What Drives the Performance of Machine Learning Factor Strategies?

Mikheil Esakia and Felix Goltz
*Scientific Beta**

November 4, 2025

Abstract

Machine learning factor models of the cross section of stock returns have produced spectacular results, which are explained by two different ingredients: expanding the information set and allowing for more flexible function forms. We disentangle the value-added of each ingredient, considering a variety of empirical settings: from highly stylised to realistic. We show that the benefit of both ingredients declines when moving from standard settings in the literature to more realistic settings that exclude microcaps, remove look-ahead bias on yet-to-be-published factors, account for transaction costs, and exclude short positions. While the value of nonlinearity disappears even before imposing transaction costs or short-sale constraints, the value of an expanded information set is more persistent. Our feature importance analysis reveals that characteristics like firm size and short-term reversal - crucial predictors in standard settings - lose most of their value once investability constraints are considered. These findings challenge claims about the universal benefits of machine learning sophistication, demonstrating that real-world implementation constraints fundamentally alter which model ingredients improve portfolio performance.

Keywords: machine learning, equity factor models, cross-section of stock returns
JEL Classification: G11, G12, C45

Acknowledgments: We thank Daniel Aguet, Giovanni Bruno, Fabrizio Ghezzi, Martial Laguerre, Ben Luyten, Jean-Michel Maeso, Vincent Milhau, Ioan Mirciov, and Antoine Naly for helpful comments and suggestions. We are grateful to Sébastien Caron and Abissi Jesus Dibi for excellent research assistance. All remaining errors are our own.

*Scientific Beta R&D, 470 Promenade des Anglais, 06200 Nice, France, Tel: +33 493 187 851, E-mail: research@scientificbeta.com

1. Introduction

A growing body of research develops machine learning (ML) approaches to predict cross-sectional return differences in the stock market. Recent work focuses on improving prediction techniques for superior out-of-sample performance, which has led to more sophisticated ML models (see, e.g., Chen, Pelger, and Zhu (2024); Didisheim et al. (2024)). We take a different perspective and ask what drives ML's predictive improvements. Is the value-added due to (i) the ability to process richer information sets (informational complexity) or (ii) the capacity to capture nonlinear and interaction effects (functional complexity)? We investigate this question across implementation settings that range from highly stylised to realistic, progressively adding real-time information constraints, microcap exclusions, and explicit trading-cost integration. We show that informational complexity consistently matters, while functional complexity matters only in unconstrained, frictionless settings, and only in conjunction with richer information.

Machine learning methods are typically motivated by their ability to exploit broader information sets and complex, nonlinear relations (see, e.g., Avramov, Cheng, and Metzker (2023); Chen, Pelger, and Zhu (2024)). However, prior evidence on these two forces typically isolates only one at a time. Gu, Kelly, and Xiu (2020) conclude that nonlinearity unambiguously improves return prediction, but do not explicitly contrast linear and nonlinear specifications when the predictor set is sparse. Similarly, Kozak, Nagel, and Santosh (2020) argue that an expanded information set should improve portfolio performance, but largely maintain linearity. Most of the ML literature either adopts a fixed information set and varies the functional form, or expands the information set while retaining linearity, leaving their relative and joint contributions insufficiently disentangled. Chen, Hanauer, and Kalsbach (2024) examine how strategy returns vary with specification choices along multiple design dimensions, but their emphasis is on documenting systematic differences across alternative ML designs. By contrast, we focus on isolating the individual and joint contributions of informational and functional complexity across different implementation settings.

A separate strand of research considers implementation realism but often addresses only subsets of the relevant constraints. Avramov, Cheng, and Metzker (2023) examine the effects of excluding hard-to-trade stocks and periods of stressed market conditions, yet transaction costs are not incorporated. Li et al. (2023) account for trading costs when evaluating ML-powered strategies, but the strategies themselves are not designed with cost considerations, simply subtracting costs ex post. Azevedo, Hoegner, and Velikov (2024) and Hanauer and Kalsbach (2023) employ turnover control mechanisms and exclude the smallest and/or most expensive stocks to trade. These contributions are valuable, but they typically consider a limited set of frictions in isolation. Furthermore, it remains unclear whether predictive gains in practice primarily reflect a broader information set, flexible functional forms, or their interaction, once real-world constraints are systematically integrated.

We address these gaps by evaluating the individual and joint contributions of informational and functional complexity in a unified design. While both forms of complexity increase the number of model parameters relative to sample observations in line with standard definitions

(e.g., Kelly, Malamud, and Zhou (2024)), we explicitly distinguish between complexity arising from an expanded information set versus complexity from more flexible functional forms that relate characteristics to returns. Specifically, we implement a 2×2 comparison of linear vs. nonlinear models across sparse vs. nonsparse predictor sets, and assess performance in three empirically relevant settings. The standard setting follows common practice in the asset pricing literature and is highly permissive, as it ignores trading costs and comes with factor hindsight. The intermediate setting removes hindsight by restricting predictors to publicly known factors and excludes microcaps, which are hard to trade without undue price impact. These adjustments are common in the literature and allow us to benchmark our results against prior findings. The realistic setting incorporates trading costs into both portfolio construction and evaluation, providing insights most relevant in practice.

Unlike prior work that either subtracts costs ex post or focuses solely on turnover reduction, we implement cost-aware portfolio construction that selects trades ex ante based on expected returns net of transaction costs. Testing cost-aware strategies is essential, since simply subtracting transaction costs from cost-ignorant strategies cannot distinguish between worthless and valuable signals that are poorly executed. Our approach is in a portfolio-sorting spirit, adapted to incorporate explicit trading-cost considerations into both construction and evaluation, rather than proposing an optimal solution to the portfolio choice problem as in Jensen et al. (2024). This cost-aware methodology materially improves net performance while lowering transaction costs through mechanisms beyond merely reducing turnover, and allows us to properly evaluate the value of informational and functional complexity under the realistic conditions that investors face in practice.

Our main findings reveal that the value of informational and functional complexity varies substantially across settings. In the standard setting, both informational and functional complexity deliver economically large and statistically significant gains: moving from a sparse to a nonsparse predictor set increases strategy returns by about 1.20 percentage points per month, while allowing for nonlinearity adds roughly 1.0 percentage point per month. In the intermediate setting, after eliminating hindsight and excluding microcaps, the incremental value of nonlinearity largely disappears: return improvements from nonlinearity are marginally significant for nonsparse models (about 0.40% per month) and statistically insignificant for sparse models. The benefit from an expanded information set persists, though attenuated, with monthly gains declining from about 1.20% to approximately 0.70%. Under realistic conditions, where costs shape both construction and evaluation, the overall return level declines substantially, from about 1.45% gross in the intermediate setting to about 0.79% net per month. When short-sale constraints are imposed, ML-based strategy still outperforms the broad market index, but these gains arise primarily from the joint application of both information and functional complexity.

Overall, our evidence shows that nonlinearity and an expanded information set do not uniformly add value. Their contributions depend on the implementation setting. The value added by an expanded information set persists across different settings, though its magnitude varies substantially with the level of realism. In contrast, the value of functional complexity mostly disappears in cost-ignorant strategies once we remove microcaps and hindsight on yet-

to-be-published factors. In a realistic setting, the benefit of nonlinearity persists only when paired with a nonsparse information set. Under long-only constraints, neither component alone provides meaningful gains, and improvements arise mainly from their joint application.

These results have practical implications for specification choices in cross-sectional factor models. If predictive gains mainly stem from using a broader information set, ML-powered portfolios can retain the interpretability of linear models without sacrificing performance. If, instead, nonlinearity is the key driver, more complex models are warranted, and improving their tractability becomes essential.

We also examine how frictions reshape feature importance in models with a nonsparse information set. We evaluate the reduction in out-of-sample returns when restricting models to smaller predictor subsets and show that excluding microcaps and incorporating trading costs materially alter the relative importance of characteristics. In particular, variables associated with firm size and those with low persistence that induce high turnover, such as short-term reversal, become substantially less influential for out-of-sample performance once trading frictions are recognised.

The rest of the paper proceeds as follows. Section 2 details our data and methodology. Section 3 examines the performance of long–short ML-powered strategies across different settings. Section 4 repeats the analysis with short-sale restrictions, and Section 5 concludes.

2. Data & Methodology

2.1. Data

Our equity universe consists of common shares[1] listed on the NYSE, AMEX and NASDAQ - taken from the CRSP database. The data for 94 firm-level characteristics come from Gu, Kelly, and Xiu (2020).[2] Our sample starts in June 1963 and ends in December 2021.[3]

The characteristics are cross-sectionally ranked each month and scaled to range between -1 and 1. Missing values are replaced by the cross-sectional median for the respective month. To avoid predictions based on very few characteristics, firm-month observations lacking data for more than two-thirds of the characteristics are excluded from nonsparse models, whereas those missing more than one-third are excluded from sparse models prior to imputation. Additionally, any characteristic for which more than two-thirds of the historical firm-month observations are missing is omitted from the training at a given time.

2.2. Model training

We predict the cross-section of stock returns over the next month using a pooled regression

[1]Share codes 10 and 11.

[2]The data was retrieved from Dacheng Xiu's homepage, available at: https://dachxiu.chicagobooth.edu/download/datashare.zip

[3]The original dataset from Gu, Kelly, and Xiu (2020) starts in 1957. However, most characteristics based on Compustat database become available in 1963.

$$R_{i,t+1} - \bar{R}_{t+1} = f\left(X_{i,t},\, \theta\right) + \epsilon_{i,t+1} \tag{1}$$

where on the left-hand side, we have the return for stock i at time $t+1$ relative to the average return of all stocks over the same period. On the right-hand side, we have a function of characteristics for stock i at time t, with parameters θ. We train our models to minimize the average squared prediction error

$$\min_{\theta} \left\{ \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \epsilon_{i,t+1}^2 \right\} \tag{2}$$

Although we follow Gu, Kelly, and Xiu (2020) in most model specifications, a key divergence lies in the dependent variable: whereas Gu, Kelly, and Xiu (2020) regress excess returns, we predict returns relative to the cross-sectional average. This choice removes the need to model the time-varying mean level of returns—an element that is notoriously hard to forecast (see, e.g., Welch and Goyal (2008)). Because our equity strategies exploit only cross-sectional dispersion, relative forecasts suffice. In fact, targeting cross-sectionally demeaned returns aligns precisely with our objective, while absolute returns are irrelevant to our objective. Consistent with this view, Chen, Hanauer, and Kalsbach (2024) show that predicting returns relative to the market produces superior strategy performance compared with raw excess-return.[4]

We closely follow Gu, Kelly, and Xiu (2020) in how we train and tune our models. In particular, we use expanding windows, employing all available historical data up to the most recent 12 years to train the models. The latest 12-year period serves as a validation sample for optimal hyperparameter selection. Although a rolling-window scheme with shorter validation blocks is also common (see, e.g., Azevedo, Hoegner, and Velikov (2024)), Chen, Hanauer, and Kalsbach (2024) report higher strategy returns when predictions are generated with expanding windows rather than rolling windows.

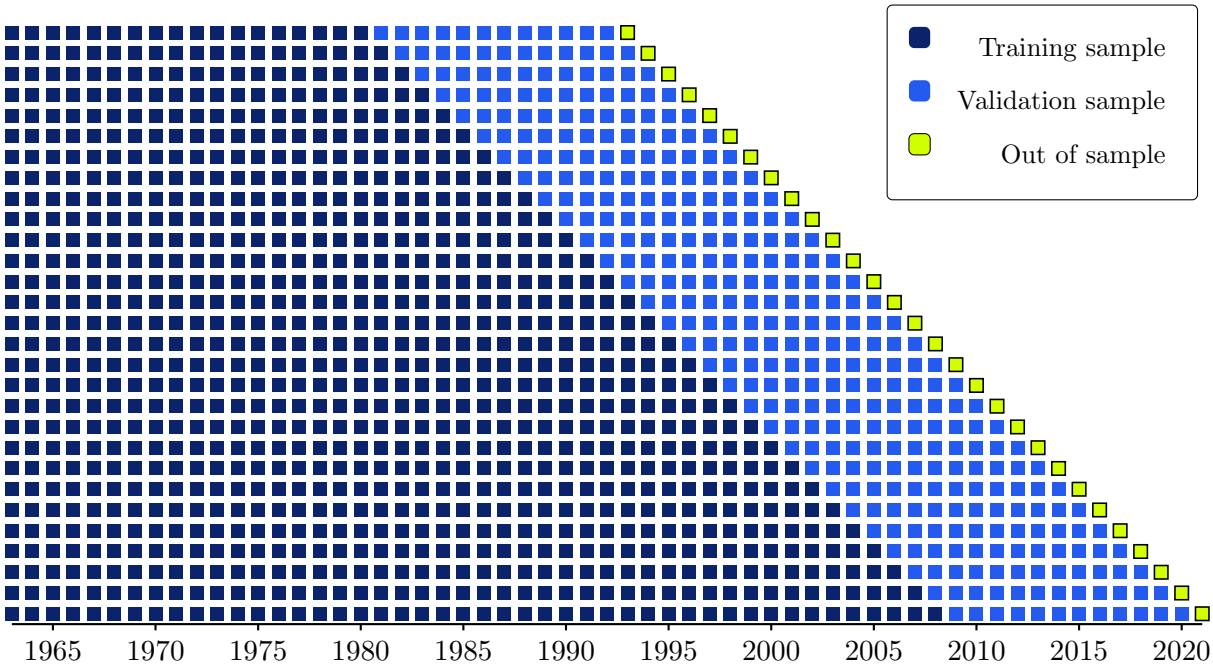
Figure 1 illustrates how models are trained, validated, and evaluated. The models are updated once a year, in June, when characteristics based on annual financial statement are refreshed for most firms.[5]

Following Gu, Kelly, and Xiu (2020), we employ early stopping when training neural networks to avoid overfitting, terminating training when validation loss begins to rise. Early stopping acts as implicit shrinkage, such as L2 regularization (see Raskutti, Wainwright, and Yu (2014)). For iterative algorithms applied to least-squares loss function, halting optimization early limits the effective weight norm, yielding solutions similar to applying an L2 penalty whose strength depends on the learning rate. We also employ an ensemble method, training the model multiple times with varying initializations and averaging the resulting predictions.

[4] Chen, Hanauer, and Kalsbach (2024) compare models that forecast (i) raw excess returns, (ii) CAPM-adjusted returns, and (iii) returns relative to the market. They find that the third specification yields the highest average strategy returns.

[5] The dataset assumes that annual financial statements are available within six months of fiscal year-end. Since most firms have December fiscal year-ends, annual firm-level characteristics are updated in June for most of the firms.

Figure 1: Training and validation of return prediction model



Finally, we apply L1 regularization to all weights to control the level of sparsity.[6] We provide detailed specifications for parameter and hyperparameter choices in the appendix.

2.3. Model specifications

Our main goal is to understand the drivers of the return differences of strategies across two dimensions: functional complexity and informational complexity. By varying parameters θ and the information set in X, we compare linear vs. nonlinear models and sparse vs. nonsparse models.

Note that, unlike the standard definition of complexity as the number of parameters relative to the sample size (see Kelly, Malamud, and Zhou (2024)), our approach distinguishes between two types of complexity. Informational complexity arises from extending the information set by adding observations on additional variables. Functional complexity stays with a given information set but allows for a flexible functional form. Both informational complexity and functional complexity increase the number of parameters, in line with the standard definition of complexity as a high number of parameters relative to the number of observations in a sample.

To see how informational complexity conceptually differs from functional complexity, consider the case of a model that predicts life expectancy of 40-year-olds. A sparse model might use easily observable variables such as smoking status, alcohol consumption, gender, and BMI. We can add informational complexity by including additional variables, such as blood pressure, cholesterol levels, diabetes status, sleep quality, exercise habits, stress levels, and family history.

[6] Although L1 regularization is associated with sparsity, in neural networks it primarily induces sparsity in the parameters (weights), not necessarily in the input features. Unlike linear models, zeroing a weight does not automatically exclude a feature, since features can influence outputs through multiple paths and layers.

These additional variables require more extensive data collection but provide richer information about the individuals in the sample.

In contrast, functional complexity involves using the same sparse set of variables but allowing for more flexible relationships. A linear model specifies that life expectancy depends additively on smoking, alcohol, gender, and BMI. A nonlinear model with the same four variables can capture interactions, for instance, that the combined effect of smoking and alcohol consumption is worse than the sum of their individual effects, and nonlinear relationships, such as an inverted U-shaped effect of BMI where both underweight and overweight individuals face reduced life expectancy.

