Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Machine Learning vs. Economic Restrictions: Evidence from Stock Return Predictability

Doron AvramovSi ChengLior Metzker

informs.
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 69, No. 5, May 2023, pp. 2587–2619
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Machine Learning vs. Economic Restrictions: Evidence from Stock Return Predictability

Doron Avramov[a], **Si Cheng**[b,*] **Lior Metzker**[c]

[a] Arison School of Business, Reichman University (IDC Herzliya), Herzliya 4610101, Israel; [b] Whitman School of Management, Syracuse University, Syracuse, New York 13244; [c] School of Business Administration, The Hebrew University of Jerusalem, Jerusalem 9190501, Israel
*Corresponding author

Contact: doron.avramov@idc.ac.il (DA); scheng24@syr.edu, https://orcid.org/0000-0002-2109-7636 (SC); lior.metzker@mail.huji.ac.il (LM)

Received: April 22, 2021
Revised: August 17, 2021
Accepted: September 26, 2021
Published Online in Articles in Advance: June 29, 2022

https://doi.org/10.1287/mnsc.2022.4449

Abstract. This paper shows that investments based on deep learning signals extract profitability from difficult-to-arbitrage stocks and during high limits-to-arbitrage market states. In particular, excluding microcaps, distressed stocks, or episodes of high market volatility considerably attenuates profitability. Machine learning-based performance further deteriorates in the presence of reasonable trading costs because of high turnover and extreme positions in the tangency portfolio implied by the pricing kernel. Despite their opaque nature, machine learning methods successfully identify mispriced stocks consistent with most anomalies. Beyond economic restrictions, deep learning signals are profitable in long positions and recent years and command low downside risk.

History: Accepted by Kay Giesecke, finance.

Funding: D. Avramov acknowledges the Israel Science Foundation (Grant 288/18) for financial support. S. Cheng acknowledges the General Research Fund of the Research Grants Council of Hong Kong [Project 14502318] for financial support.

Supplemental Material: The data files and online appendix are available at https://doi.org/10.1287/mnsc.2022.4449.

Keywords: machine learning • return prediction • neural networks • financial distress • Fintech

1. Introduction

Financial economists have uncovered a plethora of firm characteristics that predict stock returns in the cross-section. However, recent work has challenged the credibility of the evidence on stock return predictability. Harvey et al. (2016) examine 296 published significant factors and conclude that 80 to 158 of them are likely to be false discoveries. Hou et al. (2020) further show that 82% of 452 anomalies become statistically insignificant after excluding microcap stocks and when using value-weighted returns. There is also mounting evidence that anomalies tend to concentrate in distressed stocks and extract most of their profitability from short positions.[1] Notably, it is increasingly difficult to exploit anomalies in recent years because of increased market liquidity and arbitrage activity.

Counter to this "anomaly challenging" strand of literature, there has been an emerging body of work that reports phenomenal investment profitability based on signals generated by machine learning methods.[2] Applying machine learning routines to financial data has been implicitly motivated by the American Finance Association (AFA) presidential address of Cochrane (2011), who suggests that in the presence of a vast collection of noisy and highly correlated return predictors, there is a need for other methods beyond cross-sectional regressions and portfolio sorts. Indeed, machine learning offers a natural way to accommodate a high-dimensional predictor set and flexible functional forms, and it uses "regularization" methods to select models, mitigate overfitting biases, and uncover complex patterns and hidden relationships.

A close look at these two strands of literature, namely, diminishing *individual* anomalies in contrast to the outstanding profitability of machine learning signals that aggregate *multiple* anomalies, suggests that our understanding of the economic significance of machine learning signals is inconclusive. Although vast evidence shows that individual anomalies concentrate in difficult-to-arbitrage stocks and high limits-to-arbitrage market episodes, machine learning methods could indicate a promising direction to identify mispricing when trading frictions attenuate because of their advanced mechanism of aggregating individual anomalies. Thus, a priori, it is impossible to make clear inferences about the economic significance of machine learning signals based on the knowledge we have gained from individual anomalies.

This paper aims to fill this gap by comprehensively examining whether investors can harvest extra profits generated by various machine learning signals that are detailed in the following paragraph. To do so, we first impose several economic restrictions. In the cross-section, we limit the universe of stocks to those that are relatively cheap to trade by excluding microcaps or distressed firms. In the time series, we examine whether investment profitability is more pronounced during high limits-to-arbitrage market states, such as high volatility and low liquidity. We then assess the turnover and the corresponding transaction costs associated with implementing machine learning-based investments and finally explore the economic foundations of trading strategies advocated by seemingly opaque machine learning methods.

We consider a comprehensive set of linear and nonlinear models, study both spread (long-minus-short) and tangency portfolios, and account for both the stochastic discount factor (SDF) and beta pricing formulations.[3] We first implement a neural network with three hidden layers (NN3) as in Gu et al. (2020) (GKX) and then follow Chen et al. (2020) (CPZ) to incorporate a no-arbitrage condition into multiple connected neural networks, including feed-forward networks (FFNs), recurrent neural networks (RNNs) with long short-term memory (LSTM) cells, and a generative adversarial network (GAN). We then analyze two conditional beta pricing models in the machine learning universe, namely, instrumented principal component analysis (IPCA) per Kelly et al. (2019) (KPS) and the conditional autoencoder (CA) extension by Gu et al. (2021).

All these machine learning methods may differ in sample coverage, input variables, and optimization objectives; hence, we do not aim to identify the best method for trading purposes or conduct a rigorous comparison across the models. Instead, we examine whether the machine learning-based investment payoff could still extend to economic significance in the presence of plausible restrictions on the investment universe. The empirical experiments are based on a large sample of U.S. stocks from 1987 to 2017. The analysis proceeds as follows. To set the stage, we replicate the results reported in the original papers. The value-weighted long-short portfolio return across all stocks is 1.56% (2.18%, 0.95%, 1.16%) per month based on the GKX (CPZ, IPCA, CA) signal, and the corresponding Fama-French six-factor (FF6)-adjusted return is 0.92% (1.87%, 0.62%, 0.75%). Such large and significant figures reflect the impressive success of machine learning techniques in generating outstanding performance relative to traditional methods such as nonregulated regressions and portfolio sorts based on individual anomalies.

Imposing economic restrictions reveals that the predictability of deep learning methods (i.e., GKX, CPZ, and CA) weakens considerably. Relative to the full sample evidence across all stocks, the value-weighted FF6-adjusted return based on the GKX (CPZ, CA) signal is 66% (71%, 48%) lower after excluding microcaps, 53% (77%, 75%) lower after excluding firms that do not have credit rating coverage, and 78% (69%, 94%) lower after excluding financially distressed firms that face further deteriorating credit conditions. None of the deep learning methods generates significant value-weighted FF6-adjusted returns at the 5% level after excluding distressed firms. For perspective, we apply the same economic restrictions to traditional methods and find a similar proportional magnitude of performance deterioration. Although IPCA underperforms deep learning models for the full sample, its performance only deteriorates modestly for subsamples that consist of cheap-to-trade stocks. Unlike deep learning models that facilitate nonlinearities, IPCA draws on the linear dependence between stock returns and firm characteristics. The evidence is thus consistent with the concept that accounting for nonlinearities is especially useful for difficult-to-value and difficult-to-arbitrage stocks.

Our findings are robust to alternative ways to implement neural networks, such as imposing economic restrictions on the training and validation samples (rather than using the entire universe of stocks), considering an alternative loss function that value weights (rather than equal weights) forecast errors, and predicting risk-adjusted (rather than raw) returns.

We then show that all four machine learning signals generate portfolio turnover that is considerably higher than most individual anomalies. The monthly turnover in the long-short portfolio is at least 87% (163%, 113%, 148%) for the GKX (CPZ, IPCA, CA) method, respectively. Altogether, in the presence of reasonable trading costs, the machine learning-based investments that we analyze would struggle to achieve a statistically and economically meaningful risk-adjusted payoff.

The previously described findings are further confirmed in an investment universe consisting of equity portfolios rather than individual stocks. In particular, we use the approach of Kozak et al. (2020) (KNS) to estimate the SDF. Because the SDF slope coefficients correspond to weights of the mean-variance efficient (MVE) portfolio (Hansen and Jagannathan 1991) and the MVE portfolio tends to take extreme stock positions associated with poor estimates of mean returns and the covariance matrix (Merton 1980, Green and Hollifield 1992), we examine the implications of economic restrictions on the SDF-implied MVE portfolio in terms of performance and portfolio weights. We find that the performance of the SDF-implied portfolio also deteriorates in subsamples with economic restrictions. Moreover, conditional on a predetermined level of market volatility, the SDF is estimated based on rather extreme portfolio positions. For instance, the SDF implies a −234% (−91%) short position in an individual anomaly portfolio at the 5th (25th) percentile and a 190% (96%) long position at the 95th (75th) percentile. Thus, the pricing kernel might be inadmissible from a real-time investment perspective. Restricting the investment universe to relatively cheap-to-trade stocks considerably mitigates equity positions.

We next examine whether and how machine learning-based performance varies with market conditions. Economic theory implies that fewer trading frictions and more arbitrage activity should improve price efficiency.[4] Consistent with the economic notion of limits to arbitrage, we demonstrate that the investment strategy based on GKX and CPZ signals (across all stocks) is considerably more profitable during periods of high investor sentiment, high market volatility, and low market liquidity. For instance, the monthly value-weighted FF6-adjusted return based on the GKX signal is statistically insignificant at 0.22% at times of low *VIX* and increases dramatically to 1.66% at times of high *VIX*, whereas the full sample average is 0.92%. In contrast, IPCA and CA models display relatively modest time series variation in trading profits and could even outperform in low limits-to-arbitrage market states.

We also examine the most recent years and find that unlike traditional anomaly-based trading strategies, all four machine learning signals continue to predict cross-sectional stock returns in the post-2001 period across all stocks. This finding supports the concept that machine learning techniques combine multiple weak sources of information into a meaningful composite signal. Consistent with our main results, anomalous return patterns in recent years are also confined to difficult-to-arbitrage stocks.

As emphasized by Karolyi and Van Nieuwerburgh (2020), it is imperative to examine the economic foundations of seemingly opaque machine learning methods. We propose two experiments to gauge the economic foundations. We first examine whether stocks with similar machine learning signals share common characteristics that are known to predict returns. The evidence shows that all four machine learning methods successfully identify mispriced stocks consistent with most anomaly-based trading strategies. Stocks in the long positions of machine learning-based investments are typically small, value, illiquid, and old stocks with a low price, low beta, high 11-month return (medium-term winners), low asset growth, low equity issuance, low credit rating coverage, and low analyst coverage. Therefore, despite their opaque nature, machine learning signals successfully identify mispriced stocks consistent with well-established empirical facts, without preselection of truly useful characteristics and models. Our findings highlight the merits of employing machine learning methods to avoid the data snooping problem in the anomaly literature and suggest that black-box-like machine learning models are reasonably interpretable, which is essential for a robust and credible assessment of out-of-sample predictability.[5]

Second, we control for the industry benchmark in machine learning signals and decompose the unconditional payoff into two components: intra-industry and inter-industry payoffs. Taking the GKX signal as an example, we construct three trading strategies based on NN3-predicted returns for all investable stocks (beyond the extreme long and short portfolios), including (1) an unconditional strategy that takes long (short) positions on market winners (market losers); (2) an intra-industry strategy that takes long (short) positions on industry winners (industry losers); and (3) an inter-industry strategy that takes long (short) positions on winner industries (loser industries). We show that the intra-industry strategy outperforms the unconditional strategy and the inter-industry strategy, suggesting that GKX signal is more informative for stock selection than for industry rotation. Our findings are robust to similar return decompositions based on IPCA and CA signals. Consistent with the finding that machine learning signals identify mispricing in difficult-to-arbitrage stocks, adjusting for industry averages further controls for similar firm fundamentals within the same industry and thus better predicts subsequent corrections due to market frictions.

With the recent development of financial technology (Fintech), using machine learning tools to identify new signals on price movements and to develop investment systems that can outperform human fund managers is gaining popularity (FSB 2017).[6] Our findings further support the concept that machine learning-based investments could hold considerable promise for asset management. First, we find that they can mitigate the downside risk and provide a good hedge during market crises. For instance, for major episodes of market downturns (e.g., the 1987 market crash, the Russian default, the bursting of the tech bubble, and the recent financial crisis), the GKX (CPZ, IPCA, CA) method generates, on average, a monthly value-weighted return of 3.56% (0.68%, 1.49%, −0.53%) after excluding microcaps, respectively. For perspective, the contemporaneous market excess return is −6.91%. Second, although the profitability of individual anomalies is driven primarily by short positions and often disappears in recent years, machine learning signals yield considerable profit in the long positions and remain viable in the post-2001 period. The performance of machine learning signals could further improve on industry adjustment.

To conclude, our paper is the first to provide large-scale evidence on the economic significance of machine learning methods. The machine learning (especially deep learning) techniques that we analyze face many of the challenges regarding cross-sectional return predictability. In particular, anomalous return patterns characterize difficult-to-value and difficult-to-arbitrage stocks. In addition, to the extent that deep learning signals predict cross-sectional stock returns for the full sample, the trading strategy is more profitable during periods of high market volatility and low market liquidity.

Machine learning signals also involve remarkably high turnover and often require taking extreme long-short positions in the tangency portfolio implied by the pricing kernel. Beyond economic restrictions, machine learning-based trading strategies nonetheless display less downside risk, yield considerable profit in long positions, and remain viable in the post-2001 period and the crisis period. Finally, black-box-like machine learning methods still generate economically interpretable trading strategies and are more meaningful for stock selection than for industry rotation.

Our analysis initiates a protocol for future work to demonstrate the feasibility of trading profits, such as (1) constructing value-weighted portfolios after excluding difficult-to-arbitrage stocks, (2) considering portfolio turnover and the corresponding transaction costs, (3) ensuring that the proposed tangency portfolio mitigates extreme long-short positions, (4) excluding high limits-to-arbitrage market states, (5) emphasizing performance over recent years, and (6) focusing on long positions.

Some recent studies have already responded to our findings and considered these economic hurdles when advocating new machine learning tools. These studies indicate the potential promise of understanding the cross-section of equity returns through the lens of deep learning and artificial intelligence (AI). For instance, CPZ continue to find a high annual Sharpe ratio of the SDF portfolio at 1.73 after removing 40% of the smallest stocks. Cong et al. (2021) use reinforcement learning to directly construct an AlphaPortfolio that maximizes the out-of-sample Sharpe ratio. The AlphaPortfolio is robust to imposing various economic restrictions in both the cross-section and time series as we propose; hence, it could be a competent and implementable approach for practitioners and advance our understanding of cross-sectional return predictability.

Taken together, our paper enriches the academic and policy discussions surrounding the adoption of machine learning techniques in asset management, including the effectiveness and sustainability of new trading signals, the lack of transparency and economic interpretability in complex machine learning algorithms, and the potential regulatory and supervisory implications related to financial stability.

The rest of the paper is organized as follows. Section 2 describes the methodology and data. Section 3 presents evidence on return predictability and other characteristics of machine learning portfolios for the full sample and subsamples with economic restrictions. Section 4 studies the time-varying return predictability of machine learning methods. Section 5 investigates the economic foundation for machine learning methods. Section 6 concludes.

2. Methodology and Data

2.1. Methodology and Data Sources

Our empirical analysis starts with the use of two deep learning methods that have been empirically successful in predicting future stock returns. We first implement a feed-forward neural network with three hidden layers having 32, 16, and 8 neurons per layer (NN3), using batch normalization and the Lasso penalty for training.[7] According to the comparative analysis of GKX, the NN3 model displays superior out-of-sample performance relative to traditional and more advanced return forecasting benchmarks. The second approach, advocated by CPZ, combines four neural networks, including two FFNs and two RNNs, with LSTM cells. Each of the LSTMs is connected to one FFN. The two FFN outcomes interact in the loss function to formulate a minimax optimization problem, termed the GAN.

As GKX and CPZ rely on multilayer neural networks, their works are considered deep learning routines. Although GKX studies a reduced-form setup in that they do not explicitly impose economic restrictions on data, CPZ incorporates a no-arbitrage condition to estimate the SDF and its stock loadings. Specifically, CPZ uses a minimax loss minimization problem formulated as a zero-sum game. One player, the asset pricing modeler, aims to choose the best-performing model, while the other player, the adversary, attempts to choose conditions under which the model delivers the worst performance. Therefore, CPZ uses an adversarial approach to select moment conditions that lead to the largest mispricing, consistent with the findings in the seminal paper by Hansen and Richard (1987).

We next consider two conditional beta pricing setups in the machine learning universe, namely, instrumented principal component analysis (IPCA) and the conditional autoencoder (CA). As proposed by KPS, IPCA formulates stock returns as a linear function of latent factors and allows factor loadings to vary with observable firm characteristics. In our main analysis, IPCA is estimated in a setup that imposes a zero-alpha restriction and considers six latent factors.[8] For robustness, we also report results for unrestricted IPCA. Gu et al. (2021) further relax the linearity assumption of KPS and use conditional autoencoder neural networks to model latent factor loadings as a flexible nonlinear function of firm characteristics. We focus on the autoencoder for which loadings on latent factors are modeled through neural networks with two hidden layers having 32 and 16 neurons per layer. The factors' neural network output layer has five neurons, indicating five latent factors (CA2).[9]

In addition to the previously described routines at the stock level, we also use the KNS approach to estimate the SDF using equity portfolios. The KNS approach aims to minimize the Hansen-Jagannathan distance

(Hansen and Jagannathan 1991) using ridge regression with three-fold cross-validation.

To summarize, our battery of tests consists of (1) nonlinear deep learning methods (GKX, CPZ, and CA) along with linear models (IPCA and KNS); (2) pricing kernel formulations (CPZ and KNS), beta pricing representations (IPCA and CA), and a reduced form approach (GKX); (3) routines that implement stock-level analysis (GKX, CPZ, IPCA, and CA) along with KNS that focuses on equity portfolios, and (4) models that account for expected return variations with both firm characteristics and macro conditions (GKX and CPZ) versus firm characteristics only (IPCA, CA, and KNS).

In what follows, we describe the sample used to implement the GKX, IPCA, and CA methods. The investment universe consists of all NYSE/AMEX/Nasdaq stocks, with daily and monthly stock data obtained from the Center for Research in Security Prices (CRSP). Quarterly and annual financial statement data come from the COMPUSTAT database. We construct 94 firm characteristics that have been documented as significant cross-sectional return predictors, including annually updated predictors such as absolute accruals and asset growth, quarterly updated predictors such as cash holdings and corporate investment, and monthly updated predictors such as 12-month momentum and idiosyncratic volatility.[10] To avoid forward-looking biases, monthly characteristics are delayed by at most one month, and quarterly and annual characteristics are delayed by at least four and six months, respectively. We also account for 74 industry dummies based on the first two digits of Standard Industrial Classification (SIC) codes and eight monthly macroeconomic predictors, as in Welch and Goyal (2008), including the dividend price ratio, earnings price ratio, stock variance, book-to-market ratio, net equity expansion, T-bill rate, term spread, and default yield spread.[11] We consider not only stock-level and industry-level predictors but also interactions between stock characteristics and macroeconomic state variables (94×8), resulting in 920 predictors in total.

The full sample period ranges from 1957 to 2017. The full sample for GKX and CA is then divided into three subperiods: 18 years for the training sample (1957 to 1974), 12 years for the validation sample (1975 to 1986), and the remaining 31 years (1987 to 2017) for out-of-sample testing. We train the model every year so that every year, the training sample expands. The size of the validation sample remains fixed while we roll it forward by one year. For the out-of-sample estimation, we average across an ensemble of nine models with the same neural network architecture but distinct initial values. The final sample for out-of-sample tests consists of 21,882 stocks, with the number of stocks per month ranging between 5,117 and 7,877. In addition to the expanding window scheme for the training sample, we also conduct an experiment using a rolling window for the training sample (18 years) and for the validation sample (12 years). The results of using either a rolling or an expanding window for the training sample are qualitatively identical. Hence, there is no particular effect of increased learning in the training sample. IPCA skips the validation procedure and requires at least 120 months for the in-sample estimation, and forward rolling is performed on a monthly basis.

The CPZ sample consists of all U.S. stocks from CRSP with available data on 46 firm characteristics related to past returns, investment, profitability, intangibles, value, and trading frictions.[12] CPZ further include 178 macroeconomic predictors, as well as nonlinear interactions among firm characteristics and between firm characteristics and macroeconomic states. The full sample period ranges from 1967 to 2016, and it is divided into 20 years for the training sample (1967 to 1986), 5 years for the validation sample (1987 to 1991), and the remaining 25 years (1992 to 2016) for out-of-sample tests. The final sample includes 7,904 stocks, with the number of stocks per month ranging between 1,933 and 2,755. The CPZ sample is populated with fewer stocks, particularly because of the requirement to have full data records for all firm characteristics.[13] For comparison, GKX set missing characteristics to be equal to their corresponding median values across all stocks.

Although the four machine learning methods we analyze could differ in their sample coverage, input variables, and optimization objectives, we do not take a stand on which is the most appropriate objective, and we do not aim to identify the best-performing method. Despite the distinct objectives of machine learning methods, such as estimating the expected returns, variances, and covariances or maximizing the ex ante Sharpe ratio, they all provide predictions for future stock returns that can be used to form trading strategies. If machine learning methods competently predict stock returns or estimate the pricing kernel, the predicted returns or loadings on the SDF should be sufficient to rank stocks and thus can be used to form outperforming investment strategies. The choice of objective is essentially a parameterization and implementation choice and relies on specific assumptions on the economic mechanism underlying the data generating process. The model performance depends on which assumption is closer to the true, unknown data generating process and remains an empirical question. Rather than comparing the models or assessing the corresponding objectives, we aim to comprehensively examine whether machine learning signals could provide value that extends to economic significance.

