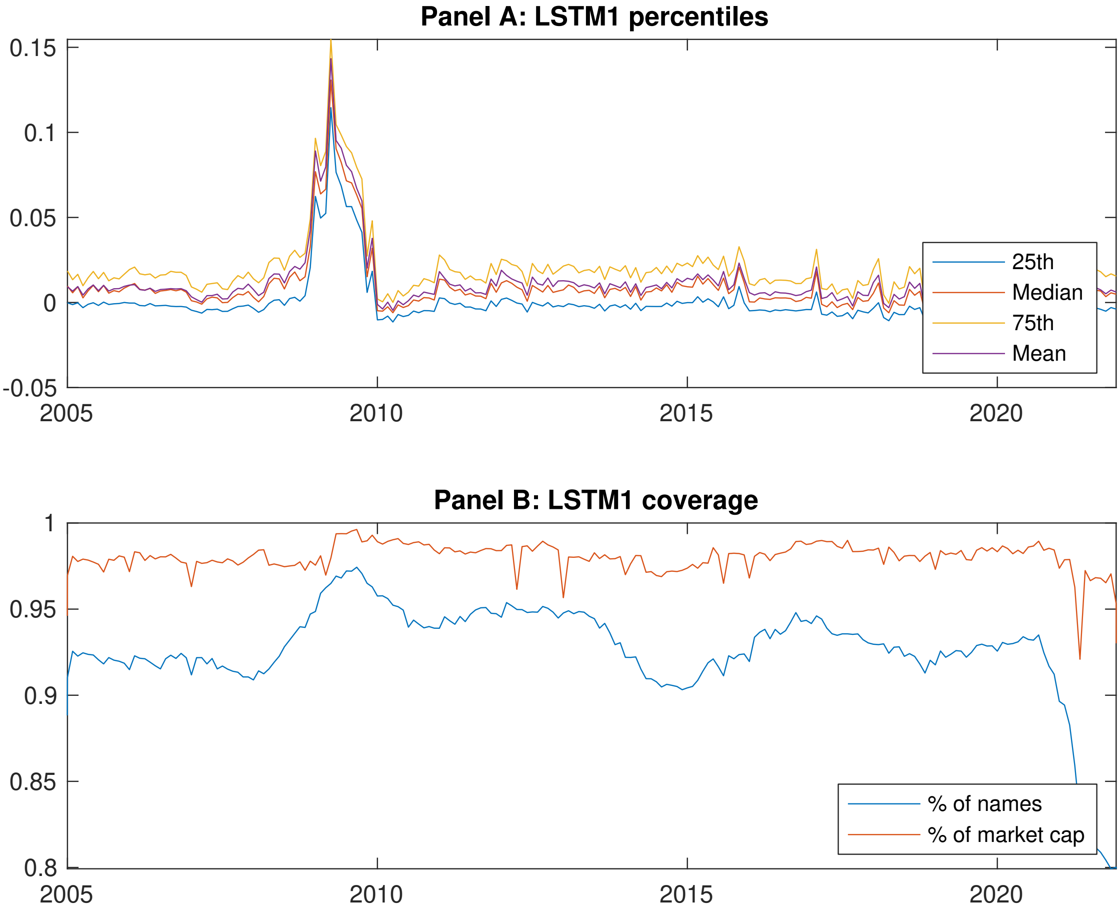


Figure 1: Times series of LSTM1 percentiles and coverage.
This figure plots descriptive statistics for LSTM1. Panel A shows cross-sectional percentiles of LSTM1 over the sample. Panel B plots the monthly coverage of LSTM1 relative to the universe of CRSP stocks with available market capitalizations.

Table 1: Basic sort: VW, quintile, NYSE-breaks

This table reports average excess returns and alphas for portfolios sorted on LSTM1. At the end of each month, we sort stocks into five portfolios based on their signal using NYSE breakpoints. Panel A reports average value-weighted quintile portfolio (L,2,3,4,H) returns in excess of the risk-free rate, the long-short extreme quintile portfolio (H-L) return, and alphas with respect to the CAPM, Fama and French (1993) three-factor model, Fama and French (1993) three-factor model augmented with the Carhart (1997) momentum factor, Fama and French (2015) five-factor model, and the Fama and French (2015) five-factor model augmented with the Carhart (1997) momentum factor following Fama and French (2018). Panel B reports the factor loadings for the quintile portfolios and long-short extreme quintile portfolio in the Fama and French (2015) five-factor model. Panel C reports the average number of stocks and market capitalization of each portfolio. T-statistics are in brackets. The sample period is 200412 to 202111.

Panel A: Excess returns and alphas on LSTM1-sorted portfolios						
	(L)	(2)	(3)	(4)	(H)	(H-L)
r^e	0.23	0.85	0.85	0.98	1.22	0.99
	[0.58]	[2.87]	[2.87]	[3.09]	[3.30]	[4.29]
α_{CAPM}	-0.76	0.07	0.07	0.16	0.28	1.04
	[-4.92]	[1.03]	[1.03]	[1.78]	[2.18]	[4.43]
α_{FF3}	-0.74	0.06	0.05	0.19	0.35	1.09
	[-4.83]	[0.93]	[0.76]	[2.18]	[3.06]	[4.78]
α_{FF4}	-0.69	0.08	0.05	0.18	0.36	1.06
	[-4.79]	[1.13]	[0.77]	[2.08]	[3.16]	[4.67]
α_{FF5}	-0.59	0.02	-0.00	0.20	0.30	0.89
	[-3.90]	[0.28]	[-0.07]	[2.29]	[2.60]	[3.95]
α_{FF6}	-0.55	0.03	-0.00	0.19	0.31	0.86
	[-3.87]	[0.45]	[-0.05]	[2.21]	[2.68]	[3.86]
Panel B: Fama and French (2018) 6-factor model loadings for LSTM1-sorted portfolios						
β_{MKT}	1.10	0.94	0.98	0.99	1.06	-0.04
	[30.17]	[54.52]	[57.04]	[43.82]	[35.37]	[-0.68]
β_{SMB}	0.03	-0.01	-0.05	0.01	0.15	0.11
	[0.54]	[-0.48]	[-1.74]	[0.15]	[2.87]	[1.15]
β_{HML}	-0.13	-0.04	-0.04	0.15	0.24	0.36
	[-2.11]	[-1.46]	[-1.49]	[4.04]	[4.77]	[3.84]
β_{RMW}	-0.40	0.11	0.13	-0.04	0.17	0.58
	[-4.89]	[2.80]	[3.40]	[-0.74]	[2.54]	[4.44]
β_{CMA}	0.02	0.03	0.04	-0.01	-0.07	-0.10
	[0.24]	[0.52]	[0.85]	[-0.23]	[-0.88]	[-0.61]
β_{UMD}	-0.19	-0.05	-0.00	0.04	-0.04	0.15
	[-5.29]	[-3.00]	[-0.29]	[1.78]	[-1.47]	[2.59]
Panel C: Average number of firms (n) and market capitalization (me)						
n	900	612	603	654	922	
me ($\$10^6$)	2916	4795	5184	4554	3052	