Both forms of complexity add parameters: informational complexity through additional variable coefficients, functional complexity through interaction terms and nonlinear transformations. Which form of complexity matters more for predictive performance is ultimately an empirical question. Our analysis systematically evaluates both dimensions across different implementation settings to determine their relative importance for investment strategies.

We use neural network to allow flexible functional forms between characteristics and future returns. Neural networks are often described as universal function approximators, as they can approximate any continuous function provided the network architecture is adequately complex. Empirical results suggest that neural network models have the strongest predictive power in the cross-section of stock returns, compared to other machine learning models (see Gu, Kelly, and Xiu (2020)).[7] We follow Gu, Kelly, and Xiu (2020) and use a neural network with three hidden layers (32–16–8 neurons).[8] For the linear specification, we estimate ridge regression. Note that linear regression is a special case of a neural network with no hidden layers and linear activation function. Hence, the primary distinction between our linear and nonlinear specifications lies in a dimensionality of the parameter vector θ.

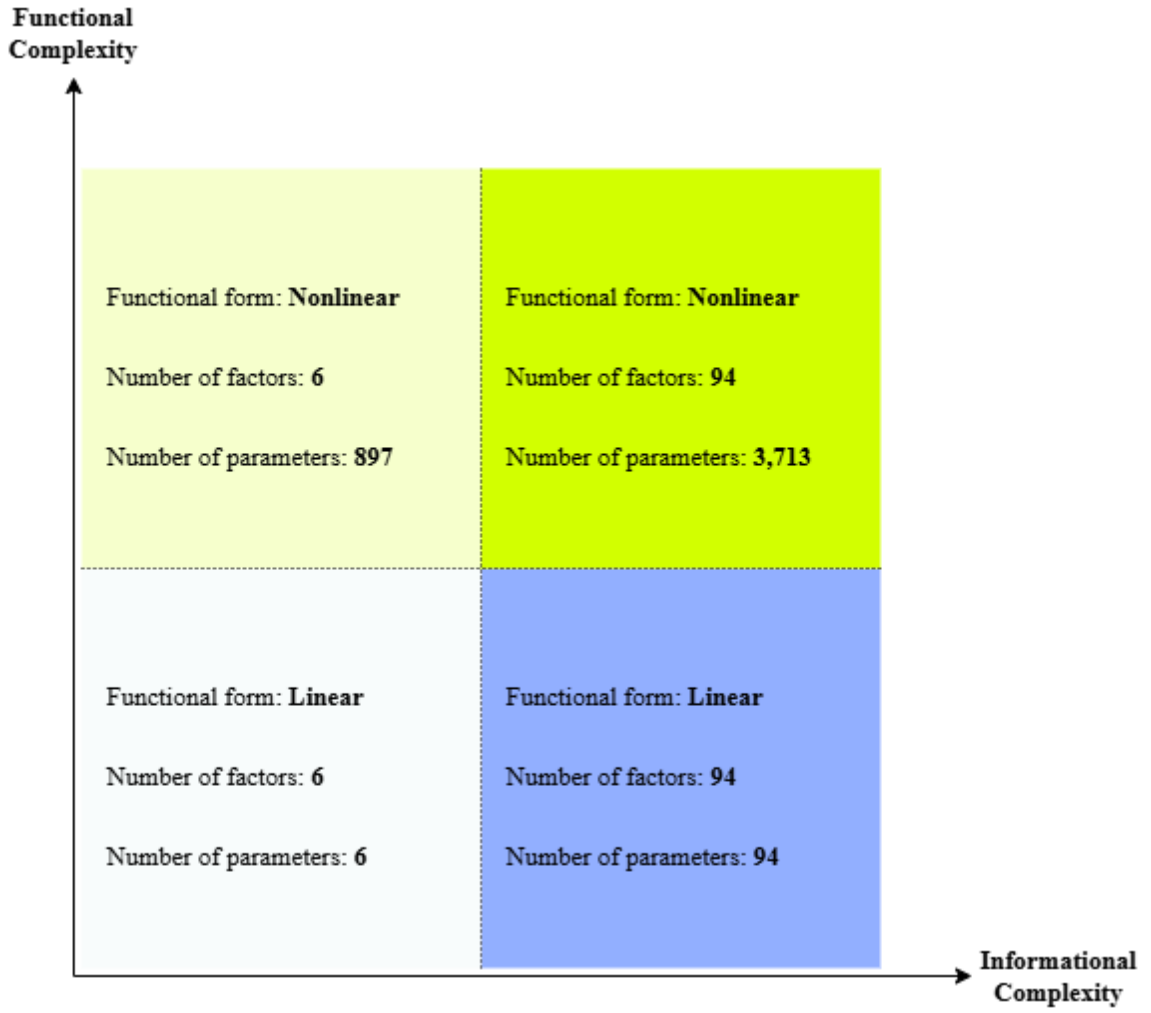
Regarding the dimensionality of inputs, we consider two sets of characteristics that define the predictor set X. The sparse model specification includes characteristics from the six-factor model of Fama and French (2018), that is, the market beta, market capitalization (size), the book-to-market ratio (value), momentum (12 months omitting the last one), operating profitability, and asset growth (investment). The nonsparse specification employs the full set of 94 characteristics.

Figure 2 illustrates the complexity of the evaluated models along two dimensions: information and functional form.

[7] While Gu, Kelly, and Xiu (2020) rely on feedforward neural networks, other studies have come up with more complex networks that also result in a very strong return predictability as well as portfolio performance. However, neural networks in general seem to perform best.

[8] There exist other methods to approximate nonlinear functions, such as random forests, boosted trees, or even using linear model with interaction and higher order terms for each predictor. We stick with a single method to estimate nonlinear relations which dominates the literature. Our objective is not to run a horse-race between different machine learning methods, but rather to compare performance of the models with simple linear specification with general nonlinear specification.

Figure 2: **Two dimensions of complexity**



Parameter count includes only learned coefficients for the predictors. It does not include tuned hyperparameters or preprocessing constants (e.g., the means/standard deviations used for standardization)

2.4. Assessing the value-added of nonlinearity and expanded information

While we provide the analysis of out-of-sample predictability for each model specification, a more relevant question from an investment perspective is how portfolios using the different return predictions perform. It is widely recognised that improving statistical fit of a return prediction model does not necessarily imply improved performance when building portfolios (see, e.g., Nagel (2021), chapter 1). For example, statistical measures like R^2 capture average prediction accuracy across all stocks, but portfolio performance may depend critically on the accuracy of rankings at the tails of the distribution. Moreover, working with portfolios reduces idiosyncratic risk and hence the impact of stock-level prediction errors. Our analysis thus focuses on the return contribution from extending traditional factor models to allow for nonlinearity and/or using broader, nonsparse set of factors.

A common practice in the literature is to sort stocks into portfolios based on model-implied expected returns, then compare the out-of-sample return differences across these portfolios. If a model reliably distinguishes between stocks with high versus low expected returns, a portfolio composed of stocks with the highest model-implied returns should outperform the one containing those with the lowest model-implied returns.

We form decile portfolios and assign equal weights to all stocks within each decile. Alternatively, one could use value-weighted decile portfolios. Our primary objective is to examine the return spread between stocks categorized as having the highest and lowest expected returns according to the model. In that sense, equal weighting provides a clearer comparison. This approach is also consistent with the objective function used when training models for return forecasts, which assigns equal weight to each firm-month observation. Moreover, to address the practical challenges associated with equally weighted portfolios, we incorporate trading costs during both the portfolio construction and evaluation process, as detailed later in this section. The results for value-weighted decile portfolios are provided in the appendix.

To assess the value of expanded information set, we compare returns of strategies based on return predictions from sparse and nonsparse models, across both linear and nonlinear specifications. The sparse model includes characteristics from the 6-factor model of Fama and French (2018). The nonsparse model includes all 94 characteristics from Gu, Kelly, and Xiu (2020). Across settings that avoid factor hindsight, we train predictive models with only those characteristics that are publicly known at a given point in time.[9] This allows avoiding hindsight that may inflate returns of a strategy. We already rely on a set of factors that have been shown to be related with expected returns. While the robustness of these relations could be debatable, we do not start from scratch[10] and use factors that we know performed well over a sample that overlaps with our sample. To avoid such hindsight from biasing our results, we exclude factors that were yet to be published.

[9] In a nonsparse factor model, factors that were published in year Y are included as a predictor starting on June of year $Y + 1$. For a sparse factor model, we use 3-factor model of Fama and French (1993) from 1993 to 1997, 4-factor model of Carhart (1997) from 1997 to 2015, and 5-factor model of Fama and French (2015) augmented with momentum factor since 2015.

[10] For example, Li et al. (2023) construct 18,000 characteristics based on various permutations of firm fundamentals.

To evaluate the incremental value of allowing for nonlinearities, we compare realized excess returns from strategies formed on linear and nonlinear return forecasts, using both sparse and nonsparse predictor sets. Our primary metric is the difference in average excess returns, which aligns with the forecasting objective of the model.

To provide a more complete assessment, we also estimate intercepts from spanning regressions that project one strategy's excess return on the other's. The intercept (alpha) from these regressions isolates the component of average return that is orthogonal to the strategy used for comparison. While our main test tells us about average return differences between strategies, the alphas show whether this difference represents truly incremental returns or simply reflects greater exposure to common sources of return variation already captured by the comparison strategy.

2.5. Different settings for evaluation

We evaluate the benefits of informational and functional complexities under three distinct settings. The first, referred to as the standard setting, aligns with common practice in the asset pricing literature. It includes all common shares from major U.S. equity exchanges and relies on factor characteristics that were not publicly known at the time.

Although this standard approach allows comparisons with existing studies, it is highly unrealistic for several reasons. First, the resulting portfolios take positions in illiquid stocks that are difficult and expensive to trade. Second, some factors used to train the model were not publicly known as predictive of future returns at the time in question.

These shortcomings are straightforward to address. It is common to restrict the set of predictors to factors that were publicly known, thereby eliminating hindsight bias (see, e.g., Li et al. (2023); Azevedo, Hoegner, and Velikov (2024)). Chen, Hanauer, and Kalsbach (2024) show that excluding factors before they were published results in a reduction of around 50 bps per month, which is statistically significant. In addition, many studies exclude microcap stocks because of the significant practical challenges they pose for replicating these strategies (see e.g. Fama and French (2008); Kozak, Nagel, and Santosh (2020)). Note that we do not train our predictive models on the reduced universe. Both linear and nonlinear models are estimated for the whole equity universe only.[11] After excluding microcaps, we are left with the largest 3000 stocks in terms of market capitalisation. Our equity universe that excludes microcaps covers 98% of the aggregate equity universe most of the times. Appendix provides comparison with other popular definitions of microcaps in the literature. We refer to this approach as the intermediate setting, because although it limits hindsight bias and removes microcaps, it does not account for the impact of trading costs.

Last but not least, we examine what we call the realistic setting. The standard setting and the intermediate setting both ignore any transaction costs arising from rebalancing portfolios as expected returns evolve. In this realistic setting, in addition to excluding microcaps and removing factor hindsight, we also incorporate trading costs into both the portfolio evaluation and

[11] Since we include market-cap and other liquidity related variables, neural networks are expected to capture relationships that are conditional of the firm size.

construction. This allows us to understand the benefits of machine learning return predictions under real-world constraints, providing the most relevant insights on the benefits of ML. Having multiple settings allows us to compare our results with existing studies, while also providing insights for investors concerned with practical constraints.

Table 1 provides the summary of each implementation setting.

Table 1: **Summary of implementation settings**

	Standard setting	Intermediate setting	Realistic setting
Exclude Microcaps	✗	✓	✓
Remove factor hindsight	✗	✓	✓
Account for trading costs	✗	✗	✓

2.6. Estimating trading costs

Providers of commercial investment products would often disclaim that past performance does not guarantee future performance. Indeed, predictability of equity returns is weak, whether in the time-series or in the cross-section, especially at shorter horizon. On the other hand, trading costs are much more reliably predictable. However, a large body of literature ignores trading costs when evaluating portfolio performance.

We explicitly account for firm-level transaction costs every time portfolios are rebalanced, that is, each month. Following Chung and Zhang (2014), we estimate monthly effective bid-ask spread as a daily average of closing quoted spreads[12]

$$c_{i,t} = \frac{1}{S}\sum_{s=1}^{S}\frac{P_{i,s}^{A} - P_{i,s}^{B}}{P_{i,s}^{A} + P_{i,s}^{B}} \tag{3}$$

where $P_{i,s}^{A}$ and $P_{i,s}^{B}$ are closing ask and bid prices on day s, respectively, and S is the number of trading days at month t. Hence, $c_{i,t}$ is the half effective bid-ask spread expressed as a percentage of the mid-price. We use trading costs estimated at time t when rebalancing portfolios at the end of time t. Therefore, our analysis is free of lookahead bias.

This low frequency proxy of effective bid-ask spread is among the most reliable proxies both in terms of cross-sectional and time-series correlation with effective spreads from high-frequency data.

Monthly spread estimates are sometimes unavailable for some stocks due to poor coverage of daily data. In such cases, we use the average spread across 20 stocks with the smallest distance[13] from a given stock. The distance between stocks i and j is defined as follows:

[12] We follow Chung and Zhang (2014) and exclude daily observations when spread is greater than 50%. Moreover, we require at least 5 daily observations for a given stock-month to compute the spread.

[13] When at least one firm-level characteristic is missing, we replace missing spread with the median spread of the stocks from the same market-cap decile. Such cases are very rare, and most spreads are replaced based on all three characteristics.

$$d_{ij} = \sqrt{(cap_i - cap_j)^2 + (ivol_i - ivol_j)^2 + (volume_i - volume_j)^2} \tag{4}$$

where *cap* refers to market capitalization, *ivol* is idiosyncratic volatility,[14] and volume is the monthly dollar trading volume.[15]

This approach significantly improves accuracy compared to a common approach of using the single nearest neighbour (see e.g. Novy-Marx and Velikov (2016)). The appendix provides a detailed comparison between the two approaches.

2.7. Building cost-aware strategies

Instead of selecting stocks with the highest ex-ante gross returns, we account for the transaction cost borne by each change in portfolio composition. The cost-aware strategies correspond to equally weighted portfolios that select stocks with the highest/lowest ex-ante expected returns over the following month, net of trading costs. Note that we aim to build strategies equivalent to simple equally weighted portfolios, while considering trading costs, rather than proposing an optimal solution to the portfolio choice problem.

At each rebalancing, selected stocks are assigned equal weights. Replacing stock i with stock j has the following impact on the expected return of a portfolio, net of transaction costs:

$$\underset{j \to i}{\Delta} \mathbb{E}[r_p] = \underbrace{\frac{1}{N}\left(\mathbb{E}[r_j] - \mathbb{E}[r_i]\right)}_{\substack{\text{difference in} \\ \text{expected returns}}} - \underbrace{\left(\frac{1}{N}c_j + w_i c_i\right)}_{\substack{\text{cost of} \\ \text{replacement}}} + \underbrace{\left|\frac{1}{N} - w_i\right| c_i}_{\substack{\text{avoided} \\ \text{rebalancing} \\ \text{cost}}} \tag{5}$$

where N is the number of stocks in the portfolio, r is the return, c is the cost of trading, and w_i is the actual weight of stock i before rebalancing. Note that we compare two scenarios: (i) replacing stock i with stock j, and (ii) keeping stock i and rebalancing its weight to $1/N$. Therefore, the net gain from the swap equals the gross expected return differential minus the replacement costs, plus the avoided rebalancing cost. This operation is pairwise and additive, and its impact is independent of the rest of the portfolio.

Figure 3 provides visual example of how each trade is evaluated. In this example, gain from replacing stock i by stock j is 0.06% in terms of portfolio-level expected return net of trading costs.