2.2. Subsamples with Economic Restrictions

To set the stage, we replicate all machine learning methods using the full sample, as in the original studies. We then move on to restrict the sample to the universe of relatively cheap-to-trade stocks. In particular, the extant literature highlights that anomalous patterns in the cross-section of asset returns are concentrated in microcap stocks.[14] For instance, Novy-Marx and Velikov (2016) show that microcaps display high gross Sharpe ratios in most anomalies relative to other size groups, but the difference is much smaller in net Sharpe ratios after accounting for transaction costs. In the same vein, Hou et al. (2020) document that 65% (82%) of anomalies are statistically insignificant after excluding microcaps based on NYSE breakpoints and when using value-weighted returns and the cutoff t statistic of 1.96 (2.78), suggesting that capital markets might be more efficient than previously thought.

Thus, on the one hand, the collective evidence indicates that microcaps are costly to trade to the extent that anomalies in those firms could easily become unprofitable for marginal investors. On the other hand, machine learning techniques are particularly useful for uncovering complex patterns and hidden relationships and for combining multiple weak sources of information into a composite signal, and they are often more effective than linear regressions in handling multicollinearity (Rasekhschaffe and Jones 2019, Gu et al. 2020, Karolyi and Van Nieuwerburgh 2020). Altogether, a natural question to explore is whether machine learning techniques can predict cross-sectional stock returns for stocks other than microcaps. Thus, our first subsample excludes microcaps.

In the same vein, our second subsample includes only rated firms, that is, firms with data on Standard & Poor's (S&P) long-term issuer credit ratings.[15] In a given month, approximately 90% of the rated firms are above the 20th NYSE size percentile. Unreported results also show that rated firms tend to be large, value firms with considerably higher past returns and liquidity, lower idiosyncratic volatility, and more analyst coverage than nonrated firms. Thus, we focus on relatively cheap-to-trade stocks by excluding nonrated firms.

The third subsample imposes an additional filter on the universe of rated firms. In particular, Avramov et al. (2013, 2018) show that market anomalies among stocks and corporate bonds tend to concentrate in financially distressed firms and particularly around credit rating downgrades. Their suggested mechanism is straightforward. Distressed firms display extreme values of predictive characteristics such as more negative past returns, high idiosyncratic volatility, a high fraction of negative earnings surprises, and high analyst dispersion; thus, they are sorted into the short leg of anomaly portfolios. The sluggish response to financial distress by retail and institutional investors leads to a wide range of anomalous patterns in the cross-section of stock and bond returns. Collectively, on the one hand, investors tend to overprice financially distressed stocks. On the other hand, credit rating downgrades are associated with substantially elevated trading frictions, and therefore, overpricing cannot be easily arbitraged away. Thus, the third subsample excludes distressed firms around credit rating downgrades. Specifically, among rated firms, we further exclude stock-month observations from 12 months before to 12 months after an issuer credit rating downgrade.

We report the number of stocks in each year for the full sample and three subsamples in the online appendix, Table IA1. Notably, applying sensible restrictions significantly reduces the number of stocks. The average monthly GKX (CPZ) sample is reduced by 49% (43%) after excluding microcaps, 78% (71%) after excluding nonrated firms, and 83% (78%) after excluding financially distressed firms around credit rating downgrades. Consequently, the existing evidence on machine learning-based investments could be dominated by stocks that are plentiful albeit low in aggregate market value and difficult to value and arbitrage.[16] This preliminary finding further motivates us to explore whether machine learning methods can clear common economic restrictions in empirical finance.

3. Return Predictability in the Presence of Economic Restrictions

We assess return predictability using conventional portfolio sorts. In particular, at the end of each month t, we construct portfolios using the four proposed machine learning signals: (1) the one-month-ahead out-of-sample stock return prediction using the NN3 model (GKX); (2) the risk loadings on the SDF estimated from a combination of deep neural networks, as noted earlier (CPZ); (3) the one-month-ahead out-of-sample stock return prediction using the IPCA model (KPS), and (4) the one-month-ahead out-of-sample stock return prediction using the CA model (Gu et al. 2021). A higher value of predicted returns and risk loadings indicates higher expected returns in the holding period. We sort stocks into decile portfolios based on predicted returns or risk loadings and evaluate portfolio returns in month $t+1$. The bottom (top) decile consists of stocks with the lowest (highest) expected returns in the next month. We compute equal-weighted and value-weighted holding period returns for each decile portfolio. We also implement the zero-cost trading strategy by taking long positions in the top decile of stocks (highest expected returns) and selling short stocks in the bottom decile (lowest expected returns). The payoff of the long-short investment strategy is computed as the high (top decile) minus low (bottom decile) portfolio return (HML).[17]

In addition to raw portfolio returns, we report risk-adjusted returns from (1) the CAPM, that is, only adjusting for the market factor (MKT, defined as the excess return on the value-weighted CRSP market index over the one-month T-bill rate); (2) the Fama-French-Carhart four-factor plus liquidity factor model (FFC+PS) consisting of the market factor (MKT), the size factor (SMB, defined as small minus big firm return premium), the book-to-market factor (HML, defined as high book-to-market minus low book-to-market return premium) (Fama and French 1993), Carhart (1997) momentum factor (MOM, defined as winner minus loser return premium), and Pástor and Stambaugh (2003) liquidity factor; (3) the Fama-French five-factor model (FF5) consisting of the market factor (MKT), the size factor (SMB), the book-to-market factor (HML), the profitability factor (RMW, defined as robust minus weak return premium), and the investment factor (CMA, defined as conservative minus aggressive return premium) (Fama and French 2015); (4) the Fama-French six-factor model (FF6) that adds the momentum factor (MOM) to FF5 (Fama and French 2018), and (5) the Stambaugh-Yuan four-factor model (SY) consisting of the market factor (MKT), the size factor (SMB), and two mispricing factors arising from the cluster of anomalies related to firm management (MGMT) and performance (PERF) (Stambaugh and Yuan 2017).[18] We rely on a single time series regression to estimate the out-of-sample alpha, whereas our main findings remain unchanged if we estimate time-varying beta for the long-short portfolio using a five-year rolling window. The standard errors in all estimations are corrected for autocorrelation with four lags using the method of Newey and West (1987).[19]

3.1. Evidence from NN3-Predicted Returns

We start by assessing the out-of-sample return predictability of the GKX method. Table 1 reports the value-weighted results. In the interest of brevity, we tabulate the equal-weighted results in the online appendix, Table IA2, and only discuss the main findings in this subsection. As shown in Panel A of Table 1, using all stocks in our sample, the value-weighted long-short portfolio return is 1.56% per month (t statistic = 4.53) over the 1987–2017 sample period, and the risk-adjusted return ranges between 0.77% and 1.89% with a t statistic above 3.03 across all factor models. In contrast to the equal-weighted results, the profits weaken considerably in value-weighted portfolios, and the average decline in economic magnitude is 48% across all performance measures. In addition, the long position on stocks with the highest expected returns generates a significant and economically larger payoff than the short position. For instance, the long leg yields a significant FF6-adjusted return of 0.77% per month, whereas the corresponding payoff on the short leg falls to a statistically insignificant −0.15%.

The evidence on the relative strength of the long versus short leg appears to be at odds with the literature documenting that the profitability of anomaly-based trading strategies is attributable primarily to the short leg of the trade (Hong et al. 2000; Stambaugh et al. 2012; Avramov et al. 2013, 2018). The evidence, however, supports the concept that machine learning routines possess superior ability to detect complex features in the data that otherwise remain unnoticed. The empirical success of long positions could be particularly valuable for institutions, such as mutual funds and pension funds, that focus primarily on long positions.

We next exclude microcaps, that is, stocks with a market capitalization smaller than the 20th NYSE size percentile at the end of the portfolio formation month t. As shown in Panel A of Table 1, the value-weighted long-short portfolio return is significant at 1.05% per month. The performance further deteriorates after adjusting for risk exposures using the FF6 model, that is, a statistically insignificant 0.31% per month, in contrast to 0.92% in the full sample. In addition, the predictive power of the GKX method in cross-sectional stock returns is adequately captured by the Stambaugh and Yuan (2017) four factors, resulting in a statistically insignificant SY-adjusted return for both equal-weighted and value-weighted long-short portfolios. Relative to the full sample, the equal-weighted (value-weighted) payoff is 65% (49%) lower after excluding microcaps across all performance measures.

The second subsample considers only rated firms, that is, firms with data on S&P long-term issuer credit ratings. We sort all rated firms into decile portfolios based on NN3-predicted returns and calculate holding period returns. We tabulate the results in Panel B of Table 1. The value-weighted long-short portfolio yields a significant return of 1.02% per month and an FF6-adjusted return of 0.43% per month. Excluding nonrated firms creates a 57% (46%) decline in economic magnitude across all equal-weighted (value-weighted) performance measures relative to the full sample.

The third subsample excludes observations on distressed firms around credit rating downgrades. In particular, among rated firms, we further exclude stock-month observations from 12 months before to 12 months after an issuer credit rating downgrade. This is not a real-time trading strategy, as we look ahead when discarding the 12-month period prior to the downgrade. However, we aim to investigate whether the trading profits generated from machine learning algorithms go beyond a small subset of firm-months (i.e.,

Table 1. Performance of Portfolios Sorted by Neural Network Predicted Returns

Panel A: Value-weighted returns to investment strategies sorted by NN3-predicted returns (full sample and microcaps excluded)

	Full sample						Nonmicrocaps					
Rank of $\hat{R}$	Return	CAPM	FFC+PS	FF5	FF6	SY	Return	CAPM	FFC+PS	FF5	FF6	SY
Low	0.328	−0.923***	−0.508***	−0.360*	−0.146	−0.033	0.364	−0.893***	−0.467***	−0.290	−0.079	0.006
	(0.78)	(−4.35)	(−3.07)	(−1.95)	(−0.99)	(−0.19)	(0.85)	(−4.09)	(−2.74)	(−1.55)	(−0.53)	(0.03)
2	0.678**	−0.346***	−0.158*	−0.163*	−0.039	0.011	0.647**	−0.390***	−0.195**	−0.156	−0.041	0.016
	(2.37)	(−3.38)	(−1.67)	(−1.76)	(−0.48)	(0.12)	(2.24)	(−3.65)	(−2.01)	(−1.63)	(−0.49)	(0.16)
3	0.855***	−0.090	−0.059	−0.102	−0.053	−0.015	0.862***	−0.090	0.006	−0.048	0.032	0.065
	(3.21)	(−1.04)	(−0.72)	(−1.19)	(−0.63)	(−0.16)	(3.24)	(−1.01)	(0.06)	(−0.52)	(0.36)	(0.66)
4	0.890***	−0.006	−0.038	−0.145**	−0.128**	−0.096	0.841***	−0.073	−0.108	−0.189**	−0.183**	−0.153
	(3.88)	(−0.08)	(−0.58)	(−2.51)	(−2.09)	(−1.23)	(3.47)	(−0.96)	(−1.47)	(−2.29)	(−2.16)	(−1.61)
5	0.917***	0.042	−0.049	−0.084	−0.109	−0.096	0.979***	0.109	0.040	−0.014	−0.019	0.006
	(3.96)	(0.52)	(−0.70)	(−1.22)	(−1.58)	(−1.33)	(4.26)	(1.18)	(0.51)	(−0.21)	(−0.28)	(0.08)
6	1.019***	0.165**	0.057	0.009	−0.029	−0.016	0.899***	0.028	−0.064	−0.125*	−0.153**	−0.155**
	(4.83)	(2.19)	(0.78)	(0.13)	(−0.44)	(−0.20)	(4.13)	(0.38)	(−0.93)	(−1.93)	(−2.33)	(−2.17)
7	1.185***	0.307***	0.167*	0.129	0.073	0.097	1.058***	0.194**	0.084	0.008	−0.027	−0.008
	(5.14)	(3.10)	(1.84)	(1.42)	(0.82)	(0.98)	(4.84)	(2.17)	(0.94)	(0.10)	(−0.34)	(−0.09)
8	1.078***	0.218**	0.055	0.054	−0.029	0.008	1.210***	0.340***	0.167**	0.171**	0.092	0.080
	(4.51)	(2.31)	(0.69)	(0.60)	(−0.36)	(0.07)	(5.20)	(3.60)	(2.26)	(2.19)	(1.26)	(0.90)
9	1.350***	0.435***	0.240**	0.275**	0.174	0.152	1.112***	0.251***	0.094	0.073	−0.002	0.045
	(5.14)	(3.02)	(2.10)	(2.37)	(1.63)	(1.20)	(4.81)	(2.73)	(1.09)	(0.75)	(−0.02)	(0.36)
High	1.883***	0.971***	0.853***	0.846***	0.770***	0.735***	1.410***	0.496***	0.304**	0.337***	0.234**	0.185
	(6.45)	(4.91)	(5.37)	(5.53)	(5.01)	(4.38)	(5.39)	(3.46)	(2.56)	(2.80)	(2.02)	(1.46)
HML	1.556***	1.894***	1.361***	1.206***	0.916***	0.769***	1.047***	1.389***	0.771***	0.627**	0.312	0.179
	(4.53)	(5.64)	(5.31)	(4.66)	(4.08)	(3.03)	(3.24)	(4.43)	(3.24)	(2.41)	(1.51)	(0.73)

Panel B: Value-weighted returns to investment strategies sorted by NN3-predicted returns (credit rating sample and downgrades excluded)

	Credit rating sample						Nondowngrades					
Rank of $\hat{R}$	Return	CAPM	FFC+PS	FF5	FF6	SY	Return	CAPM	FFC+PS	FF5	FF6	SY
Low	0.398	−0.807***	−0.419**	−0.426*	−0.178	0.019	0.783**	−0.364*	−0.027	−0.081	0.123	0.368*
	(0.94)	(−3.70)	(−2.48)	(−1.93)	(−1.10)	(0.10)	(2.07)	(−1.72)	(−0.15)	(−0.37)	(0.65)	(1.87)
2	0.759***	−0.265**	−0.108	−0.203*	−0.063	0.025	0.956***	−0.039	0.112	0.027	0.143	0.251**
	(2.61)	(−2.46)	(−1.05)	(−1.84)	(−0.62)	(0.20)	(3.58)	(−0.38)	(1.16)	(0.24)	(1.40)	(2.05)
3	0.923***	−0.017	0.067	−0.048	0.039	0.028	1.135***	0.231*	0.302**	0.206	0.268*	0.282*
	(3.35)	(−0.14)	(0.60)	(−0.40)	(0.34)	(0.21)	(4.46)	(1.84)	(2.42)	(1.50)	(1.97)	(1.93)
4	0.855***	−0.066	−0.142	−0.235**	−0.221**	−0.176	0.981***	0.095	0.021	−0.050	−0.048	0.014
	(3.30)	(−0.63)	(−1.55)	(−2.51)	(−2.23)	(−1.40)	(4.08)	(0.98)	(0.23)	(−0.57)	(−0.51)	(0.11)
5	1.096***	0.222*	0.160*	−0.018	0.007	0.058	1.186***	0.326**	0.269**	0.082	0.104	0.176
	(4.78)	(1.86)	(1.80)	(−0.25)	(0.09)	(0.62)	(5.30)	(2.56)	(2.40)	(0.88)	(1.09)	(1.48)
6	0.890***	0.024	−0.091	−0.126	−0.161**	−0.158*	1.031***	0.174*	0.049	0.017	−0.033	−0.032
	(3.96)	(0.27)	(−1.19)	(−1.56)	(−2.14)	(−1.86)	(4.78)	(1.81)	(0.56)	(0.18)	(−0.40)	(−0.36)
7	1.004***	0.171*	0.036	−0.052	−0.085	−0.054	1.127***	0.306***	0.188*	0.100	0.064	0.105
	(4.67)	(1.74)	(0.41)	(−0.66)	(−1.10)	(−0.61)	(5.43)	(3.07)	(1.86)	(1.18)	(0.74)	(1.08)

Table 1. (Continued)

Panel B: Value-weighted returns to investment strategies sorted by NN3-predicted returns (credit rating sample and downgrades excluded)

	Credit rating sample						Nondowngrades					
Rank of $\hat{R}$	Return	CAPM	FFC+PS	FF5	FF6	SY	Return	CAPM	FFC+PS	FF5	FF6	SY
8	1.216***	0.352***	0.161**	0.134*	0.056	0.043	1.313***	0.461***	0.272***	0.255***	0.167**	0.123
	(5.38)	(3.39)	(2.00)	(1.67)	(0.75)	(0.46)	(6.06)	(4.54)	(3.42)	(2.95)	(2.21)	(1.33)
9	1.070***	0.223**	0.038	0.016	−0.055	0.037	1.139***	0.306***	0.109	0.079	0.000	0.081
	(4.61)	(2.27)	(0.42)	(0.16)	(−0.60)	(0.28)	(5.04)	(2.97)	(1.18)	(0.76)	(0.00)	(0.61)
High	1.422***	0.537***	0.323***	0.358***	0.255**	0.169	1.506***	0.632***	0.418***	0.436***	0.326***	0.238
	(5.62)	(3.83)	(2.59)	(2.88)	(2.18)	(1.21)	(6.05)	(4.35)	(3.20)	(3.40)	(2.68)	(1.64)
HML	1.024***	1.344***	0.743***	0.784***	0.433**	0.150	0.723**	0.996***	0.445**	0.517*	0.204	−0.129
	(3.18)	(4.40)	(3.23)	(2.70)	(2.05)	(0.59)	(2.49)	(3.40)	(2.00)	(1.87)	(0.92)	(−0.52)

Notes. At the end of each month *t*, stocks are sorted into deciles according to their one-month-ahead out-of-sample predicted returns ($\hat{R}$) using a neural network with three hidden layers (NN3) (Gu et al. 2020). This table reports the value-weighted returns for month $t+1$ for each decile portfolio and the strategy of going long (short) on the highest (lowest) expected return stocks (HML) over the entire sample period from 1987 to 2017. Portfolio returns are further adjusted by the CAPM, Fama-French-Carhart four-factor and Pástor-Stambaugh liquidity factor model (FFC+PS), Fama-French five-factor model (FF5), Fama-French six-factor model (FF6), and Stambaugh-Yuan four-factor model (SY). Panel A reports the results for the full sample and the subsample that excludes microcaps. Panel B reports similar statistics for the subsamples that exclude nonrated firms and credit rating downgrades. Newey-West adjusted *t* statistics are shown in parentheses.

*, **, and ***Significant at the 10%, 5%, and 1% levels, respectively.

distressed firms around credit rating downgrades).[20] As shown in Panel B of Table 1, during periods of improving or stable credit conditions, the value-weighted long-short portfolio yields a significant return of 0.72% per month, whereas the FF6-adjusted return is no longer statistically significant. Moreover, only one (two) of five risk-adjusted returns remains significant at the 5% threshold for the equal-weighted (value-weighted) portfolio. Excluding distressed firms creates an 86% (70%) decline in economic magnitude across all equal-weighted (value-weighted) performance measures relative to the full sample.

For perspective, we consider two traditional methods as benchmarks. First, we apply the standard Ordinary Least Squares (OLS) regression methodology with all 920 predictors used in the GKX estimation to predict the one-month-ahead out-of-sample stock returns. For consistency, the sample starts from 1957, and the out-of-sample test ranges from 1987 to 2017. We estimate the OLS regression every month, and the sample expands over time. Next, we sort all firms into decile portfolios based on OLS-predicted returns and calculate holding period returns. Given the multidimensional challenge in the OLS regression, our second benchmark relies on portfolio sorts while combining the payoff on individual anomalies. Specifically, we sort stocks into deciles according to each of the 94 firm characteristics. We adjust the sign of those characteristics so that a higher value predicts higher future performance, in accordance with the literature. We compute the equal- and value-weighted holding period returns for each characteristic and then average the portfolio returns across 94 characteristics.

We tabulate the results in the online appendix, Table IA3, where Panel A shows the results for portfolios sorted by OLS-predicted returns and Panel B shows those for the combination of individual anomalies. We report only the performance of long-short portfolios for brevity. As expected, the GKX method substantially improves the investment payoff relative to the traditional methods. It outperforms the standard OLS by 82% (99%) across all equal-weighted performance measures for the full sample (after excluding microcaps), and the OLS method fails to deliver significant value-weighted payoffs for the full sample as well as three subsamples. Relative to the full sample, the equal-weighted payoff based on OLS-predicted returns is 66% (64%, 64%) lower after excluding microcaps (nonrated firms, distressed firms) across all performance measures, while the corresponding value-weighted payoff is 40% (43%, 76%) lower, respectively.

Moving to the combination of individual anomalies as an alternative benchmark, the GKX method generates a payoff that is approximately seven (nine) times higher across all equal-weighted (value-weighted) performance measures for the full sample and continues to outperform in all three subsamples. Relative to the full sample, the equal-weighted payoff based on individual anomalies is 44% (25%, 77%) lower after excluding microcaps (nonrated firms, distressed firms) across all performance measures, whereas the corresponding value-weighted payoff is 38% (55%, 106%) lower.

In summary, when we apply standard economic restrictions and focus exclusively on the subset of relatively cheap-to-trade stocks, the long-short portfolio performance weakens considerably in terms of both statistical significance and economic magnitude. On the one hand, machine learning substantially improves the investment payoff compared with traditional methods such as OLS regression and portfolio sorts based on individual anomalies. On the other hand, both machine learning and traditional methods deliver lower payoffs in the presence of economic restrictions, and their performance deteriorates by a similar proportional magnitude.

One may argue that because we train the NN3 model using the entire universe of stocks, this training scheme could be biased in favor of detecting predictive patterns, especially for microcaps, nonrated firms, and distressed firms. However, as long as there are shared attributes between the subsamples, training based on the comprehensive universe could benefit from the rich return structures in big data and capture subsample patterns better. Nevertheless, as a robustness check, we retrain the NN3 model in each subsample separately and assess the out-of-sample return predictability.