Table 2: Robustness to sorting methodology & trading costs

This table evaluates the robustness of the choices made in the LSTM1 strategy construction methodology. In each panel, the first row shows results from a quintile, value-weighted sort using NYSE break points as employed in Table 1. Each of the subsequent rows deviates in one of the three choices at a time, and the choices are specified in the first three columns. For each strategy construction methodology, the table reports average excess returns and alphas with respect to the CAPM, Fama and French (1993) three-factor model, Fama and French (1993) three-factor model augmented with the Carhart (1997) momentum factor, Fama and French (2015) five-factor model, and the Fama and French (2015) five-factor model augmented with the Carhart (1997) momentum factor following Fama and French (2018). Panel A reports average returns and alphas with no adjustment for trading costs. Panel B reports net average returns and Novy-Marx and Velikov (2016) generalized alphas as prescribed by Detzel et al. (2022). T-statistics are in brackets. The sample period is 200412 to 202111.

Panel A: Gross Returns and Alphas

Portfolios	Breaks	Weights	r^e	α_{CAPM}	α_{FF3}	α_{FF4}	α_{FF5}	α_{FF6}
Quintile	NYSE	VW	0.99	1.04	1.09	1.06	0.89	0.86
			[4.29]	[4.43]	[4.78]	[4.67]	[3.95]	[3.86]
Quintile	NYSE	EW	2.07	2.08	2.08	2.11	1.89	1.91
			[10.03]	[9.86]	[9.98]	[10.18]	[9.11]	[9.31]
Quintile	Name	VW	1.46	1.49	1.56	1.60	1.31	1.34
			[4.44]	[4.44]	[4.77]	[4.91]	[3.96]	[4.09]
Quintile	Cap	VW	0.74	0.73	0.78	0.75	0.68	0.65
			[3.87]	[3.71]	[4.18]	[4.07]	[3.54]	[3.45]
Decile	NYSE	VW	1.66	1.81	1.87	1.87	1.44	1.45
			[4.08]	[4.41]	[4.61]	[4.60]	[3.60]	[3.60]

Panel B: Net Returns and Novy-Marx and Velikov (2016) generalized alphas

Portfolios	Breaks	Weights	r^e_{net}	α^*_{CAPM}	α^*_{FF3}	α^*_{FF4}	α^*_{FF5}	α^*_{FF6}
Quintile	NYSE	VW	0.85	0.83	0.90	0.88	0.70	0.68
			[3.68]	[3.58]	[3.94]	[3.91]	[3.15]	[3.13]
Quintile	NYSE	EW	1.11	1.00	1.01	1.02	0.85	0.86
			[5.57]	[5.19]	[5.24]	[5.34]	[4.52]	[4.60]
Quintile	Name	VW	1.27	1.16	1.25	1.27	1.02	1.03
			[3.92]	[3.61]	[3.92]	[4.01]	[3.22]	[3.30]
Quintile	Cap	VW	0.64	0.53	0.60	0.59	0.51	0.50
			[3.32]	[2.81]	[3.26]	[3.23]	[2.76]	[2.73]
Decile	NYSE	VW	1.45	1.48	1.56	1.56	1.18	1.18
			[3.59]	[3.65]	[3.88]	[3.88]	[3.00]	[3.00]

Table 3: Conditional sort on size and LSTM1

This table presents results for conditional double sorts on size and LSTM1. In each month, stocks are first sorted into quintiles based on size using NYSE breakpoints. Then, within each size quintile, stocks are further sorted based on LSTM1. Finally, they are grouped into twenty-five portfolios based on the intersection of the two sorts. Panel A presents the average returns to the 25 portfolios, as well as strategies that go long stocks with high LSTM1 and short stocks with low LSTM1 .Panel B documents the average number of firms and the average firm size for each portfolio. The sample period is 200412 to 202111.