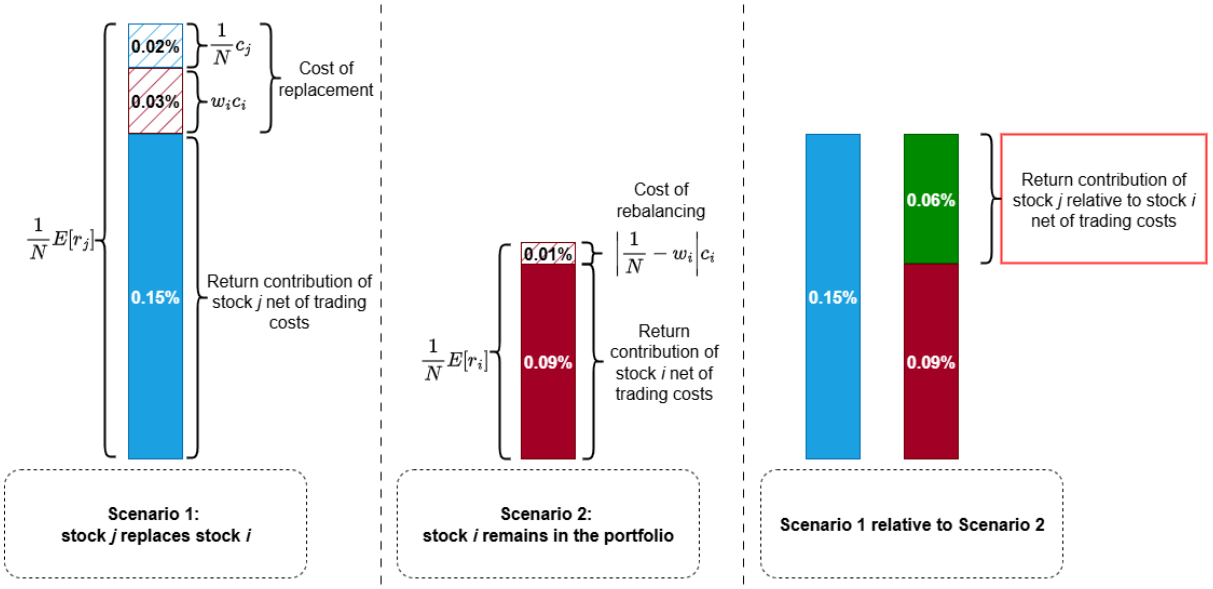
We can group stock-specific terms from (5) in the following way

$$\underset{j \to i}{\Delta} \mathbb{E}[r_p] = \underbrace{\left[\frac{1}{N}\left(\mathbb{E}[r_j] - c_j\right)\right]}_{\underset{j\to}{\Delta}\mathbb{E}[r_p]} - \underbrace{\left[\frac{1}{N}\mathbb{E}[r_i] + c_i\left(w_i - \left|\frac{1}{N} - w_i\right|\right)\right]}_{\underset{i\leftarrow}{\Delta}\mathbb{E}[r_p]} \tag{6}$$

[14] Idiosyncratic volatility is the standard deviation of the residuals in Fama-French 3-factor model estimated using daily returns over the past 90 days. We require at least 30 observations in total and at least 10 observations during the most recent month.

[15] For these characteristics, we use ranks scaled to range from zero to one.

Figure 3: **Evaluating a trade in cost-aware strategies**



Following values have been assumed in this example: $\mathbb{E}[r_j] = 2.0\%$; $\mathbb{E}[r_i] = 1.0\%$; $c_i = 0.2\%$; $c_j = 0.2\%$; $w_i = 15\%$; $N = 10$

This allows to measure the impact of each stock entering/remaining in portfolio independent of each other and other trades. We can build an equally weighted portfolio that maximizes the expected return over the next period net of trading costs as follows

- **List A**: sort portfolio constituents (denoted by i) according to $\underset{i\leftarrow}{\Delta}\mathbb{E}[r_p]$ in ascending order. Let i range from 1 to N
- **List B**: sort the remaining stocks (denoted by j) according to $\underset{j\rightarrow}{\Delta}\mathbb{E}[r_p]$ in descending order. Let j also range from 1 to N
- For all $i = j$ between 1 and N, keep replacing the i^{th} stock from **List A** by j^{th} stock from **List B** as long as $\underset{j\rightarrow}{\Delta}\mathbb{E}[r_p] - \underset{i\leftarrow}{\Delta}\mathbb{E}[r_p] > 0$

In a long-short context, we need to repeat the same procedure for the short leg with a slight adjustment. The impact on gross expected returns in a short leg will be exactly the opposite, while the impact due to trading costs will remain the same. Hence, we can define the stock-specific components in a short-leg as follows

$$\underset{j\rightarrow}{\Delta}\mathbb{E}[r_p^s] = -\frac{1}{N}\left(\mathbb{E}[r_j] + c_j\right) \tag{7}$$

$$\underset{\rightarrow i}{\Delta}\mathbb{E}[r_p^s] = -\frac{1}{N}\mathbb{E}[r_i] + c_i\left(w_i - \left|\frac{1}{N} - w_i\right|\right) \tag{8}$$

Note that we only account for proportional costs and exclude scale dependent costs. Hence, we assume that the size of trades is modest and cannot have impact on prices. This is not a

very stringent assumption because we exclude microcaps from the analysis. This ensures that trading happens in relatively more liquid stocks, mitigating possible impact on stock prices due to trading.[16]

3. What drives the returns of ML-powered strategies?

3.1. Out of sample stock-level predictability

Before analysing the strategy returns based on information from return prediction models, we look at the statistical accuracy of stock-level predictions for different models. Table 2 shows the out of sample R^2 computed at a stock-level. As expected, richer information set, as well as accounting for nonlinearities, both substantially improve the predictability of stock level returns. The predictability is noticeably lower when removing the factor hindsight. Overall, the predictability at a stock level is very low.

Table 2: **Out-of-sample R^2 of different predictive models**

	With Factor Hindsight		Without Factor Hindsight	
	Linear	Nonlinear	Linear	Nonlinear
Sparse factors	0.06%	0.19%	0.02%	0.16%
Nonsparse factors	0.20%	0.40%	0.08%	0.25%

The reported measures correspond to R^2 computed over the period from July 1993 to December 2021. Each firm-month observation is weighted equally. Since the return predictions are made for returns relative to the average return in a given month, we cross-sectionally demean realized returns each month.

3.2. What are the benefits of nonlinearity and expanded information set?

We analyse the returns of strategies using different predictive models in three different settings. The standard setting is comparable with the typical setup in the literature, such as Gu, Kelly, and Xiu (2020). In this setting, transaction costs are ignored both in strategy design and in evaluation of returns. We also allow for additional unrealistic assumptions by including microcap stocks in the analysis and by including all characteristics from the start of the analysis, even at times when the corresponding return effects had not yet been documented. Our objective is to provide a decomposition of return effects in a framework that is comparable with that used in the prior literature.

In an intermediate setting, we exclude microcaps and eliminate the hindsight on factors. Such a setting is also common in the literature (see e.g Avramov, Cheng, and Metzker (2023); Chen, Hanauer, and Kalsbach (2024)). We refer to this setting as the "intermediate setting" as it shares some of realistic adjustments that we make later.

[16]Note that when we analyse the performance of long-short strategies, we are not just looking at the difference in net returns of long and short portfolios. We are looking at the net returns of long-short strategies where the trading costs of both long and short legs reduce returns. We do not account for security lending fees and margin requirements.

Finally, we account for the hindsight and include a factor in available set only post publication, exclude microcaps to loosen the assumption of no price impact, and analyse returns net of proportional transaction costs. We refer to this setting as the "realistic setting".

Table 3 reports mean monthly returns of long-short strategies based on return predictions from various models.[17]

Considering a standard setting, we find that the strategy based on the linear sparse model delivers a monthly gross return of 1.45%, while the strategy based on the nonsparse nonlinear model delivers a monthly return of 3.63%. Moreover, the nonlinearity adds 0.91% and 0.99% to the performance under sparse and nonsparse specifications, respectively. Deviating from the sparse factor model adds 1.19% and 1.26% to the performance in linear and nonlinear models, respectively. The improvements coming from increasing both functional and informational complexity are economically large and statistically significant. Moreover, they are of the same order, implying that they both contribute meaningfully to the superior performance of ML models.

Moving to the intermediate setting, which excludes microcaps and eliminates factor hindsight, we find that the performance of long-short strategies is relatively modest. In case of a sparse linear model, we observe a mean monthly return of 0.33%, which increases to 1.45% in case of strategy powered by a nonlinear nonsparse model. The value added by nonlinearity is around 0.40% per month, but it is statistically insignificant for the sparse model, and only marginally significant for the nonsparse model. On the other hand, we find that expanding information set adds significant value when considering both linear and nonlinear models. These improvements are relatively modest compared to the standard setting, but they exceed the value added by nonlinearity in this intermediate setting, adding 0.73%, and 0.66% over sparse models, under linear and nonlinear specifications, respectively.

Moving to the realistic setting, where we rely on cost-aware strategies, we find that the returns net of trading costs are considerably lower compared to gross returns from the standard and intermediate settings. In most cases, the returns are close to zero, and statistically insignificant. The only strategy yielding positive and statistically significant returns is based on the nonlinear nonsparse factor model.

In this realistic setting, the differences in average net returns between the ML-powered strategy and a strategy based on a traditional factor model appear to stem from both informational and functional complexity. However, the contribution of expanded information set is both larger in magnitude and statistically more robust. Allowing for interactions and nonlinear functional forms offers little benefit with a sparse set of characteristics, where the return improvement is only 21 basis points per month, with a t-statistic of 0.57 — well below conventional significance thresholds. In contrast, strategies based on nonlinear nonsparse models outperform their linear nonsparse counterparts by 43 basis points per month, a marginally significant improvement (t-statistic of 1.79). Relying on a broader information delivers stronger and more consistent performance gains: 43 basis points for linear models and 65 basis points for nonlinear models, with corresponding t-statistics significant at the 10% and 1% levels, respectively. The results

[17] We report cumulative returns of the strategies from Table 3 in the appendix.

suggest that the incremental value of both informational and functional complexity remain broadly consistent across intermediate and realistic settings.

Table 3: **Mean monthly returns of long-short strategies (in %)**

	Linear	Nonlinear	Nonlinear minus Linear
Panel A: Standard Setting			
(includes microcaps, with factor hindsight, ignores trading costs)			
Sparse	1.45***	2.36***	0.91***
	(5.20)	(7.68)	(4.78)
Nonsparse	2.64***	3.63***	0.99***
	(8.44)	(8.72)	(4.14)
Nonsparse minus Sparse	1.19***	1.26***	2.18***
	(4.13)	(4.26)	(5.21)
Panel B: Intermediate Setting			
(excludes microcaps, without factor hindsight, ignores trading costs)			
Sparse	0.33	0.79*	0.45
	(1.14)	(1.95)	(1.07)
Nonsparse	1.06***	1.45***	0.39*
	(3.96)	(4.68)	(1.85)
Nonsparse minus Sparse	0.73***	0.66**	1.12***
	(2.71)	(2.24)	(3.60)
Panel C: Realistic Setting			
(excludes microcaps, without factor hindsight, accounts for trading costs)			
Sparse	-0.07	0.14	0.21
	(-0.29)	(0.38)	(0.57)
Nonsparse	0.36	0.79**	0.43*
	(1.27)	(2.56)	(1.79)
Nonsparse minus Sparse	0.43*	0.65***	0.86***
	(1.73)	(2.74)	(2.94)

The sample goes from June 1993 to December 2021. The table reports mean monthly returns (t-stats). The last row reports differences between nonlinear and linear models, and columns report differences between nonsparse and sparse predictor sets. The row-column intersection reports the mean return difference between strategies based on nonlinear nonsparse and the linear sparse models. Strategies correspond to equally weighted long-short portfolios that select stocks (10%) based on monthly return predictions of linear (ridge) and nonlinear (neural network) models estimated on both sparse and nonsparse predictor sets. Each month, microcaps are defined as stocks outside the largest 3000 by market capitalization. The sparse set of predictors include characteristics from the Fama-French 6-factor model: market beta, market cap, book-to-market ratio, return momentum (12-1), operating profitability, and asset growth. The nonsparse set includes 94 factors from Gu, Kelly, and Xiu (2020). We remove factor hindsight by only including the characteristics that were publicly known at a given point in time. T-statistics are based on Newey-West adjusted standard errors with the lag of 4. Statistical significance at the 10%, 5% and 1% levels are indicated by *, ** and ***, respectively.

The results of spanning regressions in Table 4 largely mirror the analysis of raw mean returns in Table 3. In the standard setting, both informational and functional complexity generate large, statistically significant gains. Excluding microcaps and removing factor hindsight (intermediate setting) attenuate magnitudes but preserve the ranking: informational complexity continues to add economically and statistically meaningful value, whereas the incremental benefit of functional complexity is only preserved when using nonsparse predictor set. In the realistic setting, alphas compress further. The benefits provided by expanded information set remain positive and statistically significant, while nonlinearity adds little. Overall, both analyses indicate that informational complexity is the primary driver of incremental performance of the ML-powered strategies relative to traditional linear models estimated on a sparse set of characteristics.

Table 4: **Alphas from spanning regressions of L/S strategies (in %)**

	Lin.	**Nonlin.**	**Sparse**	**Nonsparse**	
Panel A: Standard Setting					
(includes microcaps, with factor hindsight, ignores trading costs)					
Nonsparse over sparse	2.10***	2.16***			
	(5.67)	(6.86)			
Nonlinear over linear			1.12***	1.21***	
			(5.11)	(4.26)	
Nonlinear nonsparse over linear sparse					3.19***
					(6.36)
Panel B: Intermediate Setting					
(excludes microcaps, without factor hindsight, ignores trading costs)					
Nonsparse over sparse	0.95***	0.97***			
	(4.73)	(5.35)			
Nonlinear over linear			0.10	0.88**	
			(0.55)	(2.23)	
Nonlinear nonsparse over linear sparse					1.08***
					(4.43)
Panel C: Realistic Setting					
(excludes microcaps, without factor hindsight, accounts for trading costs)					
Nonsparse over sparse	0.54***	0.49***			
	(3.11)	(2.73)			
Nonlinear over linear			0.30*	0.56	
			(1.67)	(1.52)	
Nonlinear nonsparse over linear sparse					0.72***
					(3.11)

The sample goes from June 1993 to December 2021. The table reports monthly alphas (t-stats) from regressions of one strategy's returns to those of another, with a constant included. Strategies correspond to equally weighted long-short portfolios that select stocks (10%) based on monthly return predictions of linear (ridge) and nonlinear (neural network) models estimated on both sparse and nonsparse predictor sets. Each month, microcaps are defined as stocks outside the largest 3000 by market capitalization. The sparse set of predictors include characteristics from the Fama-French 6-factor model: market beta, market cap, book-to-market ratio, return momentum (12-1), operating profitability, and asset growth. The nonsparse set includes 94 factors from Gu, Kelly, and Xiu (2020). We remove factor hindsight by only including the characteristics that were publicly known at a given point in time. T-statistics are based on Newey-West adjusted standard errors with the lag of 4. Statistical significance at the 10%, 5% and 1% levels are indicated by *, ** and ***, respectively.

Overall, the value added by nonlinearity depends on the implementation setting, while the benefit of expanded information set persists across all settings with varying magnitude. In the

standard setting most often considered in the literature, both allowing nonlinearity and using broader information provide large performance improvements. In the intermediate setting, where microcaps are excluded and factor hindsight is removed, the value added by nonlinearity disappears for sparse models and is substantially weaker for nonsparse models. In contrast, expanded information set continues to provide performance benefits over sparse models for both linear and nonlinear models, though the magnitude of these benefits is reduced. These results remain largely unchanged in the realistic setting when accounting for trading costs in both strategy design and performance evaluation. However, the magnitude of returns from ML-powered strategies is generally lower than in the less realistic settings, as expected. Using the machine learning predictions that combine broader information set with nonlinear functional form boosts the monthly returns (alphas) of linear sparse models by 2.18% (3.19%) in the standard setting, 1.12% (1.08%) in the intermediate setting, and 0.86% (0.72%) in the realistic setting.

3.3. Investability of long-short strategies

We now turn to analysing the key investability metrics of the strategies considered above. We focus on strategies from the intermediate and realistic settings, which exclude microcaps and are free of factor hindsight. Table 5 provides insights on the total amount traded as well as the estimated cost of those trades. In addition, we identify stocks that are hard to trade based on historical volumes, assuming a marginal investor with limited assets tracking these strategies. Our goal is to understand (i) how does increasing number of predictors and allowing for nonlinearities affect implementability of these strategies, and (ii) the impact of integrating trading costs into portfolio construction.

The results in Table 5 indicate that strategies based on sparse models exhibit materially lower turnover than those based on nonsparse models. For cost-ignorant portfolios, average turnover increases from 30.5% to 133.1% when moving from sparse to nonsparse under a linear specification, and from 58.3% to 129.0% under a nonlinear specification. By contrast, allowing for nonlinear functional forms in return prediction results in relatively modest impact in strategy turnover (from 30.5% to 58.3%) when using sparse set of predictors, and leaves turnover essentially unchanged within nonsparse models (133.1% vs 129.0%).