The results are tabulated in Table 2. We report only the value-weighted performance of long-short portfolios for brevity and tabulate the equal-weighted results in the online appendix, Table IA4. First, we exclude microcaps in both the NN3 estimations and the portfolio sorts. Relative to the full sample, the equal-weighted (value-weighted) payoff is 54% (37%) lower after excluding microcaps across all performance measures. Moreover, there is a modest improvement in trading profits when we exclude microcaps in the training and validation samples. The value-weighted return (FF6-adjusted return) is 1.19% (0.49%) per month, in contrast to 1.05% (statistically insignificant at 0.31%) when the machine learning algorithm is trained with the full sample, whereas the SY-adjusted return remains statistically insignificant.

Next, we exclude nonrated firms from both the NN3 estimations and the portfolio sorts. Because the data on S&P long-term issuer credit ratings are sparse before December 1985, the out-of-sample test starts in 1999. The training and validation samples are defined as in GKX. To put our findings in perspective, we repeat the analysis from Table 1 during the post-1999 period and show that excluding nonrated firms creates a 50% (52%) decline in economic magnitude across all equal-weighted (value-weighted) performance

Table 2. Robustness Check: Neural Network Estimation Based on Subsamples

			Value-weighted returns to investment strategies sorted by NN3-predicted returns					
Training	Testing	Sample period	Return	CAPM	FFC+ PS	FF5	FF6	SY
Nonmicrocaps	Nonmicrocaps	1987–2017	1.191***	1.512***	0.838***	0.838***	0.486***	0.306
			(4.22)	(5.38)	(4.48)	(3.49)	(2.96)	(1.57)
Credit rating sample	Credit rating sample	1999–2017	1.021**	1.460***	1.044***	0.605	0.406	0.137
			(1.98)	(3.80)	(3.33)	(1.54)	(1.39)	(0.42)
Nondowngrades	Nondowngrades	1999–2017	0.797**	0.764**	0.263	0.147	0.035	−0.132
			(2.33)	(2.02)	(0.82)	(0.46)	(0.13)	(−0.36)
Full sample	Full sample	1987–2017	1.556***	1.894***	1.361***	1.206***	0.916***	0.769***
			(4.53)	(5.64)	(5.31)	(4.66)	(4.08)	(3.03)
Full sample	Nonmicrocaps	1987–2017	1.047***	1.389***	0.771***	0.627**	0.312	0.179
			(3.24)	(4.43)	(3.24)	(2.41)	(1.51)	(0.73)
Full sample	Full sample	1999–2017	1.753***	2.144***	1.610***	1.307***	1.154***	0.827**
			(3.44)	(5.00)	(5.16)	(4.02)	(3.98)	(2.50)
Full sample	Nonmicrocaps	1999–2017	1.232**	1.647***	1.102***	0.771**	0.609**	0.326
			(2.56)	(4.17)	(3.82)	(2.43)	(2.36)	(1.05)
Full sample	Credit rating sample	1999–2017	1.267**	1.641***	1.087***	0.907**	0.724**	0.265
			(2.59)	(4.11)	(3.55)	(2.41)	(2.50)	(0.78)
Full sample	Nondowngrades	1999–2017	0.858*	1.177***	0.667**	0.628*	0.460	−0.113
			(1.97)	(3.01)	(2.13)	(1.69)	(1.43)	(−0.33)

Notes. At the end of each month t, stocks are sorted into deciles according to their NN3-predicted returns ($\hat{R}$) (Gu et al. 2020). This table reports the value-weighted return for month $t+1$ for the strategy of going long (short) on the highest (lowest) expected return stocks. Portfolio returns are further adjusted by the CAPM, Fama-French-Carhart four-factor and Pástor-Stambaugh liquidity factor model (FFC+PS), Fama-French five-factor model (FF5), Fama-French six-factor model (FF6), and Stambaugh-Yuan four-factor model (SY). We use the full sample and subsamples that exclude microcaps, nonrated firms, and credit rating downgrades for the NN3 estimations (in the "Training" column) and the portfolio sorts (in the "Testing" column). Newey-West adjusted t statistics are shown in parentheses.

*, **, and ***Significant at the 10%, 5%, and 1% levels, respectively.

measures relative to the full sample. In addition, for investors interested in trading rated firms, training the machine learning algorithm with the full sample could increase the trading profits. When we train the NN3 model using the entire universe of stocks and exclude nonrated firms in portfolio construction, the value-weighted long-short portfolio yields a significant FF6-adjusted return of 0.72% per month. In contrast, training with the subset of rated firms generates a statistically insignificant monthly FF6-adjusted return of 0.41%.

We further exclude distressed firms around credit rating downgrades from the subsample of rated firms in both the NN3 estimations and the portfolio sorts. Excluding distressed firms creates an 82% (84%) decline in economic magnitude across all equal-weighted (value-weighted) performance measures relative to the full sample, and the FF6-adjusted return is statistically insignificant for both equal-weighted and value-weighted portfolios. Furthermore, training the machine learning algorithm with the full sample also generates higher value-weighted trading profits for nondistressed firms.

Overall, training the machine learning algorithm in subsamples with economic restrictions does not alter our main findings, that is, investment payoffs on machine learning-based trading strategies deteriorate in the presence of sensible economic restrictions. In addition, training the machine learning algorithm using a selective set of stocks of interest could adversely affect the out-of-sample return predictability potentially as a result of data scarcity.

Because we focus on value-weighted portfolio returns on a risk-adjusted basis in the out-of-sample test, we also use an alternative objective function to train the NN3 model. Instead of minimizing the equal-weighted forecast errors in predicting the raw returns, we train the model to minimize the value-weighted forecast errors in predicting the FF6-adjusted returns. This allows us to tilt the estimates toward large stocks and focus on predicting risk-adjusted returns in training and validating the machine learning algorithm.

The results are tabulated in the online appendix, Table IA5, where Panel A shows the results for portfolios sorted by NN3-predicted alphas and Panel B shows the results for portfolios sorted by NN3-predicted returns (as in GKX) using the same universe of stocks.[21] In Panel A, the monthly value-weighted FF6-adjusted return is 0.61% for the full sample, 0.38% after excluding microcaps, and becomes statistically insignificant after excluding nonrated firms or credit rating downgrades. In addition, the trading profits based on NN3-predicted alphas using the value-weighted loss function do not outperform the original GKX method (i.e., NN3-predicted returns using the equal-weighted loss function) for the full sample or for most subsamples with economic restrictions, implying that the seemingly more aligned objective function does not necessarily improve the predictive performance. Collectively, our main findings are robust to the alternative objective function that tilts toward predicting risk-adjusted returns for large stocks.

In short, our findings are robust to alternative ways to implement the neural network model, such as imposing economic restrictions on the training and validation samples (rather than using the entire universe of stocks), as well as considering an alternative loss function that value weights (rather than equal weights) forecast errors of risk-adjusted returns (rather than raw returns). As Karolyi and Van Nieuwerburgh (2020) point out, our experiments also call for more research on the best practices to make implementation choices when applying machine learning models to financial data, such as how to balance the tradeoff between completeness and relevance in sample selection and how to set an appropriate objective function.

3.2. Evidence from the Adversarial Approach

We implement the analysis from Table 1 but use the deep learning signal based on the adversarial approach of CPZ. Table 3 has the same layout as Table 1, where Panel A shows the results for the full sample and the subsample excluding microcaps, and Panel B shows those for the subsample of rated firms and the subsample excluding credit rating downgrades. For brevity, we tabulate the equal-weighted results in the online appendix, Table IA6.

As shown in Panel A, over the 1987–2016 sample period, the long-short portfolio return (FF6-adjusted return) across all stocks is highly significant and economically large, that is, 2.18% (1.87%) per month in the value-weighted portfolio. Similar to our prior findings, the economic magnitude is considerably attenuated after excluding microcaps, that is, the long-short portfolio return (FF6-adjusted return) declines to 1.08% (0.55%). Across all performance measures, the equal-weighted (value-weighted) trading profit is 60% (61%) lower for the subsample excluding microcaps relative to the full sample.

As shown in Panel B, the value-weighted long-short strategy across rated firms yields a monthly return of 0.81%, and the FF6-adjusted return is statistically insignificant. If we further exclude credit rating downgrades, the long-short portfolio return (FF6-adjusted return) is 0.92% (0.57% with t statistic = 1.82). Across all performance measures, the equal-weighted (value-weighted) trading profit is 60% (72%) lower for the subsample of rated firms relative to the full sample and 65% (64%) lower when we further exclude firms with deteriorating credit conditions. Thus, investment payoffs based on the CPZ signal also deteriorate in the presence of sensible economic restrictions.

Table 3. Performance of Portfolios Sorted by Risk Loadings on the Stochastic Discount Factor

Panel A: Value-weighted returns to investment strategies sorted by risk loadings (full sample and microcaps excluded)

	Full sample						Nonmicrocaps					
Rank of β	Return	CAPM	FFC+ PS	FF5	FF6	SY	Return	CAPM	FFC+ PS	FF5	FF6	SY
Low	0.078	−0.962***	−0.593***	−0.859***	−0.590***	−0.425*	0.342	−0.662***	−0.355**	−0.617***	−0.398**	−0.284
	(0.21)	(−4.35)	(−3.28)	(−3.58)	(−3.01)	(−1.82)	(0.99)	(−3.48)	(−2.17)	(−3.11)	(−2.25)	(−1.36)
2	0.735**	−0.205	−0.055	−0.206	−0.103	−0.077	0.673**	−0.264*	−0.093	−0.222	−0.118	−0.048
	(2.45)	(−1.35)	(−0.37)	(−1.35)	(−0.65)	(−0.43)	(2.42)	(−1.81)	(−0.65)	(−1.54)	(−0.79)	(−0.26)
3	0.812***	−0.123	0.033	−0.026	0.040	0.109	1.079***	0.163	0.330***	0.270**	0.335***	0.385**
	(2.95)	(−0.99)	(0.30)	(−0.23)	(0.33)	(0.78)	(4.10)	(1.32)	(2.77)	(2.23)	(2.63)	(2.58)
4	0.859***	−0.035	0.094	0.012	0.061	0.127	0.705**	−0.206	−0.129	−0.179*	−0.177*	−0.124
	(3.36)	(−0.33)	(0.89)	(0.12)	(0.53)	(0.93)	(2.47)	(−1.51)	(−1.21)	(−1.71)	(−1.70)	(−1.17)
5	0.965***	0.055	0.064	0.042	0.025	0.018	0.987***	0.090	0.120	0.084	0.102	0.060
	(3.82)	(0.53)	(0.68)	(0.43)	(0.26)	(0.19)	(4.03)	(0.88)	(1.20)	(0.78)	(0.94)	(0.53)
6	1.131***	0.231*	0.173	0.220**	0.167	0.156	1.227***	0.345***	0.273***	0.291***	0.251***	0.250**
	(4.40)	(1.97)	(1.61)	(2.06)	(1.63)	(1.33)	(5.16)	(3.65)	(2.95)	(3.01)	(2.66)	(2.58)
7	1.255***	0.366***	0.283***	0.241**	0.202*	0.249**	1.054***	0.171	0.107	0.113	0.065	0.061
	(5.24)	(3.83)	(2.94)	(2.54)	(1.91)	(2.33)	(4.02)	(1.25)	(0.85)	(0.93)	(0.53)	(0.46)
8	1.224***	0.355***	0.233**	0.150	0.126	0.204*	1.138***	0.235**	0.127	0.076	0.024	0.059
	(4.95)	(3.11)	(2.24)	(1.28)	(1.09)	(1.73)	(4.46)	(2.30)	(1.22)	(0.73)	(0.21)	(0.49)
9	1.476***	0.533***	0.419***	0.238	0.277*	0.395**	1.289***	0.407***	0.287**	0.192	0.191	0.315**
	(4.92)	(2.74)	(2.71)	(1.49)	(1.78)	(2.51)	(5.05)	(2.85)	(2.13)	(1.30)	(1.27)	(2.20)
High	2.261***	1.094***	1.325***	1.010***	1.277***	1.558***	1.425***	0.420**	0.300*	0.102	0.150	0.264
	(5.05)	(3.58)	(5.04)	(3.01)	(4.39)	(4.95)	(4.55)	(2.30)	(1.92)	(0.65)	(0.96)	(1.64)
HML	2.183***	2.056***	1.918***	1.869***	1.867***	1.983***	1.083***	1.083***	0.655***	0.720***	0.548**	0.548**
	(6.37)	(5.68)	(5.66)	(5.29)	(4.86)	(5.30)	(4.28)	(4.06)	(2.79)	(2.93)	(2.23)	(2.27)

Panel B: Value-weighted returns to investment strategies sorted by risk loadings (credit rating sample and downgrades excluded)

	Credit rating sample						Nondowngrades					
Rank of β	Return	CAPM	FFC+ PS	FF5	FF6	SY	Return	CAPM	FFC+ PS	FF5	FF6	SY
Low	0.504	−0.516**	−0.257	−0.579**	−0.371*	−0.205	0.711**	−0.280	−0.030	−0.324	−0.150	0.025
	(1.38)	(−2.39)	(−1.31)	(−2.48)	(−1.79)	(−0.83)	(1.98)	(−1.25)	(−0.14)	(−1.35)	(−0.68)	(0.10)
2	0.853***	−0.063	0.100	−0.052	0.056	0.161	0.965***	0.074	0.222	0.088	0.174	0.289
	(2.85)	(−0.37)	(0.61)	(−0.27)	(0.28)	(0.78)	(3.14)	(0.39)	(1.30)	(0.41)	(0.83)	(1.42)
3	0.928***	0.024	0.149	0.072	0.121	0.209	1.041***	0.145	0.249*	0.235*	0.252*	0.291*
	(3.76)	(0.21)	(1.18)	(0.57)	(0.89)	(1.28)	(4.13)	(1.10)	(1.87)	(1.74)	(1.74)	(1.83)
4	0.757***	−0.091	−0.038	−0.238**	−0.202*	−0.104	0.841***	0.004	0.067	−0.141	−0.107	−0.004
	(3.04)	(−0.79)	(−0.34)	(−2.20)	(−1.90)	(−0.84)	(3.54)	(0.03)	(0.59)	(−1.29)	(−0.99)	(−0.03)
5	0.937***	0.059	0.071	0.012	0.029	0.038	1.034***	0.160	0.166	0.133	0.139	0.128
	(3.77)	(0.49)	(0.57)	(0.10)	(0.22)	(0.29)	(4.09)	(1.22)	(1.20)	(0.95)	(0.97)	(0.89)
6	1.120***	0.249**	0.193	0.219*	0.188	0.104	1.225***	0.361***	0.289**	0.336**	0.283**	0.214
	(4.73)	(2.02)	(1.57)	(1.74)	(1.47)	(0.78)	(5.09)	(2.63)	(2.16)	(2.43)	(2.09)	(1.47)
7	1.144***	0.292**	0.197	0.158	0.108	0.154	1.278***	0.424***	0.337**	0.306**	0.249*	0.287**
	(4.44)	(2.46)	(1.60)	(1.18)	(0.82)	(1.22)	(4.94)	(3.41)	(2.58)	(2.19)	(1.82)	(2.11)
8	1.164***	0.273**	0.158	0.105	0.058	0.123	1.269***	0.393***	0.270*	0.153	0.107	0.157
	(4.63)	(2.27)	(1.28)	(0.85)	(0.43)	(0.90)	(5.12)	(2.87)	(1.90)	(1.05)	(0.69)	(1.05)
9	1.253***	0.377**	0.245*	0.111	0.111	0.196	1.342***	0.483***	0.337**	0.212	0.204	0.291**
	(4.78)	(2.35)	(1.70)	(0.77)	(0.76)	(1.38)	(5.30)	(2.99)	(2.32)	(1.45)	(1.37)	(2.00)
High	1.316***	0.290	0.164	−0.020	0.052	0.134	1.626***	0.689***	0.567***	0.372*	0.423**	0.536***
	(3.81)	(1.35)	(0.88)	(−0.10)	(0.27)	(0.69)	(5.01)	(3.02)	(2.93)	(1.89)	(2.13)	(2.85)
HML	0.812***	0.806***	0.420	0.559*	0.423	0.339	0.915***	0.969***	0.597**	0.696**	0.574*	0.511
	(2.83)	(2.68)	(1.51)	(1.88)	(1.46)	(1.18)	(2.91)	(3.01)	(2.00)	(2.14)	(1.82)	(1.64)

Notes. At the end of each month t, stocks are sorted into deciles according to their risk loadings (β) on the stochastic discount factor estimated from a combination of deep neural networks (Chen et al. 2020). This table reports the value-weighted returns for month $t + 1$ for each decile portfolio as well as the strategy of going long (short) on the highest (lowest) risk loading stocks ("HML") over the entire sample period from 1987 to 2016. Portfolio returns are further adjusted by the CAPM, Fama-French-Carhart four-factor and Pástor-Stambaugh liquidity factor model (FFC+PS), Fama-French five-factor model (FF5), Fama-French six-factor model (FF6), and Stambaugh-Yuan four-factor model (SY). Panel A reports the results for the full sample as well as the subsample that excludes microcaps. Panel B reports similar statistics for the subsamples that exclude nonrated firms and credit rating downgrades. Newey-West adjusted t statistics are shown in parentheses.

*, **, and ***Significant at the 10%, 5%, and 1% levels, respectively.

3.3. Evidence from IPCA and the CA

In this section, we use alternative machine learning signals based on the IPCA approach of KPS and the CA model of Gu et al. (2021). Both IPCA and CA extract latent factors consistent with the notion of principal component analysis (PCA). However, unlike the standard PCA, both IPCA and CA allow factor loadings to vary with predictive characteristics either linearly (IPCA) or nonlinearly through neural networks (CA).[22]

We implement similar portfolio analysis as before. Table 4 reports the results, where Panel A shows the results for portfolios sorted by IPCA-predicted returns and Panel B shows those for portfolios sorted by CA2-predicted returns. For brevity, we present only raw returns and FF6-adjusted returns for the full sample and for subsamples that exclude microcaps, nonrated firms, and credit rating downgrades.[23]

As shown in Panel A, the value-weighted long-short portfolio return (FF6-adjusted return) across all stocks is 0.95% (0.62%) per month based on the IPCA signal, which is 39% (32%) lower than that based on the GKX signal, as shown in Table 1, and 57% (67%) lower than that based on the CPZ signal, as shown in Table 3. More importantly, there is no material deterioration of performance in subsamples with economic restrictions. The long-short strategy based on the IPCA signal generates significant FF6-adjusted returns for all subsamples, ranging between 0.43% and 0.61% per month. Although the IPCA signal underperforms GKX and CPZ for the full sample, it outperforms the GKX signal by 82% and the CPZ signal by 10% based on FF6-adjusted returns across the three subsamples with economic restrictions. Note that IPCA draws on linear dependence between average returns and firm characteristics. In contrast, deep learning models already facilitate nonlinearities. The evidence is thus consistent with the concept that accounting for nonlinearities is especially useful for microcaps and difficult-to-arbitrage stocks.[24]

In Panel B of Table 6, the value-weighted long-short portfolio return (FF6-adjusted return) across all stocks is 1.16% (0.75%) per month based on the CA signal, and the performance weakens considerably in subsamples with economic restrictions. The monthly FF6-adjusted return for the long-short portfolio declines to 0.39% after we exclude microcaps and becomes statistically insignificant after we exclude nonrated firms or distressed firms.

Finally, we conduct two sets of robustness checks. First, to address the potential inconsistency between the in-sample estimations (based on the full sample) and the portfolio sorts (based on subsamples with economic restrictions), we re-estimate the IPCA for each subsample separately. When credit rating data are required for the in-sample estimation, the out-of-sample test begins in 1996. The results are reported in the online appendix, Table IA7, Panel A. Excluding microcaps from the IPCA estimation yields a value-weighted FF6-adjusted return of 0.76% per month, in contrast to 0.61% when IPCA is estimated for the full sample. If we exclude microcaps when training the NN3 model and IPCA estimation, the IPCA signal still outperforms the GKX signal (the corresponding FF6-adjusted return is 0.49%, as shown in Table 2). Moreover, estimating IPCA for the subsample of rated or nondistressed firms considerably reduces the model performance, and all long-short portfolio payoffs are no longer significant at the 5% threshold. Consistent with our early experiment on NN3 estimation (Table 2), restricting the in-sample estimation to a selective set of stocks of interest could adversely affect the out-of-sample return predictability.

Another robustness check is to estimate the unrestricted version of IPCA that allows for intercepts as functions of the instruments. As shown in the online appendix, Table IA7, Panel B, we do not find a material deterioration of performance for subsamples excluding microcaps and nonrated firms, whereas the FF6-adjusted return is no longer statistically significant once we exclude distressed firms.

Collectively, the innovative machine learning (especially deep learning) techniques that we analyze face the usual challenge of cross-sectional return predictability, and anomalous return patterns are concentrated in stocks that are relatively difficult to value and difficult to arbitrage. The trading profits based on deep learning signals in GKX, CPZ, and CA often disappear on a risk-adjusted basis after we impose common economic restrictions in empirical finance, such as excluding microcaps, nonrated firms, or distressed firms. Although IPCA underperforms nonlinear deep learning models for the full sample, it delivers consistent risk-adjusted performance even in the presence of economic restrictions. Despite the challenge of generating sizable trading profits using the existing machine learning tools, all machine learning methods substantially improve the investment payoff compared with the traditional methods with and without economic restrictions.

