Asset Pricing with Attention Guided Deep Learning *

Philippe Chatigny[1,3], Ruslan Goyenko[2,3], and Chengyu Zhang[2,3]

[1] *University of Sherbrooke*
[2] *McGill University*
[3] *Financial Innovations and Risk Management Labs, FIRM*

This Version: July 18, 2022

Abstract

Deep learning methods, which can accommodate wide ranges of various stock characteristics to identify optimal investment portfolio or stochastic discount factor (SDF), have been criticised for extracting their superior performances from difficult to arbitrage stocks, high limits-to-arbitrage market conditions or extreme turnovers. We introduce *attention-guided* deep learning, which, in a data driven way, allows identifying the most influential time-varying firm characteristics contributing to SDF. Attention dramatically improves SDF performance and attention to multiple firm characteristics reduces portfolio rebalancing costs. The attention guided SDF outperforms existing models after excluding small and micro-cap stocks, avoids extreme portfolio weights, and unlike other models, exhibits the best performance during market regimes with the highest price efficiency.

Keywords: No-arbitrage, optimal portfolio, conditional asset pricing, deep learning, attention, big data

JEL Codes: G10, G12, G13, G14

*We are grateful to Peter Caines, Stefano Giglio, Bryan Kelly, Aditya Mahajan, Andreas Neuhierl (discussant), Pablo Piantanida as well as to conference participants of INQUIRE-UK 2022 spring residential, 4th Future of Financial Information Conference-2022, 2022 China International Conference in Finance and seminar participants at Yale School of Management, University of Manitoba, and the ISS Seminar at the Centre for Intelligent Machines, McGill University, for helpful comments. The authors acknowledge financial support from SSHRC.

Corresponding author: Ruslan Goyenko; email: ruslan.goyenko@mcgill.ca

1 Introduction

The recent abundance of financial data related to various stock characteristics provides alternative sources of information desirable to consider by investors while forming optimal portfolios.[1] The recent literature reportedly confirms and shows clear economic gains to investors who use big data and the most recent advancements in machine learning (ML) to accompany their investment decisions (Gu et al. (2020, 2021); Freyberger et al. (2020); Chen et al. (2020); Cong et al. (2020))

Using ML, and especially deep learning with big data for portfolio construction can come at severe costs. As argued by Avramov et al. (2021), a spectacular performance of many ML approaches in high dimensional financial data with very low signal-to-noise ratio is not robust to simple economic restrictions, as most of profitability is driven by difficult-to-arbitrage stocks, or short-selling positions, or by periods where arbitrageurs are the most financially constrained.

In this paper we propose a novel approach to portfolio construction which uses large set of stock characteristics, deep learning, and what differentiates us from the rest of the literature - the attention guided deep learning. Our attention networks are trained to identify the most relevant firm characteristics which subsequently can the best characterize stochastic discount factor (SDF) or pricing kernel, with the final goal to construct the optimal mean-variance efficient portfolio (Hansen and Jagannathan (1991)). These attention networks in our deep learning architecture are designed to "think" like a human. For example, ideally, a portfolio manager would not be considering a myriad of substitutable firm characteristics but would be trying to identify the most fundamental, non-redundant characteristics which contribute to the long-term overall portfolio performance. Moreover, this portfolio manager would, ideally, be considering these characteristics not only on a single stock level but also their complementarity vs substitutability across all stocks in the portfolio. Our attention guided deep learning neural networks are designed to do exactly that.

Similar to the rest of the literature (Chen et al. (2020), Kelly et al. (2019)), we use no-arbitrage assumption to train our neural networks. By construction, the attention networks have a dual purpose. First, they allow to shrink the redundant characteristics and find characteristics-sparse SDF, similar to Kozak et al. (2020). As a result, these characteristics then should reflect the most fundamental risks captured by SDF, which in turn avoids identifying anomalies attributed to difficult-to-arbitrage stocks (Avramov et al. (2021)).

[1] Earlier studies advocating portfolio choices conditional either on macro or asset specific variables include Brandt (1999); Aït-sahali and Brandt (2001); Brandt et al. (2009).

Second, by explicitly focusing attention on the most important fundamental asset- and cross-asset characteristics we implicitly allow for complementarity among these characteristics to contribute to the reduction in rebalancing/turnover costs on the whole portfolio level. As argued by DeMiguel et al. (2020), combining all stock characteristics reduces trading costs on the overall portfolio level as the trades required to rebalance in one underlying stock based on different characteristics often cancel out. Since the relevance of these characteristics can change from one month to another, similar to Kelly et al. (2019) and Gu et al. (2021) we let them to be determined in the data driven way.

Our contribution to the literature is three-fold. First, we show that attention guided deep learning portfolio performs well in out-of-sample and thus provides a good empirical proxy for an observable SDF. Our SDF achieves an out-of-sample, 01/2005 to 12/2020, annualized Sharpe ratio of 2.8, and annualized alpha of 11% after using Fama-French five factors and momentum, FF6, risk adjustment. High alpha indicates that FF6 do not capture the universe of all risk factors, which is similar to conclusions of Kelly et al. (2019), Kozak et al. (2020), and Gu et al. (2021).

Second, we also examine whether the economic gains of this portfolio survive after monthly re-balancing and trading costs. Unlike Avramov et al. (2021), who re-examine previous studies and find that most of the potential profits disappear after accounting for trading costs, we find that significant gains remain. For example, when we drop from our sample bottom 10% (20%) of market capitalization stocks based on NYSE breakpoints (Hou et al. (2020)), the Sharpe ratio decreases to 1.26 (1.01), and FF6- alpha is 4% (3%) per year, and still being highly statistically significant. If we only use 3000 stocks of Russell 3000 index which automatically excludes small stocks, we obtain Sharpe ratio of 1.5, and an annualized FF6 alpha of 5.1% (t=4.94). For comparison, Sharpe ratio of S&P500 (Russell 3000) for this period is 0.56 (0.55). Therefore, excluding micro- and small-cap stocks, an investor still can substantially outperform the market. Moreover, the portfolio turnover is quite low, 35%, with the average turnover on the long position of 22%, and short position of 13%. The average trading costs measured by relative effective bid-ask spreads after excluding bottom 20% smallest stocks is 15 bps, or 27 bps for the cross-section of Russell 3000 components. Overall, we conclude that attention to multiple firm characteristics (DeMiguel et al. (2020)) helps to reduce the noise of less relevant characteristics, focus the optimization network's attention on the characteristics which define the most fundamental signals and allows an investor to achieve gains after trading costs.

Third, we contribute methodologically to the growing literature on application of ML/deep-learning in asset pricing. We adapt the architecture of Neural Basis Expansion Analysis for

Interpretable time series forecasting (N-BEATS) of Oreshkin et al. (2019) for asset pricing on the cross-section of all US publicly listed common stocks, their 94 Green et al. (2017) firm-specific characteristics and 18 macro-variables that we describe below. N-BEATS uses deep neural architecture which is well entrenched in computer vision or natural language processing but has very limited applications for financial data forecasts (with the only exception of Chen et al. (2020)). As the name suggests, N-BEATS is originally designed for time-series forecasting (M3, M4 and TOURISM data sets) where it has been shown to outperform individual or ensembles of classical statistical techniques (Oreshkin et al. (2019, 2020)).

The residual network architecture of N-BEATS makes it particularly appealing for constructing bottom-up optimal SDF portfolio. Each N-BEATS' block has two branches, backcast and forecast predictions. Each upstream block removes the portion of the signal that it can approximate well, the forecast, and the residual, the backcast, is passed on to another block, making the forecasts of downstream blocks easier. Therefore, the whole process can be thought as running a sequential analysis of the input signal in one consistent network. It makes N-BEATS more interpretable, and easier to train compared to most other deep neural networks.

The input here is firm-specific characteristics interacted with macro-variables. Economically, the first block of the network forecasts the first features-group importance, e.g. characteristics to approximate size factor. The output of this block is the forecasts based on the size factor. The backcast of this block is the residual, which for this particular example can include value and momentum related signals. The next downstream block then identifies feature-group related to, e.g., value, and produces the forecasts based on the value factor. Its backcast is the residual, momentum, which is subsequently passed to the next block to make the forecasts based on momentum factor, and so on. The final output, the prediction, is the sum of predictions across all blocks' partial forecasts. As our objective function is to minimize SDF loss function, by construction, each block of the network is intended to identify a specific risk factor. The final SDF is the linear combination of the risk factor premiums, where the factors are identified in the data driven way.

This analysis is intuitively similar to IPCA approach of Kelly et al. (2019), or autoencoder of Gu et al. (2021). The difference is that these papers are focused on identifying risk factors non-parametrically. While we identify risk factors within our network, via block structure, our output is the final SDF which is a compensation for the risks carried by the factors. The number of factors in our model is determined by the number of blocks. Kelly et al. (2019), Kozak et al. (2020) and Lettau and Pelger (2020) show that 5 or 6 factors are sufficient to span SDF. We confirm this intuition as our network with 5 blocks performs similarly to the

one which uses 6 blocks. The biggest difference from the previous literature, however, is that our factors are *attention* determined, i.e. each block which identifies a risk factor is preceded by an *attention* layer. As we show by comparing the base, no *attention*, versus *attention* guided specification, attention is critical in our model, as the base model without it, similar to all other ML models, especially deep learning models, falls under umbrella of Avramov et al. (2021) criticism.

The asset attention we implement is based on a very popular in computer science *self-attention*, which is normally used in *Transformers* architecture. *Transformers* are widely used in a variety of applications in supervised machine translation, speech recognition and natural language processing. *Attention*, the way we train it in the beginning of each block determining a risk factor, allows identifying which firm characteristics are more influential than others in defining a non-parametric risk factor. Instead of pure *Transformers* architecture which has been used by Cong et al. (2020) in a different context, we deploy FAVOR+, the most recent AI algorithm used in *Performers*, introduced by Choromanski et al. (2020). Unlike *Transformers*, *Performers* use *linear* space and time complexity that approximates *Transformers* with high accuracy but with more efficient convergence properties, i.e. an improved version of *Transformers*.

To make the argument more convincing, we condition SDF performance on the market regimes identified by high vs low market volatility or illiquidity. Theoretically, fewer trading frictions, lower volatility and more arbitrage activity should improve price efficiency. Higher price efficiency should allow our networks to better identify priced risk factors, make better asset allocation decisions, and exhibit better performance compared to low price efficiency regimes. Avramov et al. (2021) convincingly argue that the stellar performance of many ML approaches in the literature generates most of their profitability during high market volatility and illiquidity regimes, i.e. during high limits-to-arbitrage market episodes.

In contrast to the previous literature, our SDF exhibits the best performance during low limits-to-arbitrage market episodes. For example, during high market volatility episodes, measured by VIX being above its sample median, our SDF has an annualized Sharpe ratio of 2.12 vs Sharpe ratio of 4.55 during low market volatility regimes. If we consider eliminating stocks at the bottom 10% (20%) based on NYSE size breakpoints, Sharpe ratio drops to 0.868 (0.532) in high volatility regime, or to 2.24 (2.26) during low volatility regime. We obtain very similar results conditioning on high vs low market illiquidity episodes.

Moreover, our attention guided deep learning network identifies a different factor structure during high vs low market volatility regimes. Consider SDF specification based on the most liquid cross-section of stocks, after eliminating stocks at the bottom 20% of NYSE market

capitalization. During high volatility regime, SDF portfolio's average excess over risk free rate return is 3.71% , and FF6-alpha is 3.85% (t=2.97) per year. Here, the alpha and the mean returns are the same, indicating that FF6 factors do not present any investment signal during higher market turbulence, and alternative, non-parametric factors play bigger role. During low market volatility regime, the average excess return is almost twice higher, 7.24% per year. FF6 alpha is more than 3 times smaller then the excess return, 2.15% (t=2.35) per year. It suggests that during low limits-to-arbitrage market episodes, FF6 factors are important. However, given economically and statistically significant alpha, it also suggests that FF6 factors are not sufficient to span SDF even during good market conditions.

We analyze the variable importance identification by each attention layer preceding each of 5 blocks in the model for the last month in our out out-of-sample predictability, December 2020, i.e. the last time the model was re-trained for the training sample ending in November 2020, and compare it to more turbulent market regime, September 2008, the inception of financial crisis. Consistent with the alpha realizations for different volatility/illiquidity regimes, we find that majority of firm-specific characteristics identified by five attention layers are closely related to Fama-French 5 factors and momentum characteristics in normal market conditions, December 2020. In contrast, for asset allocation decisions for September 2008, a completely different set of firm and macro-economic characteristics is identified. Since the block structure of our model also allows to rank the importance of the factors, we find that in the end of August 2008, macro factor which is captured by default spread, CPI inflation, S&P500 returns, and betting against beta, BAB, funding liquidity factor (Frazzini and Pedersen (2014)) is the first in the ranking, i.e. identified by the first upstream block. The other factors are related to past returns, price jumps, the strength of the balance sheet, i.e. the ability to weather the storm, stock volatility, firm assets volatility, and firms' real estate holdings. The latter is interesting, since the model was never explicitly told that the incoming crisis is caused by real estate bubble.

Overall, unlike the limits-to-arbitrage argument of Avramov et al. (2021), we conclude that carefully constructed deep learning network guided by economic fundamentals can overcome economic restrictions and provide profitable investment opportunities. In particular, our approach advocates machine learning attention as an important instrument in financial data analysis. Another important message is that ML approaches to the optimal portfolio construction provide substantial improvements and economic gains over traditional approaches motivated by the modern portfolio theories (Markowitz, 1952). The rest of the paper is organized as follows. Section 2 overviews related literature. Section 3 introduces the modelling architecture framework. Section 4 describes the data and empirical results. Section 5 discusses

economics of *Attention* mechanism, and Section 6 concludes the paper.

2 Related Literature

Our paper is related to the recent but fast growing literature on applications of machine learning in asset pricing.[2] The fast growth of this literature was partly motivated by AFA presidential address of Cochrane (2011) who highlights the abundance of noisy and highly correlated return predictors, and the need for other methods beyond cross-sectional regressions and portfolio sorts. ML is an obvious solution, as ML methods are perfectly suited for processing big data, condensing sparsity among large sets of predictors, reducing data dimensionality, and putting emphasis on non-redundant variable selection.