This sharp rise in turnover due to expanded information set translates into markedly higher transaction costs for cost-ignorant strategies, increasing from 0.30% to 1.15% for linear model, and from 0.50% to 1.20% for nonlinear model. Moving from linear to nonlinear models also raises costs, but by much less - about 20 basis points for the sparse set and 5 basis points for nonsparse set.

While prior work documents that expanding the factor set can reduce turnover because trades partially offset across signals (see DeMiguel et al. (2020)), our setting differs. The broader predictor set includes characteristics tied to short-horizon signals, such as reversals and liquidity-related measures, which induce frequent trading. As we show later, our predictive models load heavily on these predictors. When these predictors are added to the sparse factor set, the resulting trading intensity overwhelms any cross-signal netting, leading to higher turnover and

consequently higher transaction costs.

Panel C of Table 5 highlights the implementation challenges of cost-ignorant strategies. For a portfolio worth 1 billion USD (as of December 2021) and an execution constraint of 10% of average historical daily volume, the ML-based strategy can require more than two months (56 trading days) to fully rebalance in illiquid stocks. Given monthly rebalancing, such horizons are impractical. Consistent with the turnover and trading-cost evidence, frictions bite harder when expanding the factor set (from sparse to nonsparse) compared to when relaxing functional form (from linear to nonlinear).

These patterns persist for cost-aware strategies, but with a crucial difference: incorporating costs in portfolio construction not only lowers turnover, but it selectively avoids expensive trades. Relative to cost-ignorant strategies, the turnover for cost-aware strategies is roughly 25-40% lower, whereas trading costs fall by about 66-72%, implying a much lower cost per unit of turnover. Reduction in the days-to-trade measures is even larger, exceeding 75% in all cases.

Table 5: **Investability of long-short strategies**

	Cost-ignorant strategies			Cost-aware strategies			Percent reduction Cost-aware vs. Cost-ignorant	
	Lin.	Nonlin.	Nonlin. - Lin.	Lin.	Nonlin.	Nonlin - Lin	Lin.	Nonlin. - Lin.
Panel A: Monthly Turnover (in %)								
Sparse	30.5	58.3	27.9	18.7	39.8	21.1	38.7	31.8
Nonsparse	133.1	129.0	-4.2	97.9	93.2	-4.7	26.4	27.7
Nonsparse - Sparse	102.6	70.6	98.5	79.2	53.4	74.5	-	-
Panel B: Monthly Trading Costs (in basis points)								
Sparse	30	50	20	8	15	7	72	69
Nonsparse	115	120	5	39	41	2	66	66
Nonsparse - Sparse	85	70	90	30	25	32	-	-
Panel C: Extreme Days-to-Trade (95^{th} percentile)								
Sparse	0.9	4.5	3.6	0.2	0.5	0.3	80%	89%
Nonsparse	22.6	56.0	33.4	4.3	14.2	9.9	81%	75%
Nonsparse - Sparse	21.7	51.5	55.1	4.1	13.8	14.1	-	-

The table reports time-series mean of monthly one-way turnover, transaction cost, and the days to trade over the sample from July 1993 to December 2021. The days to trade each month is computed as a 95^{th} percentile of stock-level measures of days to trade, which is how many days it takes a given stock to trade assuming no more than 10% of the average daily dollar traded volume (ADDTV) is traded. ADDTV is measured as a daily average within given month. We assume the notional amount of 1 billion USD as an initial investment in each strategy. The notional amount is deflated based on returns of CRSP value-weighted market index and is equal to 53.2 million USD as of June 30, 1993. The strategies correspond to equally weighted long-short portfolios that select stocks (10%) based on monthly return predictions of linear (RIDGE) and nonlinear models (Neural Network) with corresponding set of predictors. Two versions of strategies are considered: ones that ignore transaction costs, and ones that consider transaction costs during the stock selection. The strategies are built on investable equity universe, excluding all stocks beyond the largest 3000 stocks at each point in time. Sparse set of predictors include characteristics from Fama-French 6-factor model: market beta, log of market-cap, book-to-market ratio, return momentum (12-1), operating profitability, and asset growth. Nonsparse set includes 94 factors from Gu, Kelly, and Xiu (2020) after adjusting for hindsight by removing factors before they were published.

3.4. Which characteristics drive predictability?

Our findings indicate that the advantages of ML-based strategies stemming from model complexity and the use of broader information set vary considerably across different settings. Furthermore, imposing realistic constraints tends to substantially diminish these advantages. This section provides additional insight into the sources of return predictability in ML models applied to the cross-section of equities. In particular, we examine the contribution of individual firm characteristics in both linear and nonlinear frameworks, within a nonsparse predictor space. To this end, we rely on standard measures of feature importance based on in-sample model fit. More importantly, we also provide an analysis of the economic value added by each feature in

terms of out-of-sample strategy performance.

3.4.1 Statistical feature importance (in-sample analysis)

To quantify the statistical importance of each feature, we measure the reduction in R^2 following Gu, Kelly, and Xiu (2020).[18] Each time a model is trained, we set the values of a given characteristic to zero (the cross-sectional median) and compute the reduction[19] in R^2. For each omitted characteristic, we average the reduction in R^2 across all dates when models are trained. We then normalize this measure of feature importance so that it sums to one across all characteristics. Figure 2 presents the results for features that rank among top 10 for either linear or nonlinear models when using broad set of factors as inputs.[20]

Both linear and nonlinear models consistently identify short-term reversal as one of the most influential predictors of monthly returns. The next four most important features are all related to firm liquidity. While both models generally agree on the relevance of these liquidity-related variables, they diverge slightly in emphasis: the nonlinear model assigns greater importance to market capitalization and dollar trading volume, whereas the linear model gives more weight to share turnover and its volatility.

Beyond the top five features, momentum-related characteristics emerge as an important predictor for both models, whereas firm fundamentals—such as the book-to-market ratio and asset growth—provide limited contribution to return predictability.

These measures of feature importance apply to the return predictions generated by the two types of models – linear and nonlinear – when applied to a broad set of characteristics. When building portfolios, however, additional information becomes relevant. For example, when excluding microcap stocks, variables with large importance in the model might not be able to express themselves if they influence returns via interaction effects which only exist within microcaps. Similarly, some stocks with high return predictions might not appear attractive when trading off these returns against transaction costs. Therefore, characteristics that improve predictions but incur high transaction costs will be less influential for portfolios than their statistical importance suggests.[21] To assess the importance of variables – not to model-predicted returns but to strategy performance – we turn to an analysis of their economic importance.

3.4.2 Economic feature importance (out-of-sample analysis)

This section evaluates the economic importance of firm-level characteristics. In contrast to the prior analysis focused on statistical relevance, our objective here is to assess the economic value

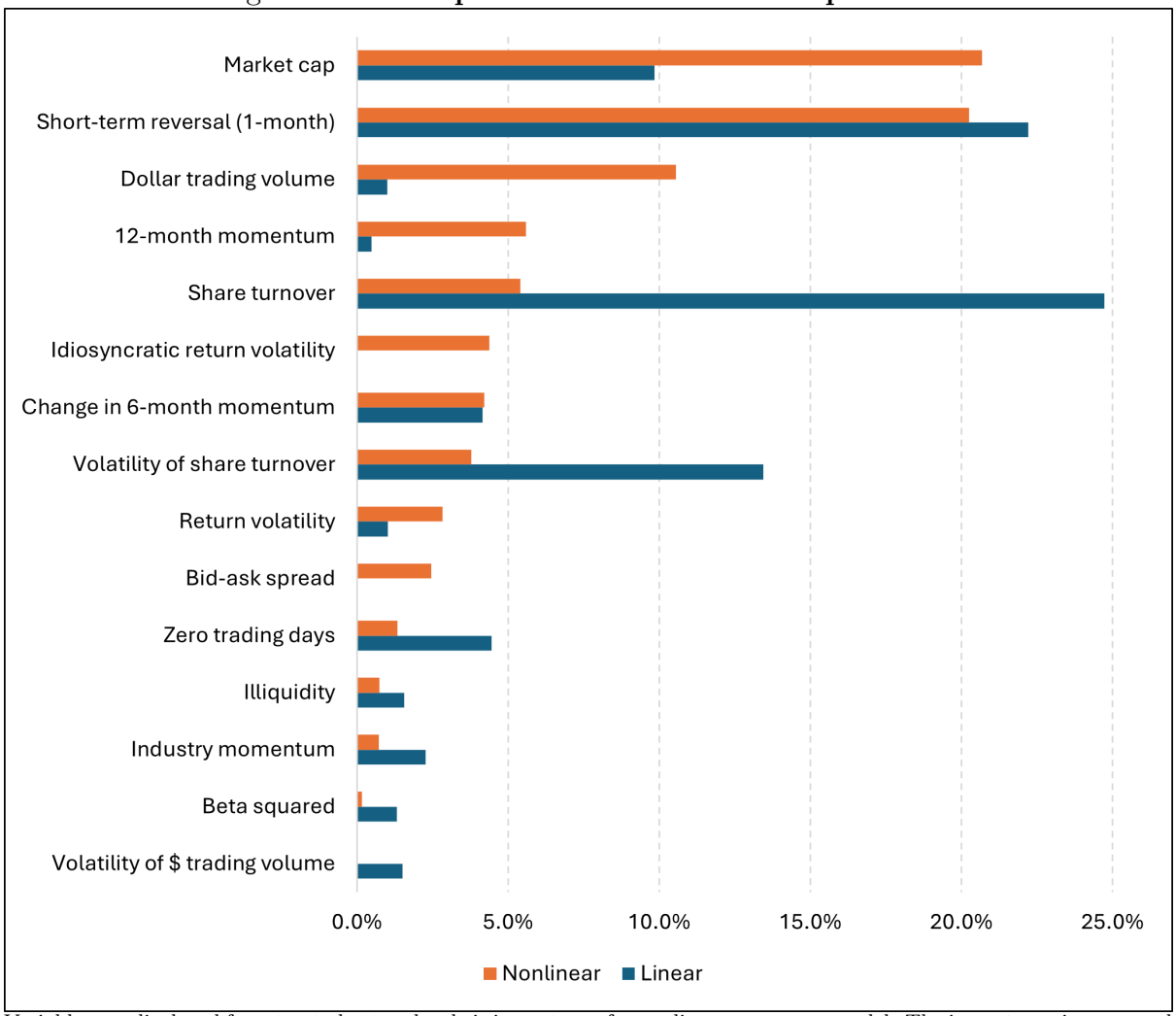
[18] We also provide feature importance measures based on the mean squared partial differential in the appendix.

[19] Ideally, one should train the model with a restricted set of characteristics, but this would require additional computing resources. Given that neural network training is computationally intensive and time-consuming, re-training for each feature subset is typically not feasible in practice.

[20] These models come with factor hindsight. We do not exclude unpublished factors to maintain the comparability across all characteristics.

[21] We do not distinguish across implementation settings when analysing statistical feature importance because we do not train separate model for the stock universe that excludes microcaps, and we do not exclude factors that were yet to be published when evaluating feature importance.

Figure 4: **In-sample statistical feature importance**



Variables are displayed from top to bottom by their importance for nonlinear nonsparse model. The importance is measured by reduction in R^2 when values of a given feature is set to cross-sectional median. Importance measures are scaled to sum up to one.

of each feature by examining the effect of its omission on out-of-sample strategy returns.

The procedure is as follows: for both linear and nonlinear models, we generate predictions while holding the value of a specific characteristic constant at its cross-sectional median at each point in time. This effectively removes the cross-sectional informational content of that feature, simulating a setting in which the model cannot differentiate stocks based on that characteristic. If the exclusion of a feature leads to a deterioration in the returns of the resulting strategy, the feature can be considered economically important. Conversely, if the omission has little to no effect on strategy performance, the feature has limited economic relevance. We look at the feature importance across the different settings employed earlier, focusing on nonsparse models.[22]

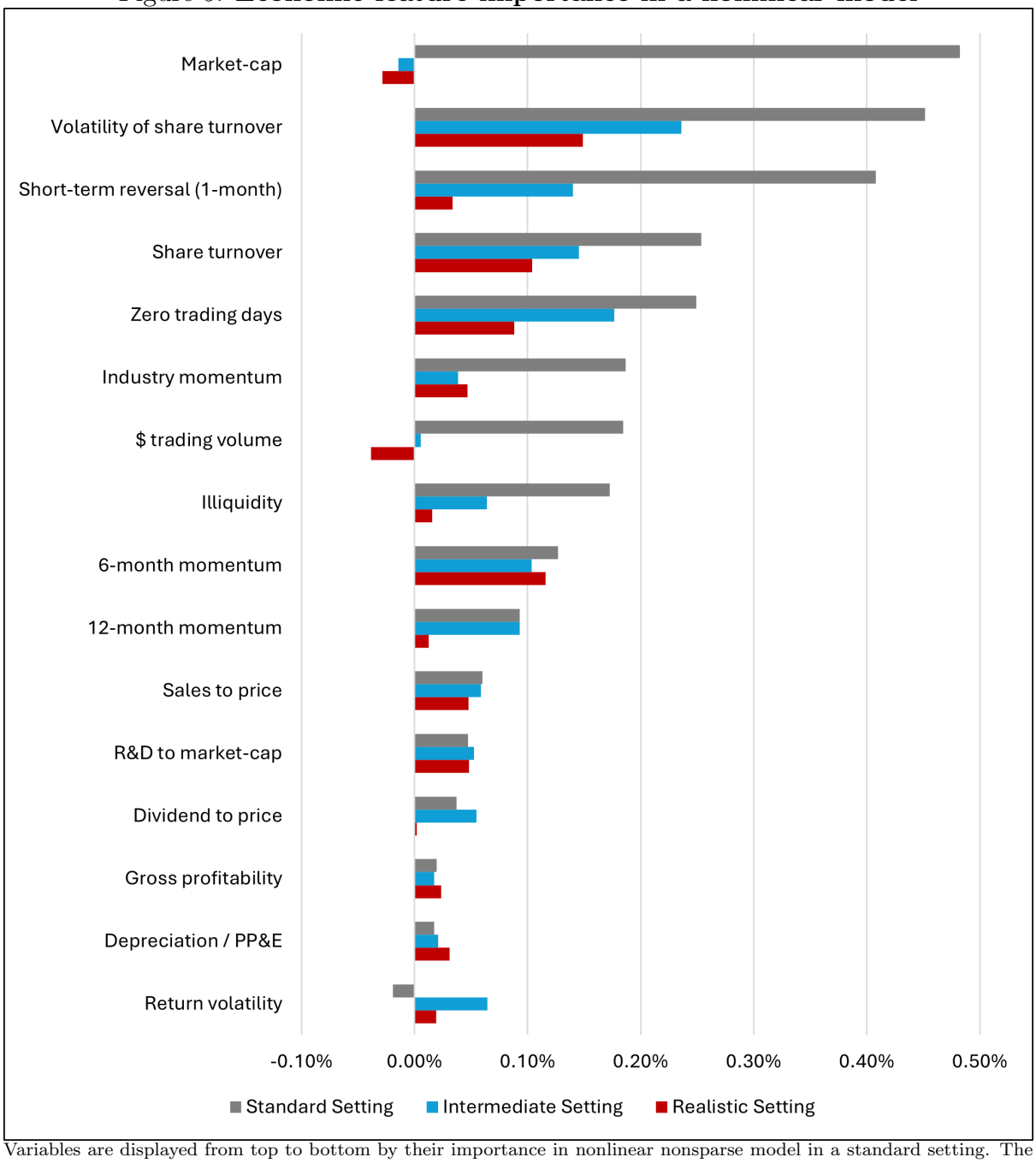
Figure 5 illustrates the reduction in out-of-sample returns when individual features are muted in nonlinear return prediction. For brevity, the results are limited to characteristics that rank among the ten most influential features in at least one setting.

The results in Figure 5 reveal meaningful differences in feature importance across implementation settings. Some variables see an important reduction in their importance when moving from the standard setting to the intermediate setting (which excludes microcaps) and the more realistic setting (which considers transaction costs). Among these variables, the reduction in importance is particularly striking for market capitalisation. While omitting this information costs 48 basis points per month in the standard setting, omission marginally increases returns in the intermediate setting without microcaps and the more realistic setting with transaction costs - as shown by negative feature importance. Similarly, considering short term reversal adds 41 basis points in the standard setting but only 14 basis points in the intermediate setting and only 3 basis points in the realistic setting. Other features, such as industry momentum and dollar trading volume, also lose importance when moving away from the standard setting.