3.4. Nonnormality, Turnover, and Trading Costs of Machine Learning Portfolios

Beyond out-of-sample return predictability, investors should be concerned with other potential risks and costs in applying investment strategies. Focusing on individual anomalies, past work shows that anomaly payoffs are prone to large drawdowns. Daniel and Moskowitz (2016) document that momentum strategies are characterized by occasional large crashes. A recent work by Arnott et al. (2019) shows that nine out of 14 popular factors are fat-tailed and asymmetric on the

Table 4. Performance of Portfolios Sorted by IPCA and Conditional Autoencoder Predicted Returns

	Full sample		Nonmicrocaps		Credit rating sample		Nondowngrades	
Rank of $\hat{R}$	Return	FF6	Return	FF6	Return	FF6	Return	FF6
Panel A: Value-weighted returns to investment strategies sorted by IPCA-predicted returns								
Low	0.454	−0.331***	0.353	−0.457***	0.357	−0.475***	0.645**	−0.217*
	(1.59)	(−3.54)	(1.17)	(−4.19)	(1.17)	(−3.76)	(2.40)	(−1.72)
2	0.763***	−0.095	0.756***	−0.075	0.782***	−0.131	0.989***	0.082
	(2.79)	(−1.14)	(2.80)	(−0.86)	(2.85)	(−1.35)	(3.87)	(0.84)
3	0.881***	−0.020	0.877***	−0.007	0.888***	−0.044	1.024***	0.094
	(3.71)	(−0.29)	(3.44)	(−0.10)	(3.41)	(−0.47)	(4.30)	(0.95)
4	1.021***	0.109*	0.871***	0.011	0.963***	0.001	1.115***	0.166*
	(4.18)	(1.84)	(3.64)	(0.14)	(4.11)	(0.02)	(5.13)	(1.74)
5	0.949***	0.009	1.057***	0.093	0.964***	0.033	1.071***	0.153*
	(4.25)	(0.12)	(4.42)	(1.34)	(3.98)	(0.44)	(4.72)	(1.91)
6	1.120***	0.124	0.993***	0.050	1.071***	0.097	1.182***	0.209**
	(4.78)	(1.62)	(4.46)	(0.67)	(4.82)	(1.19)	(5.35)	(2.31)
7	1.005***	−0.002	1.112***	0.125	1.093***	0.036	1.221***	0.169**
	(4.17)	(−0.03)	(4.67)	(1.55)	(4.72)	(0.43)	(5.73)	(1.99)
8	1.055***	0.005	1.020***	0.007	0.967***	−0.080	1.037***	−0.019
	(4.57)	(0.07)	(4.22)	(0.11)	(4.13)	(−0.98)	(4.63)	(−0.19)
9	1.298***	0.188*	1.099***	0.030	1.106***	−0.047	1.212***	0.084
	(5.05)	(1.91)	(4.79)	(0.35)	(4.64)	(−0.46)	(5.29)	(0.76)
High	1.399***	0.293**	1.254***	0.157	1.249***	0.132	1.378***	0.214*
	(5.01)	(2.15)	(4.77)	(1.49)	(4.80)	(1.17)	(5.27)	(1.72)
HML	0.945***	0.624***	0.901***	0.613***	0.893***	0.607***	0.733***	0.430**
	(5.62)	(3.31)	(5.57)	(3.71)	(5.08)	(3.38)	(4.08)	(2.37)
Panel B: Value-weighted returns to investment strategies sorted by CA2-predicted returns								
Low	−0.045	−0.506***	0.210	−0.254*	0.309	−0.241	0.683*	0.038
	(−0.11)	(−3.13)	(0.53)	(−1.75)	(0.78)	(−1.42)	(1.79)	(0.18)
2	0.588*	−0.163	0.622**	−0.110	0.673**	−0.177	0.918***	0.100
	(1.94)	(−1.49)	(2.08)	(−0.97)	(2.21)	(−1.33)	(3.40)	(0.81)
3	0.740***	−0.126	0.674***	−0.141	0.820***	−0.047	1.027***	0.182
	(2.99)	(−1.45)	(2.64)	(−1.55)	(3.19)	(−0.43)	(4.30)	(1.52)
4	0.857***	−0.046	0.809***	−0.072	0.795***	−0.136*	0.929***	0.009
	(3.75)	(−0.60)	(3.56)	(−1.11)	(3.36)	(−1.69)	(4.19)	(0.10)
5	1.041***	0.093	0.952***	0.017	0.896***	−0.075	1.010***	0.043
	(4.69)	(1.51)	(4.33)	(0.23)	(4.00)	(−1.19)	(4.74)	(0.62)
6	1.053***	0.085	1.034***	0.092	1.085***	0.069	1.154***	0.155**
	(4.27)	(1.18)	(4.49)	(1.58)	(4.95)	(1.07)	(5.49)	(2.13)
7	1.160***	0.115	1.102***	0.126	1.031***	0.020	1.184***	0.180*
	(4.45)	(1.21)	(4.35)	(1.34)	(4.09)	(0.22)	(4.89)	(1.79)
8	1.207***	0.176*	1.107***	0.066	1.187***	0.117	1.339***	0.276**
	(4.39)	(1.78)	(4.26)	(0.67)	(4.66)	(1.04)	(5.47)	(2.27)
9	1.220***	0.057	1.248***	0.180*	1.217***	0.112	1.335***	0.238*
	(3.89)	(0.50)	(4.44)	(1.68)	(4.32)	(0.88)	(5.04)	(1.81)
High	1.114***	0.241	1.316***	0.133	1.183***	−0.054	1.348***	0.086
	(2.89)	(1.37)	(4.28)	(1.23)	(3.53)	(−0.35)	(4.27)	(0.56)
HML	1.159***	0.746***	1.105***	0.387**	0.874***	0.187	0.665**	0.048
	(4.17)	(3.01)	(4.22)	(2.03)	(2.97)	(0.79)	(2.22)	(0.20)

Notes. In Panel A, at the end of each month t, stocks are sorted into deciles according to their one-month-ahead out-of-sample predicted returns ($\hat{R}$) using instrumented principal component analysis (IPCA) (Kelly et al. 2019). We report the value-weighted returns for month $t+1$ for each decile portfolio as well as the strategy of going long (short) on the highest (lowest) expected return stocks (HML) over the entire sample period from 1987 to 2017. Portfolio returns are further adjusted by the Fama-French six-factor model (FF6). We report results for the full sample as well as for subsamples that exclude microcaps, nonrated firms, and credit rating downgrades. Panel B reports similar statistics when decile portfolios are sorted by the one-month-ahead out-of-sample predicted returns ($\hat{R}$) using a conditional autoencoder model with two hidden layers and five latent factors (CA2) (Gu et al. 2021). Both IPCA and CA2 impose a zero-alpha restriction. Newey-West adjusted t statistics are shown in parentheses.

*, **, and ***Significant at the 10%, 5%, and 1% levels, respectively.

downside. In addition, transaction costs could significantly reduce the anomaly payoff. Novy-Marx and Velikov (2016) find that most low-turnover and mid-turnover strategies remain profitable after accounting for transaction costs, while no high-turnover strategy achieves significantly positive net excess return.

In this section, we explore the downside risk and turnover associated with machine learning portfolios.

As described before, we sort stocks into decile portfolios according to four machine learning signals. We compute the value-weighted holding period return for each decile portfolio and implement a zero-cost trading strategy by taking long positions in the top decile of stocks and selling short stocks in the bottom decile. We also include the market portfolio as a benchmark, and market excess return is defined as value-weighted CRSP market return in excess of the one-month T-bill rate. We report the annualized Sharpe ratio, the skewness and excess kurtosis of the monthly returns,[25] the maximum drawdown, the average monthly return during the crisis period, and the monthly turnover for the long-short machine learning portfolios and market portfolio.

In particular, we follow GKX to define the maximum drawdown of a strategy as $MDD = \max_{0 \le t_1 \le t_2 \le T} (Y_{t_1} - Y_{t_2})$, where Y_{t_1} and Y_{t_2} refer to the cumulative log return from month 0 (i.e., January 1987) to t_1 and t_2, respectively. In addition, the crisis period includes the market crash in October 1987, the Russian default in August 1998, the bursting of the tech bubble in April 2000, the Quant crisis in August 2007, and the Bear Stearns bailout, and Lehman bankruptcy during the recent financial crisis, that is, March, September, and October 2008 (Griffin et al. 2011, Cella et al. 2013).

We also follow GKX to define the turnover in month t as $TO_t = \frac{1}{2}\sum_{i\in L}\left|w_{i,t} - \frac{w_{i,t-1}(1+r_{i,t})}{\sum_i w_{i,t-1}(1+r_{i,t})}\right| + \frac{1}{2}\sum_{j\in S}\left|w_{j,t} - \frac{w_{j,t-1}(1+r_{j,t})}{\sum_j w_{j,t-1}(1+r_{j,t})}\right|$, where $i \in L$ ($j \in S$) indicates that stock i (j) belongs to the entire universe of long positions (short positions) in months and $t-1$, $w_{i,t}$ ($w_{j,t}$) refers to the weight of stock i (j) in the portfolio in month t, and $r_{i,t}$ ($r_{j,t}$) refers to the return of stock i (j) in month t. By construction, the one-side turnover for long positions or short positions ranges between zero and one, and the turnover in the long-short portfolio, that is, TO_t, ranges between zero and two.

We tabulate the results in Table 5, where Panel A shows the results for portfolios sorted by the NN3-predicted returns (GKX), Panel B shows those for portfolios sorted by the risk loadings on the SDF (CPZ), Panel C shows those for portfolios sorted by the IPCA-predicted returns (KPS), Panel D shows those for portfolios sorted by the CA2-predicted returns (Gu et al. 2021), and Panel E shows those for the market portfolio. We report results for the full sample and subsamples that exclude microcaps, nonrated firms, and credit rating downgrades. Following Arnott et al. (2019), all returns are scaled to 10% volatility per year to facilitate comparisons across various samples and methods.

First, machine learning portfolios earn an annualized Sharpe ratio ranging between 0.78 and 1.23 for the full sample compared with 0.53 for the market portfolio. Imposing economic restrictions significantly reduces the Sharpe ratio, although machine learning methods still outperform the market in most subsamples except for the one that excludes credit rating downgrades.

Second, both the GKX and CPZ methods display positive skewness and excess kurtosis for the full sample and all three subsamples, whereas the market portfolio is negatively skewed. Additionally, both IPCA and CA methods exhibit negative skewness for the full sample and most subsamples, although they are less negatively skewed than the market portfolio.

Third, machine learning methods can mitigate downside risk and protect investors from extreme crashes. The maximum drawdown for machine learning portfolios ranges between 20% and 35% for the full sample, whereas the market portfolio experiences a larger drawdown at 49%. Among most subsamples, machine learning methods also experience comparatively smaller drawdowns than the market portfolio. In addition, the average monthly returns on machine learning-based trading strategies are mostly positive during the crisis period, that is, 2.93% to 4.10% for the GKX method, −0.02% to 0.90% for the CPZ method, −0.64% to 1.49% for the IPCA method, and −1.80% to −0.05% for the CA method. All machine learning methods exhibit a significant improvement from the average market return of nearly −7% contemporaneously.[26]

Finally, all machine learning methods require high turnover in portfolio rebalancing. The monthly turnover in the long-short portfolio ranges between 87% and 98% for the GKX method, between 163% and 168% for the CPZ method, between 113% and 119% for the IPCA method, and between 148% and 157% for the CA method.[27] This creates a one-side turnover (average turnover on the long and short sides) of at least 43% for the GKX method, 81% for the CPZ method, 56% for the IPCA method, and 74% for the CA method. To put this in perspective, low-turnover strategies such as size and value typically display monthly one-side turnover of below 10%; the corresponding number is between 14% and 35% for mid-turnover strategies such as failure probability and idiosyncratic volatility and above 90% for high-turnover strategies such as short-run reversals and seasonality (Novy-Marx and Velikov 2016). Recall that the GKX (CPZ, IPCA, CA) method generates a value-weighted monthly FF6-adjusted return of 0.92% (1.87%, 0.62%, 0.75%) for the full sample and 0.31% (0.55%, 0.61%, 0.39%) after excluding microcaps; we can therefore infer a break-even transaction cost of 0.94% (1.12%, 0.53%, 0.48%) for the long-short portfolio in the full sample and 0.36% (0.34%, 0.54%, 0.26%) for those in the subsample

Table 5. Nonnormality and Turnover of Machine Learning Portfolios

	Characteristics of value-weighted machine learning portfolios					
	Sharpe ratio	Skewness	Excess kurtosis	Maximum drawdown	Return in crisis	Turnover
			Panel A: Sorted by NN3-predicted returns			
Full sample	0.944	0.631	5.222	0.350	4.100	0.976
Nonmicrocaps	0.644	0.361	7.062	0.349	3.563	0.869
Credit rating sample	0.639	0.064	7.875	0.420	3.435	0.889
Nondowngrades	0.449	0.146	8.550	0.333	2.931	0.920
			Panel B: Sorted by risk loadings			
Full sample	1.225	1.063	5.932	0.209	0.472	1.664
Nonmicrocaps	0.839	0.326	1.582	0.246	0.677	1.625
Credit rating sample	0.566	0.267	1.440	0.407	−0.023	1.652
Nondowngrades	0.602	0.344	1.675	0.447	0.903	1.678
			Panel C: Sorted by IPCA-predicted returns			
Full sample	0.967	−0.449	4.805	0.203	0.574	1.186
Nonmicrocaps	0.978	−0.267	5.369	0.234	1.493	1.130
Credit rating sample	0.880	−0.219	4.221	0.315	0.474	1.164
Nondowngrades	0.697	−0.069	3.158	0.349	−0.640	1.184
			Panel D: Sorted by CA2-predicted returns			
Full sample	0.784	−0.077	2.418	0.202	−0.047	1.565
Nonmicrocaps	0.748	0.291	4.684	0.207	−0.529	1.478
Credit rating sample	0.522	−0.471	4.119	0.252	−1.796	1.542
Nondowngrades	0.387	−0.616	4.930	0.345	−0.167	1.571
			Panel E: Market portfolio			
Full sample	0.527	−0.978	3.323	0.486	−6.954	0.089
Nonmicrocaps	0.530	−0.959	3.222	0.485	−6.907	0.086
Credit rating sample	0.543	−0.932	3.423	0.498	−6.747	0.080
Nondowngrades	0.682	−0.856	3.311	0.408	−6.615	0.084

Notes. In Panel A, at the end of each month t, stocks are sorted into deciles according to their NN3-predicted returns (Gu et al. 2020). We compute the value-weighted returns for month $t+1$ for the strategy of going long (short) on the highest (lowest) expected return stocks over the entire sample period from 1987 to 2017. For the long-short strategy, we report the annualized Sharpe ratio, the skewness and excess kurtosis of the monthly returns, the maximum drawdown, the average monthly return during the crisis period, and the monthly turnover. We report the results for the full sample as well as subsamples that exclude microcaps, nonrated firms, and credit rating downgrades. Panels B to D report similar statistics when decile portfolios are sorted by the risk loadings on the stochastic discount factor (Chen et al. 2020), IPCA-predicted returns (Kelly et al. 2019), and CA2-predicted returns (Gu et al. 2021), respectively. Panel E reports similar statistics on the value-weighted market portfolio in excess of the one-month T-bill rate. All returns on the long-short strategy and market index are scaled to 10% volatility per year.

excluding microcaps to completely absorb the FF6-adjusted return in the GKX (CPZ, IPCA, CA) method, respectively.[28]

Such break-even transaction cost estimates seem moderate relative to those in the existing literature. Novy-Marx and Velikov (2016) focus on value-weighted portfolios sorted by the signals using NYSE breakpoints between 1963 and 2013 and suggest that transaction costs account for more than 1% of the monthly one-sided turnover (equivalent to 0.5% of the long-short portfolio turnover as in our estimates) and that the statistical significance of the return spread decreases proportionately. Because Novy-Marx and Velikov (2016) sort portfolios based on NYSE breakpoints, the long-short portfolios are less likely to be dominated by small firms, and their estimates could be more relevant for our subsample excluding microcaps.[29]

Alternatively, several papers argue that investors face proportional transaction costs that decrease with firm size and over time (Brandt et al. 2009, Hand and Green 2011). We follow Brandt et al. (2009) and define the one-way transaction cost of stock i in month t as $c_{i,t} = z_{i,t} \times T_t$, where $z_{i,t} = 0.006 - 0.0025 \times ME_{i,t}$, $ME_{i,t}$ refers to the market capitalization of stock i in month t and is normalized to be between zero and one. T_t decreases linearly from 2.6 in 1987 to 1 in 2002 and remains at 1 thereafter.[30] As a result, the transaction cost is approximately 1.56% for the smallest firms and 0.91% for the largest firms in 1987 and approximately 0.6% for the smallest firms and 0.35% for the largest firms after 2002. This indicates an average transaction cost of 0.67% for the full sample and 0.64% for the subsample excluding microcaps over the entire sample period from 1987 to 2017.

For all machine learning methods, our break-even transaction costs, which range between 0.48% and 1.12% for the full sample and between 0.26% and 0.54% for the subsample excluding microcaps, appear to be mostly below or approximately the same as the estimated costs, that is, 0.5% to 0.67%. Therefore, accounting for reasonable transaction costs would make it difficult for most machine learning signals to leave alpha on the table. However, our findings do not imply that machine learning-based trading strategies are unprofitable for *all* traders after accounting for transaction costs. Instead, we show that it is challenging for an average investor to achieve statistically and economically meaningful risk-adjusted performance in the presence of reasonable transaction costs. Investors may thus need to adjust their expectations of the actual investment returns they can receive. Although we explicitly focus on the universe of cheap-to-trade stocks to assess economic significance, from the modeling perspective, investors who are sensitive to transaction costs could modify the existing off-the-shelf machine learning models and construct portfolios under constraints, hence optimizing after-trading-cost performance. Potentially promising examples are presented in Bryzgalova et al. (2020), Chen et al. (2020), Allena (2021), and Cong et al. (2021). Their findings reinforce the notion that machine learning signals based on the entire sample do not sufficiently characterize the investment universe beyond difficult-to-arbitrage stocks; thus, accounting for trading costs in the optimization improves performance.

3.5. Performance and Weights of the SDF-Implied Tangency Portfolio

In this section, we consider the method proposed by KNS. Similar to CPZ, the KNS approach incorporates a no-arbitrage condition to estimate the SDF. Although CPZ estimate the SDF using individual stocks, KNS focus on equity portfolios that represent characteristics-based trading strategies. We first form portfolios based on the 94 predictive characteristics employed by GKX.[31] We follow KNS and perform a rank transformation for each characteristic and normalize each rank-transformed characteristic. We then construct long-short portfolios and compute characteristics-weighted portfolio returns. We rely on daily returns and split the sample into two subperiods. The first period, from September 1964 to December 2004, is used for in-sample estimation.[32] The second period, from January 2005 to December 2017, establishes the out-of-sample testing period.

Focusing on the first period, we estimate market loadings and then orthogonalize portfolio returns with respect to the market. We next estimate SDF slope coefficients for each characteristic using a ridge regression. That is, we minimize the Hansen-Jagannathan distance (Hansen and Jagannathan 1991) measure subject to an L^2-penalty, whereas the penalty parameter is chosen by a three-fold cross-validation method. To incorporate our proposed economic restrictions, we separately estimate the SDF for the full sample and three subsamples that exclude microcaps, nonrated firms, and financially distressed firms.[33] Because SDF slope coefficients indicate weights of the MVE portfolio, we use SDF coefficients estimated from the pre-2005 sample to compute the implied out-of-sample MVE portfolio return (orthogonalized with respect to the market portfolio). Portfolio returns are *rescaled* to have standard deviations equal to the in-sample standard deviation of the excess return on the aggregate market index, allowing us to quantify the investment weights in the tangency portfolio conditional on a predetermined level of volatility.[34]

The SDF-implied MVE portfolio return is already CAPM-adjusted because the portfolio is orthogonal to the market. In addition, the FF6-adjusted return is estimated by regressing SDF-implied MVE portfolio returns on benchmark portfolio returns, where the benchmark portfolio return is estimated from unregularized MVE portfolio weights based on five nonmarket factors in the pre-2005 period. We also report the annualized Sharpe ratio and the quantile distribution of MVE portfolio weights (i.e., SDF slope coefficients) across the 94 characteristics.

The results are tabulated in Table 6. First and foremost, the evidence in KNS clearly applies to our universe of 94 test assets. The performance of the SDF-implied MVE portfolio is notable, with an annual Sharpe ratio of 2.32 and both CAPM-adjusted and FF6-adjusted returns exceeding 3% monthly for the full sample. Next, we confirm our finding that imposing economic restrictions reduces the performance of machine learning methods. For instance, the SDF-implied MVE portfolio yields an FF6-adjusted return of 0.90% per month after excluding microcaps, and the FF6-adjusted return is no longer significant at the 5% level after excluding nonrated firms or distressed firms. The annualized Sharpe ratio also declines to 0.98 (0.90, 0.83) after we exclude microcaps (nonrated firms, distressed firms).

Notably, portfolio weights (or pricing kernel slope coefficients) display high dispersion across the 94 predictors and often exhibit extreme, possibly infeasible, values. To achieve the same volatility as the market, the SDF-implied MVE portfolio requires taking a −199% (−91%) short position for an individual predictor at the 10th (25th) percentile and a 169% (96%) long position at the 90th (75th) percentile. Portfolio positions are obviously more extreme at the 5th and 95th percentiles, rising to −234% and 190%, respectively.[35] Thus, the pricing kernel–based optimal portfolio may be an inadmissible investment from a practical perspective. However, imposing economic restrictions significantly lowers the odds of extreme positions. For instance, the SDF-implied MVE portfolio requires taking a −24% (−14%, −22%) short position for an individual predictor at the 25th percentile and a 41% (19%, 14%) long position at the 75th percentile after microcaps (nonrated firms, distressed firms) are excluded, respectively.