This new literature has been first pioneered by Gu et al. (2020) who demonstrate the advantages of various machine learning methods for predicting the panel of individual US stock returns with multiple firm characteristics. Similarly, Bianchi et al. (2021) show the ability of ML methods to predict bond returns.

A separate strand of the literature applies ML methods to non-parametrically obtain systematic risk factors. Lettau and Pelger (2020) introduce risk premium (RP) PCA that leveraging multiple firm characteristics, incorporates information about the first and second moments of the data and yields a more efficient estimator than standard PCA. RP-PCA extract five significant factors that capture most of time series and cross-sectional variations in the data.

Kelly et al. (2019) introduce instrumental PCA, IPCA, to perform dimensionality reduction of the characteristic space. They project the individual stock returns into managed portfolios based on observed characteristics and then apply PCA to the projected data. This allows for time-varying factor loading and the linear dependence of the loading on characteristics. Gu et al. (2021) use autoencoder to allow for non-linear dependence of time varying factor loadings on firm characteristics. Both Kelly et al. (2019) and Gu et al. (2021) find that 5 to 6 risk factors are sufficient to span SDF.

Feng et al. (2021) impose a no-arbitrage constraint and estimate factor risk loadings using a deep neural network. Freyberger et al. (2020) employ factor selection with Lasso-style penalty, and assume that stochastic discount factor (SDF) has a sparse exposure to firm

[2]See for example, Heaton et al. (2017), Feng et al. (2021), Rapach et al. (2019), Rapach and Zhou (2020), Choi et al. (2020), Freyberger et al. (2020), Gu et al. (2020), Gu et al. (2021), Han et al. (2020), Bianchi et al. (2021), Cong et al. (2020), Kim et al. (2021), Kelly et al. (2019), Chen et al. (2020), Kozak et al. (2020), Lettau and Pelger (2020), Bryzgalova et al. (2020).

characteristics.

Kozak et al. (2020) construct optimal portfolio using Ridge type estimator and apply it to the candidate factors spanning SDF. This authors too find that 5 factors are sufficient to span SDF. They show however that traditional Fama-French 5 factors and momentum are not efficient enough to describe the complexity of SDF. This conclusion motivates the rest of the literature searching for more efficient non-parametric factors to capture the risk universe of current markets.

Bryzgalova et al. (2020) warn that besides unknown factors, the cross-section of assets to build them from is not well determined. The authors use asset pricing (AP) decision trees to first build a cross-section of asset returns, and then apply elastic net to this cross-section in order to construct an optimal portfolio.

More recently, Chen et al. (2020) employ deep neural networks to construct bottom-up SDF conditional on various firm characteristics and latent macro-economic states for all U.S. equities. The authors use Generative Adversarial Network (GAN) of Goodfellow et al. (2014) to optimize SDF loss function, where SDF's stock weights are determined by multiple stock characteristics and the latent macro-regimes identified by LSTM cell. Chen et al. (2020) strictly emphasize the challenges that deep neural network approaches can help to overcome: (i) the dependence of SDF on all available information implying that SDF is potentially a function of very large set of variables; (ii) the unknown and potentially complex functional form of SDF; (iii) the dynamic structure of risk exposures over time for individual assets which potentially can depend not only on macro-economic conditions but also on changing asset-specific characteristics over time.

Cong et al. (2020) build Alpha portfolio with reinforcement learning by directly optimizing Sharpe ratio. This is a different exercises from bottom up SDF construction or identifying risk factors spanning SDF, as their model is trained to maximize Sharpe ratio of Long-Short arbitrage strategy. The authors however clearly show the advantages of AI approaches and big data to asset pricing as their Alpha portfolio outperforms all existing analogs in the literature.

As ML in asset pricing literature seem to start reaching the heights of its popularity and acceptance by both academics and practitioners, Avramov et al. (2021) issue a big warning. The ML algorithms, while applying to portfolio management, are detrimental to performance after trading costs, limits-to-arbitrage or leverage considerations, or performing the best during market regimes when arbitragers are the most financially constrained. The authors examine the majority of the ML models described above, and find that they are practically not implementable.

In this paper we address Avramov et al. (2021) criticism and show that ML combined with big data to construct tradable SDF portfolio outperforms existing analogs in the traditional empirical asset pricing literature after trading costs, and in the various market regimes. Our approach is based on two pillars.

First, as argued by DeMiguel et al. (2020), incorporating a large set of firm-specific characteristics is critical to reduce trading costs on the whole portfolio level. This approach however is not new, as conditional SDF based on many firm characteristics is already implemented (Chen et al. (2020), Kozak et al. (2020)), and yet its after-trading cost, turnover and extreme positions performances are questioned by Avramov et al. (2021).

Second, the simplicity and tractability of the neural architecture in N-BEATS allows us modelling an important economic feature - Asset Attention. Asset Attention to firm specific characteristics, applied every time before rebalancing portfolio weights, enables tradability of the whole SDF portfolio.

The attention mechanism has been used by Cong et al. (2020) towards the output of their networks to rank stocks to identify the top and bottom 10% for their Long and Short portfolio positions. Instead, we apply the attention mechanism to the input, rather than the output, in our network. Therefore, the attention layer serves as the time varying filter to our deep learning network about the most meaningful features determining the risk factors. Empirically, the fundamental nature of these features allows minimizing turnover on the whole portfolio level (DeMiguel et al. (2020)).

Intuitively, our model architecture is closely related to autoencoder PCA of Gu et al. (2021). Our building blocks in the network, similar to autoencoder, allow for non-linearity among firm characteristics to characterize factor risk loadings. Similar to the authors, we identify that 5 blocks (factors) are sufficient to span SDF. Although we do not explicitly model factor loadings, it is implicit in the model architecture that their affect on the asset returns is linear. The difference from autoencoder PCA of Gu et al. (2021) and all other ML methodologies is that we determine risk factors sequentially, and with the *asset attention* which permits better identification of their time varying nature. Moreover, the upstream vs. downstream nature of the blocks in the model allows to rank the factors in the order of their relative importance. We show that not only the nature of factors changes between different market regimes, but also their ranking of importance.

3 Model

3.1 No-Arbitrage Asset Pricing

The no-arbitrage assumption implies the existence of stochastic discount factor, SDF, M_{t+1}, which for any excess return, $R^e_{t+1,i}$, satisfies the equation 1.

$$\mathbb{E}[M_{t+1}R^e_{i,t+1}] = 0 \Leftrightarrow \mathbb{E}_t[R^e_{i,t+1}] = \overbrace{\frac{\mathrm{Cov}_t(R^e_{i,t+1}, M_{t+1})}{\mathrm{Var}_t(M_{t+1})}}^{\beta_{t,i}} \cdot \overbrace{\frac{\mathrm{Var}_t(M_{t+1})}{\mathbb{E}_t[M_{t+1}]}}^{\lambda_t} = \beta_{i,t}\lambda_t \tag{1}$$

The factor loading, $\beta_{i,t}$, is exposure to the systematic risk factors, and λ_t is the risk price associated with factors. M_{t+1} is a linear function of factors, which maps to a factor model for excess returns of the following form:

$$R^e_{i,t+1} = \alpha_{i,t} + \beta'_{i,t}F_{t+1} + \epsilon_{i,t+1} \tag{2}$$

where $E_t\left[\epsilon_{i,t+1}\right] = E_t\left[\epsilon_{i,t+1}F_{t+1}\right] = 0, \quad E_t\left[F_{t+1}\right] = \lambda_t$, and most importantly, $\alpha_{i,t} = 0$ for all i in t.

Along the lines of Hansen and Jagannathan (1991), the SDF can be formulated as

$$M_{t+1} = 1 - \sum_{i=1}^{N_t} \omega_{t,i}R^e_{i,t+1} = 1 - \boldsymbol{\omega}_t^\top \boldsymbol{R}^e_{t+1}, \tag{3}$$

The fundamental Euler equation, $\mathbb{E}_t\left[R^e_{t+1}M_{t+1}\right] = 0$, implies the SDF weights

$$\omega_t = \mathbb{E}_t\left[R^e_{t+1}{R^e_{t+1}}^\top\right]^{-1}\mathbb{E}_t\left[R^e_{t+1}\right] \tag{4}$$

These are the weights for any portfolio on the mean-variance efficient frontier. The tangency, optimal, portfolio is then defined as $F_{t+1} = \omega_t^\top R^e_{t+1}$ and we denote this factor as the traded SDF.

By construction, the SDF is a single traded factor which characterizes and can price the best all cross-section of available assets. The individual assets, however, as specified in eq. (2) can have a multi-factor structure.

3.2 Conditional SDF Loss Function

The estimation of SDF weights is a multi-dimensional optimization. For each individual asset i, and time t we consider the following functional form for the weight,

$$\omega_{t,i} = \omega\left(\mathbf{I}_t, \mathbf{I}_{t,i}\right) \tag{5}$$

where $\mathbf{I}_t$ is macro-economic conditioning variables that are not asset specific, and $\mathbf{I}_{t,i}$ are firm characteristics which are asset specific, like size, momentum, book-to-market and etc.

Following intuition of Hansen and Jagannathan (1997) which was also implemented by Chen et al. (2020) in the setting similar to ours, the general form of our optimization function is

$$\omega_{t,i}^{*} = \underset{\omega_{t,i}}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} \left(\overbrace{\left(1 - \sum_{i=1}^{N_t} \omega_{t,i}\left(\mathbf{I}_t, \mathbf{I}_{t,i}\right) R_{t+1,i}^{e}\right)}^{M_{t+1}} R_{t+1,n}^{e} \right)^2 \tag{6}$$

3.3 Model Architecture

The model architecture we propose relies on the N-BEATS model from Oreshkin et al. (2019). We use its deep neural architecture, which we improve upon by adding *Asset Attention* layer described in details below. Figure 1 presents the model architecture. The model consists of several blocks the forecasts of which are assembles on a stack level.

Each block, Figure 1 (left), takes as input a vector $\mathbf{x}_{\ell,i} \in \mathbb{R}^D$ and outputs two vectors $\tilde{\mathbf{x}_{\ell,i}} \in \mathbb{R}^D$ and $\tilde{\mathbf{y}_{\ell,i}} \in \mathbb{R}^H$ The first block of the model ($\ell = 1$) considers input $\mathbf{x}_{1,i} = [\mathbf{I}_t, \mathbf{I}_{t,i}]$, i.e., a set of macroeconomic conditioning variables that are not asset specific ($\mathbf{I}_t \in \mathbb{R}^{D_1}$) and firm-specific characteristics observed at time ($\mathbf{I}_{t,i} \in \mathbb{R}^{D_2}$) such as $D = D_1 + D_2$. For any other block, its inputs $\mathbf{x}_{\ell,i}$, is the residual outputs of the previous block $\tilde{\mathbf{x}}_{\ell-1,i}$ that we refer as the "backcast". Each block produces a partial forecast $\tilde{\mathbf{y}_{\ell,i}}$ that are aggregated in a hierarchical fashion.

Each basic building block is divided into three parts: **(1)** An *Asset attention* layer with linear complexity in respect to $\mathbf{x}_\ell$ length that transform the input vector to $\tilde{\mathbf{x}}_\ell$ keeping the same dimensionality, **(2)** An FC network that projects $\tilde{\mathbf{x}}_\ell$ into a fixed higher-dimensional representation $\mathbf{z}_L \in \mathbb{R}^+$ from which we produce forward $\mathbf{c}_\ell^f$ (Forecast) and backward $\mathbf{c}_\ell^b$ (Backast) coefficients **(3)** the backward g_ℓ^b and forward g_ℓ^f that project the coefficient from

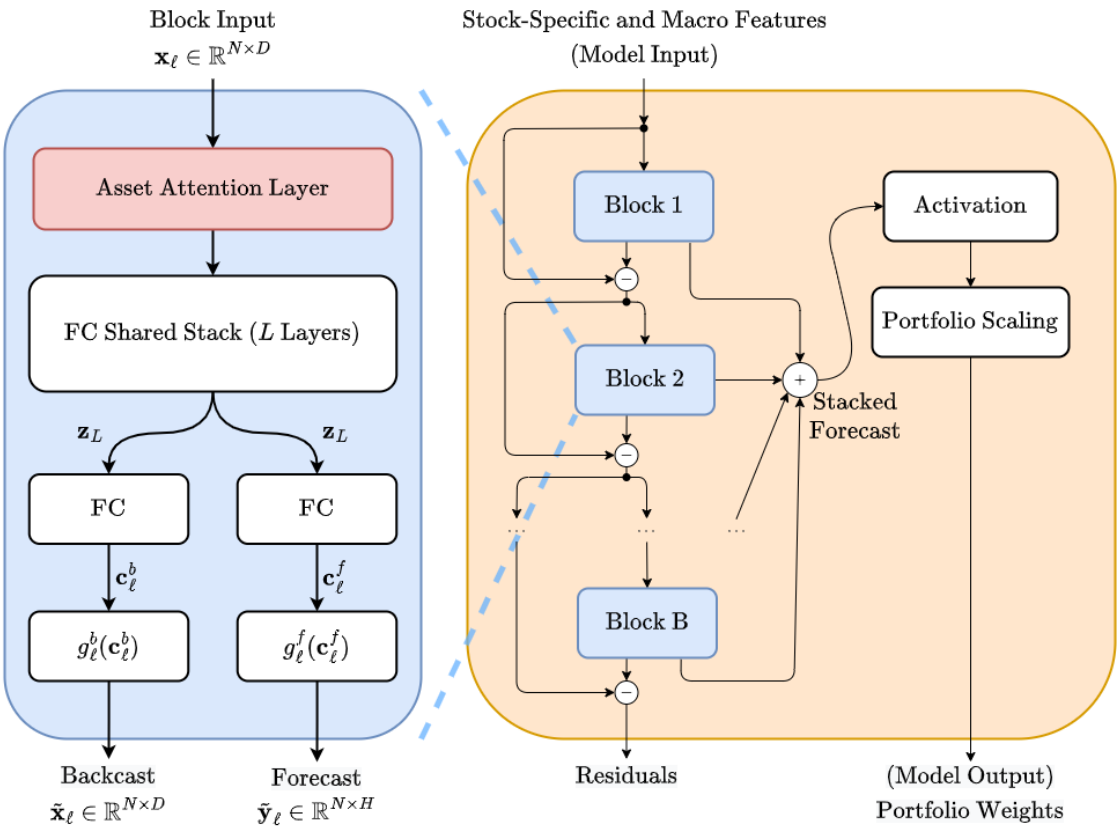


Figure 1: Illustration of the model. The basic block (left) consists of an *asset attention* module that applies *self-attention* on the feature input, followed by a multi-layer FC network with ReLu nonlinearities activation function. It predicts basis expansion coefficients both forward $\mathbf{c}_\ell^f$ (Forecast) and backward $\mathbf{c}_\ell^b$ (Backcast) that are projected over basis functions to produce the backcast $\tilde{\mathbf{x}}_\ell$ and the forecast $\tilde{\mathbf{y}}_\ell$. Forecasts are aggregated in hierarchical fashion by adding all partial forecast of each block. These partial forecasts are then passed through an activation function and portfolio weights scaling to construct the SDF.

the first part internally on the set of basis functions and produce $\tilde{\mathbf{x}}_\ell$ and $\tilde{\mathbf{y}}_\ell$.