In contrast, other features show some consistency in their importance across different settings. For instance, share turnover and its volatility remain among the most influential predictors in both cost-aware and cost-ignorant strategies, even after excluding microcap stocks. Other characteristics related to prominent factors, such as value and momentum, also maintain their importance when moving to the more realistic setting. The performance reduction from omitting 6-month momentum, sales to price, and R&D to market-cap is similar in magnitude across all three settings. However, in realistic settings — where returns are lower due to excluding microcaps and accounting for trading costs — these characteristics become more important relative to other factors, as they represent a larger fraction of the strategy's total performance.

[22]Unlike in the main analysis of strategy performance, we do not eliminate the factor hindsight. This allows comparing the impact of each feature on out of sample returns over the common period, which would not be possible if the set of predictors changed over time.

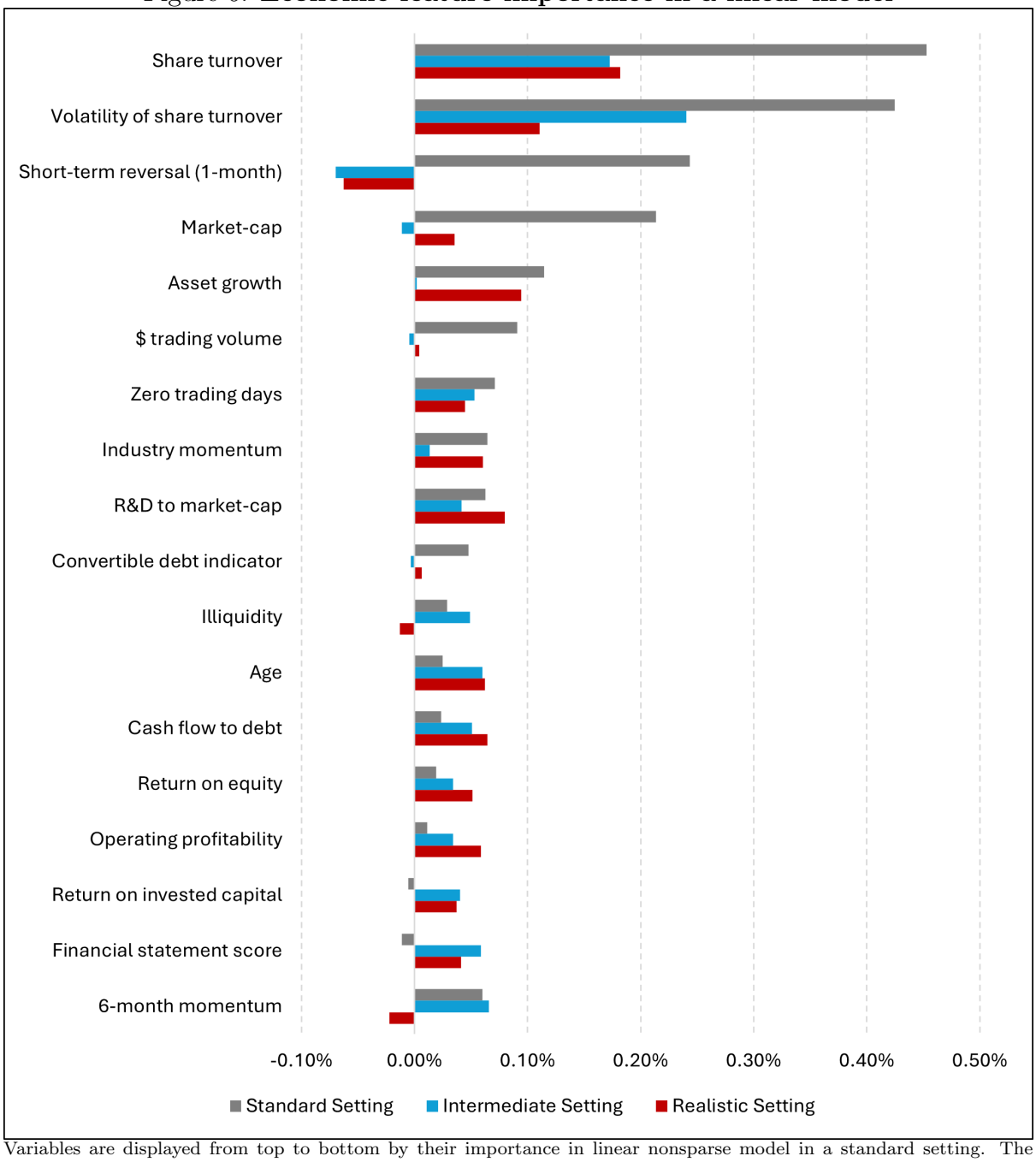
Figure 5: **Economic feature importance in a nonlinear model**



Variables are displayed from top to bottom by their importance in nonlinear nonsparse model in a standard setting. The economic feature importance is measured by reduction in long-short strategy based on return predictions when values of a given feature is set to cross-sectional median. Displayed features rank among ten most important ones in at least one setting.

The results are broadly consistent when examining the linear model. As shown in Figure 6, both short-term reversal and market capitalization lose their relevance once microcap stocks are excluded. In contrast, share turnover and its volatility remain among the most influential predictors across all three settings. Overall, we find substantial agreement between linear and nonlinear models regarding the ranking of the most important features.

Figure 6: **Economic feature importance in a linear model**



Variables are displayed from top to bottom by their importance in linear nonsparse model in a standard setting. The economic feature importance is measured by reduction in long-short strategy based on return predictions when values of a given feature is set to cross-sectional median. Displayed features rank among ten most important ones in at least one setting.

4. Do benefits of ML survive in a long-only setting?

To assess the broader applicability of these findings, we examine strategy performance under short-sale constraints - a restriction faced by a significant portion of the investment management industry. Many institutional investors, including pensions funds and endowments, are prohibited from short-selling due to regulatory or mandate restrictions. Additionally, short-selling entails costs beyond our previous analysis, including securities lending fees and margin requirements

that can further diminish returns.

The analysis in this section complements our primary results by evaluating the role of different aspects of machine learning strategies under the constraints faced by a large fraction of market participants.

4.1. What drives the returns of ML-powered long-only strategies?

The long-only portfolios considered in this section simply correspond to the long leg of the strategies discussed in the previous sections.[23] Moreover, we look at the returns relative to the broad market portfolio that includes all stocks in the reference equity universe.[24]

Table 6 shows that, overall, performance gains from both informational and functional complexity turn out to be weaker when short-selling is not used.[25]

In the standard setting, nonlinearity adds substantial value with statistically significant improvements for both sparse and nonsparse models. However, an expanded information set improves performance only in the linear model. Even in this highly stylised setting, simply removing short sales from the strategies cancels the value of an expanded information set when using nonlinear models to predict returns.

In the intermediate setting, strategies based on nonsparse models do generate returns of 45 bps and 48 bps per month for linear and nonlinear models, respectively. However, the incremental value of both nonlinearity and expanded information set is close to zero and statistically insignificant.

In the realistic setting, returns are further diminished due to trading costs. Again, we find no substantial differences between models with varying complexity levels: neither informational nor functional complexity add value on their own. However, the nonlinear nonsparse model, which jointly uses both types of complexity, still generates significant returns of 27 bps per month.

[23] Appendix provides results for value-weighted long-only strategies.

[24] In intermediate and realistic settings, microcaps are excluded both from long-only strategies as well as from the respective benchmark. The broad portfolio equally weights all stocks in a given equity universe.

[25] We report cumulative returns of the strategies from Table 6 in the appendix.

Table 6: **Mean monthly returns of long-only strategies (in excess of the market portfolio, in %)**

	Linear	Nonlinear	Nonlinear minus Linear
Panel A: Standard Setting			
(includes microcaps, with factor hindsight, ignores trading costs)			
Sparse	0.83***	1.61***	0.77***
	(5.52)	(6.48)	(3.90)
Nonsparse	1.34***	1.90***	0.55***
	(7.32)	(5.65)	(2.73)
Nonsparse minus Sparse	0.52***	0.28	1.07***
	(3.00)	(1.57)	(3.28)
Panel B: Intermediate Setting			
(excludes microcaps, without factor hindsight, ignores trading costs)			
Sparse	0.26	0.39**	0.13
	(1.61)	(2.45)	(0.71)
Nonsparse	0.45***	0.48***	0.03
	(3.40)	(3.72)	(0.34)
Nonsparse minus Sparse	0.19	0.09	0.22*
	(1.28)	(0.53)	(1.67)
Panel C: Realistic Setting			
(excludes microcaps, without factor hindsight, accounts for trading costs)			
Sparse	0.08	0.10	0.01
	(0.57)	(0.70)	(0.09)
Nonsparse	0.17	0.27**	0.10
	(1.30)	(2.23)	(0.99)
Nonsparse minus Sparse	0.09	0.17	0.19
	(0.57)	(1.30)	(1.51)

The sample goes from June 1993 to December 2021. The table reports mean monthly returns in excess of the market portfolio (t-stats). The last row reports differences between nonlinear and linear models, and columns report differences between nonsparse and sparse predictor sets. The row-column intersection reports the mean return difference between strategies based on nonlinear nonsparse and the linear sparse models. Strategies correspond to equally weighted long-only portfolios that select stocks (10%) based on monthly return predictions of linear (ridge) and nonlinear (neural network) models estimated on both sparse and nonsparse predictor sets. Each month, microcaps are defined as stocks outside the largest 3000 by market capitalization. The sparse set of predictors include characteristics from the Fama-French 6-factor model: market beta, market cap, book-to-market ratio, return momentum (12-1), operating profitability, and asset growth. The nonsparse set includes 94 factors from Gu, Kelly, and Xiu (2020). We remove factor hindsight by only including the characteristics that were publicly known at a given point in time. T-statistics are based on Newey-West adjusted standard errors with the lag of 4. Statistical significance at the 10%, 5% and 1% levels are indicated by *, ** and ***, respectively.

The spanning regressions in Table 7 confirm that neither an expanded information set nor nonlinearity add consistent value without short selling, even in the more stylised settings. After

accounting for common variation in strategy returns, the incremental contribution of each component is, in most cases, statistically indistinguishable from zero. Thus, studies on the benefits of machine learning strategies that ignore short-selling constraints likely overstate the benefits of machine learning complexity.

Table 7: **Alphas from spanning regressions of long-only strategies (in %)**

	Lin.	**Nonlin.**	**Sparse**	**Nonsparse**	
Panel A: Standard Setting					
(includes microcaps, with factor hindsight, ignores trading costs)					
Nonsparse over sparse	0.23 (1.52)	-0.11 (-0.75)			
Nonlinear over linear			0.42** (2.51)	0.07 (0.39)	
Nonlinear nonsparse over linear sparse					0.42 (1.59)
Panel B: Intermediate Setting					
(excludes microcaps, without factor hindsight, ignores trading costs)					
Nonsparse over sparse	-0.01 (-0.10)	0.02 (0.20)			
Nonlinear over linear			0.16 (1.60)	0.20 (1.61)	
Nonlinear nonsparse over linear sparse					0.14 (1.24)
Panel C: Realistic Setting					
(excludes microcaps, without factor hindsight, accounts for trading costs)					
Nonsparse over sparse	0.08 (0.69)	-0.03 (-0.28)			
Nonlinear over linear			0.21** (2.19)	0.13 (0.98)	
Nonlinear nonsparse over linear sparse					0.14 (1.35)

The sample goes from June 1993 to December 2021. The table reports monthly alphas (t-stats) from regressions of one strategy's excess returns to those of another, with a constant included. Strategies correspond to equally weighted long-only portfolios that select stocks (10%) based on monthly return predictions of linear (ridge) and nonlinear (neural network) models estimated on both sparse and nonsparse predictor sets. Each month, microcaps are defined as stocks outside the largest 3000 by market capitalization. The sparse set of predictors include characteristics from the Fama-French 6-factor model: market beta, market cap, book-to-market ratio, return momentum (12-1), operating profitability, and asset growth. The nonsparse set includes 94 factors from Gu, Kelly, and Xiu (2020). We remove factor hindsight by only including the characteristics that were publicly known at a given point in time. T-statistics are based on Newey-West adjusted standard errors with the lag of 4. Statistical significance at the 10%, 5% and 1% levels are indicated by *, ** and ***, respectively.

4.2. Investability of long-only strategies

Table 8 presents key investability metrics for long-only strategies, revealing patterns consistent with those observed for long-short strategies in Table 5. The key driver of higher turnover and higher implementation frictions is the expanded information set rather than nonlinearity. Nonsparse models generate substantially more trading and longer execution horizons than sparse models. By contrast, switching from linear to nonlinear within a given predictor set has a small incremental effect, and within the nonsparse set it is often negligible. This pattern holds for turnover, trading costs, and extreme days-to-trade.

Most notably, cost-aware strategies exhibit a proportionally larger reduction in trading costs (two-thirds to three-quarters) relative to their decrease in portfolio turnover (about one third) compared to cost-ignorant strategies. The decline in extreme days-to-trade is of similar or greater magnitude. This suggests that cost-aware algorithms strategically avoid trades in expensive-to-trade stocks.

Table 8: **Investability of long-only strategies**

	Cost-ignorant strategies			Cost-aware strategies			Percent reduction Cost-aware vs. Cost-ignorant	
	Lin.	Nonlin.	Nonlin. - Lin.	Lin.	Nonlin.	Nonlin - Lin	Lin.	Nonlin. - Lin.
Panel A: Monthly Turnover (in %)								
Sparse	17.4	27.6	10.2	9.9	17.1	7.2	43	38
Nonsparse	69.2	67.6	-1.6	49.4	45.8	-3.6	29	32
Nonsparse minus Sparse	51.8	40.1	50.2	39.5	28.8	35.9	-	-
Panel B: Monthly Trading Costs (in basis points)								
Sparse	23	25	3	5	6	1	78	76
Nonsparse	64	60	-3	20	16	-4	68	73
Nonsparse minus Sparse	41	35	38	15	10	11	-	-
Panel C: Extreme Days-to-Trade (95^{th} percentile)								
Sparse	3.8	3.6	-0.2	0.7	0.5	-0.3	80	87
Nonsparse	14.7	16.7	2.0	4.4	6.2	1.8	70	63
Nonsparse minus Sparse	10.9	13.1	12.9	3.6	5.7	5.4	-	-

The table reports time-series mean of monthly one-way turnover, transaction cost, and the days to trade over the sample from July 1993 to December 2021. The days to trade each month is computed as a 95^{th} percentile of stock-level measures of days to trade, which is how many days it takes a given stock to trade assuming no more than 10% of the average daily dollar traded volume (ADDTV) is traded. ADDTV is measured as a daily average within given month. We assume the notional amount of 1 billion USD as an initial investment in each strategy. The notional amount is deflated based on returns of CRSP value-weighted market index and is equal to 53.2 million USD as of June 30, 1993. The strategies correspond to equally weighted long-only portfolios that select stocks (10%) based on monthly return predictions of linear (RIDGE) and nonlinear models (Neural Network) with corresponding set of predictors. Two versions of strategies are considered: ones that ignore transaction costs, and ones that consider transaction costs during the stock selection. The strategies are built on investable equity universe, excluding all stocks beyond the largest 3000 stocks at each point in time. Sparse set of predictors include characteristics from Fama-French 6-factor model: market beta, log of market-cap, book-to-market ratio, return momentum (12-1), operating profitability, and asset growth. Nonsparse set includes 94 factors from Gu, Kelly, and Xiu (2020) after adjusting for hindsight by removing factors before they were published.

Comparing results in Table 8 with Table 5 also reveals asymmetries between long and short legs. While turnover is roughly halved for long-only strategies, the short leg contributes disproportionately to trading costs in cost-aware strategies based on nonlinear models. Specifically, trading costs of 0.15% and 0.41% for nonlinear cost-aware strategies are reduced to 0.06% and 0.16% when long-only constraints are applied, implying that trading costs from the short leg equal to 0.09% and 0.25%, which are approximately 50% higher than those of the long leg.

5. Conclusion

We analyse equity strategies formed from return predictions of machine learning factor models, focusing on the role of functional and informational complexity. Our results generate important findings on four issues: (1) the relative importance of nonlinearity and expanded information set in ML models, (2) the economic importance of features across different implementation settings, (3) the effectiveness of transaction cost management strategies, and (4) the impact of short-sale constraints.