Table 6. Performance and Weights of SDF-Implied Mean-Variance Efficient Portfolios

	Characteristics of SDF-implied MVE portfolios													
				SDF-implied MVE portfolio weights										
	CAPM	FF6	Sharpe ratio	Mean	Standard deviation	Minimum	5%	10%	25%	Median	75%	90%	95%	Maximum
Full sample	3.662*** (6.01)	3.338*** (5.90)	2.318	0.083	1.338	−2.994	−2.343	−1.994	−0.912	0.341	0.964	1.687	1.895	3.182
Nonmicrocaps	1.543*** (3.88)	0.895*** (2.87)	0.977	0.084	0.447	−0.863	−0.666	−0.592	−0.238	0.072	0.407	0.647	0.741	1.431
Credit rating sample	1.418*** (2.97)	0.717* (1.93)	0.898	−0.006	0.254	−0.678	−0.473	−0.382	−0.137	−0.003	0.187	0.326	0.387	0.419
Nondowngrades	1.308*** (2.92)	0.545 (1.59)	0.828	−0.022	0.248	−0.543	−0.448	−0.370	−0.217	0.004	0.135	0.293	0.376	0.595

Notes. We follow Kozak et al. (2020) to construct the SDF-implied MVE portfolio based on ridge regression in the pre-2005 sample and estimate the out-of-sample performance in the 2005 to 2017 period. We report the monthly abnormal returns adjusted by CAPM and the Fama-French six-factor model (FF6). CAPM-adjusted returns equal the SDF-implied MVE portfolio returns because the portfolio is orthogonal to the market. FF6-adjusted returns are estimated by regressing the SDF-implied MVE portfolio returns on benchmark portfolio returns, where the benchmark portfolio returns are estimated from unregularized MVE portfolio weights and five nonmarket factors in the pre-2005 period. We also report the annualized Sharpe ratio and the quantile distribution of the SDF-implied MVE portfolio weights. We report the results for the full sample and subsamples that exclude microcaps, nonrated firms, and credit rating downgrades.

*, **, and ***Significant at the 10%, 5%, and 1% levels, respectively.

The collective evidence suggests that the machine learning (especially deep learning) techniques we examine, face the usual challenge of cross-sectional return predictability. We analyze the out-of-sample trading profits using predictive signals generated from five machine learning methods advocated by the recent literature, including three deep learning signals (i.e., GKX, CPZ, CA) and the IPCA and KNS methodologies. We consider both the SDF methodology adopted by CPZ (stock-level) and KNS (portfolio-level) as well as beta pricing formulations per IPCA and CA. Once we apply NYSE breakpoints to exclude microcaps, the value-weighted FF6-adjusted return is 66% (71%, 48%) lower than that for the full sample based on the GKX (CPZ, CA) signal. Similarly, excluding distressed firms, the value-weighted FF6-adjusted return is 78% (69%, 94%) lower than that for the full sample based on the GKX (CPZ, CA) signal. In addition, the value-weighted FF6-adjusted return is significant in only one of three subsamples (i.e., excluding microcaps, nonrated firms, and distressed firms) at the 5% threshold for all three deep learning signals. Moreover, IPCA underperforms deep learning models for the full sample but does not display a material deterioration of performance in subsamples with economic restrictions. As machine learning-based trading strategies require relatively high portfolio turnover and take extreme long-short positions in the SDF-implied tangency portfolio, investors may need to further lower their expectations of achievable performance.

Our findings suggest that economic restrictions play an important role in assessing whether newly proposed machine learning methods are effective and exploitable in real time, especially for investors who are sensitive to transaction costs and prefer to avoid microcaps and distressed firms in portfolio management. In line with Arnott et al. (2018) and Karolyi and Van Nieuwerburgh (2020), our findings highlight the importance of adopting back-testing protocols (such as economic restrictions, in our context) in evaluating machine learning methods and verifying their external validity when applying to other settings, for example, different universes of stocks, markets, asset classes, and sample periods.

4. Time-Varying Return Predictability

The previous results demonstrate that the cross-sectional return predictability of machine learning signals deteriorates among relatively cheap-to-trade stocks. We investigate whether return predictability through machine learning methods varies over time. Specifically, we first link trading profits to variations in market conditions and then examine predictable patterns over the most recent years.

4.1. Analysis Based on Portfolio Returns

Economic theory implies that less trading friction and more arbitrage activity should improve price efficiency. Consequently, anomaly-based trading strategies should be more profitable in the presence of binding limits to arbitrage. The existing evidence concerning many anomalies has typically supported this expectation. First, Stambaugh et al. (2012) find stronger market anomalies following high-sentiment periods. They attribute the sentiment effect to binding short-sale constraints, which are particularly prominent during episodes of high investor sentiment, because stock prices reflect the views of more optimistic investors in the presence of heterogeneous beliefs about fundamental values (Miller 1977). Avramov et al. (2018) further show that market-wide sentiment and firm-level financial distress jointly drive overpricing among stocks and corporate bonds. Second, a great deal of theoretical work predicts that higher volatility reduces the liquidity-provision capacity of market makers because of tightened funding constraints and reduced risk appetite (Gromb and Vayanos 2002, Brunnermeier and Pedersen 2009, Adrian and Shin 2010). Thus, anomaly payoffs (especially those related to liquidity provision) could increase because of liquidity dry-up during times of financial market turmoil. Finally, Chordia et al. (2014) find that the recent regime of increased stock market liquidity is associated with the attenuation of equity return anomalies because of increased arbitrage activity.

In our empirical experiments, we examine the payoff of machine learning portfolios in subperiods depending on the state of investor sentiment, market volatility, and market liquidity. We consider the following market state variables: (1) investor sentiment, defined as the monthly investor sentiment from Baker and Wurgler (2007) (*SENT*), or the aligned investor sentiment using the partial least squares approach from Huang et al. (2015) (*PLS SENT*)[36]; (2) realized market volatility (*MKTVOL*), defined as the standard deviation of daily CRSP value-weighted index returns in a month; (3) implied market volatility (*VIX*), defined as the monthly VIX index of implied volatilities of S&P 500 index options[37]; and (4) market illiquidity (*MKTILLIQ*), defined as the value-weighted average of stock-level Amihud (2002) illiquidity for all NYSE/AMEX stocks in a month (Avramov et al. 2016). We divide the full sample into two subperiods, that is, high versus low investor sentiment (aligned investor sentiment, realized market volatility, implied market volatility), according to the median breakpoint of *SENT* (*PLS SENT*, *MKTVOL*, *VIX*) over the entire sample period. Unlike other market state variables, we obtain the median breakpoint for market illiquidity in the pre- and post-2001 periods separately, as the decimalization in January 2001 considerably reduced trading costs.

We repeat the portfolio analysis described before and sort stocks into decile portfolios according to machine learning signals. We compute the value-weighted holding period return for each decile portfolio and implement a zero-cost trading strategy by taking long positions in the top decile of stocks and shorting stocks in the bottom decile. We report the results for the full sample and subsamples that exclude microcaps, nonrated firms, and credit rating downgrades. For brevity, we present only FF6-adjusted returns in the long-short trading strategy, whereas our findings are robust to alternative performance measures, such as raw returns and various risk-adjusted returns, which are documented previously.

We tabulate the results in Table 7, where Panel A shows the results for portfolios sorted by the NN3-predicted returns (GKX), Panel B shows those for portfolios sorted by the risk loadings on the SDF (CPZ), Panel C shows those for portfolios sorted by the IPCA-predicted returns (KPS), and Panel D shows those for portfolios sorted by the CA2-predicted returns (Gu et al. 2021). Starting with Panel A, several findings are worth noting. First, the long-short trading profit across all stocks is significant at the 5% level in all subperiods except for the low-*VIX* period. The investment strategy is also more profitable during periods of high investor sentiment, high market volatility, and low market liquidity. Second, among all market state variables, both realized and implied market volatility play an important role in explaining time-varying return predictability. The value-weighted FF6-adjusted return is 0.64% (statistically insignificant at 0.22%) per month at times of low *MKTVOL* (*VIX*) and dramatically increases to 1.28% (1.66%) at times of high *MKTVOL* (*VIX*), while the full sample average is 0.92% from Table 1, Panel A. Third, if we focus on the subsamples excluding microcaps or nonrated firms, the investment strategy remains more profitable during periods of high market volatility (in terms of both realized and implied volatility). Finally, considering the subsample excluding credit rating downgrades, none of the subperiods displays significant long-short trading profit at the 5% level. This last finding reinforces the concept that after excluding distressed firms around credit rating downgrades, the GKX signal does not deliver meaningful FF6-adjusted returns, that is, the full sample average is statistically insignificant at 0.20% per month (Table 1, Panel B).

In Panel B, decile portfolios are sorted by risk loadings on the SDF (CPZ). For the full sample, we observe significant long-short trading profit at the 5% level in all subperiods, and the investment strategy outperforms during periods of high investor sentiment, high market volatility (in terms of both realized and implied volatility), and low market liquidity. In line with the full sample results shown in Table 3, return predictability deteriorates upon imposing economic restrictions, and only two subperiods display significant long-short trading profit at the 5% level.

In Panel C, the trading strategy based on the IPCA signal remains profitable at the 5% level for the full sample as well as for the subsample excluding microcaps in all subperiods except for the low-*PLS SENT* period. More importantly, the trading profit based on IPCA does not change materially during high limits-to-arbitrage market states. For instance, the monthly value-weighted FF6-adjusted return is 0.51% (0.70%) for the full sample and 0.56% (0.62%) for the subsample excluding microcaps at times of low (high) limits to arbitrage (proxied by investor sentiment, market

Table 7. Performance of Machine Learning Portfolios by Market State

	Value-weighted FF6-adjusted returns of machine learning portfolios									
	SENT		PLS SENT		MKTVOL		VIX		MKTILLIQ	
	Low	High	Low	High	Low	High	Low	High	Low	High
Panel A: Sorted by NN3-predicted returns										
Full sample	0.732***	0.879**	0.684**	1.041***	0.641**	1.283***	0.218	1.662***	0.736***	1.075***
	(3.00)	(2.40)	(2.27)	(2.82)	(2.34)	(3.96)	(0.95)	(4.16)	(2.70)	(3.36)
Nonmicrocaps	0.334	0.095	0.250	0.265	0.062	0.814***	−0.050	0.747**	0.222	0.461
	(1.45)	(0.29)	(0.89)	(0.82)	(0.26)	(2.87)	(−0.20)	(2.03)	(0.84)	(1.51)
Credit rating sample	0.389	0.274	0.335	0.452	0.095	0.896***	0.025	0.808*	0.512**	0.351
	(1.49)	(0.91)	(1.19)	(1.45)	(0.44)	(2.93)	(0.11)	(1.96)	(1.97)	(1.15)
Nondowngrades	0.355	−0.161	0.225	0.111	0.054	0.597*	−0.060	0.582	0.294	0.120
	(1.43)	(−0.50)	(0.81)	(0.33)	(0.23)	(1.70)	(−0.24)	(1.25)	(1.21)	(0.36)
Panel B: Sorted by risk loadings										
Full sample	1.372***	2.453***	1.383***	2.198***	1.352***	2.364***	0.898**	2.451***	1.150***	2.128***
	(3.83)	(3.78)	(3.50)	(3.87)	(3.64)	(4.09)	(2.30)	(3.80)	(2.73)	(4.53)
Nonmicrocaps	0.512**	0.681	0.151	0.824**	0.423	0.665	0.321	0.386	0.527	0.432*
	(2.14)	(1.60)	(0.54)	(2.44)	(1.41)	(1.62)	(1.14)	(1.01)	(1.45)	(1.76)
Credit rating sample	0.269	0.658	0.244	0.500	0.448	0.527	0.097	0.337	0.691	0.073
	(0.84)	(1.35)	(0.61)	(1.29)	(1.12)	(1.12)	(0.31)	(0.69)	(1.63)	(0.20)
Nondowngrades	0.329	0.945*	0.411	0.685*	0.453	0.699	0.031	0.651	0.847*	0.155
	(0.93)	(1.79)	(0.87)	(1.66)	(0.96)	(1.45)	(0.09)	(1.19)	(1.86)	(0.39)
Panel C: Sorted by IPCA-predicted returns										
Full sample	0.558**	0.667**	0.363*	0.753***	0.590***	0.694**	0.440**	0.763**	0.603**	0.645***
	(2.46)	(2.03)	(1.68)	(2.64)	(3.57)	(2.56)	(2.60)	(2.23)	(2.39)	(2.78)
Nonmicrocaps	0.606***	0.615**	0.387*	0.742***	0.566***	0.636**	0.478***	0.647**	0.759***	0.463**
	(3.06)	(2.15)	(1.96)	(2.94)	(3.28)	(2.59)	(2.98)	(2.25)	(3.34)	(2.23)
Credit rating sample	0.647***	0.512*	0.466**	0.667**	0.533***	0.653**	0.498**	0.649**	0.782***	0.447*
	(2.82)	(1.66)	(2.08)	(2.39)	(2.97)	(2.51)	(2.56)	(2.14)	(3.18)	(1.96)
Nondowngrades	0.535**	0.326	0.386	0.429	0.378**	0.451	0.478**	0.325	0.583**	0.292
	(2.37)	(1.10)	(1.64)	(1.55)	(2.04)	(1.60)	(2.23)	(0.96)	(2.22)	(1.26)
Panel D: Sorted by CA2-predicted returns										
Full sample	0.748***	0.617	0.673**	0.869**	0.634**	0.606	0.453*	0.910**	0.359	0.787**
	(2.70)	(1.41)	(2.41)	(2.17)	(2.34)	(1.61)	(1.75)	(2.04)	(1.51)	(2.14)
Nonmicrocaps	0.584**	0.064	0.224	0.471	0.537**	0.254	0.524**	0.122	0.292	0.275
	(2.55)	(0.21)	(0.89)	(1.50)	(2.13)	(0.79)	(2.28)	(0.33)	(1.13)	(1.03)
Credit rating sample	0.304	−0.055	0.016	0.287	0.531	−0.074	0.508*	−0.300	−0.102	0.263
	(1.15)	(−0.15)	(0.05)	(0.78)	(1.49)	(−0.19)	(1.69)	(−0.67)	(−0.30)	(0.81)
Nondowngrades	0.204	−0.218	−0.173	0.153	0.376	−0.170	0.193	−0.248	−0.331	0.245
	(0.73)	(−0.54)	(−0.55)	(0.40)	(1.30)	(−0.45)	(0.73)	(−0.54)	(−0.89)	(0.84)

Notes. In Panel A, at the end of each month t, stocks are sorted into deciles according to their NN3-predicted returns (Gu et al. 2020). We report the value-weighted Fama-French six-factor-adjusted returns for month $t+1$ for the strategy of going long (short) on the highest (lowest) expected return stocks in various subperiods, including when investor sentiment (SENT), aligned investor sentiment (PLS SENT), market volatility (MKTVOL), implied market volatility (VIX) and market illiquidity (MKTILLIQ) are high (above median) and low (below median) in month t. We obtain the median breakpoint for market illiquidity in the pre- and post-2001 periods separately. We report the results for the full sample and for subsamples that exclude microcaps, nonrated firms, and credit rating downgrades. Panels B to D report similar statistics when decile portfolios are sorted by the risk loadings on the stochastic discount factor (Chen et al. 2020), IPCA-predicted returns (Kelly et al. 2019), and CA2-predicted returns (Gu et al. 2021), respectively. Online Appendix A provides detailed definitions for each variable. Newey-West adjusted t statistics are shown in parentheses.

*, **, and ***Significant at the 10%, 5%, and 1% levels, respectively.

volatility, and market liquidity). Shifting the focus to subsamples excluding nonrated firms and credit rating downgrades, the IPCA signal outperforms in market states with low limits to arbitrage. The monthly value-weighted FF6-adjusted return is 0.59% (0.59%) when we exclude nonrated firms and 0.47% (statistically insignificant at 0.37%) when we exclude distressed firms around credit rating downgrades at times of low (high) limits to arbitrage.

In Panel D, unlike the GKX and CPZ methods, the CA signal shows mixed evidence across market states for the full sample. Considering the subsample excluding microcaps, the CA signal generates higher profit during periods of low limits to arbitrage. Unreported results further suggest that machine learning portfolios display similar turnover across various market states.

Overall, we find that GKX and CPZ signals predict cross-sectional returns across all stocks, especially during periods of high investor sentiment, high market volatility, and low market liquidity—consistent with the economic concept of limits to arbitrage. In contrast, the machine learning signals generated from conditional beta pricing models, that is, IPCA and CA models, display low time series variation in trading profits and could even outperform in low limits-to-arbitrage market states. However, restricting the investment universe to cheap-to-trade stocks attenuates return predictability across all market states, and the payoff of deep learning signals (GKX, CPZ, and CA) often becomes statistically insignificant on a risk-adjusted basis over the entire sample period and in various market states.

4.2. Time Series Regressions

We next perform regression analysis to jointly consider all market state variables. We also explicitly control for other proxies for market states and macroeconomic conditions. Because the return predictability of deep learning signals weakens considerably in subsamples with economic restrictions, we focus on the full sample including all stocks to conduct time series analysis.

First, we examine whether the payoff of machine learning portfolios varies by market state. Specifically, we estimate the following monthly time series regression:

$$\begin{aligned} HML_t = {} & \alpha_0 + \beta_1 High\ SENT_{t-1} + \beta_2 High\ MKTVOL_{t-1} \\ & + \beta_3 High\ MKTILLIQ_{t-1} + \beta_4 M_{t-1} + c_1' F_t + c_2' F_t \\ & \times High\ SENT_{t-1} + c_3' F_t \times High\ MKTVOL_{t-1} \\ & + c_4' F_t \times High\ MKTILLIQ_{t-1} + e_t \end{aligned} \tag{1}$$

where HML_t refers to the value-weighted return on the high minus low decile in machine learning portfolios across all stocks in month t. $High\ SENT_{t-1}$ refers to a dummy variable that takes a value of one if the Baker and Wurgler (2007) investor sentiment (*SENT*) is above the median over the entire sample period and zero otherwise; $High\ SENT_{t-1}$ is further replaced with $High\ PLS\ SENT_{t-1}$, defined as a dummy variable that takes a value of one if the aligned investor sentiment from Huang et al. (2015) is above the median over the entire sample period and zero otherwise; $High\ MKTVOL_{t-1}$ refers to a dummy variable that takes a value of one if the market volatility (*MKTVOL*) is above the median over the entire sample period and zero otherwise; $High\ MKTVOL_{t-1}$ is further replaced with $High\ VIX_{t-1}$, defined as a dummy variable that takes a value of one if the implied market volatility (*VIX*) is above the median over the entire sample period and zero otherwise; and $High\ MKTILLIQ_{t-1}$ refers to a dummy variable that takes a value of one if market illiquidity (*MKTILLIQ*) is above the median and zero otherwise. We obtain the median breakpoint for market illiquidity in the pre- and post-2001 periods separately. All market state variables are defined as in Table 7; M_{t-1} refers to a set of other proxies for market conditions: down market state (*DOWN*), defined as a dummy variable that takes the value of one if the CRSP value-weighted index return is negative and zero otherwise; term spread (*TERM*), defined as the difference between the average yield of 10-year Treasury bonds and three-month T-bills; and default spread (*DEF*), defined as the difference between the average yield of bonds rated BAA and AAA by Moody's. The vector F stacks the six common risk factors identified in Fama and French (2018), including the market factor (MKT), the size factor (SMB), the book-to-market factor (HML), the profitability factor (RMW), the investment factor (CMA), and the momentum factor (MOM). We also report Newey and West (1987) adjusted t statistics with four lags.

The results are presented in the online appendix, Table IA8. In Panels A and B, decile portfolios are sorted by the NN3-predicted returns (GKX) and the risk loadings on the SDF (CPZ), respectively. We find that the trading profit based on the GKX signal is higher during periods of high market volatility in terms of both realized and implied volatility after controlling for investor sentiment, market liquidity, macroeconomic variables, and risk factors (models 1 to 4). Models 5 to 8 further consider the time-varying factor risk exposures by interacting the six common risk factors with market state variables. We continue to find that the trading profit based on the GKX signal is higher during high-*VIX* periods. Similarly, the long-short portfolio payoff based on the CPZ signal is higher during periods of high investor sentiment, high implied market volatility, and low market liquidity (models 1 to 4). In addition, the implied market volatility appears to be the most informative market state variable and is highly significant in all specifications after controlling for macroeconomic variables and time-varying risk exposures. Overall, market-wide predictive variables capture a large variation in investment payoffs based on GKX and CPZ signals.

In Panels C and D, decile portfolios are sorted by the IPCA-predicted returns (KPS) and the CA2-predicted returns (Gu et al. 2021), respectively. The findings are in line with the portfolio analysis shown in Table 7, that is, the trading profits generated from conditional beta pricing models do not exhibit significant time series variation across market states after controlling for macroeconomic variables and risk factors as well as time-varying risk exposures.

Overall, this joint predictive regression specification further reinforces our finding that machine learning payoffs based on GKX and CPZ signals display substantial time variation and, more importantly, that superior performance characterizes market states that are associated with high trading frictions.

4.3. Return Predictability in Recent Years

The U.S. equity market has undergone substantial structural changes since the 2000s, such as the introduction of decimalization and improved market liquidity, the greater participation of institutional investors, better access to a broader range of data, developments in financial technology, and the adoption of advanced quantitative analysis. Chordia et al. (2014) show that most anomalies were attenuated after decimalization in January 2001, with the average return and Sharpe ratio from a trading strategy consisting of 12 anomalies more than halved.

In the presence of the growing popularity of exploiting big data and quantitative models in asset management, we examine whether machine learning techniques have remained meaningful in recent years. We repeat the portfolio analysis described before and sort stocks into decile portfolios according to machine learning signals. We implement a zero-investment trading strategy by taking long positions in the top decile of stocks and shorting stocks in the bottom decile and compute the holding period return in the post-2001 period. We report only the value-weighted performance of long-short portfolios in Table 8 for brevity, where Panel A shows the results for portfolios sorted by the NN3-predicted returns (GKX), Panel B shows those for portfolios sorted by the risk loadings on the SDF (CPZ), Panel C shows those for portfolios sorted by the IPCA-predicted returns (KPS), and Panel D shows those for portfolios sorted by the CA2-predicted returns (Gu et al. 2021).