Panel A: portfolio average returns and time-series regression results

		LSTM1 Quintiles					LSTM1 Strategies					
		(L)	(2)	(3)	(4)	(H)	r^e	α_{CAPM}	α_{FF3}	α_{FF4}	α_{FF5}	α_{FF6}
Size quintiles	(1)	-0.34	0.27	0.90	1.47	2.31	2.65	2.58	2.60	2.70	2.30	2.39
		[-0.61]	[0.62]	[2.13]	[3.31]	[3.89]	[7.75]	[7.42]	[7.60]	[8.35]	[6.74]	[7.45]
	(2)	0.22	0.82	1.02	1.09	1.39	1.17	1.28	1.34	1.34	1.04	1.05
		[0.41]	[1.90]	[2.50]	[2.58]	[2.90]	[4.25]	[4.64]	[5.17]	[5.16]	[4.22]	[4.22]
	(3)	0.44	0.83	1.04	0.98	1.13	0.70	0.76	0.76	0.68	0.49	0.41
		[0.92]	[2.12]	[2.67]	[2.46]	[2.63]	[2.72]	[2.93]	[2.98]	[2.84]	[1.94]	[1.78]
	(4)	0.47	0.79	0.85	1.10	1.14	0.67	0.74	0.77	0.72	0.58	0.53
		[1.09]	[2.32]	[2.46]	[3.09]	[2.87]	[3.11]	[3.38]	[3.65]	[3.53]	[2.76]	[2.65]
	(5)	0.43	0.92	0.77	0.96	1.09	0.66	0.62	0.69	0.65	0.63	0.59
		[1.32]	[3.22]	[2.63]	[3.21]	[3.28]	[3.13]	[2.89]	[3.34]	[3.21]	[2.95]	[2.84]

Panel B: Portfolio average number of firms and market capitalization

		LSTM1 Quintiles					LSTM1 Quintiles				
		Average n					Average market capitalization ($\$10^6$)				
		(L)	(2)	(3)	(4)	(H)	(L)	(2)	(3)	(4)	(H)
Size quintiles	(1)	384	384	384	384	384	68	76	74	69	45
	(2)	123	123	123	123	123	118	121	122	121	119
	(3)	87	87	87	87	87	202	208	208	207	205
	(4)	75	75	75	75	75	435	448	445	442	436
	(5)	69	69	69	69	69	2917	3157	3514	3416	3328

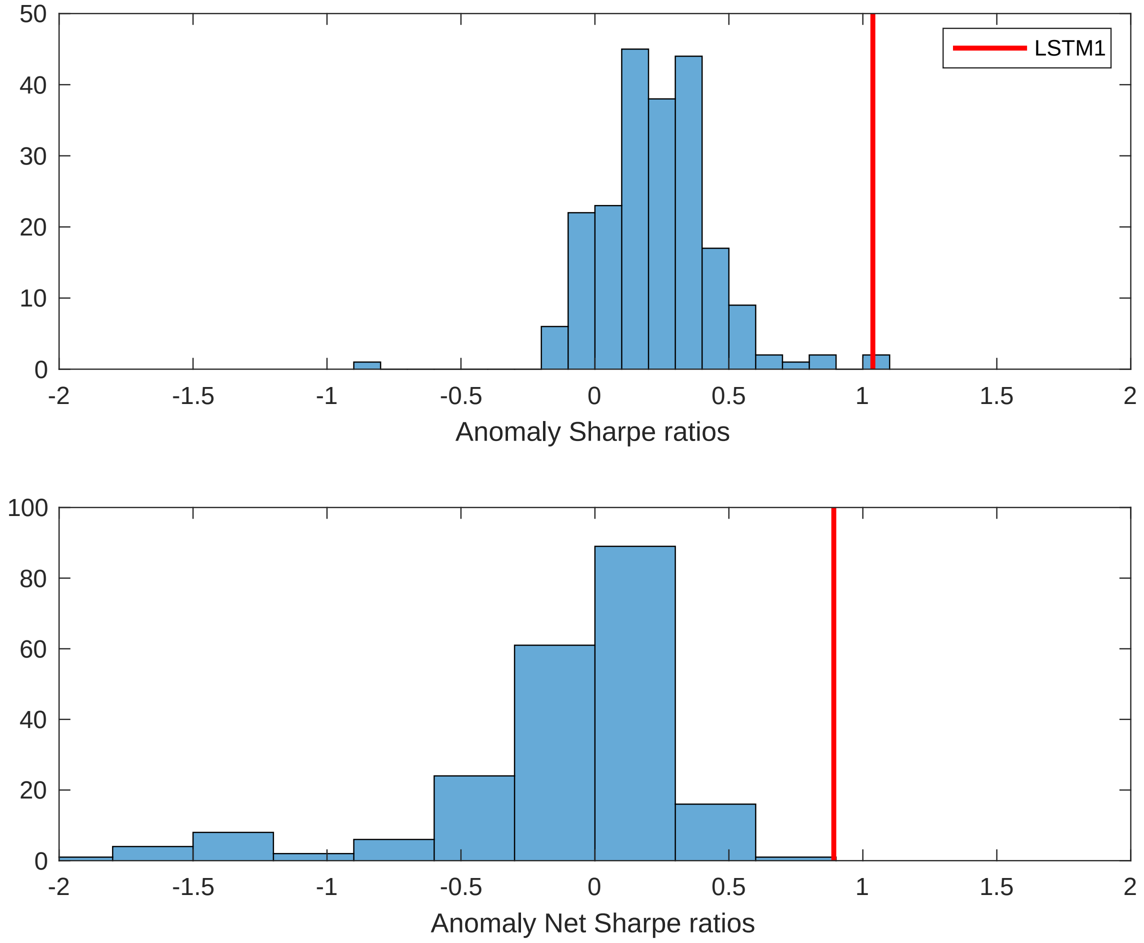


Figure 2: Distribution of Sharpe ratios.
This figure plots a histogram of Sharpe ratios for 212 anomalies, and compares the Sharpe ratio of the LSTM1 with them (red vertical line). Panel A plots results for gross Sharpe ratios. Panel B plots results for net Sharpe ratios.