First, on the role of functional versus information complexity, we find that which ingredients matter is strongly dependent on the setting used to evaluate strategies. In a standard setting using gross returns across the full equity universe, both nonlinearity and expanded information set yield similar levels of added value. However, this conclusion does not hold in more realistic settings. When we exclude microcaps and remove factor hindsight, functional complexity ceases to show meaningful performance gains when using sparse information sets. Informational complexity proves more robust, retaining value even as we move to realistic settings, though this value is substantially diminished. Importantly, combining nonlinearity with a broader information set still delivers meaningful gains in realistic settings, suggesting that the joint application of these ingredients to machine learning models remains valuable even under practical constraints.

Second, we provide a measure of economic feature importance that captures which characteristics contribute most to out of sample returns. Here too, the implementation setting fundamentally alters which predictors drive ML strategy performance. Size and short-term reversal, two dominant features in the standard setting, lose their importance once microcaps are excluded and transaction costs are considered. This finding highlights how results from highly stylised settings may provide misleading guidance about which firm characteristics matter for investment performance.

Third, our analysis reveals interesting findings on transaction cost management. We find that cost-aware strategies using ML signals achieve reductions in transaction costs that far exceed the reduction in turnover. This suggests that approaches that directly target a reduction in turnover are not suitable for implementing ML return predictions. In contrast, cost-aware strategies only reduce unproductive turnover, rather than any type of turnover.

Fourth, we find that the benefits of using ML-based return signals differ markedly between long-short and long-only strategies. Under short-sale constraints and in a realistic setting, the strategy based on ML signals still delivers meaningful outperformance over a market benchmark, but gains are markedly muted compared to the long-short case.

Overall, the benefits of machine learning signals are not as ubiquitous as advertised. The benefits of allowing for nonlinear relations are substantially reduced when leaving the highly stylised settings that are standard in the literature. While expanded information sets prove more robust, even their benefits diminish considerably under realistic constraints, especially when short-selling is not allowed. Careful use of these signals, particularly the combination of expanded information sets with nonlinearity, can still deliver meaningful improvements in

achievable performance. Investors need to evaluate these techniques under their specific implementation constraints rather than assume that more complex models will universally outperform.

References

Avramov, D., S. Cheng, and L. Metzker (2023). "Machine learning vs. economic restrictions: Evidence from stock return predictability". In: *Management Science* 69.5, pp. 2587–2619.

Azevedo, V., C. Hoegner, and M. Velikov (2024). "The Expected Returns on Machine-Learning Strategies". Working paper.

Carhart, M. M. (1997). "On persistence in mutual fund performance". In: *The Journal of finance* 52.1, pp. 57–82.

Chen, L., M. Pelger, and J. Zhu (2024). "Deep learning in asset pricing". In: *Management Science* 70.2, pp. 714–750.

Chen, M., M. Hanauer, and T. Kalsbach (2024). "Design choices, machine learning, and the cross-section of stock returns". Working paper.

Chung, K. H. and H. Zhang (2014). "A simple approximation of intraday spreads using daily data". In: *Journal of Financial Markets* 17, pp. 94–120.

DeMiguel, V., A. Martin-Utrera, F. J. Nogales, and R. Uppal (2020). "A transaction-cost perspective on the multitude of firm characteristics". In: *The Review of Financial Studies* 33.5, pp. 2180–2222.

Didisheim, A., S. B. Ke, B. T. Kelly, and S. Malamud (2024). *Apt or "aipt"? the surprising dominance of large factor models*. Tech. rep. National Bureau of Economic Research.

Dimopoulos, Y., P. Bourret, and S. Lek (1995). "Use of some sensitivity criteria for choosing networks with good generalization ability". In: *Neural Processing Letters* 2, pp. 1–4.

Fama, E. F. and K. R. French (2008). "Dissecting anomalies". In: *The journal of finance* 63.4, pp. 1653–1678.

— (1993). "Common risk factors in the returns on stocks and bonds". In: *Journal of Financial Economics* 33.1, pp. 3–56.

— (2015). "A five-factor asset pricing model". In: *Journal of Financial Economics* 116.1, pp. 1–22.

— (2018). "Choosing Factors". In: *Journal of Financial Economics* 128.2, pp. 234–252.

Gu, S., B. Kelly, and D. Xiu (2020). "Empirical Asset Pricing via Machine Learning". In: *The Review of Financial Studies* 33.5, pp. 2223–2273.

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

Jensen, T. I., B. T. Kelly, S. Malamud, and L. H. Pedersen (2024). "Machine Learning and the Implementable Efficient Frontier". In: *Review of Financial Studies*. Forthcoming.

Kelly, B., S. Malamud, and K. Zhou (2024). "The virtue of complexity in return prediction". In: *The Journal of Finance* 79.1, pp. 459–503.

Kozak, S., S. Nagel, and S. Santosh (2020). "Shrinking the cross-section". In: *Journal of Financial Economics* 135.2, pp. 271–292.

Li, B., A. Rossi, X. Yan, and L. Zhen (2023). "Real-time Machine Learning in the Cross-Section of Stock Returns". In: *Journal of Financial Economics*. Conditionally accepted.

Nagel, S. (2021). *Machine Learning in Asset Pricing*. Princeton University Press.

Novy-Marx, R. and M. Velikov (2016). "A taxonomy of anomalies and their trading costs". In: *The Review of Financial Studies* 29.1, pp. 104–147.

Raskutti, G., M. J. Wainwright, and B. Yu (2014). "Early stopping and non-parametric regression: an optimal data-dependent stopping rule". In: *The Journal of Machine Learning Research* 15.1, pp. 335–366.

Welch, I. and A. Goyal (2008). "A comprehensive look at the empirical performance of equity premium prediction". In: *The Review of Financial Studies* 21.4, pp. 1455–1508.

Appendix A: Parameters of neural network

Table A1: Parameters of neural network

Parameter	**Values**
Number of nodes in hidden layers	{32, 16, 8}
Activation function	ReLU
Loss function	MSE
Optimization algorithm	Adam
Maximum number of epochs	100
Early stopping patience	5
Batch normalization	TRUE
Batch size	10,000
Number of initializations	10
Learning rate	{0.001; 0.01}
L1 penalty	[1e-5; 1e-3]
Number of trials (tuning)	20
Number of initializations (tuning)	3

Appendix B: Performance of decile-sorted portfolios

Table B1: Monthly gross performance of equal-weighted decile portfolios

	Linear						Nonlinear					
	Sparse			Nonsparse			Sparse			Nonsparse		
	mean	std	S.R.	mean	std	S.R.	mean	std	S.R.	mean	std	S.R.
Panel A: Cost-ignorant strategies, including microcaps, with hindsight												
D1	0.55	7.55	0.17	-0.12	6.40	-0.17	0.42	8.96	0.09	-0.56	8.41	-0.31
D2	0.86	6.37	0.37	0.52	5.84	0.20	0.73	7.06	0.27	0.53	6.45	0.19
D3	0.99	6.19	0.45	0.77	5.71	0.36	0.91	6.22	0.41	0.80	5.55	0.39
D4	0.99	6.04	0.46	0.92	5.60	0.45	0.95	5.91	0.45	0.92	5.09	0.50
D5	1.06	5.95	0.51	1.05	5.54	0.54	0.99	5.55	0.50	1.12	5.14	0.63
D6	1.16	5.80	0.58	1.21	5.61	0.63	1.00	5.15	0.55	1.24	5.06	0.72
D7	1.30	5.79	0.67	1.39	5.89	0.71	1.16	4.98	0.68	1.40	5.25	0.80
D8	1.29	5.73	0.67	1.53	6.35	0.73	1.37	5.20	0.79	1.51	5.42	0.85
D9	1.53	5.76	0.81	1.95	6.75	0.91	1.41	5.33	0.80	1.69	6.06	0.86
D10	2.00	6.19	1.02	2.52	7.76	1.04	2.79	8.15	1.11	3.07	9.85	1.01
D10-D1	1.45	4.96	1.01	2.64	4.99	1.83	2.36	5.14	1.59	3.63	5.72	2.20
Panel B: Cost-ignorant strategies, excluding microcaps, without hindsight												
D1	0.99	6.07	0.46	0.45	6.96	0.13	0.66	9.09	0.18	0.09	8.64	-0.04
D2	1.06	6.17	0.49	0.85	6.04	0.38	0.90	6.72	0.37	0.73	6.90	0.27
D3	0.93	6.05	0.42	0.88	5.60	0.43	1.00	5.99	0.47	0.93	6.02	0.43
D4	0.97	6.03	0.45	1.00	5.58	0.51	0.99	5.49	0.51	1.03	5.53	0.53
D5	0.99	6.05	0.46	1.10	5.47	0.58	1.01	5.62	0.51	1.14	5.32	0.62
D6	1.09	5.92	0.53	1.14	5.46	0.61	0.97	5.79	0.47	1.16	5.23	0.65
D7	1.05	5.83	0.52	1.17	5.57	0.61	1.05	5.33	0.56	1.25	5.21	0.71
D8	1.11	5.70	0.56	1.27	5.99	0.63	1.28	5.53	0.69	1.31	5.23	0.75
D9	1.09	5.63	0.56	1.21	6.35	0.56	1.28	5.54	0.69	1.40	5.47	0.77
D10	1.32	6.00	0.65	1.51	6.98	0.66	1.45	6.17	0.71	1.54	6.41	0.73
D10-D1	0.33	4.22	0.27	1.06	5.02	0.73	0.79	6.77	0.40	1.45	5.24	0.96
Panel C: Cost-aware strategies, excluding microcaps, without hindsight												
D1	1.10	5.81	0.55	0.61	6.88	0.22	0.85	8.58	0.27	0.22	8.70	0.01
D10	1.12	5.90	0.55	1.36	6.81	0.60	1.14	6.11	0.54	1.42	6.10	0.70
D10-D1	0.01	3.78	0.01	0.74	4.88	0.53	0.29	5.98	0.17	1.20	5.18	0.80

Table B2: Monthly net performance of equally weighted decile portfolios

	Linear						Nonlinear					
	Sparse			Nonsparse			Sparse			Nonsparse		
	mean	std	S.R.	mean	std	S.R.	mean	std	S.R.	mean	std	S.R.
Panel A: Cost-ignorant strategies, including microcaps, with hindsight												
D1	0.27	7.58	0.04	-1.37	6.72	-0.80	0.04	8.99	-0.05	-2.31	8.77	-0.99
D2	0.37	6.42	0.10	-0.94	6.12	-0.63	0.06	7.11	-0.06	-0.92	6.67	-0.57
D3	0.36	6.24	0.10	-0.78	5.97	-0.56	0.14	6.28	-0.02	-0.54	5.72	-0.44
D4	0.22	6.10	0.02	-0.70	5.83	-0.53	0.13	5.98	-0.03	-0.40	5.26	-0.39
D5	0.15	6.02	-0.02	-0.64	5.78	-0.49	0.08	5.62	-0.06	-0.21	5.29	-0.26
D6	0.07	5.89	-0.06	-0.57	5.83	-0.45	0.00	5.23	-0.12	-0.13	5.23	-0.21
D7	0.10	5.87	-0.05	-0.47	6.07	-0.38	0.10	5.07	-0.06	-0.02	5.41	-0.13
D8	-0.02	5.82	-0.12	-0.44	6.52	-0.33	0.26	5.27	0.05	-0.02	5.60	-0.13
D9	0.17	5.85	-0.01	-0.12	6.89	-0.15	0.25	5.42	0.04	-0.12	6.28	-0.17
D10	1.14	6.23	0.53	0.50	7.77	0.14	1.55	8.12	0.58	0.19	9.88	0.00
D10-D1	0.28	5.05	0.19	-0.60	4.95	-0.42	0.70	5.10	0.48	-0.95	5.87	-0.56
Panel B: Cost-ignorant strategies, excluding microcaps, without hindsight												
D1	0.93	6.07	0.42	-0.06	7.02	-0.12	0.44	9.10	0.10	-0.50	8.67	-0.27
D2	0.94	6.18	0.42	0.26	6.08	0.04	0.58	6.73	0.20	0.12	6.94	-0.03
D3	0.76	6.06	0.33	0.26	5.64	0.04	0.63	6.01	0.26	0.31	6.06	0.07
D4	0.77	6.04	0.34	0.35	5.62	0.10	0.59	5.51	0.26	0.41	5.58	0.14
D5	0.78	6.05	0.34	0.44	5.52	0.16	0.60	5.65	0.26	0.51	5.35	0.21
D6	0.85	5.92	0.39	0.45	5.50	0.17	0.53	5.82	0.21	0.53	5.27	0.23
D7	0.79	5.83	0.36	0.47	5.61	0.18	0.61	5.36	0.27	0.62	5.24	0.29
D8	0.81	5.71	0.38	0.56	6.03	0.21	0.86	5.54	0.42	0.69	5.26	0.33
D9	0.82	5.64	0.39	0.50	6.38	0.17	0.89	5.55	0.44	0.78	5.50	0.38
D10	1.09	6.01	0.52	0.87	6.97	0.34	1.19	6.17	0.57	0.94	6.41	0.41
D10-D1	0.03	4.23	0.03	-0.09	5.09	-0.06	0.29	6.82	0.15	0.25	5.37	0.16
Panel C: Cost-aware strategies, excluding microcaps, without hindsight												
D1	1.08	5.81	0.53	0.44	6.89	0.13	0.77	8.58	0.24	-0.02	8.69	-0.08
D10	1.07	5.90	0.52	1.16	6.80	0.49	1.08	6.11	0.51	1.26	6.09	0.61
D10-D1	-0.07	3.78	-0.06	0.36	4.87	0.25	0.14	6.00	0.08	0.79	5.22	0.52