These operation are described by the following set of equations (7):

$$\begin{aligned}
\mathbf{z}_{1,i} &= \mathrm{FC}_\ell(\widehat{\mathrm{Att}_{\leftrightarrow}}(\mathbf{x}_{\ell,i})) & \mathbf{z}_{l,i} &= \mathrm{FC}_\ell(\mathbf{z}_{l-1,i}) \\
\mathbf{c}_{\ell,i}^f &= \mathrm{FC}_{T_w}^f(\mathbf{z}_{L,i}) & \mathbf{c}_{\ell,i}^b &= \mathrm{FC}_{T_w}^b(\mathbf{z}_{L,i}) \\
\tilde{\mathbf{y}}_{\ell,i} &= g_\ell^f(\mathbf{c}_\ell^f) = \sum_{j=1}^{\dim(\mathbf{c}_\ell^f)} \mathbf{c}_{\ell,i}^f \mathbf{v}_j^f & \tilde{\mathbf{x}}_{\ell,i} &= g_\ell^b(\mathbf{c}_\ell^b) = \sum_{j=1}^{\dim(\mathbf{c}_\ell^b)} \mathbf{c}_{\ell,i}^b \mathbf{v}_j^b
\end{aligned} \tag{7}$$

Here, $l \in \mathbb{N}^+, 1 < l < L, l_l \in L = \{l_1, \cdots l_L\}$ is the number of layers in the ℓ-th block and $\mathbf{z}_{l,i}$ corresponds to the embedding computed at the l-th hidden layers for the i-th asset. FC corresponds to a fully connected layer with ReLU non-linearity activation Nair and Hinton (2010), and $\mathbf{v}^f$ and $\mathbf{v}^b$ are forecast and backcast basis vectors. These vectors are set to be

learnable parameters in our case, but could be set to a specific functional forms that are fixed prior training the model.

For the *self-attention* module, we rely on the *Fast Attention Via positive Orthogonal Random features* (FAVOR+) method of Choromanski et al. (2020) instead of the regular *self-attention* (Vaswani et al. (2017)) for scalability purposes.

The traditional *self-attention* considers a sequence $\tilde{\mathbf{x}}_\ell = [\mathbf{x}_\ell^{(1)}, \cdots \mathbf{x}_\ell^{(D)}]; d \in \{1 \cdots D\}$ with $\mathbf{x}_\ell^{(d)} \in \mathbb{R}^{L_{\text{dim}}}$ such that L_{dim} is the size of input sequence of observed tokens. Then regular *dot-product attention* is a mapping of tokens to *queries, keys and values* matrices $\boldsymbol{K}, \boldsymbol{Q}, \boldsymbol{V} \in \mathbb{R}^{L_{\text{dim}} \times r}$ where r is the dimension of each token [3]. Eq.8 demonstrates how *dot-product attention* is usually done where $\boldsymbol{A} \in \mathbb{R}^{D \times D}$, $\mathbf{1}_D$ is the all-ones vector of length D, $\exp(\cdot)$ is the exponential function applied element wise and $\text{diag}(\cdot)$ is a function that returns the resulting diagonal matrix of the matrix provided as input.

$$\text{Att}_{\leftrightarrow}(\boldsymbol{K}, \boldsymbol{Q}, \boldsymbol{V}) = \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{V}, \quad \boldsymbol{A} = \exp(\boldsymbol{Q}\boldsymbol{K}^\top/\sqrt{r}) \quad \boldsymbol{D} = \text{diag}(\boldsymbol{A}\mathbf{1}_D) \tag{8}$$

Instead FAVOR+ relies on a *random feature map* $\phi(\boldsymbol{u}) \in \mathbb{R}_+^{r'}$ for $r' > 0$ given a kernel $K : \mathbb{R}^r \to \mathbb{R}_+$ defined for the mapping $\phi : \mathbb{R}^r \to \mathbb{R}_+^{r'}$ such as $K(\boldsymbol{x}, \boldsymbol{y}) = [\phi(\boldsymbol{x})^\top \phi(\boldsymbol{y})]$[4]. We impose $\boldsymbol{A}$ to follow the form of $\boldsymbol{A}(i,j) = K(\boldsymbol{q}_i^\top, \boldsymbol{k}_j^\top)$ with $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$ being respectively the i-th and j-th query and key row-vector in $\boldsymbol{Q}$ and $\boldsymbol{K}$. Thus given $\boldsymbol{Q}', \boldsymbol{K}' \in \mathbb{R}^{D \times r'}$ with rows given respectively by $\phi(\boldsymbol{q}_i^\top)^\top$ and $\phi(\boldsymbol{k}_i^\top)^\top$ the self attention is approximate by eq. 9.[5] Overall, the aim of the *self-attention* layer is to evaluate the relative importance of each covariate in $\mathbf{x}_\ell$ with respect to others and adjust the input vector of the block based on features that matter the most. The self-attention is applied asset-wise, hence the term: *Asset attention layer*.

$$\widehat{\text{Att}_{\leftrightarrow}}(\boldsymbol{K}, \boldsymbol{Q}, \boldsymbol{V}) = \widehat{\boldsymbol{D}}(\boldsymbol{Q}'((\boldsymbol{K}')^\top \boldsymbol{V}), \quad \widehat{\boldsymbol{D}} = \text{diag}(\boldsymbol{Q}'((\boldsymbol{K}')^\top \mathbf{1}_D)) \tag{9}$$

FAVOR+ has a critical advantage over traditional *self-attention*. As can be seen from eq. 9, unlike regular *self-attention*, it has a linear architecture. The linear architecture provides strong theoretical guarantees to achieve unbiased estimation of the attention matrix, uniform convergence and lower variance of approximation. This makes FAVOR+ an ideal candidate to

[3] For brevity, the notation of the *queries, keys and values* and *attention matrix* $\boldsymbol{A}$ for the i-th ts and ℓ-th block is dropped without loss of generality. All matrices denoted by a bold capital letter is associated to one-and-only i-th asset for the ℓ-th block; e.g: $\boldsymbol{K} \Leftrightarrow \boldsymbol{K}_{\ell,i}$

[4] Here, $\boldsymbol{x}$ and $\boldsymbol{y}$ are input vector of the mapping ϕ and are not to be confused with the input and output of the model.

[5] For more details and explanation of the FAVOR+ method, we refer readers to Choromanski et al. (2020).

apply self-attention, or *Asset Attention* as we name it, to larger sets of features efficiently without a significant increase in computational costs.

Eqs. in 7 are then repeated iteratively for ℓ blocks. The individual blocks are stacked using two residual branches. The first branch, illustrated in Fig. 1 (right), runs over the backcast signal produced by each block and iteratively decomposes the input vector such that the subsequent block considers the residuals of the preceding block. The partial forecasts of each block are aggregated on the stack level by simply adding them:

$$\tilde{\mathbf{y}}_i = \sum_{l}^{\ell} \tilde{\mathbf{y}}_{l,i} \tag{10}$$

To simplify the notation from earlier, we consider eq. 11 as the function that establishes the contribution of the i-th asset $\mathbf{x}_i = [\mathbf{I}_t, \mathbf{I}_{t,i}]$ to M_t, where θ_f is the set of all parameters of all blocks.

$$\tilde{\omega}_i = \tilde{\mathbf{y}}_i = f_{\theta_f}(\mathbf{x}_i) \tag{11}$$

Given the functional form for the weights, eq. 6 can now be re-written as

$$\omega_i^* = \underset{\theta_f^*}{\operatorname{argmin}} \frac{1}{N} \sum_{n=0}^{N} \left(\frac{T_n}{T} \sum_{t=1}^{T_n} \overbrace{\left(1 - \sum_{i=1}^{N_t} f_{\theta_f}(\mathbf{x}_{t,i}) R_{t,i}^e \right)}^{M_t} R_{t,n}^e \right)^2 \tag{12}$$

where T is total sample size of the training sample, T_n is the number of month for a particular stock in the training sample, given that the panel is unbalanced, and N is the total number of stocks. To run minimization problem in Eq. 12, we use a stochastic gradient descent (SGD) optimization with Adam algorithm of Kingma and Ba (2015), fixed-sized batches of uniformly sampled assets with replacement and a three-step learning rate schedule over a fixed set of iterations. Similar to Oreshkin et al. (2019) and Chen et al. (2020), we perform a bagging procedure of Breiman (1996) by considering 10 models in the ensemble which are trained with different random initialization each month. Additional details of the procedure and hyper-parameters are specified in Appendix B.

4 Empirical Results

4.1 Data and Empirical Design

Our sample period is from 1990-01 to 2020-12 and it is defined by the availability of CBOE VIX index which we use in the set of macroeconomic variables. We consider all universe of US publicly listed common stocks as well as the cross-section of 3000 components of Russell 3000 index[6]. The monthly stock returns are obtained from *CRSP*.

Similar to Gu et al. (2020) we rely on a large set of 94 stock-level predictors of Green et al. (2017). The definitions of these characteristics are provided in the Appendix, Table A.1.[7]

We also use 12 macroeconomic series following the variable definitions detailed in Welch and Goyal (2008), including dividend-price ratio (dp), dividend yield (dy), earnings-price ratio (ep), stock variance (svar), book-to-market ratio (bm), net equity expansion (ntis), Treasury-bill rate (tbl), long term rate of returns (ltr), term spread (tms), default spread (dfy), default return spread (dfr), and Consumer Price Index (infl)[8]. To the set of these variables we also add the returns on S&P500 index, and VIX to proxy for overall market volatility and sentiment, as well as market wide illiquidity measured by Amihud (2002) ILLIQ ratio (amihud) and bid-ask spread measure of Corwin and Schultz (2012). We first compute these measures on the stock level, and then aggregate them on the market level. To proxy for the market-wide funding illiquidity we use TED spread and Frazzini and Pedersen (2014) betting against beta, BAB, factor. The data on TED and BAB are from FRED and AQR websites respectively. All macroeconomic variables are summarized in Table A.2.

We use the sample from 01-1990 to 12-2004, 180 months, as the first training sample. Our first out-of-sample, *OOS* prediction for SDF weights is therefore for January, 2005. We then roll the training sample by 1 month, keeping its 180-month length fixed, and our next *OOS* prediction is for February, 2005, and so on. Overall, our *OOS* period is from 01-2005 to 12-2020 for a total of 192 months. All results reported below are for our *OOS* forecast period.

[6]We are grateful to Paul Schultz for providing the data on Russell 3000 constituents

[7]To construct the variables we use the SAS code available from Jeremiah Green's Web site. As it is a common practice with big data, similar to Green et al. (2017) and Gu et al. (2020) we replace missing characteristics with the cross-sectional median at each month for each stock.

[8]The monthly data for these variables are from Amit Goyal's website

	All Stocks		q>10%		q>20%		Russell 3000 Stocks	
	Attention	**Base**	**Attention**	**Base**	**Attention**	**Base**	**Attention**	**Base**
Return	0.0099	0.0106	0.0050	0.0057	0.0046	0.0058	0.0057	0.0062
Std.Dev.	0.0121	0.0145	0.0138	0.0193	0.0157	0.0237	0.0133	0.0180
Alpha	0.0088	0.0095	0.0031	0.0026	0.0022	0.0017	0.0042	0.0039
Alpha t-stat	8.060	7.657	3.694	2.967	2.616	2.058	4.936	4.145
Sharpe	2.834	2.537	1.255	1.016	1.006	0.842	1.486	1.196
InfRatio	2.639	2.389	0.910	0.682	0.625	0.468	1.214	0.933
MaxDD	0.033	0.060	0.073	0.147	0.103	0.243	0.065	0.131
Max 1M Loss	-0.032	-0.030	-0.051	-0.083	-0.058	-0.107	-0.048	-0.083

Notes: The table presents summary statistics for *OOS*, 01-2005 to 12-2020, SDF performance for different cross-sections of stocks. Two sets of results refer to attention guided deep learning model, denoted *Attention*, and the base model, without asset attention, denoted *Base*. The statistics are average monthly portfolio returns and standard deviations, monthly portfolio alpha, Alpha, after FF6 factors risk adjustment, and its robust for autocorrelation and heteroscedasticity t-statistics, annualized Sharpe and Information ratios, followed by maximum draw-down, MaxDD, and maximum 1-month portfolio loss.

Table 1: *OOS* SDF Performance

4.2 *OOS* Performance

Table 1 presents *OOS* SDF portfolio performance statistics for all stocks, and separately for sub-samples after removing the bottom 10% or 20% of stocks by market cap using NYSE market capitalization break points (Hou et al. (2020)), as well as focusing explicitly on the cross-section of 3000 largest stocks which are components of Russell 3000. Note, we train the model on all available stocks, and than for each separate cross-section of stocks, we do not re-estimate the model, but re-scale the portfolio weights such that they add up to 1.

The table reports average monthly portfolio returns and standard deviations, FF6 Alpha and its t-statistics adjusted for autocorrelation and heteroscedasticity. The table also reports the annualized *OOS* Sharpe ratios and Information ratios. Sharpe ratio is defined as the *OOS* mean excess portfolio return divided by its standard deviation, and Information ratio is the intercept, Alpha, from regressing excess portfolio returns on FF6 factors, divided by the residual mean squared error of this regression. The other statistics report maximum one month loss, and maximum draw-down (MaxDD). MaxDD is simply the highest drop in monthly returns between two consecutive months.