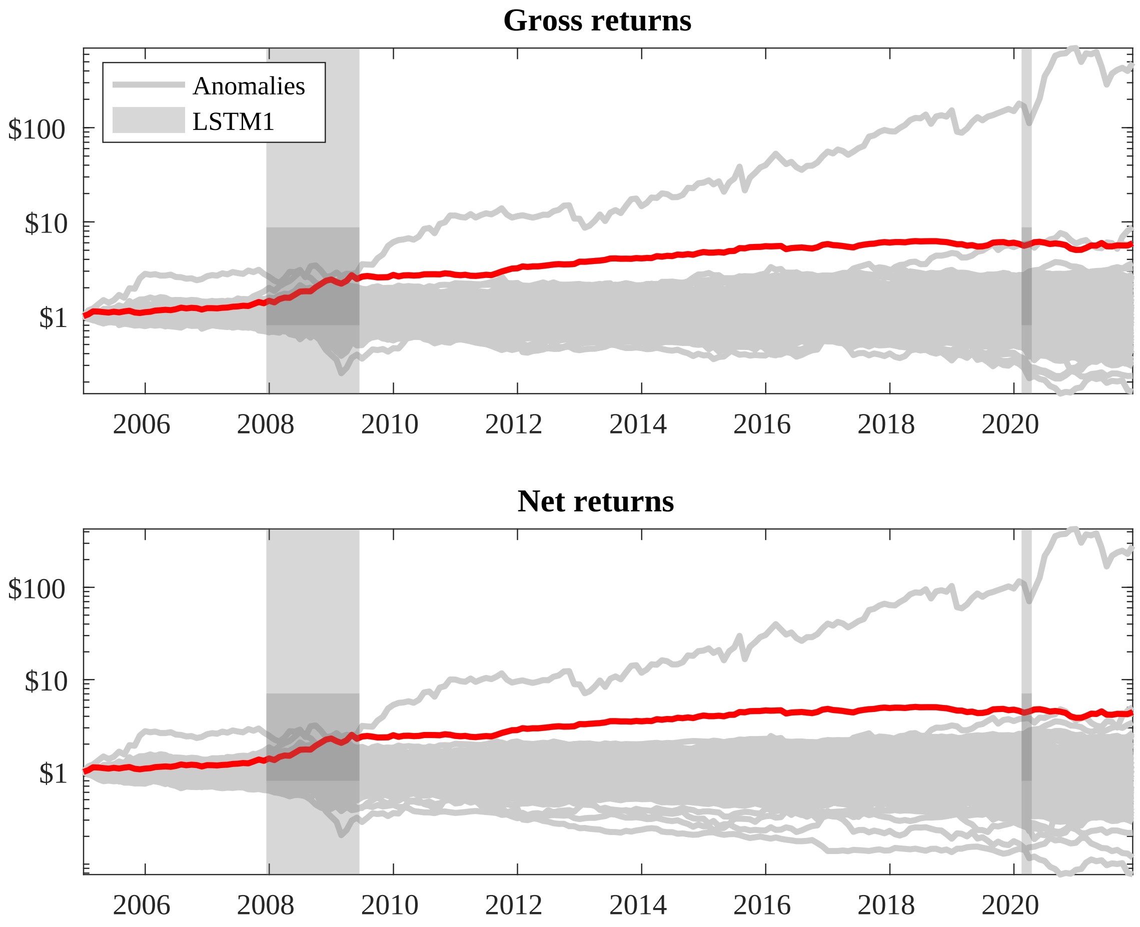
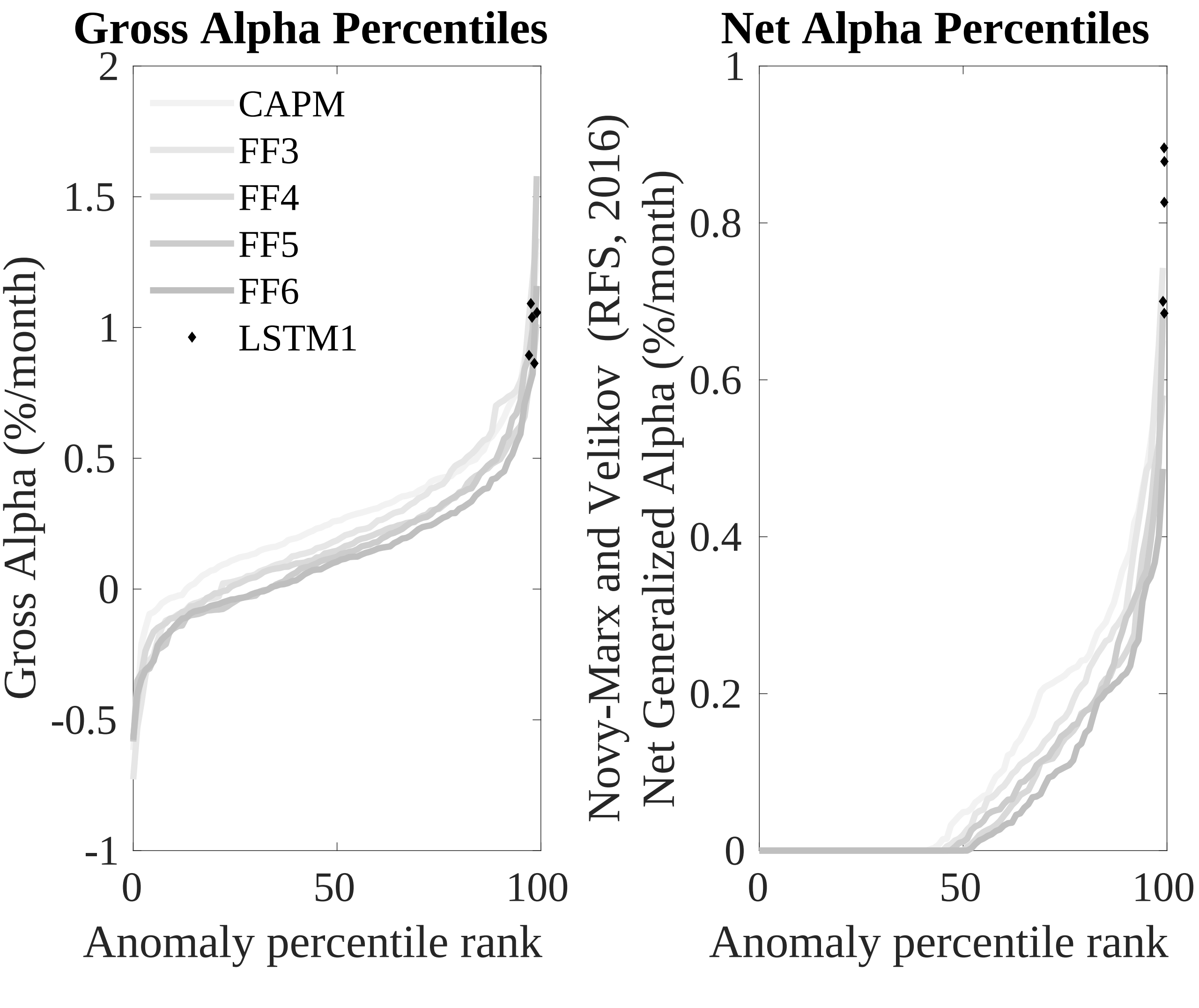