Table B3: Monthly gross performance of value-weighted decile portfolios

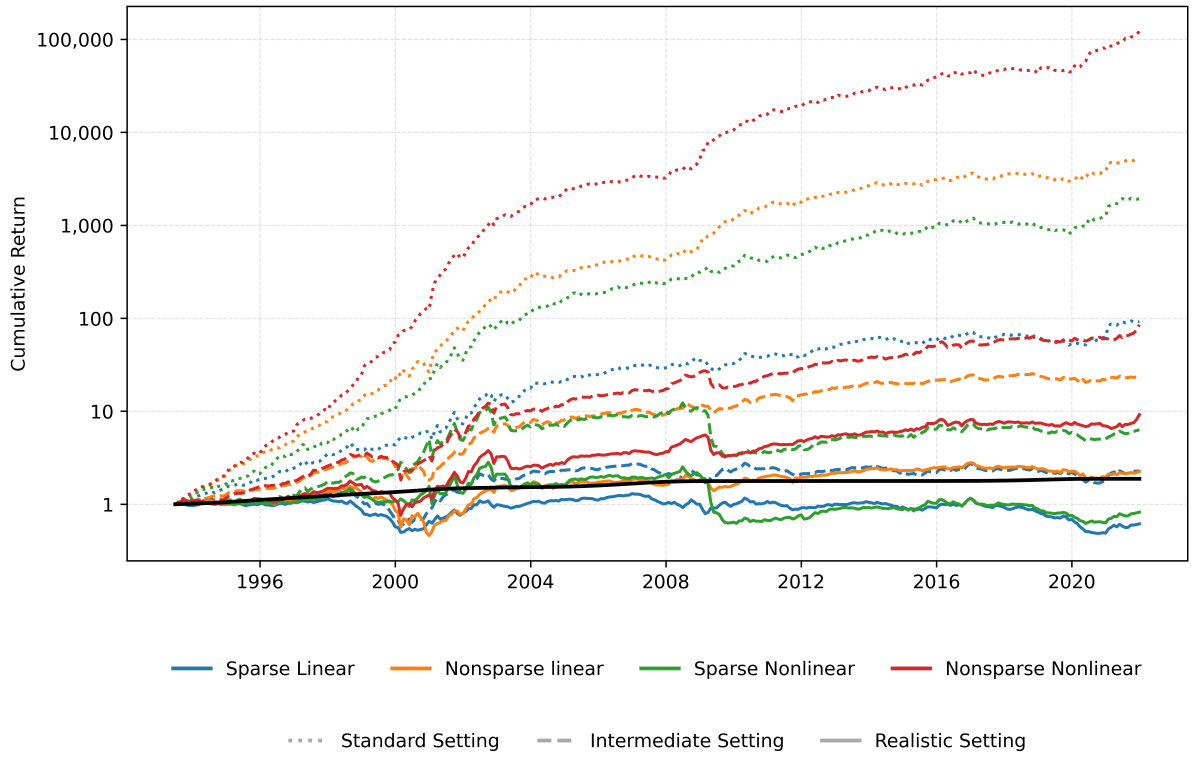
	Linear						Nonlinear					
	Sparse			Nonsparse			Sparse			Nonsparse		
	mean	std	S.R.	mean	std	S.R.	mean	std	S.R.	mean	std	S.R.
Panel A: Cost-ignorant strategies, including microcaps, with hindsight												
D1	0.87	5.37	0.44	0.58	5.00	0.27	0.52	6.94	0.17	0.29	7.48	0.05
D2	0.96	4.66	0.58	0.87	4.59	0.52	0.89	5.26	0.47	0.70	5.12	0.35
D3	1.05	4.30	0.70	0.98	4.52	0.61	0.84	4.59	0.50	0.79	4.49	0.46
D4	1.00	4.43	0.64	0.95	4.54	0.59	1.00	4.49	0.63	1.00	4.20	0.68
D5	0.91	4.36	0.58	1.12	4.59	0.71	1.07	4.31	0.71	1.19	4.50	0.77
D6	1.10	4.56	0.70	1.09	4.51	0.69	1.07	4.31	0.72	1.05	4.47	0.67
D7	1.13	4.80	0.68	1.05	5.08	0.59	0.96	4.60	0.58	1.16	4.68	0.72
D8	1.11	5.18	0.62	1.34	5.46	0.73	1.31	4.99	0.78	1.32	4.95	0.80
D9	1.17	5.78	0.59	1.25	5.78	0.64	1.29	5.12	0.75	1.41	5.44	0.78
D10	1.57	6.30	0.76	1.38	6.15	0.68	1.61	5.87	0.84	2.17	7.50	0.92
D10-D1	0.70	5.25	0.46	0.80	4.82	0.58	1.08	5.64	0.67	1.88	6.71	0.97
Panel B: Cost-ignorant strategies, excluding microcaps, without hindsight												
D1	1.08	4.70	0.66	0.55	5.22	0.24	0.84	7.49	0.28	0.28	7.16	0.05
D2	1.03	4.46	0.66	0.84	4.83	0.47	0.89	5.49	0.45	0.62	5.28	0.29
D3	0.98	4.43	0.62	1.07	4.44	0.69	0.97	4.75	0.57	0.81	4.65	0.47
D4	0.85	4.50	0.51	0.93	4.41	0.59	1.09	4.45	0.70	0.90	4.43	0.56
D5	1.02	5.33	0.54	1.03	4.68	0.63	0.98	4.69	0.59	1.04	4.39	0.68
D6	0.95	4.97	0.53	1.09	4.51	0.70	1.02	4.73	0.61	1.17	4.50	0.76
D7	0.96	5.12	0.53	1.07	4.84	0.63	0.87	4.64	0.51	1.17	4.41	0.77
D8	1.05	5.06	0.59	1.27	5.44	0.69	1.19	5.23	0.67	1.20	4.61	0.76
D9	1.02	5.65	0.51	1.00	5.62	0.50	1.15	4.99	0.67	1.37	4.94	0.83
D10	1.13	6.07	0.54	1.38	6.23	0.66	1.09	5.82	0.54	1.62	5.78	0.86
D10-D1	0.05	4.85	0.04	0.83	5.20	0.55	0.25	6.70	0.13	1.34	5.63	0.82

Table B4: Monthly net performance of value weighted decile portfolios

	Linear						Nonlinear					
	Sparse			Nonsparse			Sparse			Nonsparse		
	mean	std	S.R.	mean	std	S.R.	mean	std	S.R.	mean	std	S.R.
Panel A: Cost-ignorant strategies, including microcaps, with hindsight												
D1	0.79	5.38	0.39	0.33	5.04	0.10	0.41	6.95	0.11	-0.03	7.52	-0.10
D2	0.84	4.67	0.49	0.59	4.63	0.30	0.73	5.26	0.36	0.44	5.15	0.17
D3	0.89	4.32	0.57	0.68	4.56	0.38	0.65	4.60	0.35	0.53	4.50	0.27
D4	0.82	4.44	0.50	0.63	4.57	0.34	0.78	4.51	0.46	0.72	4.23	0.44
D5	0.70	4.37	0.41	0.80	4.62	0.46	0.84	4.34	0.52	0.88	4.52	0.53
D6	0.87	4.57	0.52	0.73	4.54	0.42	0.86	4.32	0.54	0.71	4.49	0.41
D7	0.87	4.81	0.49	0.68	5.10	0.33	0.73	4.61	0.41	0.77	4.67	0.43
D8	0.77	5.19	0.39	0.93	5.47	0.47	1.06	4.98	0.61	0.86	4.97	0.47
D9	0.74	5.78	0.33	0.81	5.78	0.38	0.95	5.13	0.52	0.79	5.49	0.38
D10	1.16	6.32	0.54	0.88	6.11	0.40	1.06	5.86	0.52	0.85	7.53	0.31
D10-D1	0.21	5.27	0.14	0.05	4.78	0.04	0.39	5.58	0.24	0.24	6.79	0.12
Panel B: Cost-ignorant strategies, excluding microcaps, without hindsight												
D1	1.06	4.70	0.65	0.29	5.26	0.07	0.72	7.50	0.25	-0.02	7.19	-0.10
D2	0.99	4.46	0.62	0.56	4.85	0.27	0.75	5.51	0.36	0.35	5.31	0.11
D3	0.92	4.44	0.57	0.79	4.47	0.47	0.80	4.76	0.45	0.55	4.67	0.27
D4	0.79	4.51	0.46	0.64	4.44	0.35	0.89	4.47	0.54	0.63	4.44	0.35
D5	0.94	5.34	0.49	0.73	4.71	0.40	0.77	4.71	0.43	0.77	4.41	0.46
D6	0.85	4.98	0.46	0.78	4.54	0.46	0.79	4.76	0.44	0.87	4.52	0.53
D7	0.86	5.13	0.45	0.75	4.86	0.40	0.65	4.66	0.34	0.87	4.43	0.54
D8	0.93	5.07	0.51	0.93	5.46	0.47	0.97	5.23	0.52	0.90	4.63	0.53
D9	0.90	5.65	0.44	0.65	5.63	0.28	0.93	5.00	0.51	1.06	4.95	0.61
D10	0.99	6.08	0.46	1.00	6.21	0.45	0.93	5.82	0.45	1.27	5.78	0.65
D10-D1	-0.12	4.86	-0.08	0.20	5.15	0.13	-0.03	6.71	-0.02	0.70	5.60	0.43

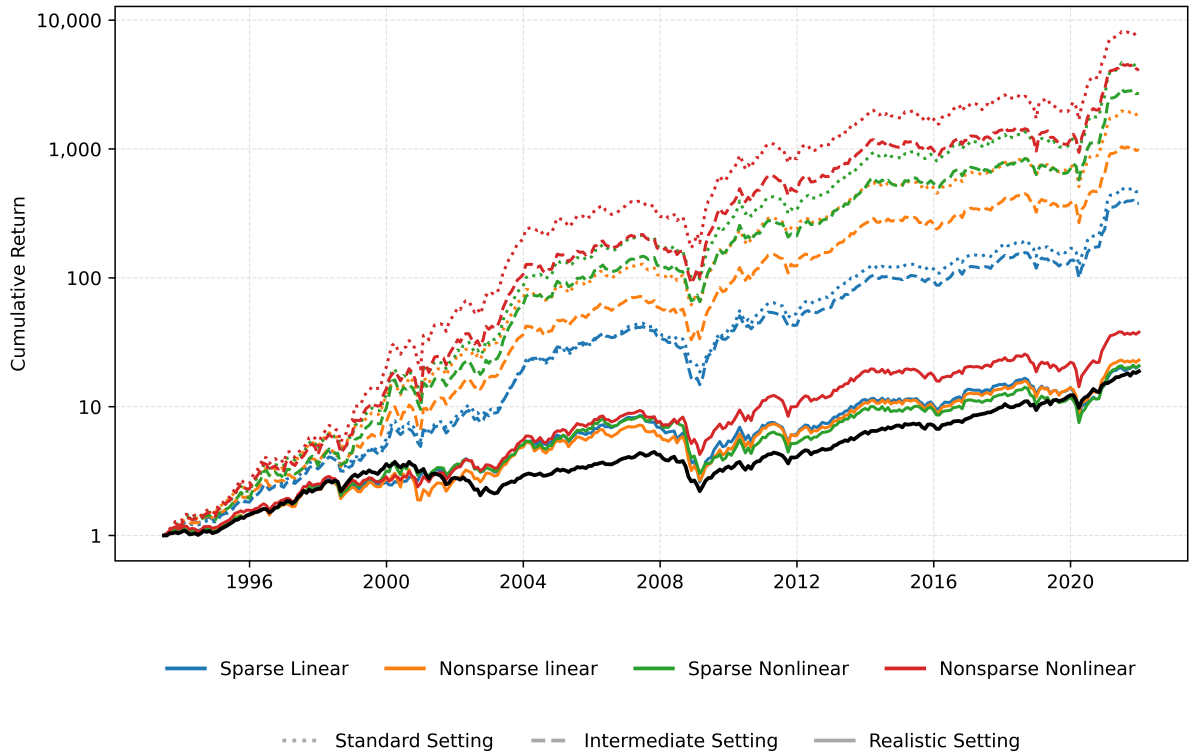
Appendix C: Cumulative Returns of ML-powered strategies

Figure C1: Cumulative returns of L/S strategies across different settings



The charts shows cumulative returns of long-short strategies based on different predictive models across different settings. The solid black line is the cumulative return on 1-month T-bill.

Figure C2: Cumulative returns of L/O strategies across different settings



The charts shows cumulative returns of long-only strategies based on different predictive models across different settings. The solid black line is the cumulative return on CRSP value-weighted market portfolio (gross).

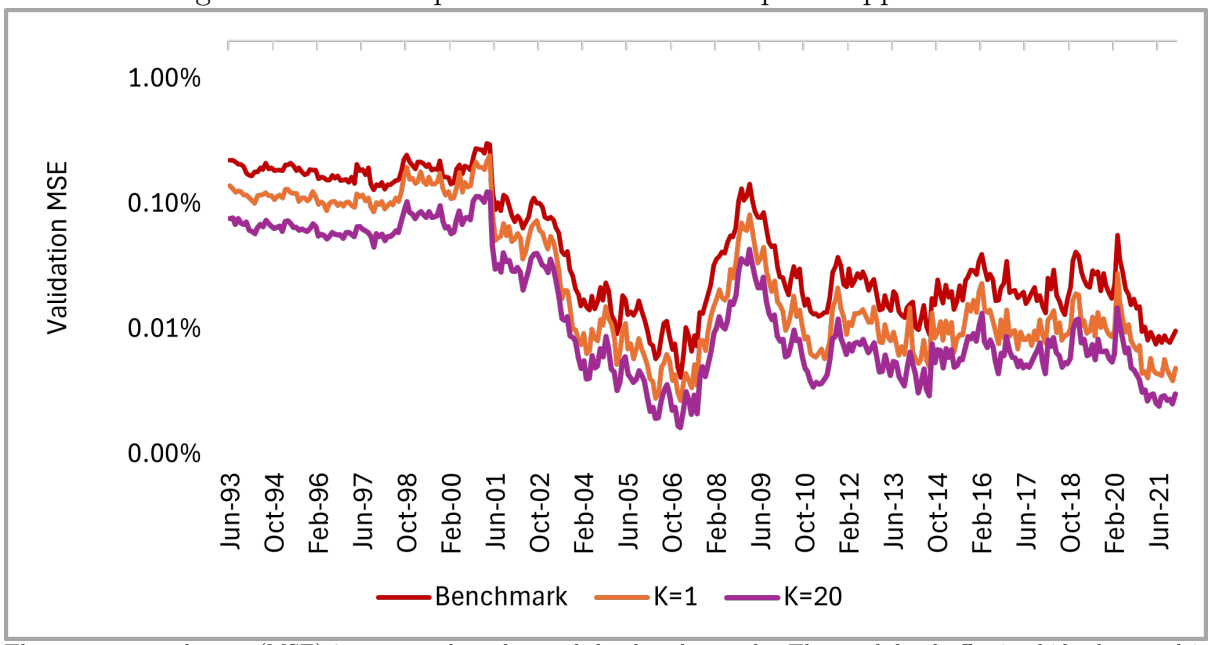
Appendix D: Dealing with missing data for effective spreads

In some cases, the effective spread cannot be estimated due to missing data for closing bid and ask prices. The common approach is to replace missing values with the closest neighbor based on Euclidean distance, in terms of market capitalization and idiosyncratic volatility (see Novy-Marx and Velikov (2016)). Instead, we replace missing spreads with the average of 20 closest neighbors, where the distance is measured by ranks of market cap, idiosyncratic volatility and volume.

We find that using the average of multiple closest neighbors leads to more accurate approximation of stock-level effective spreads. We also find that adding trading volume as additional dimension further improves the accuracy of approximation. We replace spreads with the median of market-capitalization decile when data for either volume or idiosyncratic volatility is not available. However, such cases are rare, and most replacements are based on all three characteristics.

Figure D1 displays the mean squared error when the effective spread is approximated using a K-nearest-neighbor approach, with the ranks of market capitalization, idiosyncratic volatility (from the Fama-French three-factor model), and dollar trading volume as features. The benchmark is the approach of Novy-Marx and Velikov (2016), which uses only two features (excluding trading volume) and sets $K = 1$.

Figure D1: Mean squared error of bid-ask spread approximation

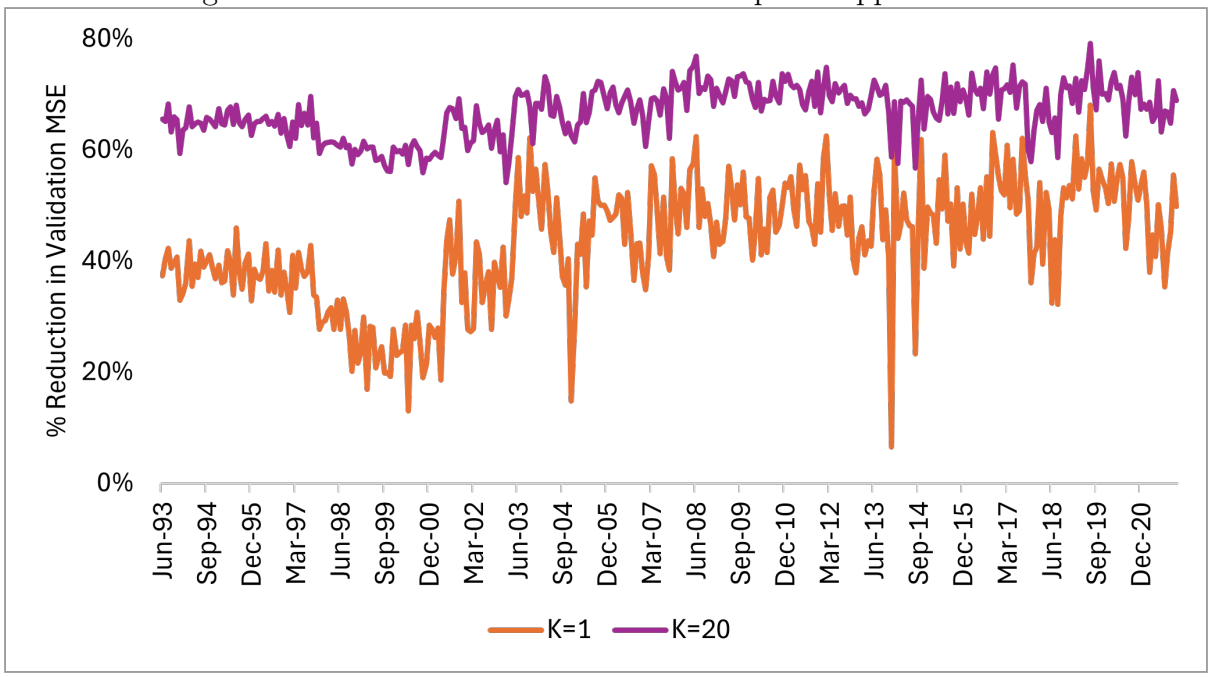


The mean squared error (MSE) is computed at the stock level each month. The stock-level effective bid-ask spread is approximated using three approaches. The benchmark follows Novy-Marx and Velikov (2016). The other two methods, $K = 1$ and $K = 20$, use a K-nearest-neighbour approach based on Euclidean distance, with ranks of market capitalization, idiosyncratic volatility (from the Fama-French three-factor model), and dollar trading volume as features. The MSE is computed based on all stocks for which an effective bid-ask spread estimate is available from daily closing prices.