Our main findings from 1987 to 2017 remain intact in the post-2001 period for the GKX, IPCA, and CA signals. The value-weighted long-short portfolio return based on the GKX (IPCA, CA) signal across all stocks is significant at 1.57% (0.88%, 1.65%) per month and 1.18% (0.69%, 1.12%) after adjusting for the FF6 model. The economic magnitude is also comparable with that for the entire sample period, that is, 1.56% (0.95%, 1.16%) for raw returns and 0.92% (0.62%, 0.75%) for the FF6-adjusted returns in Tables 1 and 4. Moreover, the machine learning signal continues to predict the cross-section of stock returns among non-microcaps and rated firms in recent years. Shifting the focus to the most restrictive subsample that excludes credit rating downgrades, the FF6-adjusted return is significant only at the 5% threshold for the IPCA method.[38]

As shown in Panel B of Table 8, the value-weighted long-short portfolio return based on the CPZ signal across all stocks is significant at 1.86% per month and 1.02% after adjusting for the FF6 model. The economic magnitude is slightly lower than that for the 1987 to 2016 period, that is, 2.18% for raw returns and 1.87% for FF6-adjusted returns (Table 3, Panel A), but the recent performance is comparable with other deep learning signals. Imposing economic restrictions further weakens out-of-sample return predictability in recent years, and we do not detect significant FF6-adjusted returns in any of the three subsamples.[39]

The overall evidence suggests that machine learning signals continue to predict cross-sectional stock returns in recent years for the full sample. Unlike for individual anomalies, there is no vast drop in trading profits of machine learning signals. This supports the concept that machine learning methods can combine multiple, presumably weak, signals into a meaningful set of information. Conversely, anomalous return patterns are still confined within difficult-to-arbitrage stocks, and thus, practitioners should remain cautious in using machine learning algorithms for real-time trading.

5. Economic Foundations of Machine Learning

Cochrane (2011) points out that traditional regression analysis and portfolio sorts could be insufficient for handling a large number of predictive variables. In response, machine learning offers a natural way to accommodate a high-dimensional predictor set and flexible functional forms and employs "regularization" methods to select models and mitigate overfitting biases. However, deep learning models are opaque in nature and are often referred to as "black boxes." As emphasized by Karolyi and Van Nieuwerburgh (2020), understanding the relevant economic mechanisms is essential for machine learning tools, especially if the goal is robust and credible out-of-sample predictability.

In this section, we focus on the full sample and provide evidence on the economic driving forces of return predictability in machine learning methods. Specifically, we examine whether stocks with similar machine learning signals also share other characteristics that predict future returns. Examining the common features of stocks selected by machine learning methods allows us to determine whether the investment decision is economically interpretable. We further control for industry benchmarks and investigate the source of return predictability.

5.1. Stock Characteristics of Machine Learning Portfolios

At the end of each month t, stocks are sorted into deciles per machine learning signal. We then compute the equal-weighted average of a comprehensive set of stock characteristics at the end of month t for each portfolio. Most stock characteristics come from the firm-level predictors used in GKX, including *Absolute Accruals*, *Log(Age)*, *Assets Growth*, *Beta*, *Book-to-Market*, *ΔShares Outstanding*, *Corporate Investment*, *Dividend-to-Price*,

Table 8. Performance of Machine Learning Portfolios in Recent Years

	Return	CAPM	FFC+PS	FF5	FF6	SY
Panel A: Value-weighted returns to investment strategies sorted by NN3-predicted returns						
Full sample	1.568***	2.012***	1.555***	1.096***	1.181***	0.699**
	(3.07)	(4.63)	(5.20)	(3.95)	(4.32)	(2.27)
Nonmicrocaps	1.106**	1.583***	1.116***	0.592**	0.687***	0.276
	(2.27)	(3.97)	(4.28)	(2.19)	(2.88)	(0.93)
Credit rating sample	1.182**	1.635***	1.165***	0.700**	0.803***	0.288
	(2.37)	(4.22)	(4.04)	(2.36)	(3.14)	(0.87)
Nondowngrades	0.796*	1.188***	0.768***	0.454	0.545*	−0.037
	(1.82)	(3.15)	(2.67)	(1.58)	(1.90)	(−0.11)
Panel B: Value-weighted returns to investment strategies sorted by risk loadings						
Full Sample	1.864***	1.759***	1.428***	1.060**	1.021**	1.407***
	(3.36)	(2.98)	(3.33)	(2.27)	(2.40)	(2.78)
Nonmicrocaps	0.954**	0.968**	0.625**	0.157	0.216	0.338
	(2.40)	(2.35)	(2.08)	(0.54)	(0.75)	(1.06)
Credit Rating Sample	0.759*	0.725*	0.400	0.114	0.165	0.237
	(1.87)	(1.67)	(1.13)	(0.33)	(0.48)	(0.63)
Nondowngrades	0.765*	0.786*	0.453	0.168	0.203	0.347
	(1.87)	(1.82)	(1.30)	(0.48)	(0.59)	(0.90)
Panel C: Value-weighted returns to investment strategies sorted by IPCA-predicted returns						
Full Sample	0.881***	0.967***	0.775***	0.643**	0.691***	0.321
	(3.95)	(4.32)	(3.81)	(2.58)	(3.31)	(1.19)
Nonmicrocaps	0.732***	0.862***	0.682***	0.532**	0.578***	0.216
	(3.47)	(4.19)	(3.76)	(2.38)	(3.06)	(0.92)
Credit Rating Sample	0.830***	0.953***	0.763***	0.661***	0.714***	0.279
	(3.68)	(4.31)	(3.65)	(2.61)	(3.28)	(1.03)
Nondowngrades	0.643***	0.724***	0.505**	0.475*	0.531**	0.088
	(2.86)	(3.05)	(2.34)	(1.76)	(2.35)	(0.31)
Panel D: value-weighted returns to investment strategies sorted by CA2-predicted returns						
Full Sample	1.651***	1.753***	1.301***	1.065***	1.122***	0.840**
	(3.96)	(4.08)	(4.22)	(3.11)	(3.38)	(2.46)
Nonmicrocaps	1.388***	1.591***	1.083***	0.647**	0.760***	0.419
	(3.30)	(3.67)	(3.95)	(2.01)	(2.84)	(1.40)
Credit Rating Sample	1.245***	1.358***	0.784***	0.436	0.566*	0.179
	(2.85)	(3.01)	(2.64)	(1.12)	(1.80)	(0.50)
Nondowngrades	0.964**	1.072**	0.566**	0.265	0.363	0.095
	(2.35)	(2.54)	(1.99)	(0.92)	(1.40)	(0.30)

Notes. In Panel A, at the end of each month t, stocks are sorted into deciles according to their NN3-predicted returns (Gu et al. 2020). We report the value-weighted returns for month $t+1$ for the strategy of going long (short) on the highest (lowest) expected return stocks in the post-2001 period. Portfolio returns are further adjusted by the CAPM, Fama-French-Carhart four-factor and Pástor-Stambaugh liquidity factor model (FFC+PS), Fama-French five-factor model (FF5), Fama-French six-factor model (FF6), and Stambaugh-Yuan four-factor model (SY). We report the results for the full sample as well as for subsamples that exclude microcaps, nonrated firms, and credit rating downgrades. Panels B to D report similar statistics when decile portfolios are sorted by the risk loadings on the stochastic discount factor (Chen et al. 2020), IPCA-predicted returns (Kelly et al. 2019), and CA2-predicted returns (Gu et al. 2021), respectively. Newey-West adjusted t statistics are shown in parentheses.

*, **, and *** Significant at the 10%, 5%, and 1% levels, respectively.

Gross Profitability, idiosyncratic return volatility (*IdioVol*), *Log(Illiquidity)*, *Leverage*, *12M Momentum*, return on assets (*ROA*), and return on equity (*ROE*). We also consider other firm characteristics, such as *Log(Price)*, *Log(Size)*, *1M Return*, the percentage of rated firms (*%Rated*), *Credit Rating*, *Analyst Coverage*, *Analyst Dispersion*, and standardized unexpected earnings (*SUE*). We obtain analyst forecast data from the Institutional Brokers' Estimate System (I/B/E/S). Online Appendix A provides detailed definitions of each variable.

Although most stock characteristics examined in this section are also used as input variables for the machine learning models, whether the same features are preserved in machine learning portfolios is unclear. For instance, if machine learning portfolios capture mainly the nonlinearities and interaction effects, the stocks identified by machine learning signals may not align with those in standard anomaly-based trading strategies relying on a single characteristic.

The results are presented in Table 9, where Panel A shows the results for portfolios sorted by the NN3-predicted returns (GKX) and the risk loadings on the SDF (CPZ) and Panel B shows the results for portfolios sorted by the IPCA-predicted returns (KPS) and the CA2-predicted returns (Gu et al. 2021). For brevity, we tabulate only stock characteristics for the bottom and top decile portfolios, as well as the difference in values between high- and low-decile portfolios (HML). We report adjusted *t* statistics from Newey and West (1987) with four lags. Several findings are worth noting. First, all machine learning methods identify stocks in line with most anomaly-based trading strategies. For instance, stocks in the long positions of a machine learning-based trading strategy are typically small, value, illiquid, and old stocks with low price, low beta, high 11-month return (medium-term winners), low asset growth, low equity issuance, low credit rating coverage, and low analyst coverage. Therefore, despite their opaque nature, machine learning techniques successfully identify mispriced stocks with solid economic foundations.[40]

Second, there are two notable incidences in which machine learning signals trade in the opposite direction of individual anomaly characteristics. First, all machine learning methods take long positions in stocks with high corporate investment, which, on an individual basis, predicts lower future returns on average (Titman et al. 2004). Second, all machine learning methods except for CPZ take long positions in stocks with high idiosyncratic volatility. Future returns may not be linear in either corporate investment or idiosyncratic volatility and return predictability could be affected by other related firm characteristics or macro conditions. As shown by Titman et al. (2004), the negative investment-return relation is more prominent among firms with higher cash flows and lower debt ratios. Stambaugh et al. (2015) also document that the idiosyncratic volatility-return relation is negative for overpriced stocks but turns positive for underpriced stocks. Such complex and often ambiguous patterns in the cross-section highlight the merits of using machine learning techniques because they can distill information from a large set of correlated characteristics.

An advantage of machine learning methods is that they accommodate high dimensionality and complex patterns in the data without preselection of truly useful characteristics and models, hence avoiding the data snooping problem that challenges the credibility of the anomaly literature (Harvey et al. 2016, McLean and Pontiff 2016, Harvey 2017, Hou et al. 2020). Our findings support this concept by showing that machine learning signals indeed identify mispriced stocks that are in line with well-established empirical facts without requiring any prior knowledge. Despite their opaque nature, machine learning models generate economically interpretable trading strategies, which is essential for robust and credible out-of-sample predictability (Karolyi and Van Nieuwerburgh 2020).

Notably, several other papers also investigate the economic interpretability of machine learning models. For instance, Kelly et al. (2019) and Gu et al. (2020) identify important characteristics according to their contribution to overall model fit. Cong et al. (2021) propose an "economic distillation" procedure that projects complex AI models onto linear modeling or natural language spaces to identify the dominant characteristics (including their higher-order terms and interaction terms). Similar to our approach, Sak et al. (2021) use a characteristic sort to identify dominant features. These studies aim to identify the influential covariates for the cross-sectional return predictability and rank all predictors by their importance. In contrast, we do not implement a horse race across all predictors in the machine learning model. Instead, we focus on the ex post interpretability of machine learning signals and associate them with a list of widely adopted and economically motivated trading signals. Although the important characteristics identified in prior work are model specific, we complement their results by providing broader implications for the economic interpretability of machine learning methods.

5.2. Intra-Industry vs. Inter-Industry Return Predictability

Returns on firms within the same industry are highly correlated, as they could be affected by common technological shocks, changes in operational and regulatory environments, and industry-specific demand and supply for certain products and services. Our prior findings suggest that deep learning signals mostly predict future returns in difficult-to-arbitrage stocks. If such trading signals capture temporary mispricing and subsequent correction due to market frictions, matching similar firms within the same industry provides a natural framework to control for firm fundamentals and understand the source of return predictability. Controlling for industry benchmarks can also clarify whether machine learning methods specialize in stock picking or industry rotation, that is, whether they cluster similar firms and trade the industry portfolio.

To pursue the analysis, we first implement an unconditional trading strategy based on the NN3-predicted return of stock *i* in month *t* (GKX), denoted by $\hat{R}_{i,t}$. We take a \$1 long position on stocks that are expected to outperform the market (market winners),

Table 9. Stock Characteristics of Machine Learning Portfolios

Panel A: Stock characteristics of machine learning portfolios

	Sorted by NN3-predicted returns				Sorted by risk loadings			
Stock characteristics	Low	High	HML	*t* statistic	Low	High	HML	*t* statistic
Log (price)	2.092	1.386	−0.706***	(−14.63)	1.670	1.304	−0.366***	(−12.89)
Log (size)	5.483	3.613	−1.870***	(−25.33)	4.646	3.524	−1.123***	(−29.82)
Book-to-market	0.780	1.391	0.612***	(10.62)	0.914	1.353	0.440***	(13.15)
Log (illiquidity)	1.220	3.791	2.571***	(22.70)	2.760	3.911	1.150***	(18.94)
Beta	1.330	1.068	−0.263***	(−6.87)	1.307	1.096	−0.211***	(−11.74)
1M return	0.027	−0.012	−0.039***	(−10.22)	0.145	−0.118	−0.263***	(−37.18)
12M momentum	−0.124	0.185	0.309***	(15.64)	−0.246	0.083	0.329***	(20.49)
IdioVol	0.076	0.084	0.008***	(4.19)	0.081	0.082	0.001	(1.55)
Absolute accruals	0.103	0.107	0.003	(1.08)	0.108	0.097	−0.011***	(−9.83)
Log (age)	2.171	2.463	0.292***	(9.93)	2.700	2.716	0.016**	(2.48)
Assets growth	0.586	0.014	−0.572***	(−11.76)	0.164	0.048	−0.116***	(−12.76)
ΔShares outstanding	0.389	0.070	−0.319***	(−10.92)	0.158	0.064	−0.094***	(−15.58)
Corporate investment	−0.107	0.022	0.129***	(9.03)	−0.026	0.001	0.027***	(4.72)
Dividend-to-price	0.012	0.009	−0.003***	(−3.61)	0.009	0.011	0.002***	(6.68)
Gross profitability	0.329	0.354	0.025	(1.57)	0.387	0.420	0.032***	(8.11)
Leverage	1.308	1.832	0.524***	(4.18)	1.340	1.388	0.048	(1.16)
ROA	−0.024	−0.011	0.013***	(5.41)	−0.019	−0.009	0.010***	(12.04)
ROE	−0.045	−0.017	0.028***	(5.87)	−0.039	−0.017	0.022***	(12.77)
%Rated	0.236	0.093	−0.142***	(−13.48)	0.196	0.102	−0.094***	(−17.53)
Credit rating	11.350	12.047	0.697**	(2.44)	12.547	13.036	0.490***	(4.38)
Analyst coverage	4.142	1.432	−2.710***	(−14.99)	3.022	1.441	−1.581***	(−17.34)
Analyst dispersion	0.049	0.057	0.008	(0.68)	0.050	0.083	0.033***	(2.61)
SUE	−0.019	−0.009	0.010***	(3.17)	−0.023	−0.016	0.006***	(2.79)

Panel B: Stock characteristics of machine learning portfolios

	Sorted by IPCA-predicted returns				Sorted by CA2-predicted returns			
Stock characteristics	Low	High	HML	*t* statistic	Low	High	HML	*t* statistic
Log (price)	2.301	1.853	−0.449***	(−21.81)	1.636	0.951	−0.685***	(−21.54)
Log (size)	5.748	4.067	−1.681***	(−67.77)	4.690	3.255	−1.435***	(−31.52)
Book-to-market	0.871	1.348	0.477***	(15.70)	0.888	1.240	0.352***	(7.01)
Log (illiquidity)	1.108	2.880	1.772***	(50.20)	2.347	4.447	2.100***	(25.37)
Beta	1.142	1.056	−0.086***	(−4.89)	1.423	1.150	−0.273***	(−11.84)
1M return	0.030	−0.028	−0.058***	(−24.57)	−0.012	0.015	0.027***	(10.50)
12M momentum	−0.085	0.285	0.370***	(21.26)	−0.132	−0.056	0.077***	(4.62)
IdioVol	0.068	0.074	0.006***	(8.48)	0.088	0.093	0.004***	(4.40)
Absolute accruals	0.107	0.082	−0.025***	(−19.71)	0.109	0.111	0.003*	(1.72)
Log (age)	2.279	2.476	0.197***	(21.21)	2.151	2.284	0.133***	(11.50)
Assets growth	0.542	0.016	−0.526***	(−19.13)	0.423	0.114	−0.310***	(−9.60)
ΔShares outstanding	0.343	0.050	−0.294***	(−18.38)	0.291	0.139	−0.151***	(−9.96)
Corporate investment	−0.068	0.006	0.074***	(9.50)	−0.069	−0.012	0.058***	(7.01)
Dividend-to-price	0.014	0.010	−0.004***	(−12.93)	0.011	0.007	−0.004***	(−5.67)
Gross profitability	0.316	0.383	0.066***	(9.95)	0.330	0.329	−0.001	(−0.15)
Leverage	1.054	1.830	0.777***	(11.37)	1.565	1.808	0.243**	(2.38)
ROA	−0.019	0.005	0.024***	(22.44)	−0.025	−0.026	−0.001	(−0.59)
ROE	−0.038	0.017	0.055***	(24.02)	−0.047	−0.048	−0.001	(−0.62)
%Rated	0.290	0.119	−0.171***	(−35.38)	0.170	0.075	−0.095***	(−12.73)
Credit rating	10.573	11.009	0.436***	(4.46)	12.929	13.920	0.991***	(6.40)
Analyst coverage	4.071	1.930	−2.141***	(−24.91)	3.248	1.136	−2.112***	(−13.47)
Analyst dispersion	0.085	0.070	−0.014	(−1.24)	0.057	0.010	−0.047*	(−1.97)
SUE	−0.012	−0.001	0.012***	(12.45)	−0.025	−0.022	0.002	(0.98)

Notes. In Panel A, at the end of each month *t*, stocks are sorted into deciles according to their NN3-predicted returns (Gu et al. 2020). We report, for the bottom and top decile portfolios, the contemporaneous equal-weighted average Log(price), Log(size), Book-to-market, Log(illiquidity), Beta, 1M return, 12M momentum, IdioVol, Absolute accruals, Log(age), Assets growth, ΔShares outstanding, Corporate investment, Dividend-to-price, Gross profitability, Leverage, ROA, ROE, %Rated, Credit rating, Analyst coverage, Analyst dispersion, and SUE, as well as the difference in values between high and low decile portfolios (HML). We also report similar statistics when decile portfolios are sorted by the risk loadings on the stochastic discount factor (Chen et al. 2020). Panel B reports similar statistics when decile portfolios are sorted by the IPCA-predicted returns (Kelly et al. 2019) and CA2-predicted returns (Gu et al. 2021). Online Appendix A provides detailed definitions for each variable. Newey-West adjusted *t* statistics are shown in parentheses.

*, **, and *** Significant at the 10%, 5%, and 1% levels, respectively.

that is, $\hat{R}_{i,t} - \hat{R}_{m,t} > 0$ and \$1 short position on stocks that are expected to underperform the market (market losers), that is, $\hat{R}_{i,t} - \hat{R}_{m,t} < 0$, where $\hat{R}_{m,t}$ refers to the equal-weighted average of $\hat{R}_{i,t}$ across all stocks in the market. That is, $\hat{R}_{m,t} = \frac{1}{N_t}\sum_{i=1}^{N_t}\hat{R}_{i,t}$, where N_t refers to the number of stocks in the market. We hold the portfolio over the next month. The winner minus loser profit from the unconditional strategy at month $t+1$, denoted as WML_{t+1}, is given by

$$WML_{t+1} = \frac{1}{H_t}\sum_{i=1}^{N_t}(\hat{R}_{i,t} - \hat{R}_{m,t})R_{i,t+1}, \tag{2}$$

$$H_t = \frac{1}{2}\sum_{i=1}^{N_t}|\hat{R}_{i,t} - \hat{R}_{m,t}|, \tag{3}$$

where $R_{i,t+1}$ refers to the return of stock i in month $t+1$, and all other variables are previously defined. The portfolio weighting scheme is similar to that in Nagel (2012), and the long-short portfolio includes all stocks in the investment universe. The weight of each stock is proportional to the stock's NN3-predicted return on a market-adjusted basis, with higher weights for better performers in the long leg and more negative weights for worse performers in the short leg. The investment in each security is scaled by the inverse of the sum of absolute deviations of stock returns from the market average so that the strategy is \$1 long in market winner stocks and \$1 short in market loser stocks. This unconditional strategy also provides a robustness check of our main results, where we focus only on extreme decile portfolios to construct a long-short strategy rather than on all investable stocks.