Figure 3: Dollar invested.
This figure plots the growth of a $1 invested in 212 anomaly trading strategies (gray lines), and compares those with the LSTM1 trading strategy (red line). The strategies are constructed using value-weighted quintile sorts using NYSE breakpoints. Panel A plots results for gross strategy returns. Panel B plots results for net strtaegy returns.


Gross Alpha Percentiles
Gross Alpha (%/month)
2
1.5
1
0.5
0
-0.5
-1
0
50
100
Anomaly percentile rank
CAPM
FF3
FF4
FF5
FF6
LSTM1
Net Alpha Percentiles
Novy-Marx and Velikov (RFS, 2016)
Net Generalized Alpha (%/month)
1
0.8
0.6
0.4
0.2
0
0
50
100
Anomaly percentile rank

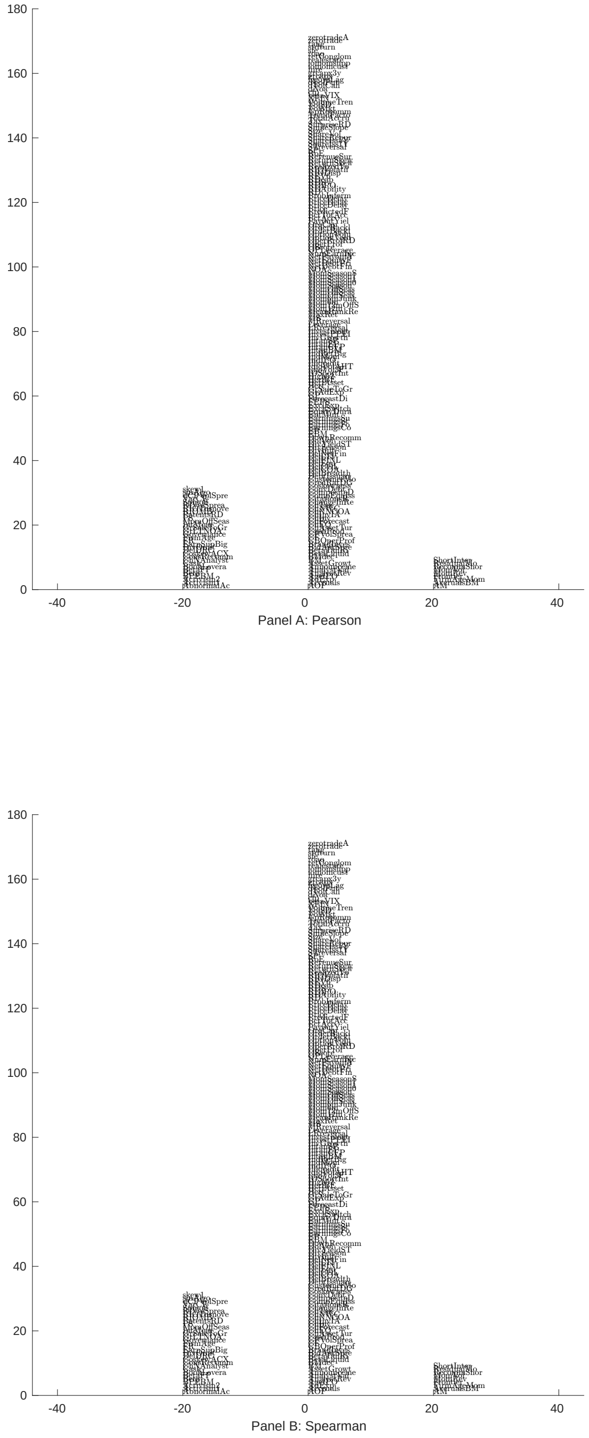


Figure 5: Distribution of correlations.
This figure plots a name histogram of correlations of 211 filtered anomaly signals with LSTM1. The correlations are pooled. Panel A plots Pearson correlations, while Panel B plots Spearman rank correlations.