The table reports two sets of results, for our base model, without asset attention, denoted *Base*, and our attention guided deep learning model, denoted *Attention*. All statistics are monthly except Sharpe and Information ratios which are annualized for the easier comparison with the rest of the literature.

Consider first Sharpe ratios. For the cross-section of *All Stocks*, *Attention* model has the highest Sharpe ratio of 2.834. It is substantially higher than the corresponding Sharpe

ratio of *Base* model, 2.537. The Information ratios are also the highest ones observed in the literature, 2.639 and 2.389 respectively. This clearly shows the advantage of asset attention and economic gains associated with it. Moreover, *Attention* models provides one of the highest, or probably the highest, Sharpe ratio reported in this literature (see Avramov et al. (2021) for the overview).

The portfolio returns and FF6 Alphas are very close in magnitudes, suggesting than FF6 risk factors do not span *Attention* determined SDF, i.e. very similar result to Kozak et al. (2020), Kelly et al. (2019), and Chen et al. (2020). The alphas are quite impressive as well, 88bps and 95bps per month for *Attention* and *Base* models respectively, and are highly statistically significant.

The other risk management statistics are also reassuring, as the maximum 1 month loss is 3% for both models, and the MaxDD is substantially smaller for *Attention* model, 3.3%, than *Base* model, 6%. Note that our *OOS* covers 2008-2009 financial crisis, and 2020 COVID-19 market turmoil wit substantially higher draw-down magnitudes on the market wide levels.

Avramov et al. (2021) warn that the high abnormal performance of ML models is driven by small or micro-cap stocks. The next four columns compare the performance of *Attention* vs. *Base* models after deleting stocks at the bottom 10% or 20% by NYSE market cap break-points. Consistent with Avramov et al. (2021) criticism, we also observe significant drops in Sharpe ratios to 1.255 or 1.006 respectively for *Attention* model, vs. 1.006 or 0.842 for *Base* model. However, while Sharpe ratios decrease, they still remain economically large for *Attention* model, especially considering Sharpe ratio of S&P500 of 0.56 for the similar time period. Importantly, *Attention* model continues dominate *Base* model by a large margin, highlighting the importance of attention guided deep learning architecture we introduce in this paper.

FF6 Alphas also decrease substantially but they are now more in line with abnormal performances typically reported in asset pricing literature. *Attention* model has the highest Alpha of 3.7% or 2.6% per year after removing bottom 10% or 20% of stocks by market cap respectively, and they are highly statistically significant.

The limits-to-arbitrage problem really appears for *Base* model using the largest stocks, i.e. after removing bottom 20% of small-cap stocks. The Alpha of 17 bps is only marginally significant, t=2.058. The maximum draw down, MaxDD, is 24.3% which is practically inadmissible from the real portfolio management perspective, especially for the portfolio constructed from the cross-section of largest stocks. It also has the smallest Sharpe ratio, 0.842, across all specifications in Table 1, which is not far from the Sharpe ratio of S&P500 for the same period, 0.56. Providing high risk, measured by MaxDD, marginal significance in

both economic and statistical terms of portfolio Alpha, and low Sharpe ratio, an investor is just better off passively holding the market index, rather than relying on *Base* model to construct the large-cap stocks' bottom-up portfolio.

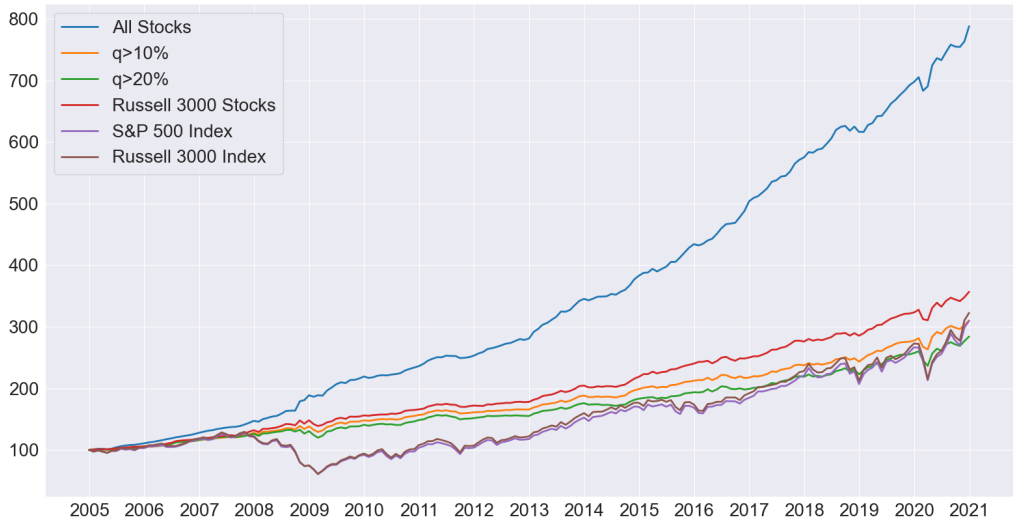
For the cross-section of stocks which are constituents of Russell 3000 index, *Attention* model dominates even by a bigger margin with Sharpe ratio of 1.49 vs Sharpe ratio of 1.196 of *Base* model. The Sharpe ratio of Russell 3000 is 0.55 for this sample period. FF6 alphas are even more impressive for this cross-section, averaging 5% across both models, and being highly statistically significant.

The results so far suggest that: (i) deep learning and multiple firms characteristics can improve substantially the empirical performance of SDF; (ii) unlike the previous literature, our SDF retains its significant economic gains after eliminating small and micro-cap stocks, except the *Base* model for the cross-section of the largest stocks, $q > 20\%$; (iii) *Attention* model substantially outperforms *Base* model.

Figure 2 depicts cumulative returns of various *Attention* model specifications (*All stocks*, or specifications after removing bottom 10% or bottom 20% by market cap, $q > 10\%$ and $q > 20\%$ respectively, or using only 3000 largest stocks of Russell 3000 index components, *Russell 3000 Stocks*) vis a vis S&P500 and Russell 3000 indexes. The figure presents the value of USD \$100 invested in the beginning of *OOS* period, 01-2005 and holding it till the end of 12-2020. The specification which uses *All stocks* exhibits the best performance where the original investment appreciates almost 8 times. The second best performance is observed for the largest 3000 stocks' specification, where the original investment appreciates almost 3 times, and outperforms the cumulative performances of S&P500 and Russell 3000 indexes.

$q > 10\%$ and $q > 20\%$ specifications, while outperforming S&P500 and Russell 3000 during major 2008-2009 financial crisis, and through out the first half of the sample, achieve similar to the market indexes performances towards the end of the sample. These portfolios however have substantially lower volatility compared to the market indexes. Moreover, these portfolios require trading in all stocks with long and short positions while the market indexes hold long positions only. In the next sub-section we consider the performances of portfolios invested only in the Long, top holdings, positions, which would be more direct comparison to the market indexes.

To complement results of Table 1, Figure 3 presents inter-temporal Sharpe ratios, estimated every two years for the *OOS* period, 01-2005 to 12-2020, for *Attention* model specifications similar to those in Figure 2. The specification with *All stocks* can achieve an annualized Sharpe ratio of almost 4.5 in the middle of the sample. The specification using *Russell 3000 Stocks* obtains Sharpe ratio of almost 3 during 2013-2014 period. The lowest Share ratios are



Notes: The figure reports cumulative *OOS* Returns of *Attention* model using *All stocks*, or specifications after removing bottom 10% or bottom 20% by market cap, $q > 10\%$ and $q > 20\%$ respectively, or using only 3000 largest stocks, Russell 3000 index components, *Russell 3000 Stocks*. The figure also plots cumulative returns of S&P500 and Russell 3000 indexes for the same *OOS* period, 01-2005 to 12-2020.

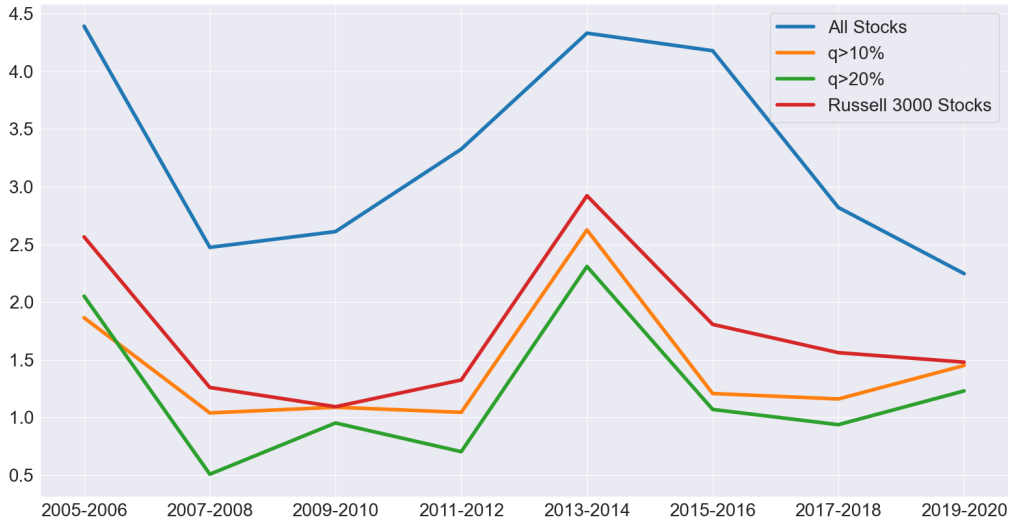
Figure 2: Cumulative *OOS* Returns of *Attention* model vs S&P500 and Russell 3000 indexes.

observed during 2008-2009 financial crisis, and during 2019-2020 Covid-19 market turmoil.

4.3 Performance of Individual Stocks in SDF Portfolio

The fundamental moment equation $\mathbb{E}_t\left[M_{t+1}R^e_{t+1,i}\right] = 0$ implies the following factor representation $R^e_{t+1,i} = \beta^{\text{SDF}}_{t,i}F_{t+1} + \epsilon_{t+1,i}$. Although we do not explicitly model SDF beta, but rather estimate each stocks' SDF weights, the single factor representation implies risk-reward trade-off for individual stocks. That is stocks which contribute the most to SDF, i.e. with the higher weights, are also expected to have higher exposure to SDF, i.e. higher factor loadings. Higher SDF loadings should be associated with higher expected returns.

In this section we examine the performance of individual stocks conditional on their SDF weights. Every month we sort stocks into decile portfolios based on their SDF weights. That is we use stocks' SDF weights to place stocks into high vs low weights' portfolios, as the high portfolio, with higher beta stocks is expected to outperform the low beta portfolio. We then estimate these decile portfolios next month value-weighted returns, using each stock's



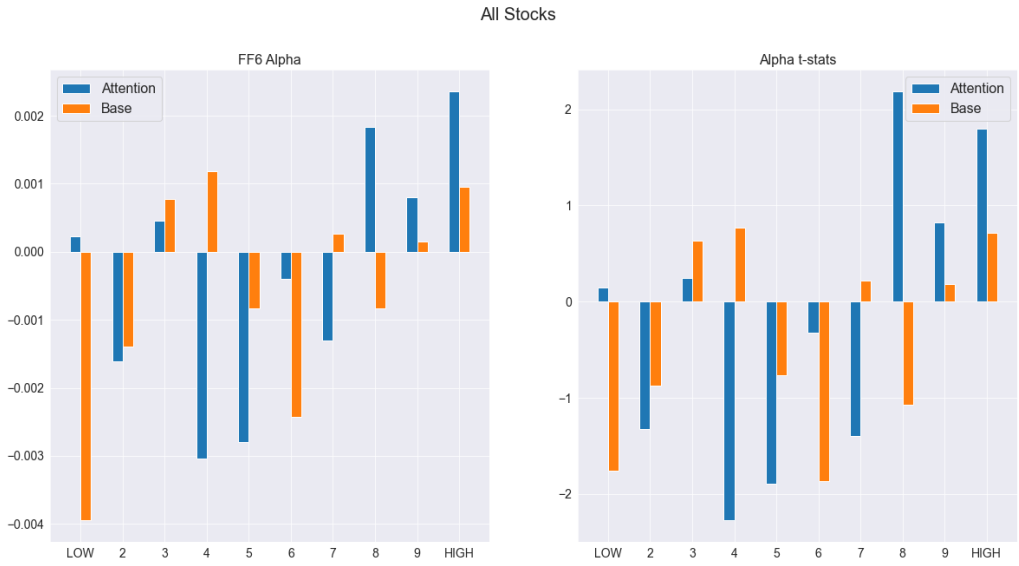
Notes: The figure reports bi-annual *OOS* Sharpe ratio of *Attention* model using *All stocks*, or specifications after removing bottom 10% or bottom 20% by market cap, $q > 10\%$ and $q > 20\%$ respectively, or using only 3000 largest stocks which are components of Russell 3000 index, *Russell 3000 Stocks*.

Figure 3: Bi-annual *OOS* Sharpe ratio, *Attention* model

market cap as the weight. We also estimate portfolio risk-adjusted FF6 alphas, and portfolio level illiquidity. We measure illiquidity with relative dollar volume-weighted effective bid-ask spreads computed from intra-day TAQ data. We fist compute effective spreads for each stock daily, and then average these estimates for a month. The portfolio level illiquidity is value-weighted effective spreads of stocks in the portfolio.

Figure 4 reports FF6 risk adjusted alphas and corresponding t-statistics adjusted for autocorrelation and hederoscedasticity for portfolios formed on the weights assigned by *Attention* vs *Base* models. While we do not expect to observe the monotonicity in portfolio alphas from Low to High deciles, this plot allows to see which cross-section drives profitability on a portfolio level. There is a significant contrast between the two models. While *Attention* model has both significant positive and negative decile portfolios' Alphas, *Base* model has only significantly negative decile portfolios' Alphas which points to limits of arbitrage, i.e. short-selling restrictions (Avramov et al. (2021)), for this specification.

Figure 5 provide similar graphs for the cross-section of 3000 largest stocks which are components of Russell 3000 index, *Russell 3000 Stocks*. Here, the results are similar, as the *Base* model has only one significant negative Alpha for Low decile portfolio, while *Attention*



Notes: The figure reports FF6 risk adjusted *Alphas* and their corresponding t-statistics for *Attention* and *Base* models using *All stocks*.