Next, as in Hameed and Mian (2015), we decompose the unconditional strategy into two components. In particular, Equation (2) can be rewritten as follows:

$$\begin{aligned} WML_{t+1} &= \frac{1}{H_t}\sum_{i=1}^{N_t}(\hat{R}_{i,t} - \hat{R}_{j,t} + \hat{R}_{j,t} - \hat{R}_{m,t})R_{i,t+1} \\ &= \frac{1}{H_t}\sum_{i=1}^{N_t}(\hat{R}_{i,t} - \hat{R}_{j,t})R_{i,t+1} \\ &\quad + \frac{1}{H_t}\sum_{i=1}^{N_t}(\hat{R}_{j,t} - \hat{R}_{m,t})R_{i,t+1} \\ &= \frac{1}{H_t}\sum_{i=1}^{N_t}(\hat{R}_{i,t} - \hat{R}_{j,t})R_{i,t+1} \\ &\quad + \frac{1}{H_t}\sum_{j=1}^{L_t}(\hat{R}_{j,t} - \hat{R}_{m,t})N_{j,t}R_{j,t+1} \end{aligned} \tag{4}$$

where $\hat{R}_{j,t}$ refers to the equal-weighted average of $\hat{R}_{i,t}$ across all stocks in industry j. That is, $\hat{R}_{j,t} = \frac{1}{N_{j,t}}\sum_{i=1}^{N_{j,t}}\hat{R}_{i,t}$, where $N_{j,t}$ refers to the number of stocks in industry j. L_t refers to the number of industries, and $R_{j,t+1}$ refers to the equal-weighted average of stock returns across all stocks in industry j in month $t+1$. That is, $R_{j,t+1} = \frac{1}{N_{j,t}}\sum_{i=1}^{N_{j,t}}R_{i,t+1}$. All other variables are defined as in Equation (2).[41]

The first term in Equation (4) represents returns to an intra-industry strategy that buys stocks that are expected to outperform the industry portfolio (industry winners), that is, $\hat{R}_{i,t} - \hat{R}_{j,t} > 0$, and sells stocks that are expected to underperform the industry portfolio (industry losers), that is, $\hat{R}_{i,t} - \hat{R}_{j,t} < 0$. The second term represents returns to an inter-industry strategy that buys the industry portfolio if the industry is expected to outperform the overall market (winner industries), that is, $\hat{R}_{j,t} - \hat{R}_{m,t} > 0$, and sells the industry portfolio if the industry is expected to underperform the overall market (loser industries), that is, $\hat{R}_{j,t} - \hat{R}_{m,t} < 0$. To scale the investment in each component to \$1 long and \$1 short, we multiply the profits by the factor of proportionality as follows:

$$\begin{aligned} WML_{t+1} &= \frac{H_t^{INTRA}}{H_t}\frac{1}{H_t^{INTRA}}\sum_{i=1}^{N_t}(\hat{R}_{i,t} - \hat{R}_{j,t})R_{i,t+1} \\ &\quad + \frac{H_t^{INTER}}{H_t}\frac{1}{H_t^{INTER}}\sum_{j=1}^{L_t}(\hat{R}_{j,t} - \hat{R}_{m,t})N_{j,t}R_{j,t+1} \\ &= \frac{H_t^{INTRA}}{H_t} \times WML_{t+1}^{INTRA} + \frac{H_t^{INTER}}{H_t} \times WML_{t+1}^{INTER}, \end{aligned} \tag{5}$$

$$H_t^{INTRA} = \frac{1}{2}\sum_{i=1}^{N_t}|\hat{R}_{i,t} - \hat{R}_{j,t}|, \tag{6}$$

$$H_t^{INTER} = \frac{1}{2}\sum_{j=1}^{L_t}|\hat{R}_{j,t} - \hat{R}_{m,t}|N_{j,t}, \tag{7}$$

where all variables are defined as in Equations (2) and (4). In particular, the winner minus loser profit from the intra-industry strategy at month $t+1$, denoted as WML_{t+1}^{INTRA}, is given by

$$WML_{t+1}^{INTRA} = \frac{1}{H_t^{INTRA}}\sum_{i=1}^{N_t}(\hat{R}_{i,t} - \hat{R}_{j,t})R_{i,t+1}. \tag{8}$$

Similarly, the winner minus loser profit from the inter-industry strategy at month $t+1$, denoted as WML_{t+1}^{INTER}, is given by

$$WML_{t+1}^{INTER} = \frac{1}{H_t^{INTER}}\sum_{j=1}^{L_t}(\hat{R}_{j,t} - \hat{R}_{m,t})N_{j,t}R_{j,t+1}. \tag{9}$$

As shown in Equation (5), the unconditional reversal profit (WML_{t+1}) is a weighted average of WML_{t+1}^{INTRA} and WML_{t+1}^{INTER}, and the weights depend on the scaling factors for the different strategies, that is, $\frac{H_t^{INTRA}}{H_t}$ and $\frac{H_t^{INTER}}{H_t}$.

The results are tabulated in Table 10, where Panel A shows the results for the full sample and subsample excluding microcaps and Panel B shows those for the subsamples excluding nonrated firms or credit rating downgrades. In Panels A1 and B1, we report the

Table 10. Performance of Machine Learning Portfolios and their Attribution

Panel A: Returns to investment strategies sorted by NN3-predicted returns (full sample and microcaps excluded)												
	Full sample						Nonmicrocaps					
Rank	Return	CAPM	FFC+PS	FF5	FF6	SY	Return	CAPM	FFC+PS	FF5	FF6	SY
Panel A1: Unconditional payoff												
Loser	0.239	−0.907***	−0.510***	−0.458***	−0.204*	−0.103	0.450	−0.787***	−0.369***	−0.272*	−0.029	0.058
	(0.60)	(−4.61)	(−3.99)	(−2.64)	(−1.71)	(−0.58)	(1.12)	(−4.08)	(−2.91)	(−1.67)	(−0.30)	(0.35)
Winner	2.050***	1.093***	1.268***	1.207***	1.351***	1.496***	1.368***	0.433***	0.300***	0.183**	0.153**	0.251***
	(5.23)	(3.87)	(5.35)	(4.70)	(5.03)	(5.04)	(5.07)	(2.87)	(3.66)	(2.53)	(2.08)	(2.70)
WML	1.812***	2.000***	1.778***	1.665***	1.554***	1.599***	0.918***	1.220***	0.669***	0.455**	0.182	0.194
	(8.12)	(8.55)	(8.24)	(8.43)	(7.24)	(7.07)	(3.50)	(4.76)	(3.70)	(2.42)	(1.38)	(0.92)
Panel A2: Intra-industry payoff												
Loser	0.310	−0.842***	−0.533***	−0.466***	−0.267**	−0.219	0.541	−0.691***	−0.389***	−0.325***	−0.145*	−0.094
	(0.77)	(−4.45)	(−4.18)	(−3.28)	(−2.38)	(−1.34)	(1.38)	(−4.35)	(−3.63)	(−2.70)	(−1.80)	(−0.68)
Winner	1.996***	1.025***	1.286***	1.212***	1.407***	1.584***	1.293***	0.329**	0.288***	0.174***	0.194***	0.310***
	(5.05)	(3.66)	(5.19)	(4.31)	(4.91)	(4.91)	(4.79)	(2.54)	(4.13)	(2.64)	(3.01)	(3.93)
WML^{INTRA}	1.686***	1.867***	1.819***	1.678***	1.675***	1.802***	0.752***	1.021***	0.678***	0.499***	0.339***	0.404**
	(8.75)	(9.90)	(8.62)	(8.73)	(7.58)	(7.68)	(3.69)	(5.81)	(4.70)	(4.23)	(3.32)	(2.57)
WML^{INTRA} × H^{INTRA}/H	1.519***	1.681***	1.642***	1.516***	1.516***	1.637***	0.643***	0.870***	0.588***	0.429***	0.300***	0.372***
	(8.78)	(9.98)	(8.57)	(8.69)	(7.53)	(7.61)	(3.69)	(5.82)	(4.64)	(4.29)	(3.33)	(2.71)
Panel A3: Inter-industry payoff												
Loser	0.670**	−0.330*	0.042	−0.015	0.256	0.472**	0.678**	−0.394**	−0.086	−0.076	0.133	0.282*
	(1.99)	(−1.80)	(0.31)	(−0.06)	(1.50)	(2.12)	(2.08)	(−2.37)	(−0.76)	(−0.47)	(1.26)	(1.95)
Winner	1.353***	0.403*	0.358**	0.362**	0.372**	0.436**	1.184***	0.230	0.068	−0.015	−0.059	0.011
	(3.93)	(1.86)	(2.47)	(2.55)	(2.44)	(2.53)	(4.24)	(1.49)	(0.81)	(−0.19)	(−0.75)	(0.11)
WML^{INTER}	0.683***	0.733***	0.316*	0.376	0.116	−0.036	0.506**	0.625***	0.154	0.061	−0.192	−0.271
	(3.46)	(3.49)	(1.91)	(1.61)	(0.67)	(−0.16)	(2.42)	(2.70)	(0.95)	(0.31)	(−1.37)	(−1.41)
WML^{INTER} × H^{INTER}/H	0.293***	0.319***	0.136*	0.149	0.038	−0.037	0.275**	0.350***	0.080	0.026	−0.119	−0.179
	(3.27)	(3.27)	(1.76)	(1.52)	(0.49)	(−0.37)	(2.46)	(2.77)	(0.88)	(0.24)	(−1.43)	(−1.59)
Panel B: Returns to investment strategies sorted by NN3-predicted returns (credit rating sample and downgrades excluded)												
	Credit rating sample						Nondowngrades					
Rank	Return	CAPM	FFC+PS	FF5	FF6	SY	Return	CAPM	FFC+PS	FF5	FF6	SY
Panel B1: Unconditional payoff												
Loser	0.337	−0.911***	−0.493***	−0.653***	−0.308**	−0.055	1.161***	0.040	0.305***	0.154	0.377***	0.642***
	(0.78)	(−4.22)	(−3.78)	(−2.67)	(−2.30)	(−0.25)	(3.44)	(0.25)	(2.91)	(0.91)	(3.28)	(3.63)
Winner	1.332***	0.386**	0.278**	0.057	0.090	0.271**	1.639***	0.738***	0.573***	0.380***	0.365***	0.505***
	(4.39)	(2.14)	(2.46)	(0.53)	(0.89)	(2.04)	(6.21)	(4.62)	(5.98)	(4.81)	(4.54)	(4.47)

Table 10. (Continued)

Panel B: Returns to investment strategies sorted by NN3-predicted returns (credit rating sample and downgrades excluded)												
	Credit rating sample						Nondowngrades					
Rank	Return	CAPM	FFC+PS	FF5	FF6	SY	Return	CAPM	FFC+PS	FF5	FF6	SY
WML	0.995***	1.296***	0.770***	0.710***	0.398**	0.326	0.478**	0.698***	0.269**	0.226	−0.012	−0.137
	(3.76)	(5.40)	(4.20)	(3.14)	(2.48)	(1.43)	(2.42)	(3.81)	(2.13)	(1.35)	(−0.10)	(−0.72)
Panel B2: Intra-industry payoff												
Loser	0.509	−0.713***	−0.414***	−0.601***	−0.320**	−0.103	1.228***	0.127	0.286**	0.131	0.291**	0.513***
	(1.22)	(−3.68)	(−3.25)	(−2.88)	(−2.47)	(−0.48)	(3.77)	(0.88)	(2.57)	(0.93)	(2.56)	(3.04)
Winner	1.227***	0.241	0.230**	0.007	0.097	0.308**	1.581***	0.649***	0.555***	0.367***	0.396***	0.569***
	(3.92)	(1.42)	(2.25)	(0.05)	(0.99)	(2.48)	(5.83)	(4.37)	(6.28)	(4.29)	(4.77)	(5.34)
$\mathrm{WML}^{\mathrm{INTRA}}$	0.718***	0.954***	0.644***	0.607***	0.417***	0.411**	0.353**	0.522***	0.269**	0.236**	0.104	0.056
	(3.56)	(5.57)	(4.20)	(3.89)	(3.08)	(2.13)	(2.41)	(3.92)	(2.45)	(2.00)	(0.94)	(0.36)
$\mathrm{WML}^{\mathrm{INTRA}} \times \mathrm{H}^{\mathrm{INTRA}}/\mathrm{H}$	0.608***	0.805***	0.550***	0.511***	0.356***	0.363**	0.304**	0.443***	0.238***	0.204**	0.099	0.068
	(3.65)	(5.69)	(4.35)	(3.92)	(3.19)	(2.32)	(2.57)	(4.10)	(2.72)	(2.15)	(1.13)	(0.56)
Panel B3: Inter-industry payoff												
Loser	0.499	−0.631***	−0.310***	−0.491**	−0.200	0.075	1.179***	0.135	0.349***	0.156	0.369***	0.642***
	(1.34)	(−3.30)	(−2.64)	(−2.30)	(−1.59)	(0.42)	(3.74)	(0.83)	(3.07)	(0.90)	(2.89)	(3.95)
Winner	1.195***	0.244	0.093	−0.126	−0.110	0.053	1.486***	0.574***	0.400***	0.196**	0.179**	0.311**
	(4.02)	(1.37)	(0.86)	(−1.30)	(−1.15)	(0.38)	(5.71)	(3.55)	(3.74)	(2.40)	(2.14)	(2.51)
$\mathrm{WML}^{\mathrm{INTER}}$	0.695***	0.876***	0.403**	0.365*	0.090	−0.022	0.307*	0.439***	0.051	0.040	−0.189	−0.332**
	(3.30)	(4.17)	(2.48)	(1.73)	(0.60)	(−0.12)	(1.85)	(2.71)	(0.39)	(0.23)	(−1.40)	(−2.05)
$\mathrm{WML}^{\mathrm{INTER}} \times \mathrm{H}^{\mathrm{INTER}}/\mathrm{H}$	0.387***	0.491***	0.221**	0.199*	0.042	−0.037	0.174*	0.255***	0.031	0.022	−0.112	−0.205**
	(3.27)	(4.11)	(2.36)	(1.68)	(0.47)	(−0.33)	(1.82)	(2.72)	(0.41)	(0.23)	(−1.47)	(−2.15)

Notes. This table reports the monthly returns from the unconditional strategy (WML) and its decomposition into intra-industry ($\mathrm{WML}^{\mathrm{INTRA}}$) and inter-industry ($\mathrm{WML}^{\mathrm{INTER}}$) returns components, as depicted in Equation (5). In Panel A1, we report the returns for WML, when the strategy takes long (short) positions in the winner (loser) stocks that are expected to outperform (underperform) the market according to the NN3-predicted returns (Gu et al. 2020). In Panel A2, we report the returns for $\mathrm{WML}^{\mathrm{INTRA}}$ when the strategy takes long (short) positions in the stocks that are expected to outperform (underperform) the industry portfolio. $\mathrm{H}^{\mathrm{INTRA}}$ and H are the scaling factors for the unconditional and intra-industry strategies. Panel A3 reports the returns for $\mathrm{WML}^{\mathrm{INTER}}$ when the strategy takes long (short) positions in the industry portfolios that are expected to outperform (underperform) the market portfolio. $\mathrm{H}^{\mathrm{INTER}}$ is the scaling factor for the inter-industry strategy. Portfolio returns are further adjusted by the CAPM, Fama-French-Carhart four-factor and Pástor-Stambaugh liquidity factor model (FFC+PS), Fama-French five-factor model (FF5), Fama-French six-factor model (FF6), and Stambaugh-Yuan four-factor model (SY). Panel A reports the results for the full sample and the subsample that excludes microcaps. Panel B reports similar statistics for the subsamples that exclude nonrated firms and credit rating downgrades. Newey-West adjusted *t* statistics are shown in parentheses.

*, **, and *** Significant at the 10%, 5%, and 1% levels, respectively.

returns for winner and loser portfolios obtained using the unconditional strategy, where winners (losers) consist of stocks that are expected to outperform (underperform) the market average according to the NN3-predicted returns in GKX. We also implement the trading strategy by taking long positions in winner stocks and shorting loser stocks. The zero-investment trading profit is computed as the winner minus loser portfolio returns (WML). Over the 1987–2017 sample period, the average long-short portfolio return is a significant 1.81% per month across all stocks and 1.55% after adjusting for the FF6 model. In addition, the raw return is 0.92% (0.48%) per month, and the FF6-adjusted return is statistically insignificant for the subsample excluding microcaps (credit rating downgrades). This analysis, based on *all* stocks, confirms our main findings from the extreme decile portfolios that machine learning signals weaken drastically in the subset of cheap-to-trade stocks.

In Panels A2 and B2, we report the returns for winner and loser portfolios from the intra-industry strategy, where winners (losers) consist of stocks that are expected to outperform (underperform) the industry average according to the NN3-predicted returns. We also implement the trading strategy by taking long positions in winner stocks and shorting loser stocks. The zero-investment trading profit is computed as the winner minus loser portfolio returns ($\mathrm{WML}^{\mathrm{INTRA}}$). In Panels A3 and B3, we report the returns for winner and loser portfolios from the inter-industry strategy, in which winners (losers) consist of industries that are expected to outperform (underperform) the market average according to the NN3-predicted returns. We implement the trading strategy by taking long positions in winner industries and shorting loser industries. The zero-investment trading profit is computed as the winner minus loser portfolio returns ($\mathrm{WML}^{\mathrm{INTER}}$). We also report the scaled results, that is, $\mathrm{WML}^{\mathrm{INTRA}} \times \mathrm{H}^{\mathrm{INTRA}}/\mathrm{H}$ and $\mathrm{WML}^{\mathrm{INTER}} \times \mathrm{H}^{\mathrm{INTER}}/\mathrm{H}$, and they add up to WML, as shown in Equation (5).

Several findings are worth noting. First, the intra-industry strategy delivers substantially higher returns than the inter-industry strategy. As shown in Panels A2 and A3, the intra-industry strategy ($\mathrm{WML}^{\mathrm{INTRA}} \times \mathrm{H}^{\mathrm{INTRA}}/\mathrm{H}$) accounts for 84% (i.e., 1.52% of 1.81%) of the unconditional payoff in raw returns and 93% of that in risk-adjusted returns across all performance measures for the full sample. Meanwhile, the inter-industry strategy ($\mathrm{WML}^{\mathrm{INTER}} \times \mathrm{H}^{\mathrm{INTER}}/\mathrm{H}$) accounts for the remaining 16% (i.e., 0.29% of 1.81%) of the unconditional payoff in raw returns and 7% of that in risk-adjusted returns across all performance measures. All risk-adjusted returns are highly significant for the intra-industry strategy, whereas only one of five risk-adjusted returns is significant at the 5% level for the inter-industry strategy. A similar pattern also holds for the three subsamples with economic restrictions. This finding implies that the GKX method emphasizes stock selection more than industry rotation.

Second, the payoff of the intra-industry strategy ($\mathrm{WML}^{\mathrm{INTRA}}$) is higher than $\mathrm{WML}^{\mathrm{INTRA}} \times \mathrm{H}^{\mathrm{INTRA}}/\mathrm{H}$ for the full sample and the subsamples because of a lower cross-sectional dispersion in industry-adjusted returns, that is, $\mathrm{H}^{\mathrm{INTRA}}/\mathrm{H} < 1$. Improvements using an intra-industry strategy are particularly important for nonmicrocaps. As shown in Panel A2, the FF6-adjusted return regains significance by controlling for the industry benchmark, that is, 0.34% per month (t statistic = 3.32), as opposed to a statistically insignificant 0.18% in the unconditional strategy (Panel A1). Similarly, the monthly SY-adjusted return is statistically insignificant at 0.19% for the unconditional strategy and more than double for the intra-industry strategy, that is, 0.40% (t statistic = 2.57). The outperformance of the intra-industry strategy confirms that the GKX signal identifies mispricing in difficult-to-arbitrage stocks.

As a robustness check, we repeat the analysis using IPCA-predicted returns and CA2-predicted returns. For brevity, we present only the raw return and FF6-adjusted return for the full sample and three subsamples in the online appendix, Table IA9. We confirm that the intra-industry strategy not only delivers substantially higher returns than the inter-industry strategy, but also outperforms the unconditional strategy on a risk-adjusted basis for most subsamples with economic restrictions. Hence, adjusting the portfolio weights by the industry average further controls for firm fundamentals and better predicts the subsequent correction because of market frictions. From a practitioner's perspective, the out-of-sample performance of machine learning portfolios can be further enhanced via a simple industry adjustment.

6. Conclusion

This paper provides large-scale evidence on the economic significance of machine learning methods. The deep learning techniques that we analyze face the usual challenge of cross-sectional return predictability. In particular, the anomalous return patterns concentrate in difficult-to-value and difficult-to-arbitrage stocks. In addition, to the extent that deep learning signals predict cross-sectional stock returns for the full sample, the trading strategy is more profitable during periods of high market volatility and low market liquidity. Machine learning signals also involve remarkably high turnover and often require taking extreme long-short positions for predetermined portfolio volatility in the tangency portfolio implied by the pricing kernel. Beyond economic restrictions, machine learning-based trading strategies nonetheless display smaller downside risk, yield considerable profit in the long positions, and remain viable in the post-2001 period and the crisis period. Finally, black-box-like machine learning methods generate economically interpretable trading strategies and are more informative for stock selection than for industry rotation.

Our findings provide timely evidence to help understand machine learning applications and propose a list of back-testing protocols for academic research and asset management. When assessing machine learning methods, it is imperative to consider common economic restrictions in both the cross-section and the time series and incorporate the trading costs due to portfolio turnover. Similarly, it is essential to estimate the SDF based on admissible stock positions. It is also important to confirm the external validity of machine learning models before applying them to different universes of stocks, markets, asset classes, and sample periods.

Our paper also suggests an important avenue for future research. In particular, the optimization routines that mix various anomalies could inherently display a high turnover of stocks in extreme long-short portfolios and thus generate high trading costs. Similarly, pricing kernel estimates could rely on rather extreme long and short positions that are inadmissible in real time. Thus, it would be useful to extend machine learning methods to endogenously account for trading costs and further impose plausible portfolio constraints. Machine learning methods can also be used to evaluate practical transaction costs of strategic trading through relaxing the assumed functional forms of dependence between the trading costs and the invested capital and introducing a nonlinear time series model of trading costs of a patient trader. These and other topics are left for future research.

Acknowledgments

The authors thank Kay Giesecke (editor), the associate editor, two anonymous referees, Bryan Kelly, Michael Halling, Serhiy Kozak, Andy Neuhierl, Shujing Wang, Dacheng Xiu, and conference and seminar participants at the 2021 American Finance Association Annual Meeting, the 3rd World Symposium on Investment Research, the 2020 Midwest Finance Association Annual Meeting, the 2019 Bank of America Merrill Lynch Asia Quant Conference, the 5th International Conference on Fintech Development, the 3rd Conference on Intelligent Information Retrieval in Accounting and Finance, the 2019 China Fintech Research Conference, Baruch College, Chinese University of Hong Kong, Hong Kong Polytechnic University, Singapore Management University, University of Hong Kong, and University of Piraeus for helpful comments.