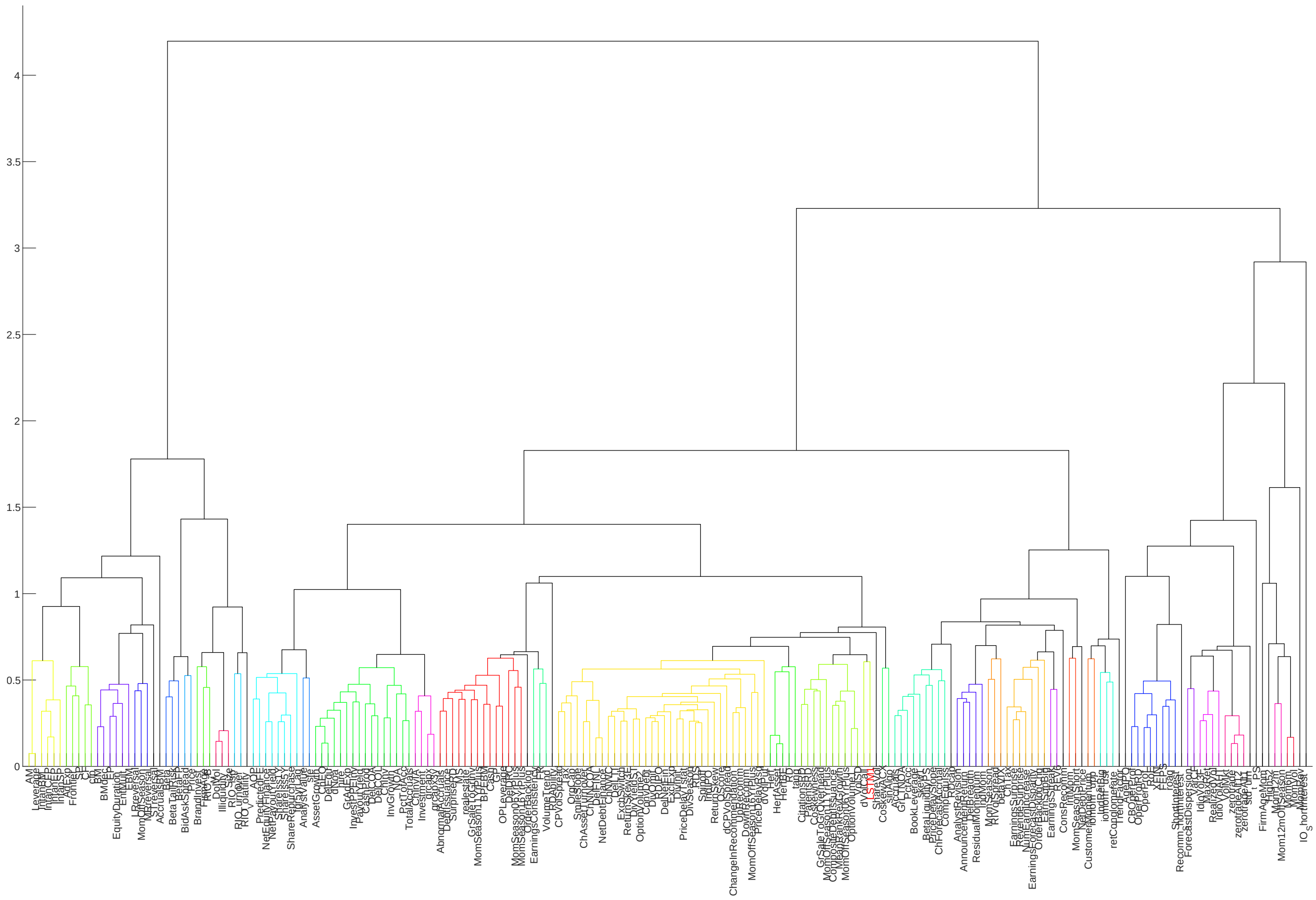


Figure 6: Agglomerative hierarchical cluster plot
This figure plots an agglomerative hierarchical cluster plot using Ward's minimum method and a maximum of 10 clusters.