Figure 4: *Alpha* and t-stats of decile portfolios, *Attention* vs *Base* model, all stocks.

has both positive and negative significant alphas across decile portfolios. Similar results are observed for the cross-section where the bottom 20% of stocks by market cap are removed. Only for $q > 10\%$ cross-section, *Base* model has a positive and significant alpha decile portfolio.

Table 2 below presents all numerical results in details. First, consider Panel A, *Attention* model. As expected, the portfolios excess returns are increasing from Low to High, from 69 bps to 93 bps per month. After FF6 factors risk adjustment, portfolio alphas are too increasing from Low to High, with only High portfolio alpha being marginally significant, 24 bps per month (t=1.80). The High-minus-Low portfolio statistics are reported for pure expositional purposes as they do not incorporate SDF investable weights but rather use stocks' market cap as the weights. These statistics only demonstrate the relative performance of extreme, Long vs Short SDF positions. For *Attention* model these differences are not statistically significant using *All Stocks*.

For *Base* model these differences are marginally significant, with High portfolio outperforming Low portfolio. A closer look at the portfolios alphas suggests that this outperfomance is mostly driven by the short position, with the alpha of -39 bps per month (t=-1.76), and raw excess return of 20 bps per month. A further closer look at the trading costs for the

Notes: The figure reports FF6 risk adjusted *Alphas* and their corresponding t-statistics for *Attention* and *Base* using only 3000 largest stocks which are components of Russell 3000 index.

Figure 5: *Alpha* and t-stats of decile portfolios, *Attention* vs *Base* model, *Russell 3000 Stocks*.

lowest decile portfolio shows that both the excess returns and the alpha are dominated by the average effective spreads, 44 bps. Therefore, ignoring shorting costs themselves, the potential profits from these short positions are completely dominated by the trading costs of executing them. This is essentially the criticism of Avramov et al. (2021) to all existing ML approaches in asset pricing. We observe it for our *Base* model as well. *Attention* model however shows substantially smaller portfolio trading costs compared to the raw excess returns or alphas.

Next, consider Panel B, which excludes the bottom 10% of smallest stocks based on NYSE market capitalization breakpoints every formation month. The results for *Attention* model begin to look better, while they look worse for *Base* model. The performance of High-minus-Low portfolio Alpha becomes marginally significant for *Attention* model, 31 bps per month (t=1.85), with the most of gains being driven by the long position, 10th decile, alpha of 23 bps per month (t=2.32), while the alpha of short position is practically zero. Moreover, the average trading costs of the long position are relatively low, 14 bps, compared to the portfolio alpha or excess raw return of 96 bps per month.

For *Base* model, the results for High-minus-Low strategy might look impressive, with the alpha of 51 bps per month (t=2.11). Interestingly, here again, the whole performance is driven by the short position, with alpha of -35 bps (t=-1.74), while the alpha of the long

position is statistically insignificant. For the short position, 1st decile, the average trading costs, 28 bps, are very close to the raw portfolio excess return, 31 bps, suggesting the limits to arbitrage (Avramov et al. (2021)).

The results for *Base* model look better and similar to those of *Attention* model in Panel C which excludes the bottom 20% of smallest stocks. They however reverse dramatically in Panel D, where we only analyse the largest 3000 stocks which are constituents of Russell 3000 index. Here *Base* delivers outstanding High-minus-Low portfolio alpha of 56 bps (t=2.40). This alpha is almost entirely driven by the short position, -41 bps (t=-2.21). The excess raw return of this portfolio, 26 bps, is lower than its average trading costs, 28 bps.

The results are completely different for *Attention* model. The High-minus-Low portfolio alpha is 38 bps (t=1.96), and the most of gains are coming from the long position with the alpha of 26 bps (t=2.49). The raw excess returns is 1.01% which far exceeds the average trading costs of 14 bps.

Each panel also reports the average market cap of each portfolio. While both *Attention* and *Base* models take long positions in the large cap stocks, *Attention* model consistently uses larger stocks on the short positions compared to *Base* models.

Overall, we observe two sets of results. First, our *Base* model exhibits inconsistent performance across different cross-sections of stocks, and its gains are driven largely by short positions with extremely high trading costs compared to the mean portfolio returns. This confirms the results of Avramov et al. (2021) that ML methods tend to identify difficult to arbitrage trading strategies.

Second, our *Attention* model's performance overturns the above criticism and clearly demonstrates that carefully constructed ML networks driven by economic assumptions demonstrate exemplary performance across various robustness verification tests.

Table 2: Performance of Stocks' Portfolios Sorted on their SDF weights

Notes: The table reports the performance of value-weighted portfolios in month $t+1$ sorted on individual stocks' SDF weights in month t, for *OOS*, 01-2005 to 12-2020. Two sets of results refer to attention guided deep learning model, denoted *Attention*, and the base model, without asset attention, denoted *Base*. The statistics are value-weighted monthly portfolio returns, portfolio alphas, Alpha, after FF6 factors risk adjustment, and their robust for autocorrelation and heteroscedasticity t-statistics. Eff. Spread is the value-weighted relative effective bid-ask spread estimated across all stocks in a portfolio. Panel A presents the results for *All stocks*, Panels B and C for the sub-samples of stocks after removing bottom 10% and 20% of stocks by market capitalization based on NYSE size break points respectively, and Panel D for the cross-section of 3000 stocks, Russell 3000 constituents.

Panel A. All Stocks											
	Attention										
	Low	2	3	4	5	6	7	8	9	High	H-L
Return	0.0069	0.0064	0.0084	0.0049	0.0059	0.0079	0.0076	0.0097	0.0094	0.0093	0.0024
Return t-stat	1.6303	1.4021	1.9920	1.0781	1.3983	1.9651	2.2009	3.1258	3.0098	3.3151	0.8278
Alpha	0.0002	-0.0016	0.0005	-0.0030	-0.0028	-0.0004	-0.0013	0.0018	0.0008	0.0024	0.0021
Alpha t-stat	0.1437	-1.3209	0.2494	-2.2733	-1.8885	-0.3182	-1.3942	2.1855	0.8223	1.8001	1.0516

Table 2 continued from previous page

Avg Size ($mln)	2,244.49	3,148.35	3,430.40	3,819.16	4,384.29	5,388.05	6,861.29	8,732.45	9,644.39	5,331.77	
Eff. Spread	0.0023	0.0016	0.0015	0.0015	0.0021	0.0012	0.0012	0.0011	0.0016	0.0032	

Base

	Low	2	3	4	5	6	7	8	9	High	H-L
Return	0.0020	0.0060	0.0078	0.0106	0.0075	0.0066	0.0084	0.0077	0.0093	0.0082	0.0063
Return t-stat	0.4209	1.2822	1.7549	2.3036	1.8174	1.5181	2.1131	2.3978	2.9117	2.9097	1.8084
Alpha	-0.0039	-0.0014	0.0008	0.0012	-0.0008	-0.0024	0.0003	-0.0008	0.0002	0.0010	0.0049
Alpha t-stat	-1.7580	-0.8761	0.6350	0.7674	-0.7599	-1.8619	0.2228	-1.0675	0.1828	0.7173	1.6850
Avg Size ($mln)	1,156.20	2,054.62	2,535.11	3,027.61	3,722.53	4,898.52	6,624.22	9,373.95	12,205.22	7,388.22	
Eff. Spread	0.0044	0.0027	0.0023	0.0016	0.0014	0.0013	0.0013	0.0011	0.0012	0.0019	

Panel B. q>10%

Attention

	Low	2	3	4	5	6	7	8	9	High	H-L
Return	0.0058	0.0081	0.0061	0.0053	0.0069	0.0078	0.0085	0.0089	0.0090	0.0096	0.0038
Return t-stat	1.3629	1.8618	1.3440	1.2046	1.6451	2.0052	2.4715	2.8381	2.8237	3.4522	1.4865
Alpha	-0.0008	0.0003	-0.0027	-0.0025	-0.0019	-0.0007	0.0001	0.0010	0.0004	0.0023	0.0031
Alpha t-stat	-0.6200	0.1645	-2.0417	-2.1396	-1.4081	-0.5781	0.1604	1.0594	0.3898	2.3152	1.8528
Avg Size ($mln)	4,557.27	5,272.07	5,435.67	5,872.06	6,495.55	7,561.96	9,445.85	11,338.26	14,183.75	14,289.04	
Eff. Spread	0.0016	0.0014	0.0013	0.0014	0.0019	0.0012	0.0011	0.0011	0.0011	0.0014	

Base

	Low	2	3	4	5	6	7	8	9	High	H-L
Return	0.0031	0.0066	0.0101	0.0079	0.0077	0.0070	0.0077	0.0081	0.0089	0.0092	0.0061
Return t-stat	0.6679	1.5241	2.2779	1.7169	1.7525	1.7899	2.1677	2.5406	2.7606	3.4168	1.9151
Alpha	-0.0035	-0.0008	0.0021	-0.0007	-0.0013	-0.0010	-0.0008	-0.0001	0.0000	0.0017	0.0051
Alpha t-stat	-1.7441	-0.6345	1.4460	-0.4660	-1.0746	-0.8084	-0.8463	-0.1074	-0.0510	1.5881	2.1184
Avg Size ($mln)	2,749.60	3,746.66	4,231.32	5,005.93	5,827.22	7,125.34	9,044.02	12,193.82	16,521.67	18,008.33	
Eff. Spread	0.0028	0.0019	0.0020	0.0013	0.0013	0.0013	0.0011	0.0010	0.0012	0.0010	

Panel C. q>20%

Attention

	Low	2	3	4	5	6	7	8	9	High	H-L
Return	0.0062	0.0079	0.0067	0.0044	0.0076	0.0079	0.0084	0.0089	0.0089	0.0098	0.0036
Return t-stat	1.4911	1.7869	1.4648	1.0040	1.8773	2.0838	2.6395	2.7824	2.8157	3.5496	1.3995
Alpha	-0.0003	-0.0007	-0.0018	-0.0036	-0.0009	-0.0010	0.0006	0.0008	0.0003	0.0025	0.0028
Alpha t-stat	-0.2636	-0.3741	-1.3160	-2.5438	-0.6941	-0.8818	0.7409	0.8725	0.2845	2.5177	1.6652
Avg Size ($mln)	6,334.18	6,965.16	6,966.71	7,487.46	8,350.49	9,596.02	11,550.64	13,712.75	16,749.51	17,376.31	
Eff. Spread	0.0015	0.0013	0.0012	0.0018	0.0014	0.0011	0.0010	0.0010	0.0014	0.0012	

Base

	Low	2	3	4	5	6	7	8	9	High	H-L
Return	0.0051	0.0072	0.0092	0.0066	0.0082	0.0074	0.0067	0.0090	0.0087	0.0097	0.0046
Return t-stat	1.0516	1.8072	1.9207	1.4650	1.9468	2.0059	1.9227	2.7920	2.6842	3.6806	1.3687
Alpha	-0.0016	-0.0002	0.0004	-0.0019	-0.0002	-0.0006	-0.0016	0.0006	-0.0002	0.0023	0.0039
Alpha t-stat	-0.8204	-0.1458	0.2871	-1.5984	-0.1680	-0.5148	-1.6559	0.7599	-0.2868	2.0305	1.6140
Avg Size ($mln)	4,220.60	5,310.87	5,820.29	6,775.24	7,558.80	9,246.53	11,268.11	14,569.20	19,088.96	21,235.74	
Eff. Spread	0.0023	0.0021	0.0013	0.0012	0.0011	0.0012	0.0011	0.0012	0.0009	0.0010	

Panel D. Russell 3000 Stocks

Attention

	Low	2	3	4	5	6	7	8	9	High	H-L
Return	0.0055	0.0088	0.0062	0.0049	0.0061	0.0080	0.0081	0.0092	0.0087	0.0101	0.0046
Return t-stat	1.2267	2.0368	1.3497	1.0884	1.4590	2.0408	2.3368	2.9108	2.6966	3.5929	1.5760
Alpha	-0.0011	0.0007	-0.0023	-0.0032	-0.0025	-0.0004	-0.0006	0.0014	0.0000	0.0026	0.0038
Alpha t-stat	-0.7683	0.4119	-1.9236	-2.2248	-1.9623	-0.3239	-0.6042	1.4333	-0.0454	2.4926	1.9575
Avg Size ($mln)	3,635.25	4,450.01	4,630.47	4,982.73	5,604.76	6,609.07	8,417.97	10,207.08	13,039.84	11,847.36	
Eff. Spread	0.0017	0.0015	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0010	0.0014	

Base

	Low	2	3	4	5	6	7	8	9	High	H-L
Return	0.0026	0.0066	0.0102	0.0086	0.0073	0.0066	0.0082	0.0074	0.0087	0.0092	0.0066

Table 2 continued from previous page

Return t-stat	0.5590	1.5204	2.2198	1.8355	1.6655	1.6610	2.2196	2.3000	2.7053	3.4115	2.0425
Alpha	-0.0041	-0.0006	0.0017	0.0001	-0.0014	-0.0016	0.0000	-0.0011	-0.0002	0.0015	0.0056
Alpha t-stat	-2.2068	-0.4871	1.1335	0.0455	-1.0787	-1.4249	0.0119	-1.5155	-0.2208	1.4092	2.4033
Avg Size ($mln)	2,024.99	3,092.91	3,475.42	4,225.46	4,905.21	6,115.03	7,983.04	11,089.78	15,357.77	15,156.44	
Eff. Spread	0.0028	0.0018	0.0016	0.0014	0.0013	0.0012	0.0011	0.0010	0.0010	0.0011	

To shed further light on extreme portfolio performances identified by *Attention* model, Figure 6 presents cumulative returns of High vs Low portfolios of Table 2 for the whole *OOS* period, 01-2005 to 12-2020 using the two most liquid cross-section of stocks' specifications: $q > 20\%$, removing the bottom 20% smallest stocks by NYSE market cap breakpoints, and *Russell 3000 Stocks*, the largest 3000 stocks which are components of Russell 3000 index. For expositional purposes, we entertain the same investment strategy of investing USD$100 in the beginning of 01-2005 and holding it through the end of 12-2020. As expected, regardless of the cross-section of stocks, High portfolio significantly outperforms Low portfolio, and the original investment appreciates almost 6 times. High portfolio consists of stocks with the highest long positions in our SDF, while Low portfolio - with the smallest, short positions. An important insight is that the directions, long vs short, are correct, as the short position significantly underperforms compared to the long position. Finally, unlike in Figure 2, the top Long only holdings of $q > 20\%$ portfolio substantially outperform all passive market indexes.