Figure D2 displays the reduction in mean squared error relative to the benchmark. Results

suggest that trading volume provides useful information for the effective spreads, while using the average of the multiple "similar" stocks further improves the approximation relative to using the single stock.

Figure D2: Reduction in MSE of bid-ask spread approximation

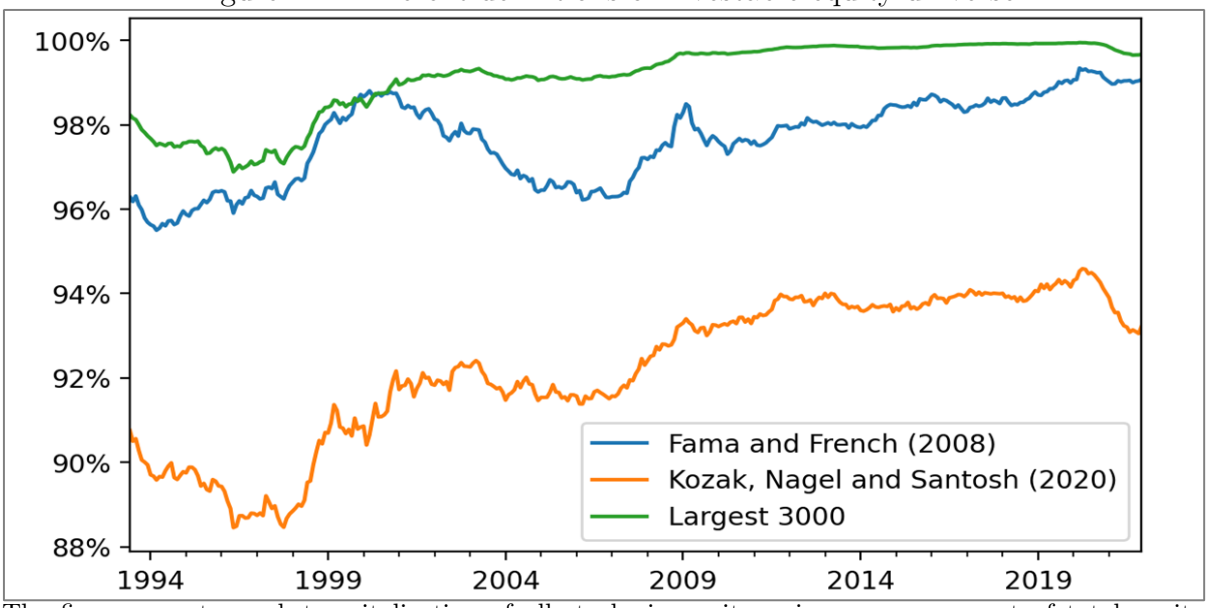


The percentage reduction in mean squared error (MSE) is computed relative to the benchmark (Novy-Marx and Velikov (2016)).

Appendix E: Different ways of defining microcaps

Figure E1 shows the market capitalization of equity universe as a percent of aggregate US equity market after excluding microcaps, where microcaps are defined following Fama and French (2008), Kozak, Nagel, and Santosh (2020), or simply as all stocks beyond the largest 3000 stocks in each US.

Figure E1: Different definitions of investable equity universe



The figure reports market-capitalization of all stocks in equity universe as a percent of total equity universe after excluding microcaps following Fama and French (2008), Kozak, Nagel, and Santosh (2020), or beyond largest 3000 stocks. We only include common shares from CRSP. Fama and French (2008) omits stocks with market capitalization below 20th percentile of NYSE stocks. Kozak, Nagel, and Santosh (2020) omit stocks with market capitalization less than 0.1% of the aggregate (all stocks) market capitalization. The aggregate equity universe is limited to common shares (share code 10 and 11) on major exchanges (NYSE, AMEX, NASDAQ).

Appendix F: In-sample statistical feature importance based on MSPD

Another commonly used measure of feature importance is the squared partial differential of Dimopoulos, Bourret, and Lek (1995). For the k^{th} feature from the feature set X, the mean squared partial differential is defined as follows

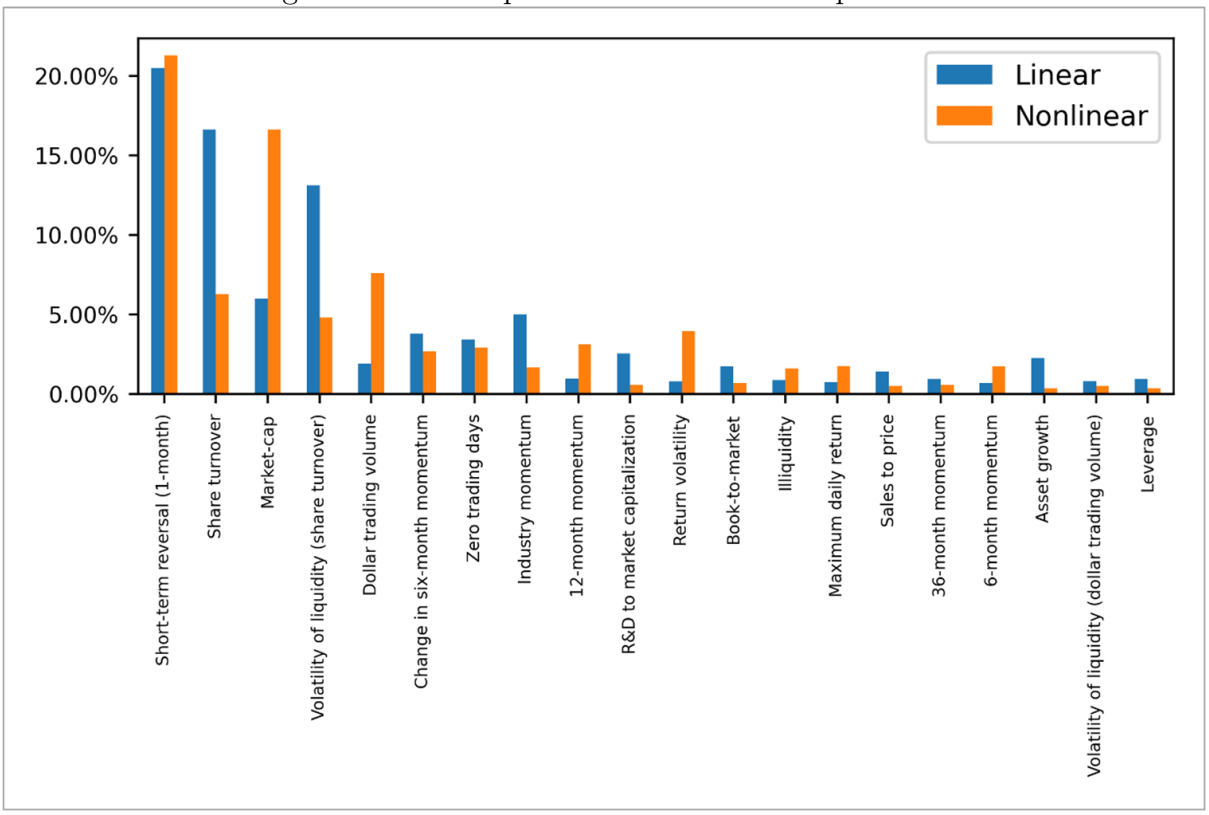
$$MSPD_k = \frac{1}{N\,T}\sum_{i=t}^{T}\sum_{i=1}^{N}\left(\frac{\delta\, f(X_{it},\,\theta)}{\delta\, X_{it}^{k}}\right)^2$$

MSPD captures the average local sensitivity of the model output with respect to each feature. After computing the MSPD for each year from 1993 to 2021, we average these values over time and rescale them so that their total sums to one. Figure F1 presents the top 20 features for both linear and nonlinear models when using nonsparse factors as inputs.[26]

Both linear and nonlinear models consistently identify short-term reversal as the most influential predictor of monthly returns. The next four most important features are all related to firm liquidity. While both models generally agree on the relevance of these liquidity-related variables, they diverge slightly in emphasis: the nonlinear model assigns greater importance to market capitalization and dollar trading volume, whereas the linear model gives more weight to share turnover and its volatility. Beyond the top five features, momentum-related characteristics emerge as the most significant predictors, whereas firm fundamentals—such as the book-to-market ratio and asset growth—contribute marginally to return predictability.

[26]These models come with factor hindsight. We do not exclude unpublished factors to maintain the comparability across all characteristics.

Figure F1: In-sample statistical feature importance

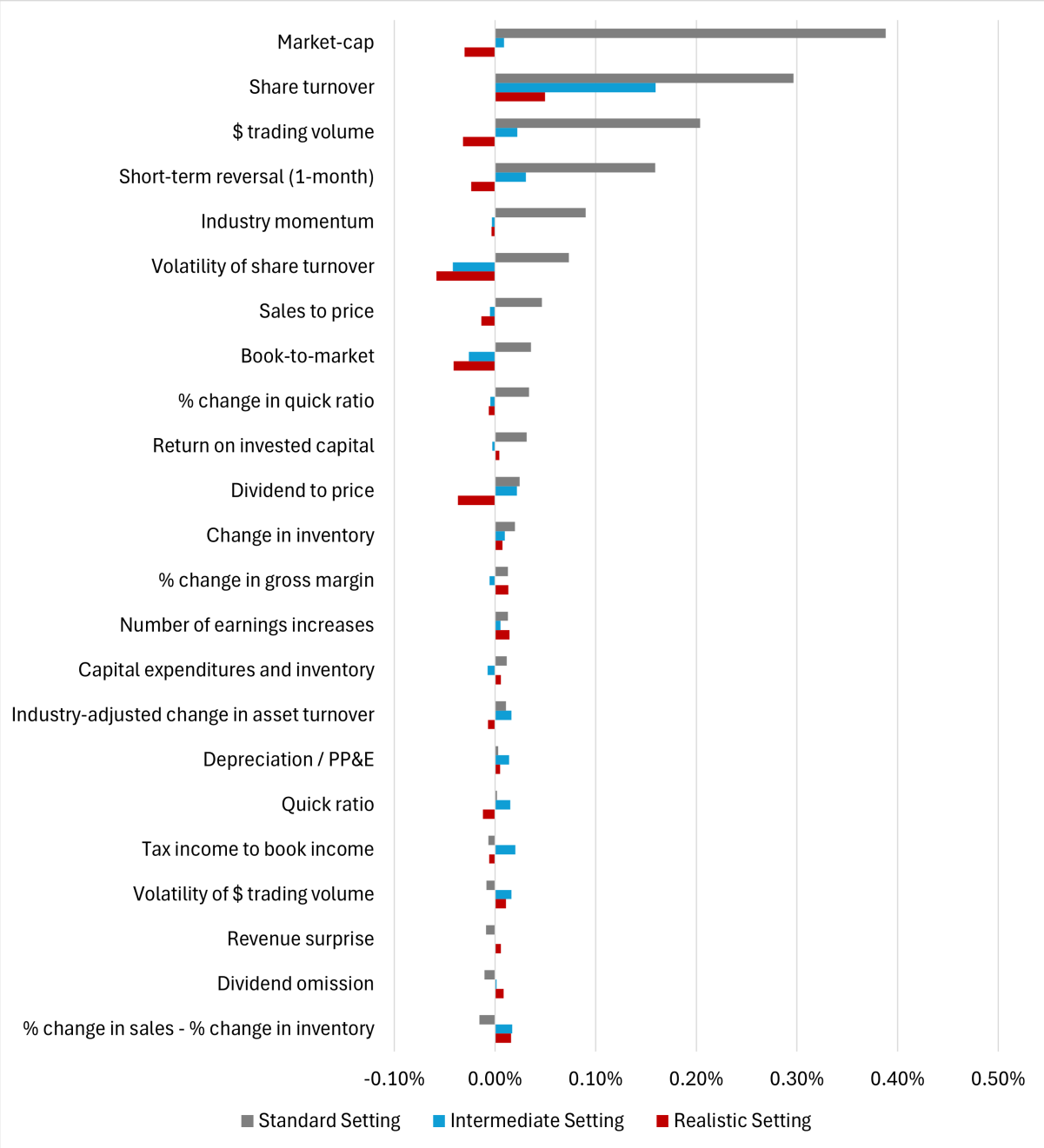


.

Appendix G: Out-of-sample economic feature importance in long-only setting

Figure G1 displays the economic feature importance, measures as a reduction in returns of long-only strategies due to excluding a given characteristic from the nonlinear return model that uses nonsparse set of predictors.

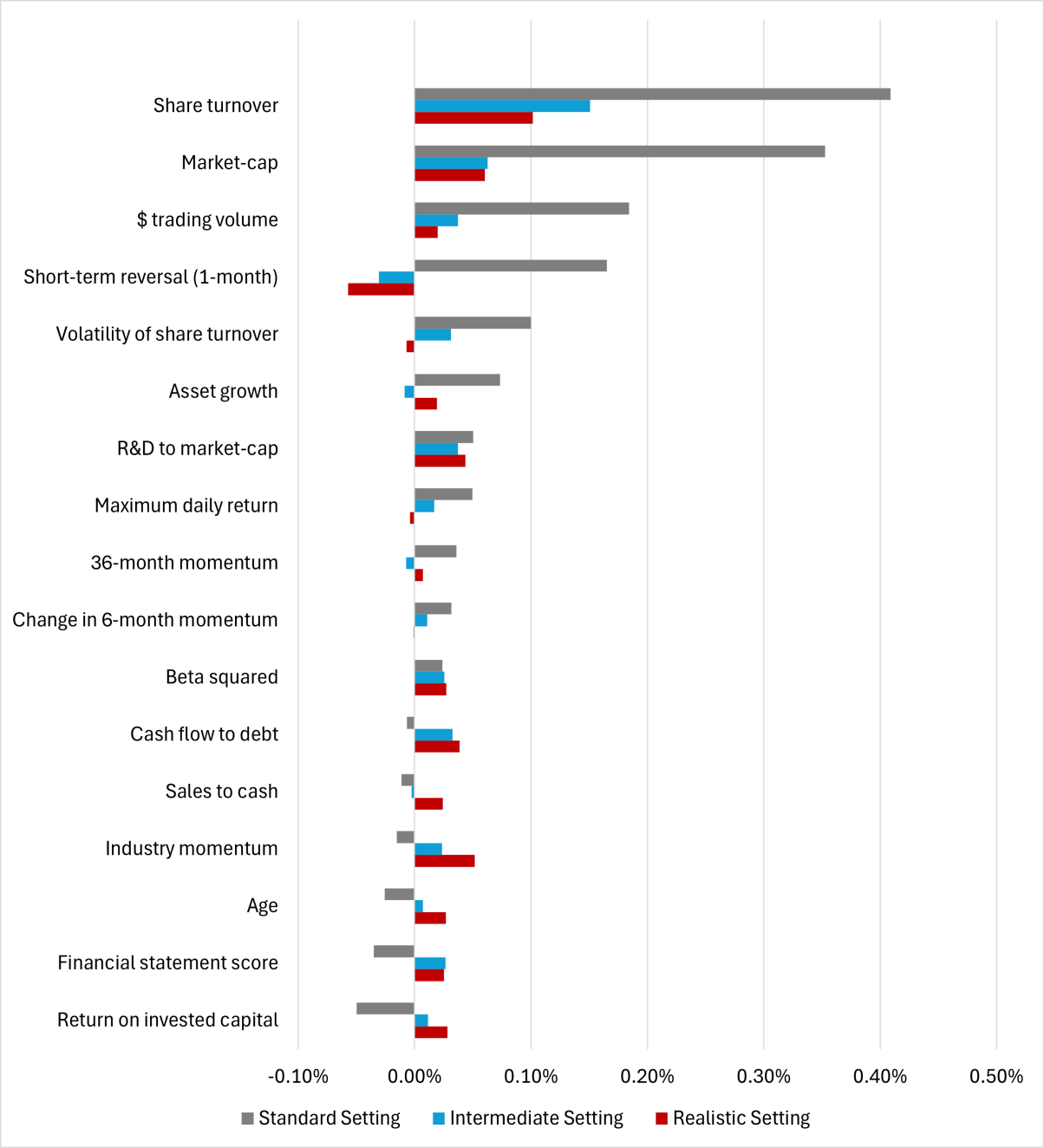
Figure G1: Economic feature importance in nonsparse nonlinear model



.

Figure G2 repeats the analysis for a linear model.

Figure G2: Economic feature importance in nonsparse linear model



.