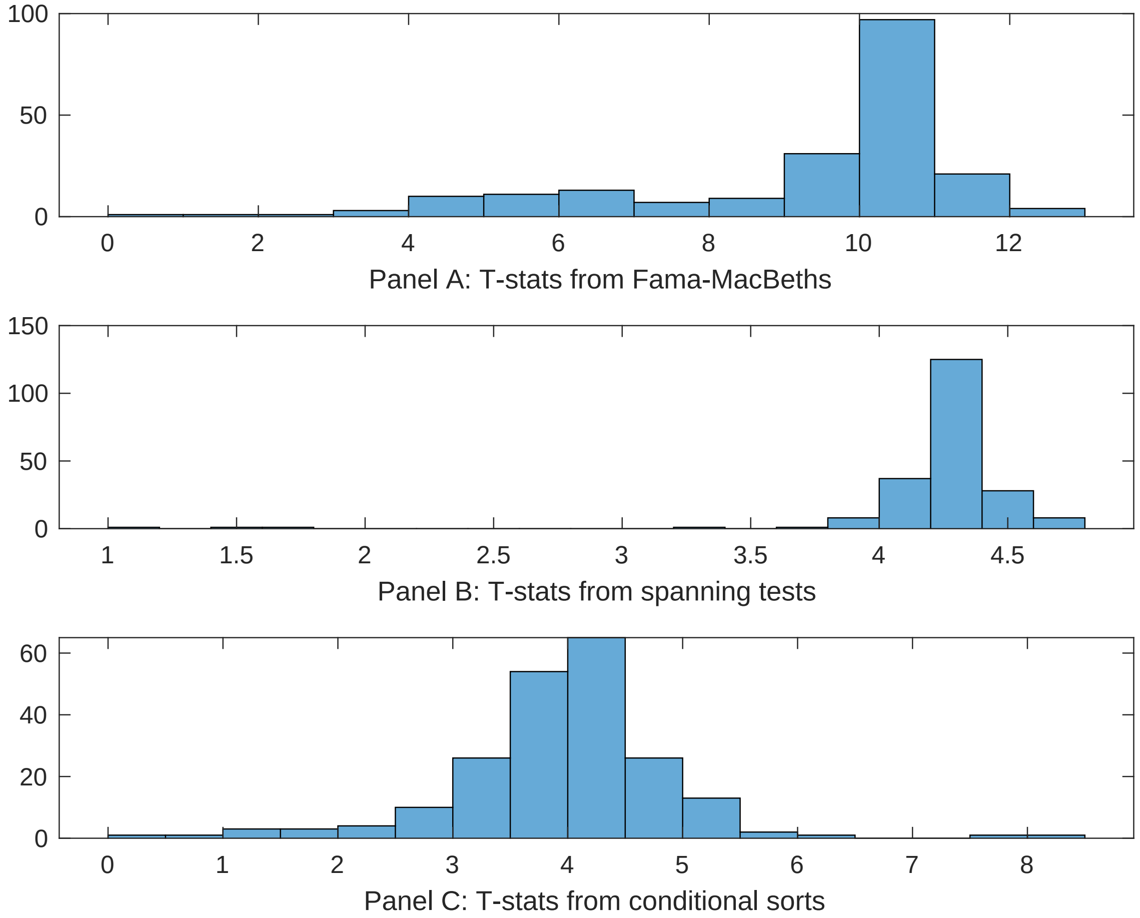


Figure 7: Distribution of t-stats on conditioning strategies
This figure plots histograms of t-statistics for predictability tests of LSTM1 conditioning on each of the 211 filtered anomaly signals one at a time. Panel A reports t-statistics on β_{LSTM1} from Fama-MacBeth regressions of the form $r_{i,t} = \alpha + \beta_{LSTM1} LSTM1_{i,t} + \beta_X X_{i,t} + \epsilon_{i,t}$, where X stands for one of the 211 filtered anomaly signals at a time. Panel B plots t-statistics on α from spanning tests of the form: $r_{LSTM1,t} = \alpha + \beta r_{X,t} + \epsilon_t$, where $r_{X,t}$ stands for the returns to one of the 211 filtered anomaly trading strategies at a time. The strategies employed in the spanning tests are constructed using quintile sorts, value-weighting, and NYSE breakpoints. Panel C plots t-statistics on the average returns to strategies constructed by conditional double sorts. In each month, we sort stocks into quintiles based one of the 211 filtered anomaly signals at a time. Then, within each quintile, we sort stocks into quintiles based on LSTM1. Stocks are finally grouped into five LSTM1 portfolios by combining stocks within each anomaly sorting portfolio. The panel plots the t-statistics on the average returns of these conditional double-sorted LSTM1 trading strategies conditioned on each of the 211 filtered anomalies.

Table 4: Fama-MacBeths controlling for most closely related anomalies
This table presents Fama-MacBeth results of returns on LSTM1. and the six most closely related anomalies. The regressions take the following form: $r_{i,t} = \alpha + \beta_{LSTM1} LSTM1_{i,t} + \sum_{k=1}^{s} ix\beta_{X_k} X_{i,t}^{k} + \epsilon_{i,t}$. The six most closely related anomalies, X, are Momentum based on FF3 residuals, Net external financing, Share issuance (1 year), Equity Duration, Days with zero trades, Earnings surprise streak. These anomalies were picked as those with the highest combined rank where the ranks are based on the absolute value of the Spearman correlations in Panel B of Figure 5 and the R^2 from the spanning tests in Figure 7, Panel B. The sample period is 200412 to 202111.

Intercept	0.34	0.35	0.31	0.72	0.27	0.37	0.72
	[0.85]	[0.86]	[0.76]	[1.88]	[0.65]	[0.87]	[1.77]
LSTM1	0.45	0.43	0.46	0.44	0.47	0.41	0.33
	[10.02]	[9.99]	[10.69]	[10.27]	[10.62]	[8.45]	[6.96]
Anomaly 1	-0.13						-0.29
	[-0.57]						[-1.21]
Anomaly 2		0.67					0.56
		[1.62]					[1.36]
Anomaly 3			0.19				0.21
			[1.70]				[1.90]
Anomaly 4				0.23			0.10
				[1.82]			[0.84]
Anomaly 5					0.21		0.47
					[1.04]		[2.10]
Anomaly 6						0.14	0.10
						[4.33]	[3.32]
# months	203	203	203	203	203	203	203
$\bar{R}^2(\%)$	1	1	1	1	1	1	0

Table 5: Spanning tests controlling for most closely related anomalies

This table presents spanning tests results of regressing returns to the LSTM1 trading strategy on trading strategies exploiting the six most closely related anomalies. The regressions take the following form: $r_t^{LSTM1} = \alpha + \sum_{k=1}^{6} \beta_{X_k} r_t^{X_k} + \sum_{j=1}^{6} \beta_{f_j} r_t^{f_j} + \epsilon_t$, where X_k indicates each of the six most-closely related anomalies and f_j indicates the six factors from the Fama and French (2015) five-factor model augmented with the Carhart (1997) momentum factor. The six most closely related anomalies, X, are Momentum based on FF3 residuals, Net external financing, Share issuance (1 year), Equity Duration, Days with zero trades, Earnings surprise streak. These anomalies were picked as those with the highest combined rank where the ranks are based on the absolute value of the Spearman correlations in Panel B of Figure 5 and the R^2 from the spanning tests in Figure 7, Panel B. The sample period is 200412 to 202111.

Intercept	0.82	0.78	0.81	0.84	0.82	0.70	0.77
	[3.78]	[3.60]	[3.70]	[3.83]	[3.72]	[3.15]	[3.63]
Anomaly 1	33.46						34.21
	[3.67]						[3.88]
Anomaly 2		44.39					33.42
		[3.33]					[2.20]
Anomaly 3			34.09				14.03
			[2.91]				[1.02]
Anomaly 4				31.79			17.57
				[2.61]			[1.42]
Anomaly 5					18.36		4.61
					[2.45]		[0.52]
Anomaly 6						21.39	10.32
						[2.57]	[1.24]
mkt	-2.03	2.43	1.52	-1.50	5.86	-0.41	7.87
	[-0.37]	[0.42]	[0.26]	[-0.26]	[0.88]	[-0.07]	[1.24]
smb	10.54	29.58	19.07	13.91	26.70	12.28	22.91
	[1.09]	[2.80]	[1.94]	[1.42]	[2.47]	[1.25]	[2.15]
hml	26.10	32.79	25.45	-6.16	26.31	30.94	2.75
	[2.78]	[3.51]	[2.65]	[-0.36]	[2.72]	[3.28]	[0.17]
rmw	52.81	39.45	44.93	57.70	54.49	55.73	24.09
	[4.09]	[2.75]	[3.21]	[4.44]	[4.12]	[4.25]	[1.68]
cma	-16.54	-35.83	-19.59	2.75	-22.21	-16.40	-26.19
	[-1.04]	[-2.08]	[-1.21]	[0.16]	[-1.35]	[-1.01]	[-1.47]
umd	-1.18	9.69	14.70	17.05	10.86	7.27	-7.24
	[-0.17]	[1.74]	[2.69]	[3.05]	[1.92]	[1.19]	[-0.97]
# months	203	203	203	203	203	203	203
$\bar{R}^2$(%)	19	18	17	16	16	16	26