As the last robustness check, we examine the portfolio, SDF, weights distribution for both models reported in Table 3 below. As before, we report two sets of results, for *Attention* and *Base* models, and for the same cross-section of stocks as in Table 2, all stocks, Panel A, removing the bottom 10% and 20% by market capitalization stocks, Panels B and C respectively, and cross-section of Russell 3000 components, Panel D. The portfolio weights are reported in percentages.

The first observation is that the individual stock weights are very small, and the means are similar across the two models for all stocks, Panel A, 0.0093% and 0.0088% for *Attention* and *Base* respectively. The extreme, min (max), positions are very similar -0.12% (0.26%) and -0.16% (0.22%) for *Attention* and *Base* respectively. *Base* model however is more concentrated with median weights being much higher, and this difference only increases with the size of the cross-section. For example the median weight for the cross-section of Russell 3000 stocks, Panel D, for *Base*, 0.0128%, almost twice exceeds that of *Attention*, 0.0056%, i.e., *Attention* model provides more diversification to investors.

Considering all the results and discussions above, and the limitations of *Base* model to avoid limits-to-arbitrage criticism, in the rest of our analysis we focus on *Attention* model.

	Mean	Std.Dev.	Min	5%	10%	25%	Median	75%	90%	95%	Max
Panel A. All Stocks											
Attention	0.0093	0.0418	-0.1230	-0.0434	-0.0304	-0.0121	0.0035	0.0227	0.0525	0.0842	0.2648
Base	0.0088	0.0395	-0.1625	-0.0492	-0.0320	-0.0101	0.0067	0.0247	0.0492	0.0748	0.2238
Panel B. q>10%											
	Mean	Std.Dev.	Min	5%	10%	25%	Median	75%	90%	95%	Max
Attention	0.0109	0.0612	-0.2263	-0.0817	-0.0573	-0.0227	0.0077	0.0420	0.0811	0.1101	0.3889
Base	0.0161	0.0571	-0.2627	-0.0796	-0.0511	-0.0138	0.0172	0.0491	0.0809	0.1034	0.3003
Panel C. q>20%											
	Mean	Std.Dev.	Min	5%	10%	25%	Median	75%	90%	95%	Max
Attention	0.0163	0.0756	-0.2758	-0.0994	-0.0694	-0.0262	0.0129	0.0567	0.1045	0.1389	0.4660
Base	0.0263	0.0685	-0.3030	-0.0885	-0.0548	-0.0109	0.0278	0.0673	0.1054	0.1313	0.3503
Panel D. Russell 3000 Stocks											
	Mean	Std.Dev.	Min	5%	10%	25%	Median	75%	90%	95%	Max
Attention	0.0090	0.0549	-0.1967	-0.0726	-0.0511	-0.0207	0.0056	0.0355	0.0711	0.0987	0.3366
Base	0.0124	0.0520	-0.2333	-0.0731	-0.0476	-0.0142	0.0128	0.0409	0.0703	0.0921	0.2689

Notes: The table reports SDF weights, in percentages, for two model specifications: attention guided deep learning model, denoted *Attention*, and the base model, without asset attention, denoted *Base*. The reported statistics are for *OOS*, 01-2005 to 12-2020.

Table 3: SDF portfolio weights distribution

4.4 Trading Costs and Turnover

As noted by Avramov et al. (2021), most of SDF factors constructed in the literature using ML comes with extreme turnovers, which under reasonable trading costs makes the candidate SDF portfolios practically not implementable.

We estimate turnover on the SDF portfolio level using the methodology outlined in Chen et al. (2020):

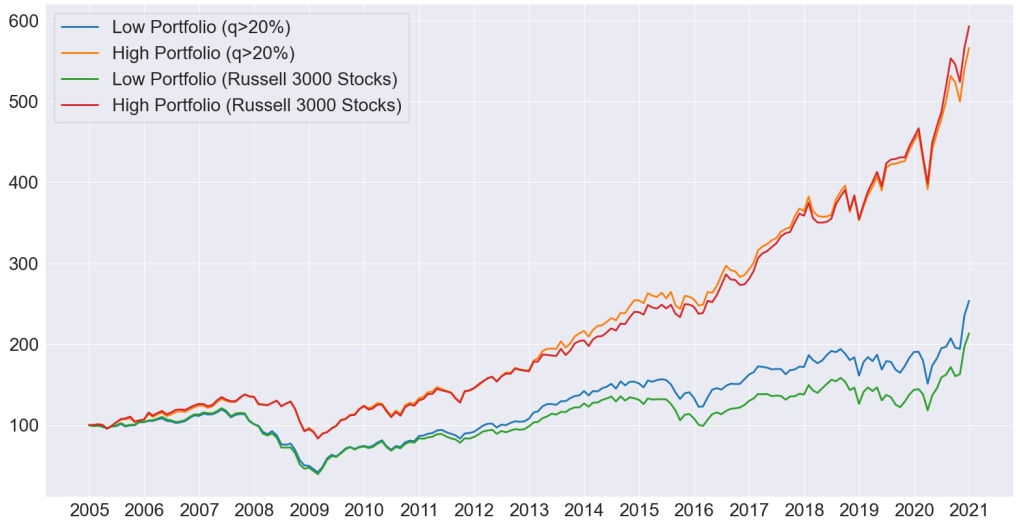
$$\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{i}\left|\left(1+R_{P,t+1}\right)w_{i,t+1}-\left(1+R_{i,t+1}\right)w_{i,t}\right|\right)$$

where $w_{i,t}$ is the weight of a stock i in the portfolio at time t, and $R_{P,t+1}=\sum_i R_{i,t+1}w_{i,t}$ is the corresponding portfolio return. Long and short positions are calculated separately, and the portfolio weights are normalized to $\|w_t\|_1=1$. To avoid double-counting, we divide the turnover by 2.[9]

Table 4 presents the results for all stocks, the sub-sample after removing the bottom 10% or 20% by market cap, and separately for Russell 3000 stocks.

The first row of the table report average monthly portfolio turnover, *Turnover All*, across

[9]Following Koijen et al. (2018) and Freyberger et al. (2020), we also recompute turnover as Turnover $_t=\frac{1}{2}\sum_i\left|w_{t-1}^i\left(1+r_t^i\right)-w_t^i\right|$, and obtain similar results



Notes: The figure present cumulative value-weighted returns of High vs Low portfolio deciles, where portfolios are formed based on individual stocks' *Attention* model SDF weights. *Attention* model is specified for the most liquid stocks: $q > 20\%$, removing the bottom 20% smallest stocks by NYSE market cap breakpoints, and *Russell 3000 Stocks*, the largest 3000 stocks which are components of Russell 3000 index. The details of portfolio constructions and the average summary statistics are presented and described in Table 2, Panels C and D respectively.

Figure 6: Cumulative, *OOS* Value-Weighted Returns of High vs Low portfolio deciles, where portfolios are formed based on individual stocks' *Attention* model SDF weights.

	All Stocks	**q>10%**	**q>20%**	**Russell 3000 Stocks**
Turnover All	0.348	0.360	0.352	0.360
Turnover Long	0.227	0.221	0.220	0.223
Turnover Short	0.121	0.139	0.132	0.137
Eff. Spreads All	0.0115	0.0033	0.0025	0.0043
Eff. Spread Long	0.0111	0.0032	0.0024	0.0040
Eff. Spread Short	0.0120	0.0035	0.0026	0.0046
Avg Size Long ($mln)	7,070.83	10,981.45	13,132.47	9,756.96
Avg Size Short ($mln)	3,345.91	5,472.13	7,103.02	4,633.49

Notes: The table presents summary statistics of trading costs and turnover for *OOS*, 01-2005 to 12-2020, SDF portfolio constructed with *Attention* model for different cross-sections of stocks. The main statistics are the average portfolio turnover and volume weighted relative effective bid-ask spreads obtained from high frequency trade and quote, TAQ, data.

Table 4: Trading Costs Analysis for *OOS* SDF Performance

long and short positions. Regardless the cross-section of stocks, the turnover is low, and ranges between 35% to 36%. Turnover of the *Long* position is slightly higher, 22%, then turnover of the *Short* position, ranging between 12% to 14%. These are the lowest turnovers observed for the tradable SDF (see Avramov et al. (2021) for review).

We next examine directly the average trading costs of stocks included into the portfolio. We measure trading costs by the average dollar volume weighted relative effective spreads obtained from high frequency intra-day trading data, TAQ, available on WRDS. More specifically, for each stock in a portfolio we first compute the average dollar volume weighted relative effective spreads per day, and than average it across all days in the month for *OOS* period 01-2005 to 12-2020. To estimate the average Effective Spread on a portfolio level, we use the individual stock's portfolio weights in the estimated SDF.

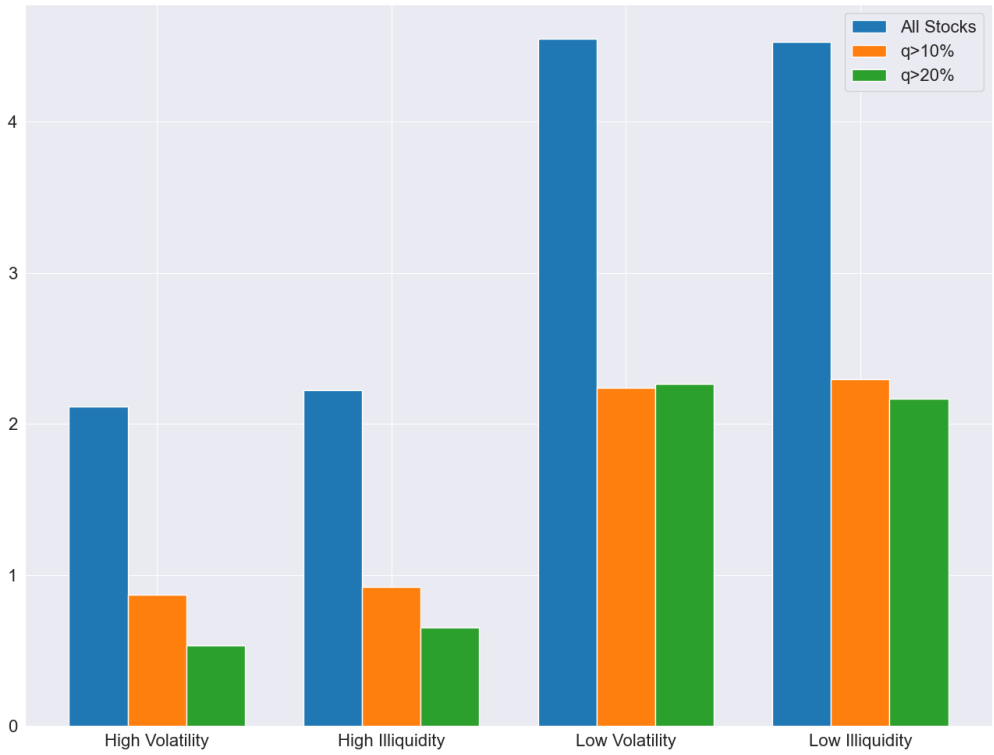
The first column reports average effective spreads for all stocks in the portfolio. Similar to conclusions of Avramov et al. (2021) we also find that on average the stocks are quite illiquid as the average portfolio Effective Spread is 1.55%. The stocks associated with the *Long* position on average are more illiquid, Eff. Spread of 1.01%, than the stocks in the *Short* position with the average spread of 54 bps. The average size in the *Long* position is $ 7bln, and $ 3.3 bln in the *Short* position.

The results in the next three columns for $q > 10\%$, $q > 20\%$ and Russell 3000 index constituents look quite different. Here, the average effective spreads on a portfolio level are 39 bps, 28 bps, and 53 bps respectively which are associated with quite liquid and easy to trade stocks. The spreads of the *Long* position on average are higher, 24 bps, 18 bps, and 32 bps respectively, than the spreads of the *Short* position, 15 bps, 11 bps, and 21 bps respectively. Both positions however use the most liquid stocks with quite large market caps reported in the last two rows.

Therefore, our tradable SDF specification is robust even to the most strict examination of trading costs.

4.5 Market Regimes

An important part of a portfolio performance analysis is the sensitivity to different market regimes. As Avramov et al. (2021) fairly note, the abnormal performance of most ML models tend to be generated during market conditions characterized by high volatility and illiquidity, i.e. the low price efficiency regimes. These regimes are also characterized by higher trading costs and higher arbitrage risks which prevents the most efficient asset allocation on the portfolio level. Moreover, our approach to SDF construction is intended to identify the



Notes: The figure present annualized *OOS* Sharpe Ratios by the market regimes conditioned on High vs Low Volatility or Illiquidity for all stocks, as well as for more liquid stocks:$q > 10\%$ and $q > 20\%$, after removing the bottom 10% or 20% smallest stocks by NYSE market cap breakpoints.

Figure 7: *OOS* Sharpe Ratios by High vs Low Volatility/Illiquidity Market Regimes .

systematic risks, the risks which are easier identifiable in the high price efficiency regimes. Therefore, our hypothesis here is that the SDF performance should be better during low market volatility and illiquidity regimes, i.e. the market conditions where risk premiums have positive realizations.

For *OOS*, 01-2005 to 12-2020, we identify Low (High) volatility or illiquidity regimes when the median value of VIX or Illiquidity, measured by bottom-up aggregated average Effective bid-ask spread, computed on a stock level earlier, is higher (lower) than its sample median value. Therefore, while the SDF performance is out-of-sample, the regime identification here is ex-post, similar to Avramov et al. (2021).

Figure 7 presents bar charts comparing annualized *OOS* Sharpe Ratios of High vs Low

Volatility and Illiquidity regimes for all stocks, as well as for more liquid stocks: $q > 10\%$ and $q > 20\%$, after removing the bottom 10% or 20% smallest stocks by NYSE market cap breakpoints. Regardless the cross-section, the Sharpe ratios during low volatility or illiquidity regimes are always almost twice higher. These are the market states with the highest price efficiency and when the arbitrageurs are the least financially constrained. Therefore the opportunities identified by *Attention* model are practically exploitable, and SDF is realistic from the investment viewpoint.