Endnotes

[1] See Stambaugh et al. (2012) and Avramov et al. (2013, 2018).

[2] See Rapach et al. (2013), Heaton et al. (2017), Feng et al. (2018, 2019), Rapach et al. (2019), Choi et al. (2020), Freyberger et al. (2020), Gu et al. (2020, 2021), Han et al. (2020), Rapach and Zhou (2020), Bianchi et al. (2021), Chinco et al. (2021), Cong et al. (2021), and Kim et al. (2021). Others focus on high-dimensionality cross-sectional asset pricing models: Kelly et al. (2019), Chen et al. (2020), Kozak et al. (2020), and Lettau and Pelger (2020a, b).

[3] Although both formulations are equivalent theoretically, their empirical performance could be different (Kan and Zhou 1999, Cochrane 2001, Jagannathan and Wang 2002).

[4] Prior research documents that the profitability of anomaly-based trading strategies is higher during periods of high investor sentiment (Stambaugh et al. 2012, Avramov et al. 2018), high market volatility (Nagel 2012), and low market liquidity (Chordia et al. 2014).

[5] Such interpretability does not imply that we attempt to explain *why* firm characteristics predict future returns. Instead, we confirm that machine learning generates economically interpretable trading strategies that are in line with the most common return predictors.

[6] See, Wigglesworth (2016), and Zuckerman and Hope (2017).

[7] We adopt the adaptive moment estimation algorithm (Adam) for the stochastic gradient descent used in the optimization.

[8] Kelly et al. (2019) show that IPCA models with five or six latent factors fail to reject the null hypothesis that characteristics explain expected returns only because they proxy for systematic risk exposures. In addition, the total R^2 and predictive R^2 tend to converge in the restricted and unrestricted IPCA models, and the incremental explanatory power from observable factors is negligible if we include five or six latent factors. We consider six latent factors to be in line with the most recent Fama-French six-factor model (Fama and French 2018).

[9] We thank Shihao Gu, Bryan Kelly, and Dacheng Xiu for generously sharing the data on stock-level predicted returns from 1987 to 2016.

[10] Details on each of the 94 firm characteristics can be found in the appendix in Green et al. (2017) and online appendix, table A.6, in Gu et al. (2020). We thank Jeremiah Green for making the SAS code used by Green et al. (2017) available via his website, https://sites.google.com/site/jeremiahrgreenacctg/home.

[11] We thank Amit Goyal for making the data available via his website, http://www.hec.unil.ch/agoyal/.

[12] We thank Markus Pelger for generously sharing the data on stochastic discount factor and stock-level factor loadings from 1967 to 2016. Details on each of the 46 firm characteristics can be found in appendix C in Chen et al. (2020).

[13] Unreported results confirm our main findings in an overlapping sample for all machine learning methods.

[14] Fama and French (2008) recognize microcaps as stocks with a market capitalization smaller than the 20th NYSE size percentile.

[15] We obtain the monthly S&P long-term issuer credit ratings from the COMPUSTAT database. We follow Avramov et al. (2009) in creating the numeric rating score, which transforms the S&P ratings into ascending numbers as follows: AAA = 1, AA+ = 2, AA = 3, AA– = 4, A+ = 5, A = 6, A– = 7, BBB+ = 8, BBB = 9, BBB– = 10, BB+ = 11, BB = 12, BB– = 13, B+ = 14, B = 15, B– = 16, CCC+ = 17, CCC = 18, CCC– = 19, CC = 20, C = 21, and D = 22. A higher credit rating score implies higher credit risk.

[16] Although some prior work account for size-based adjustments, the market cap considerations in existing studies are not unified. For instance, GKX report equal-weighted performance when microcaps are excluded (online appendix, table A.10) and value-weighted performance without market cap restrictions (Table 7); CPZ consider the market cap subsample threshold as a parameter; KPS partition the sample into large and small stocks, where the former is

the top 1,000 stocks in terms of market cap and the latter consists of the rest; and KNS restrict the investment universe to stocks with market caps greater than 0.01% of the aggregate stock market capitalization. Our approach is unified across all methods, that is, excluding microcaps and value weighting, as proposed by Hou et al. (2020) in the context of individual anomalies. We also consider restrictions that go beyond the market cap in both the cross-section and time series. Resorting to such a unified approach delivers new and previously undocumented evidence on economic significance.

[17] We follow the existing literature and report performance based on decile portfolios, for example, table 7 in GKX, figure 7 and table II in CPZ, and table 3 in Gu et al. (2021) for both IPCA and CA methods. In addition, we report out-of-sample Sharpe ratios in Tables 5 and 6 for all machine learning signals in both the full sample and three subsamples. As a robustness check, we also report investment payoff based on *all* stocks (instead of the extreme deciles) in Table 10 and the online appendix, Table IA9.

[18] We thank Kenneth French and Robert Stambaugh for making the common factor returns available via their websites: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html; https://fnce.wharton.upenn.edu/profile/stambaug/. The results using Stambaugh and Yuan (2017) factors end in 2016 due to data availability.

[19] Newey and West (1987) suggest using lags corresponding to $N^{1/4}$, where N refers to the number of observations. Our sample includes 372 monthly observations; thus, we use four lags.

[20] The stock-month observations from 12 months before to 12 months after an issuer credit rating downgrade account for 23% of the credit rating sample.

[21] Given that we require at least 24 observations for alpha estimation, the sample size is slightly smaller than the main results. However, as shown in the online appendix, Table IA5, Panel B, the findings are qualitatively and quantitatively comparable to those in Table 1.

[22] Gomes et al. (2003) formulate an equilibrium model in which beta could vary with firm-level predictors, such as size and book-to-market, whereas Avramov and Chordia (2006) show empirically that such conditional formulations improve the pricing ability of the model.

[23] Stocks that are plentiful but low in aggregate market value may dominate equal-weighted portfolio returns, and value-weighting accurately captures the total wealth effect experienced by investors (Fama 1998); therefore, we focus on value-weighted results in all subsequent analyses. Unreported results indicate similar findings for equal-weighted portfolios.

[24] Without an exhaustive analysis of candidate machine learning models, it is premature to identify the best-performing method in the presence of economic restrictions. We leave this issue for future research.

[25] For perspective, the skewness and excess kurtosis are equal to zero under a normal distribution.

[26] For perspective, we assess the performance of other long-short portfolios using the FF6 model for the crisis period. Both the size factor (SMB) and the momentum factor (MOM) yield negative returns, that is, −3.16% for SMB and −0.20% for MOM, whereas some other factors provide a good hedge during the crisis, that is, 2.44% for the value factor (HML), 3.21% for the profitability factor (RMW), and 3.46% for the investment factor (CMA). All returns are scaled to 10% annual volatility.

[27] Unreported results show that the monthly turnover in the equal-weighted long-short portfolio ranges between 79% and 82% for the GKX method, between 141% and 147% for the CPZ method, between 99% and 103% for the IPCA method, and between 130% and 136% for the CA method.

[28] We compute the break-even transaction cost as the payoff to the long-short investment strategy divided by its turnover. For instance, the break-even transaction cost for GKX method for the full sample equals $0.916\%/0.976 = 0.94\%$, where 0.916% is the value-weighted FF6-adjusted return (Table 1, Panel A), and 0.976 is the corresponding turnover (Table 5, Panel A). Unreported results show that if we exclude microcaps from the training sample (as shown in Table 2), the monthly turnover in the value-weighted long-short portfolio remains at 99% for the GKX method, which further creates a break-even transaction cost of 0.49%. In addition, when we use NN3 with the value-weighted loss function to predict alpha (as shown in the online appendix, Table IA5, Panel A), the monthly turnover in the value-weighted long-short portfolio in the full sample remains at 106%, which further creates a break-even transaction cost of 0.57%.

[29] One caveat of applying the transaction costs estimates from Novy-Marx and Velikov (2016) is that the transaction costs are based on effective bid-ask spread and discard the price impact and therefore underestimate the true costs faced by large traders in the market. On the other hand, the transaction costs are computed from aggregate data including multiple trade types and traders; hence, they overestimate the actual trading costs of a sophisticated institutional trader (Frazzini et al. 2018).

[30] Brandt et al. (2009) define T_t so that transaction costs in 1974 are four times larger than those in 2002. We follow their approach to compute a starting value of 2.6 in 1987.

[31] KNS rely on two sets of characteristics. One includes 50 firm characteristics underlying common anomalies, and the other is based on 80 predictive characteristics consisting of 68 financial ratios from WRDS and 12 variables based on past monthly returns. They supplement the two sets of raw characteristics with characteristics based on second and third powers and linear first-order interactions of characteristics. We adopt the same set of 94 firm characteristics as in GKX to be consistent with other results reported in this paper.

[32] The in-sample estimation ranges from September 1964 to December 2004 for the full sample and for the subsample excluding microcaps and ranges from January 1986 to December 2004 for the subsamples excluding nonrated firms and distressed firms due to data availability.

[33] KNS exclude stocks with market caps below 0.01% of aggregate stock market capitalization at each point in time. For perspective, this sample selection criterion excludes all microcaps and approximately 60% of the nonmicrocaps in our sample. We nevertheless implement their analysis for the full sample as well as three subsamples to demonstrate the role of economic restrictions. In addition, we also examine the important role of financial distress at the firm level and business conditions in the aggregate.

[34] This rescaling exercise addresses the concern of SDF scalability. Otherwise, portfolio weights are not uniquely identified because they are *proportional* to the inverse of the covariance matrix times the mean vector.

[35] A plausible way to mitigate extreme stock positions would be to add an L^1-penalty (as in the least absolute shrinkage and selection operator, namely, LASSO) in addition to L^2. Although the tuning parameter in LASSO can be found through a model selection criterion or cross-validation, its optimal value can still indicate extreme positions. To mitigate that concern, the tuning parameter value must be bounded. The addition of an L^1-penalty could challenge the nonsparse representation advocated by KNS.

[36] We thank Jeffrey Wurgler and Guofu Zhou for making the index of investor sentiment available via their websites,

http://people.stern.nyu.edu/jwurgler/; http://apps.olin.wustl.edu/faculty/zhou/zpublications.html.

[37] We obtain the monthly VIX index from the CBOE website, http://www.cboe.com/products/vix-index-volatility/vix-options-and-futures/vix-index/vix-historical-data.

[38] Unreported results show that the machine learning-based investments remain profitable in the post-2013 period, i.e., the last five years in our sample, and the monthly value-weighted long-short portfolio return is 0.72% (1.30%, 0.98%) across all stocks and 0.46% (1.31%, 1.26%) after excluding microcaps for the GKX (IPCA, CA) signal. In addition, unrestricted IPCA also predicts the cross-section of stock returns for the full sample and for subsamples with economic restrictions in the post-2001 period (online appendix, Table IA7, Panel C).

[39] In unreported tests, we run Fama and MacBeth (1973) regressions of realized excess returns on predictive signals derived from all four machine learning methods. Evidence indicates considerably lower slope coefficients in the post-2001 period. Thus, considering all stocks (beyond the extreme long and short portfolios), machine learning-based predictability seems to decrease in recent years.

[40] Unreported results confirm that our main findings are robust in the post-2001 period and various market states such as investor sentiment, market volatility, and market liquidity.

[41] Given the nature of this decomposition, we analyze machine learning signals based only on predicted returns in this subsection, that is, GKX, IPCA, and CA.

References

Adrian T, Shin HS (2010) Liquidity and leverage. *J. Financial Intermediary* 19(3):418–437.

Allena R (2021) Confident risk premia: Economics and econometrics of machine learning uncertainties. Working paper, University of Houston, Houston, TX.

Amihud Y (2002) Illiquidity and stock returns: Cross-section and time-series effects. *J. Financial Marketing* 5(1):31–56.

Arnott R, Harvey CR, Kalesnik V, Linnainmaa J (2019) Alice's adventures in factorland: Three blunders that plague factor investing. *J. Portfolio Management* 45(4):18–36.

Arnott R, Harvey CR, Markowitz H (2018) A backtesting protocol in the era of machine learning. Working paper, Duke University, Durham, NC.

Avramov D, Chordia T (2006) Asset pricing models and financial market anomalies. *Rev. Financial Stud.* 19(3):1001–1040.

Avramov D, Cheng S, Hameed A (2016) Time-varying liquidity and momentum profits. *J. Financial Quant. Anal.* 51(6):1897–1923.

Avramov D, Chordia T, Jostova G, Philipov A (2009) Dispersion in analysts' earnings forecasts and credit rating. *J. Financial Econom.* 91(1):83–101.

Avramov D, Chordia T, Jostova G, Philipov A (2013) Anomalies and financial distress. *J. Financial Econom.* 108(1):139–159.

Avramov D, Chordia T, Jostova G, Philipov A (2018) Bonds, stocks, and sources of mispricing. Working paper, George Mason University, Fairfax, VA.

Baker M, Wurgler J (2007) Investor sentiment in the stock market. *J. Econom. Perspective* 21(2):129–151.

Bianchi D, Büchner M, Tamoni A (2021) Bond risk premiums with machine learning. *Rev. Financial Stud.* 34(2):1046–1089.

Brandt MW, Santa-Clara P, Valkanov R (2009) Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *Rev. Financial Stud.* 22(9):3411–3447.

Brunnermeier MK, Pedersen LH (2009) Market liquidity and funding liquidity. *Rev. Financial Stud.* 22(6):2201–2238.

Bryzgalova S, Pelger M, Zhu J (2020) Forest through the trees: Building cross-sections of stock returns. Working paper, London Business School, London.

Carhart MM (1997) On persistence in mutual fund performance. *J. Finance* 52(1):57–82.

Cella C, Ellul A, Giannetti M (2013) Investors' horizons and the amplification of market shocks. *Rev. Financial Stud.* 26(7):1607–1648.

Chen L, Pelger M, Zhu J (2020) Deep learning in asset pricing. Working paper, Stanford University, Stanford, CA.

Chinco A, Neuhierl A, Weber M (2021) Estimating the anomaly base rate. *J. Financial Econom.* 140(1):101–126.

Choi D, Jiang W, Zhang C (2020) Alpha go everywhere: Machine learning and international stock returns. Working paper, The Chinese University of Hong Kong, Hong Kong.

Chordia T, Subrahmanyam A, Tong Q (2014) Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? *J. Accounting Econ.* 58(1):51–58.

Cochrane JH (2001) A rehabilitation of stochastic discount factor methodology. NBER Working Paper No. 8533, National Bureau of Economic Research, Cambridge, MA.

Cochrane JH (2011) Discount rates. *J. Finance* 66(4):1047–1108.

Cong LW, Tang K, Wang J, Zhang Y (2021) AlphaPortfolio for investment and economically interpretable AI. Working paper, Cornell University, Ithaca, New York.

Daniel K, Moskowitz T (2016) Momentum crashes. *J. Financial Econom.* 122(2):221–247.

Fama EF (1998) Market efficiency, long-term returns, and behavioral finance. *J. Financial Econom.* 49(3):283–306.

Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. *J. Financial Econom.* 33(1):3–56.

Fama EF, French KR (2008) Dissecting anomalies. *J. Finance* 63(4):1653–1678.

Fama EF, French KR (2015) A five-factor asset pricing model. *J. Financial Econom.* 116(1):1–22.

Fama EF, French KR (2018) Choosing factors. *J. Financial Econom.* 128(2):234–252.

Fama EF, MacBeth J (1973) Risk, return, and equilibrium: Empirical tests. *J. Political Econom.* 81(3):607–636.

Feng G, He J, Polson NG (2018) Deep learning for predicting asset returns. Working paper, City University of Hong Kong, Hong Kong.

Feng G, Polson NG, Xu J (2019) Deep learning in characteristics-sorted factor models. Working paper, City University of Hong Kong, Hong Kong.

Frazzini A, Israel R, Moskowitz TJ (2018) Trading costs. Working paper, Yale University, New Haven, CT.

Freyberger J, Neuhierl A, Weber M (2020) Dissecting characteristics nonparametrically. *Rev. Financial Stud.* 33(5):2326–2377.

FSB (2017) Artificial intelligence and machine learning in financial services. Financial Stability Board.

Gomes J, Kogan L, Zhang L (2003) Equilibrium cross-section of returns. *J. Political Econom.* 111(4):693–732.

Green J, Hand JRM, Zhang XF (2017) The characteristics that provide independent information about average U.S. monthly stock returns. *Rev. Financial Stud.* 30(12):4389–4436.

Green RC, Hollifield B (1992) When will mean-variance efficient portfolios be well diversified? *J. Finance* 47(5):1785–1809.

Griffin JM, Harris JH, Shu T, Topaloglu S (2011) Who drove and burst the tech bubble? *J. Finance* 66(4):1251–1290.

Gromb D, Vayanos D (2002) Equilibrium and welfare in markets with financially constrained arbitrageurs. *J. Financial Econom.* 66(2–3):361–407.

Gu S, Kelly B, Xiu D (2020) Empirical asset pricing via machine learning. *Rev. Financial Stud.* 33(5):2223–2273.

Gu S, Kelly B, Xiu D (2021) Autoencoder asset pricing models. *J. Econometrics* 222(1):429–450.

Hameed A, Mian GM (2015) Industries and stock return reversals. *J. Financial Quant. Anal.* 50(1–2):89–117.

Han Y, He A, Rapach D, Zhou G (2020) Firm characteristics and expected stock returns. Working paper, University of North Carolina at Charlotte, Charlotte, NC.

Hand JRM, Green J (2011) The importance of accounting information in portfolio optimization. *J. Accounting Auditing Finances* 26(1):1–34.

Hansen LP, Jagannathan R (1991) Implications of security market data for models of dynamic economies. *J. Political Econom.* 99(2):225–262.

Hansen LP, Richard SF (1987) The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. *Econometrica* 55(3):587–613.

Harvey CR (2017) The scientific outlook in financial economics. *J. Finance* 72(4):1399–1440.

Harvey, CR, Liu Y, Zhu H (2016) …and the cross-section of expected returns. *Rev. Financial Stud.* 29(1):5–68.

Heaton JB, Polson NG, Witte JH (2017) Deep learning for finance: Deep portfolios. *Appl. Stochastic Models Bus. Industry* 33(1):3–12.

Hong H, Lim T, Stein JC (2000) Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *J. Finance* 55(1):265–295.

Hou K, Xue C, Zhang L (2020) Replicating anomalies. *Rev. Financial Stud.* 33(5):2019–2133.

Huang D, Jiang F, Tu J, Zhou G (2015) Investor sentiment aligned: A powerful predictor of stock returns. *Rev. Financial Stud.* 28(3): 791–837.

Jagannathan R, Wang Z (2002) Empirical evaluation of asset-pricing models: A comparison of the SDF and beta methods. *J. Finance* 57(5):2337–2367.

Kan R, Zhou G (1999) A critique of the stochastic discount factor methodology. *J. Finance* 54(4):1221–1248.

Karolyi GA, Van Nieuwerburgh S (2020) New methods for the cross-section of returns. *Rev. Financial Stud.* 33(5):1879–1890.

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

Kim S, Korajczyk RA, Neuhierl A (2021) Arbitrage portfolios. *Rev. Financial Stud.* 34(6):2813–2856.

Kozak S, Nagel S, Santosh S (2020) Shrinking the cross-section. *J. Financial Econom.* 135(2):271–292.

Lettau M, Pelger M (2020a) Estimating latent asset-pricing factors. *J. Econometrics* 218(1):1–31.

Lettau M, Pelger M (2020b) Factors that fit the time series and cross-section of stock returns. *Rev. Financial Stud.* 33(5):2274–2325.

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

Merton RC (1980) On estimating the expected return on the market: An exploratory investigation. *J. Financial Econom.* 8(4):323–361.

Miller EM (1977) Risk, uncertainty, and divergence of opinion. *J. Finance* 32(4):1151–1168.

Nagel S (2012) Evaporating liquidity. *Rev. Financial Stud.* 25(7):2005–2039.

Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55(3):703–708.

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

Pástor L, Stambaugh R (2003) Liquidity risk and expected stock returns. *J. Political Econom.* 111(3):642–685.

Pontiff J, Woodgate A (2008) Share issuance and cross-sectional returns. *J. Finance* 63(2):921–945.

Rapach DE, Zhou G (2020) Time-series and cross-sectional stock return forecasting: New machine learning methods. Jurczenko E, ed. *Machine Learning for Asset Management* (John Wiley & Sons, Hoboken, NJ), 1–33.

Rapach DE, Strauss JK, Zhou G (2013) International stock return predictability: What is the role of the United States? *J. Finance* 68(4):1633–1662.

Rapach DE, Strauss JK, Tu J, Zhou G (2019) Industry return predictability: A machine learning approach. *J. Financial Data Sci.* 1(3): 9–28.

Rasekhschaffe KC, Jones RC (2019) Machine learning for stock selection. *Financial Anal. J.* 75(3):70–88.

Sak H, Huang T, Chng MT (2021) Exploring the factor zoo with a machine-learning portfolio. Working paper, Hong Kong University of Science and Technology, Hong Kong.

Stambaugh RF, Yu J, Yuan Y (2012) The short of it: Investor sentiment and anomalies. *J. Financial Econom.* 104(2):288–302.

Stambaugh RF, Yu J, Yuan Y (2015) Arbitrage asymmetry and the idiosyncratic volatility puzzle. *J. Finance* 70(5):1903–1948.

Stambaugh RF, Yuan Y (2017) Mispricing factors. *Rev. Financial Stud.* 30(4):1270–1315.

Titman S, Wei K, Xie F (2004) Capital investments and stock returns. *J. Financial Quant. Anal.* 39(4):677–700.

Welch I, Goyal A (2008) A comprehensive look at the empirical performance of equity premium prediction. *Rev. Financial Stud.* 21(4):1455–1508.

Wigglesworth R (2016) Money managers seek AI's 'deep learning'. *Financial Times*.

Zuckerman G, Hope B (2017) The Quants run Wall Street now. *Wall Street Journal*.