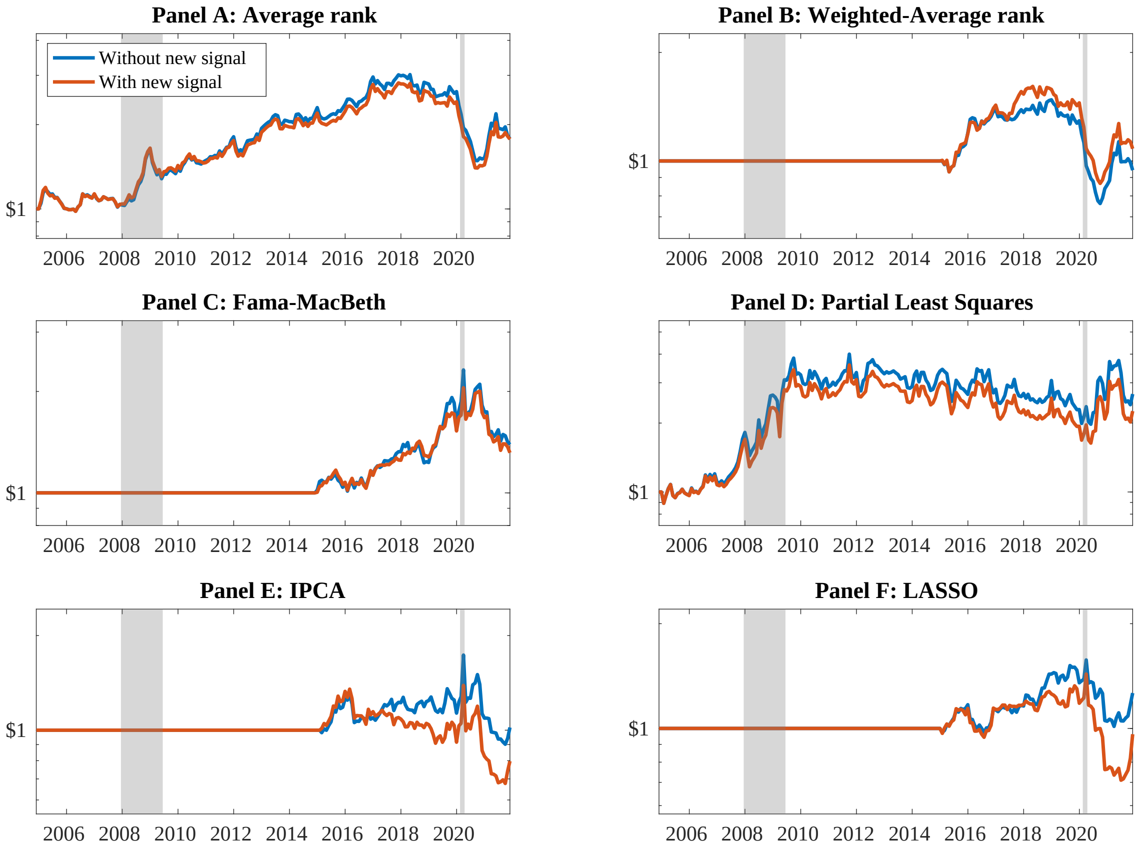


Figure 8: Combination strategy performance
This figure plots the growth of a $1 invested in trading strategies that combine multiple anomalies following Chen and Velikov (2022). In all panels, the blue solid lines indicate combination trading strategies that utilize 162 anomalies. The red solid lines indicate combination trading strategies that utilize the 162 anomalies as well as LSTM1. Panel A shows results using "Average rank" as the combination method. Panel B shows results using "Weighted-Average rank" as the combination method. Panel C shows results using "Fama-MacBeth" as the combination method. Panel D shows results using "Partial Least Squares" as the combination method. Panel E shows results using "IPCA" as the combination method. Panel F shows results using "LASSO" as the combination method. See Section 6 for details on the combination methods.

References

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance*, 52:57–82.

Chen, A. and Velikov, M. (2022). Zeroing in on the expected returns of anomalies. *Journal of Financial and Quantitative Analysis*, Forthcoming.

Chen, A. Y. and Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review*, 27(2):207–264.

Detzel, A., Novy-Marx, R., and Velikov, M. (2022). Model comparison with transaction costs. *Journal of Finance, Forthcoming.*

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Fama, E. F. and French, K. R. (2018). Choosing factors. *Journal of Financial Economics*, 128(2):234–252.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: empirical tests. *Journal of Political Economy*, 81(3):607–636.

Haugen, R. A. and Baker, N. L. (1996). Commonality in the determinants of expected stock returns. *Journal of Financial Economics*, 41(3):401–439.

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.

Lewellen, J. (2015). The cross-section of expected returns. *Critical Finance Review*, 4(1):1–44.

Light, N., Maslov, D., and Rytchkov, O. (2017). Aggregation of information about the cross section of stock returns: A latent variable approach. *Review of Financial Studies*, 30:1339–1381.

Novy-Marx, R. and Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *Review of Financial Studies*, 29(1):104–147.

Novy-Marx, R. and Velikov, M. (2023). Assaying anomalies. *Working paper*.