Table 5 presents the detailed numerical results for the high volatility/illiquidity regimes in Panel A, and for the low volatility/illiquidity regimes in Panel B. The structure of the table is similar to Table 1 above, except we drop the cross-section of Russell 3000, as the results for the sub-sample of $q > 20\%$ are more conservative.

The biggest difference from other existing ML models in the literature is that the best performance of SDF portfolio is observed during Low, Panel B, volatility and illiquidity regimes. Here, in Low volatility regimes, an annualized Sharpe ratio is 4.55 for *All Stocks*, or 2.24 and 2.26 for $q > 10\%$ and $q > 20\%$ respectively. The corresponding Sharpe ratios are almost half of the magnitudes in High volatility regimes, 2.117, 0.868 and 0.532 respectively.

Besides Sharpe ratios, portfolio FF6 risk adjusted alphas are notable as well. First, for the high volatility regime, Panel A, for the most liquid cross-section, $q > 20\%$, the monthly alpha is 32 bps (t=2.97), which is similar to the raw excess return of 31 bps. It implies that FF6 factors do not span SDF during these times. Similar conclusions about FF6 factor efficiency for a representative SDF are also discussed by Kozak et al. (2020), Kelly et al. (2019), Chen et al. (2020).

During low volatility regime, FF6 alphas are substantially smaller than raw returns but yet economically meaningful. For example, for the most liquid cross-section, $q > 20\%$, the monthly alpha is 18 bps (t=2.35), which is almost three times smaller than the mean portfolio return of 60 bps. Here, FF6 factors play bigger role in spanning SDF, yet they are not enough to fully represent it. The results are qualitatively similar for High vs Low Illiquidity regimes.

Similar to Table 4, Table 6 presents trading costs analysis of portfolios under different regimes of Table 5. Unlike Table 4, here we cannot estimate turnovers due to the sample split between High and Low regimes. The important message of Table 6 is that regardless of High or Low Volatility/Illiquidity regimes, SDF portfolio continues using stocks with very similar illiquidity, i.e. trading costs, and market capitalizations.

Overall, we conclude that SDF performance is not driven by the hard-to-arbitrage market states, and the best performance is obtained during the regimes of higher price efficiency, low

Panel A. High Volatility/Illiquidity Regime						
	All Stocks	**q>10%**	**q>20%**	**All Stocks**	**q>10%**	**q>20%**
	High Volatility			**High Illiquidity**		
Return	0.0092	0.0044	0.0031	0.0098	0.0047	0.0039
Std.Dev.	0.0150	0.0174	0.0201	0.0153	0.0178	0.0205
Alpha	0.0082	0.0043	0.0032	0.0083	0.0037	0.0028
Alpha t-stat	6.318	3.948	2.974	5.584	2.717	2.025
Sharpe	2.117	0.868	0.532	2.224	0.921	0.654
InfRatio	1.934	1.003	0.748	1.942	0.854	0.637
Max 1M Loss	-0.032	-0.051	-0.058	-0.032	-0.051	-0.058
Panel B. Low Volatility/Illiquidity Regime						
	All Stocks	**q>10%**	**q>20%**	**All Stocks**	**q>10%**	**q>20%**
	Low Volatility			**Low Illiquidity**		
Return	0.0106	0.0056	0.0060	0.0099	0.0052	0.0052
Std.Dev.	0.0080	0.0087	0.0092	0.0076	0.0079	0.0084
Alpha	0.0070	0.0023	0.0018	0.0075	0.0022	0.0017
Alpha t-stat	7.825	2.873	2.346	11.108	2.803	2.033
Sharpe	4.550	2.239	2.264	4.531	2.295	2.165
InfRatio	3.594	1.113	0.843	3.928	1.182	0.907
Max 1M Loss	-0.011	-0.017	-0.019	-0.011	-0.016	-0.017

Notes: The table presents summary statistics for *OOS*, 01-2005 to 12-2020, *Attention* model SDF performance for the cross-sections of *All Stocks*, and after removing the bottom 10% or 20% of smallest stocks based on NYSE market capitalization's breakpoints. High (Low) volatility or illiquidity regime is determined when the value of VIX or Illiquidity, respectively, is above (below) its median sample value. The statistics are average monthly portfolio returns and standard deviations, monthly portfolio alpha, Alpha, after FF6 factors risk adjustment, and its robust for autocorrelation and heteroscedasticity t-statistics, annualized Sharpe and Information ratios, followed by maximum draw-down, MaxDD, and maximum 1-month portfolio loss.

Table 5: SDF Performance in High vs Low Market Volatility/Illiquidity regimes

volatility and illiquidity regimes.

4.6 Which Covariates Get More Attention?

Which firm characteristics, on average, contribute the most to SDF identification? We use the *saliency map* technique from Simonyan et al. (2013) that attributes feature importance for a single prediction by taking the absolute value of the partial derivative of the output with respect to the input features. The absolute value of the gradient points to input features that can be perturbed the least in order for the target output to change the most: the higher

Panel A. High Volatility/Illiquidity Regime						
	All Stocks	**q>10%**	**q>20%**	**All Stocks**	**q>10%**	**q>20%**
	High Volatility			**High Illiquidity**		
Eff. Spreads All	0.0130	0.0036	0.0025	0.0128	0.0034	0.0024
Eff. Spread Long	0.0118	0.0033	0.0024	0.0116	0.0031	0.0022
Eff. Spread Short	0.0146	0.0040	0.0028	0.0143	0.0038	0.0027
Avg Size Long ($mln)	6,639.26	10,083.87	12,082.02	7,377.65	11,039.78	13,134.43
Avg Size Short ($mln)	2,992.19	5,052.31	6,606.73	3,077.36	5,088.76	6,652.82
Panel B. Low Volatility/Illiquidity Regime						
	All Stocks	**q>10%**	**q>20%**	**All Stocks**	**q>10%**	**q>20%**
	Low Volatility			**Low Illiquidity**		
Eff. Spreads All	0.0099	0.0030	0.0025	0.0102	0.0032	0.0026
Eff. Spread Long	0.0103	0.0030	0.0025	0.0105	0.0032	0.0026
Eff. Spread Short	0.0093	0.0030	0.0024	0.0096	0.0032	0.0026
Avg Size Long ($mln)	7,502.40	11,879.03	14,182.93	6,764.01	10,923.12	13,130.52
Avg Size Short ($mln)	3,699.64	5,891.95	7,599.31	3,614.46	5,855.49	7,553.22

Notes: The table reports average trading costs of SDF portfolio for *OOS*, 01-2005 to 12-2020, *Attention* model for the cross-sections of *All Stocks*, and after removing the bottom 10% or 20% of smallest stocks based on NYSE market capitalization's breakpoints. High (Low) volatility or illiquidity regime is determined when the value of VIX or Illiquidity, respectively, is above (below) its median sample value. *Eff. Spreads* are relative dollar-volume weighted effective bid-ask spreads estimated using intraday trades and quotes data, TAQ. They are first computed daily and then aggregated monthly on the stock level. Then, using stock's portfolio weights, these estimates are computed on the portfolio level.

Table 6: Trading Costs Analysis of SDF Portfolio in Various Market Regimes.

it is for a variable, the more it plays a role in the forecast. By producing the saliency map for all assets over the whole training sample, we can rank what features are the most important based on the overall importance-score $S(\mathbf{x})$ computed for each feature.

More specifically, given the i-th $\mathbf{x}_{t,i} \in \mathbb{R}^D$ observed at time t, the sensitivity (the importance) $S(\mathbf{x}_{i,t}) \in \mathbb{R}^D$ of the D's features can be estimated with eq.13.

$$S(\mathbf{x}_{t,i}) = \left| \frac{\partial f_\theta(\mathbf{x}_{t,i})}{\partial \mathbf{x}_{t,i}} \right| \qquad (13) \qquad S(\mathbf{x}) = \texttt{pool}\left(\sum_{t=1}^{T} \sum_{i=1}^{N_T} S(\mathbf{x}_{t,i}) \right) \qquad (14)$$

where t is an *OOS* month, T is the number of *OOS* months, 192, and N_T is the number of stocks available at time t.

The overall ranking can be computed from eq. 14 where `pool` is a pooling function. We consider the median to pool all importance-scores assigned to the feature across all training samples.

We compute the saliency map for all predictions made by *Attention* model, over the total

OOS evaluation periods of 192 months, aggregate the result using the median and present the normalized importance for firm-specific features described in Appendix A, A.1. Note that while an individual feature importance can vary from month to month, we report its aggregated rank across all test samples to demonstrate its persistence.

Figure 8 presents the ranking score results, where all ranks are normalized to add up to 1 for the cross-section of *All Stocks*.[10]

First, among firm-specific characteristics, panel (a) , short term reversals (mom1ret) are always ranked as the top feature. This very similar to the results of Gu et al. (2019), who also find that mom1ret is always ranked as one of the top features across all 5 factors in their autoencoder model. However, the short-term reversals measured by mom1ret do not always imply reversals. Medhat and Schmeling (2022) find a short-term momentum for stocks with high trading volume and high mom1ret, while reversals are observed for only thinly traded stocks and high mom1ret. This short-term momentum is as profitable and as persistent as conventional price momentum (Medhat and Schmeling (2022)). Given this perspective, it is not suprising that ML methods identify this variable as an important predictor.

It further followed by the volatility of liquidity variable - standard deviation of shares turnover (std_turn) and then by industry adjusted cash flow to price ratio (cfp_ia) and industry sales concentration (herf). The latter is interesting, since the variables related to the firm cash flows are not often ranked as highly. Normally, momentum related variables (mom6m, mom12m) or volatility and market betas are ranked at the top (Gu et al. (2020, 2019); Kelly et al. (2019). In our analysis, these characteristics are ranked below top 20, except *maxret*, maximum daily return for the month, which is, similar to mom1ret, related to subsequent reversals.

The top characteristics that our model identifies are heavily dominated by firms' fundamentals: earnings announcement returns (ear), industry adjusted percentage change in capex ($pchcapx\_ai$), Piotrosky financial statements score (ps) and Mohanram financial statement score (ms), R&D to market capitalization ratio (rd_mve), return on assets (roaq). These variables further followed by other important fundamentals: asset growth (agr), book-to-market (bm), change in tax expenses (chtx), percent accruals (pctacc), or abnormal earnings announcement volume (aeavol). Trading activity variables, dollar volume or turnover, are also ranked in the top 20, but their combined share is substantially smaller than those of firm fundamental variables.

Another interesting observation is that idiosyncratic volatility (idiovol), which is often

[10]We present the ranking for all stocks as they are similar across all other cross-sections.

identified as the main predictor of stock returns (Ang et al. (2006)) is ranked at the bottom. The recent explanation of idiosyncratic volatility puzzle is that it is not the risk but rather market sentiment and limits-to-arbitrage that explain the predictability (Stambaugh et al. (2012)). Confirming this evidence, our model ranks idiovol at the bottom of importance.

Overall, our SDF captures information related to firms' fundamentals and accounting balance sheet variables. This further supports our statement that our SDF represent fundamental, cash flow related systematic factors.

Macro-economic variables' importance is presented separately in Figure 8, panel (b). Betting against Beta, BAB, factor (Frazzini and Pedersen (2014)) is at the top of macro-features' importance, followed by S&P500 index, and then followed by a long term rate of returns (ltr), and TED spreads. The variables related to market wide liquidity, amihud and cs_baspread are not ranked as high. However, BAB and TED spread are commonly used measures of funding liquidity. Therefore, on the market-wide level, funding liquidity conditions are of the first order of importance (Brunnermeier and Pedersen (2009)).

Figure 9 presents the variable importance for all, firm-specific and macro-features together. The top firm specific variables dominate macro-variables in their relative importance. However, BAB is still in the top 10 predictors after controlling for firm specific predictors. S&P500 just falls shortly into 11th importance rank. Most of other macro-variables, except TED spread and Amihud's illiquidity, are at the bottom of importance. Thus, not all macro-variables are equally important after firm specific characteristics.

4.7 The Incremental Importance of Macro-economic Variables

To understand the incremental importance of macro-economic variables, we re-estimate *Attention* model without macro-features and compare the portfolio performance summary statistics to those reported in Table 1. Table 7 below presents the results and Figure 10 plots *OOS* Sharpe ratios obtained with *Attention* model using macro and without macro-variables. First, as can be seen from Figure 10, macro-variables substantially improve Sharpe ratios across all cross-section of stocks. Therefore, incremental predictions obtained with BAB and other macro-factors are quite important.

The numerical results in Table 7 provide more detailed differences. The magnitudes of Sharpe ratios on average 0.30 lower compared to those reported in Table 1, which is economically meaningful. Maximum draw-downs and 1-month losses are also higher in Table 7, while alphas are on average 20bps lower.

	All Stocks	q>10%	q>20%	Russell 3000 Stocks
Return	0.0080	0.0036	0.0033	0.0043
Std.Dev.	0.0111	0.0143	0.0165	0.0140
Alpha	0.0069	0.0018	0.0010	0.0028
t-stat	8.001	2.043	1.118	3.196
Sharpe	2.494	0.866	0.700	1.067
InfRatio	2.491	0.575	0.315	0.894
MaxDD	0.071	0.141	0.204	0.131
Max 1M Loss	-0.050	-0.076	-0.097	-0.068

Notes: The table reports SDF performance summary statistics for *Attention* model, similar to those reported in Table 1, for the model estimated without inclusion macro-variables.

Table 7: *Attention* model, no macro-variables

The macro-features have the highest impact on the portfolio performance constructed with only large stocks, $q > 20\%$. The annualized Sharpe ratio here is only 0.7, which is very close to the Sharpe ratio of S&P500 for the same period, 0.56, and the portfolio Alpha is statistically insignificant. Therefore, withholding macro-economic information from portfolio optimization aimed on holding only large stocks makes almost no difference from passively holding S&P500 index. Maximum draw-down of 20.4% makes this strategy extremely risky.

We conclude that, besides stock specific characteristics, inclusion of macro-economic information into portfolio optimization is crucial.

5 Attention and Turnover

5.1 Turnover

What do we pay attention to and how it affects turnover? DeMiguel et al. (2020) suggest that incorporating more characteristics is better and leads to lower turnover on the portfolio level as required trades on multiple characteristics for the same underlying stock can cancel each other out. We do not directly control for the turnover in our optimization, we only "pay attention" to multiple characteristics.

The direct test of the theory's implications is a test of whether "paying attention" to fewer characteristics results in higher turnover. Here we run the following experiment. Instead of using all 94 stock characteristics for each stock-month as before, we only use 20 characteristics. We also use forward looking information and pre-select top and bottom 20 characteristics

for the variable importance ranking, Figure 8, panel (a). Here, using the forward, ex-post information is not an issue as we want to understand the following: (i) whether reducing the number of characteristics leads to an increase in turnover; (ii) whether the quality, or informativeness (top 20 vs. bottom 20) of smaller set of characteristics makes a difference for an increase in turnover.

Table 8 presents two sets of results using top 20 characteristics, the first column, and bottom 20 characteristics, the second column. Here, to compare the results to the base case with all characteristics for all stocks, Table 1 - the first column, we also report the portfolio performance statistics. One important observation is that decreasing the number of features leads to a significant drop in the performance. The annualized Sharpe ratio of 2.83 which uses all data drops to 0.81 and 0.66 when we use top and bottom 20 features respectively. Therefore the big data approach for portfolio construction really matters and extra information of the whole set of 94 characteristics makes significant difference for ex-post performance.

Further, for smaller set of predictors, the quality of predictors matters. Top 20 features and macro-variables lead to higher Sharpe ratio compared to the similar optimization using bottom 20 features. Moreover the Sharpe ratio of the latter, 0.66, is very close to the market Sharpe ratio for the similar period, 0.56. Its portfolio Alpha of 38 bps per month is only marginally significant, t=2.35, while the Alpha of top-20 feature portfolio is 0.77, and highly significant, t=5.299. Note that the latter is very close to the base-line model's Alpha of 0.88 bps per month. Therefore, an investor who does not use big data in portfolio construction, or who does not know ex-ante the optimal set of best predictors, is better off just investing in a passive market index portfolio.

The turnover statistics reported below portfolio performance statistics provide a direct evidence to out testable hypotheses. Here, we compare these turnovers to those of base-line model reported in the first column of Table 4, where the overall portfolio turnover is 0.348, with turnovers on the Long and Short positions of 0.227 and 0.121 respectively. For both, top and bottom 20 features, turnovers increase significantly, with overall portfolio turnover is slightly smaller, 0.424, for top 20 features, compared to 0.438 for bottom 20 features.

To make an exposition easier, the last three rows present the percentage increase in turnover compared to the base-line model. For the top 20 features the overall, *Turnover All* increases by 22% versus 26% increase for the bottom 20 features. More importantly, the highest turnover increase is observed on the short positions, 40% and 48% for top and bottom features respectively, where the limits-to-arbitrage are the highest. The turnover on the long position increases less, 12% and 14% for the top and bottom features respectively.

	All Stocks	
	TOP 20 Features	**Bottom 20 Features**
Return	0.00647	0.00573
Std.Dev	0.02769	0.03025
Alpha	0.00769	0.00376
t-stat	5.299	2.348
Sharpe	0.809	0.656
InfRatio	0.291	0.141
MaxDD	0.481	0.484
Max 1 M Loss	-0.250	-0.129
Turnover All	0.424	0.438
Turnover Long	0.254	0.259
Turnover Short	0.169	0.179
Turnover All % increase	22%	26%
Turnover Long % increase	12%	14%
Turnover Short % increase	40%	48%

Notes: The table reports SDF performance summary statistics for *Attention* model, similar to those reported in Table 1, first column, All Stocks, for the model estimated using only top 20 or bottom 20 stock-specific characteristics. The relative feature-characteristics importance and rankings are determined ex-post. The table also report summary of portfolio turnovers and their percentage increases compared to the base-line model reported in the first column of Table 4.

Table 8: *Attention* model, fewer stock characteristics/features (small data)

Overall, two important conclusions emerge. First, the number of characteristics matters for portfolio turnover, and fewer characteristics lead to significantly higher turnover (DeMiguel et al. (2020)). Second, the quality or the relative importance of characteristics matters, and, everything else equal, more important characteristics do lead to lower portfolio level turnover. Since ex-ante an investor does not know which characteristics can be important from one month to another, then "paying attention" to large set of all available characteristics is the way to reduce the overall portfolio level turnover.

5.2 Attention

Our main innovation is to show that machine *Attention* is a viable instrument which allows to decrease the noise of financial data, establish persistent causalities and improve performance of deep learning models from being not really relevant from real investment perspective (Avramov et al. (2021)) to been fully tradable and implementable/scalable portfolio management strategies. All our results so far point to enormous benefits of using machine *Attention* in

finance applications.

Here, we present the analysis of what the attention mechanism that we model is actually "paying attention" to. To remind, each of five blocks in our deep learning model, Figure 1, begins with multi-head attention layer, i.e. all the data input first goes through attention embedding. Thus, after the model being trained, *Attention* pre-identifies the feature importance before all features go further into an optimization process of each block. Therefore, *Attention* first helps optimization routing by decreasing the noise across multiple features, emphasizing more persistent causalities/features, and, second, the optimization withing each block only further refines the feature importance, pre-identified by attention layer, which should result in better forecasts.

This is the closest analog from machine translations to a real life portfolio manager decision making process. A portfolio manager would first be aware of multiple firm characteristics which could potentially matter. However, her professional experiences taught her that only a hand-full among them really makes a difference. Therefore, among the myriad of many variables she only will be paying attention to what historically she has observed is more important for the future portfolio performance. She then will use these data to make asset allocation decisions.

Our modeling set up with *Attention* replicates this portfolio manager's approach. The only difference is that the portfolio manager has a pre-specified set of rules given her experiences, and it will take her some time to analyze ex-post and introduce a new rule if there are structural changes in the market data, i.e. changes in markets behaviour.

Machine *Attention* learns the rules from the market data and identifies the best rules leading to better and consistent performance. It is very similar to what the portfolio manager does with only two exceptions. First, machine *Attention* is not limited to a relatively small set of variables/signals that human can physically analyze, but covers the whole universe of all possible signals and interactions between them. This therefore enables machine *Attention* to identify rules that human might not be aware of in real time. Second, it also makes machine *Attention* to react faster to incorporate new rule into "decision making", optimization process, while for a human it might take a few months or even years to collect enough of historical data in order to establish statistical significance of new signals.

As an example, here we demonstrate two episodes of very different market conditions and what *Attention* ex-ante identifies as important signals for portfolio construction. We use the last *OOS* month of out sample, December 2020, i.e. the last time we refit the model in our sample, as a low volatility market regime, and we compare it to September 2008, an outburst of financial crisis, as an example of extreme volatility regime. We identify *Attention* feature

importance ex-ante for each month-episode, November 2020 and August 2008 respectively.

Before we proceed, we first briefly overview what *Attention* embedding is and what it indicates in natural language processing. *Attention* embedding is a matrix where diagonal elements represent the probability of a word representing itself, and off-diagonal elements represent probabilities of other words in a sentence to describe the object in the sentence. The magnitudes of the higher off-diagonal probabilities describe how closely the other words are attributed to the objective word. Figure 11 provides an example from textual analysis. After the model is trained on the sentence "The quick brown fox", the 4x4 *Attention* matrix in the figure shows the diagonal elements with high probabilities, i.e. each word has very high its own representation. What is important here is off-diagonal elements, where the object of the sentence is "fox", and "quick" and "brown" are close characterizations of the "fox". Projecting this example into our setting, *Attention* essentially allows to cluster characteristics around the objectively important main characteristics. These main characteristics should provide economic characterization of the fundamental factors.

The *Attention* feature importance is computed using eq. 14, except that for each element D of *Attention* layer output we compute the absolute value of the gradient for each input variable, and we do it D times to obtain DxD attention metrics. Instead of a simple, diagonal, variable importance, here we report *Attention* matrices which show the importance of off-diagonal interaction effects for each feature, i.e. similar to the column "fox" in the above example.

Figure 12 and Figure 13 present attention matrices for 12/2020 and 09/2008 respectively, with each figure having 5 sub-figures corresponding to the attention layer of each of 5 blocks. The block structure of the model, where each block corresponds to a fundamental factor, also ranks the factors in the order of importance, from Block 1, the most important, to Block 5, the last in the order of importance. The rankings of the factors and the factors themselves can change from one month to another. The color scheme has standard importance identifications with the darker purple colors indicating higher importance, and lighter orange colors indicating low-to-no importance at all.

Consider first Figure 12, 12/2020. Panel (a) presents the results for the first block, block1, of the stack. The macro-economic variables get less attention in the first block, as they all have mostly lighter colors and are grouped in the end of the ordering, the far right corner of the matrix on the horizontal axis. The darker purple colors really appear for: turnover, earnings announcement returns, Piotroski financial statement score and Quick ratio (firm's financial and liquidity scores), size, price delay, and reversals, mom1ret. Besides the importance of these variables on their own, individually, for the first block, the interaction

effects of these variables with most of macro-variables is as important. This can be seen by the darkness of vertical purple lines centered on each of this variables which is gradually fading out from the bottom, where all macro-features are centered on the vertical axis, to the top. Besides macro-variables, all of these variables are also importantly interact with all firm-specific sales-based indicators, which are located just above macro-variables on the vertical axis. Thus, the cross-feature interactions are as important signals as an individual feature importance.

Panel (b) of Figure 12 presents feature importance for the block 2. It identifies a completely different set of important variables. First, two macro-indicators become important: ltr and dfr, long-term rate and default spread respectively. Second, a different set of firm specific characteristics appears of the first order of importance: bm (book-to-market), divi (dividend initiation), pchsale_pchxsga (Percentage change in sales less percentage change in SG&A), and saleinv (Sales to inventory ratio). The interaction effects are also different, and mostly limited to the bottom 10 macro-economic indicators on the vertical axis.

Block 3, panel (c), brings attention to again completely different set of variables. Similar to block 1, it ignores macro-features. The firm-specific characteristics which are most important for block 3 are: aeavol (Abnormal earnings announcement volume), age, cash, cashdebt (Cash flow to debt), lev (leverage), mve_ai (Industry-adjusted size), rd_mve (R&D to market capitalization), rd_sales (R&D to sales), sale_cash (Sales to cash), and sp (Sales to price). As far as cross- feature interaction importance is concerned, these variables are highly effectively interact with all macro-features, and just a hand-full of firm-specific characteristics located just above macro-indicators on the vertical axis.

Unlike other blocks, Block 4, panel (d), gives a lot of importance to macro-indicators of market-wide volatility and illiquidity: svar_macro (market variance), VIX and Amihud illiquidity. Among stock-specific features, leverage appears to be important again, lev, together with the new features: cfp (Cash flow to price ratio), pchsale_pchrect (Percentage change in sales less percentage change in A/R), stdcf (Cash flow volatility), and tb (Tax income to book income).

Finally, block 5, panel (e), stands apart from all other blocks, since only one sock-specific variable is important here: mom1m.

To summarize, each block identifies different signals from different variables. Upstream blocks highlight the most important factors and downstream blocks the rest of factors which follow in order of importance. For 12/2020 predictions, the last time we re-train the model, the factor importance can be summarized as follows, from high, 1, to low, 5: 1. trading activity and financial, firm level liquidity; 2. value characteristics; 3. cash flows; 4. macro-

and cash flow volatility; 5. short-term momentum and reversals. Note that the first three factors have a close link with Size, Value and Profitability Fama-French factors respectively.

Figure 13 provides a completely different variable importance representation for 09/2008, the beginning of financial crisis. Unlike for 12/2020, Block 1 *Attention* is mostly focused on individual feature importance of the macro-variables: dfr (Default Return Spread), infl (CPI inflation) , spvw (S&P 500 Index Returns), and bab (betting-against beta, funding liquidity proxy), and they have important significant inter-action effects with all other macro-variables. Among stock-specific features, only one, std_turn (Volatility of liquidity (share turnover)), stands out. Thus the fist block is mostly macro factor which is unarguable of the first order importance during inception of global crisis.

Block 2, panel (b), highlights only one stock-specific variable importance, mom1m. In the previous example, mom1m was the fifth in the order of importance factor. It is easy to expect that during market downturn, past stock returns are as important as global market-wide factors.

Block 3, panel (c), pays attention to only few firm specific features: pchcapx_ai (Industry adjusted percentage change in capital expenditures), maxret (maximum daily return), and chatoia (Industry-adjusted change in asset turnover). Thus, price jumps during past month, maxret, as well as the ability to weather the storm, i.e. company specific capex expenditures and assets turnarounds, are also unarguably important signals during the crisis.

Block 4, panel (d), brings up the importance of mostly two stock-specific characteristics: absacc (Absolute Accruals) and realestate (Real estate holdings), and less so to salerec (Sales to receivables). Identifying real estate holdings' importance at the inception of the crisis caused by the real estate bubble cannot be interpreted as coincidental.

Finally, Block 5 portrays the importance of the fallowing stock-specific features: stdacc (Accrual volatility), sin (sin stocks), betasq (beta squared), chcssho (Change in shares outstanding). This set of variables is again completely different from everything above, and aims to capture uncertainty on the firm level. Dividend yield and TED spread are important macro-features here. The latter is expected, as TED spread measures funding liquidity constraints which became binding at that time.

The factor importance here can be summarized in the following ranking: 1. macroeconomic conditions; 2. reversal or continuation (past month returns); 3. stock price jumps and the strength of firm's level balance sheets; 4. real estate holdings vs. expected cash inflows; 5. stock level volatility and firm level uncertainty. Therefore, unlike for 12/2020, a completely different set of fundamentals is ex-ante identified to drive the performance in 09/2008. We

can agree ex-post that these fundamentals did play a significant role in determining future price movements.

This evidence also echoes the results reported in Table 5 for different volatility regimes. For low volatility regimes, after excluding small cap stocks, Panel B, FF6 Alphas are 2 to 3 times smaller than raw returns. This suggests that FF6 factors are important during low volatility, stable market conditions. The factor characteristics identified for 12/2020 have significant overlap with FF6 factor characterizations (size, value, profitability). In contrast, the factor characteristics identified during the crisis, for 09/2008, have very little to do with FF6 factors, except, perhaps, the one related to price trends.

6 Conclusion

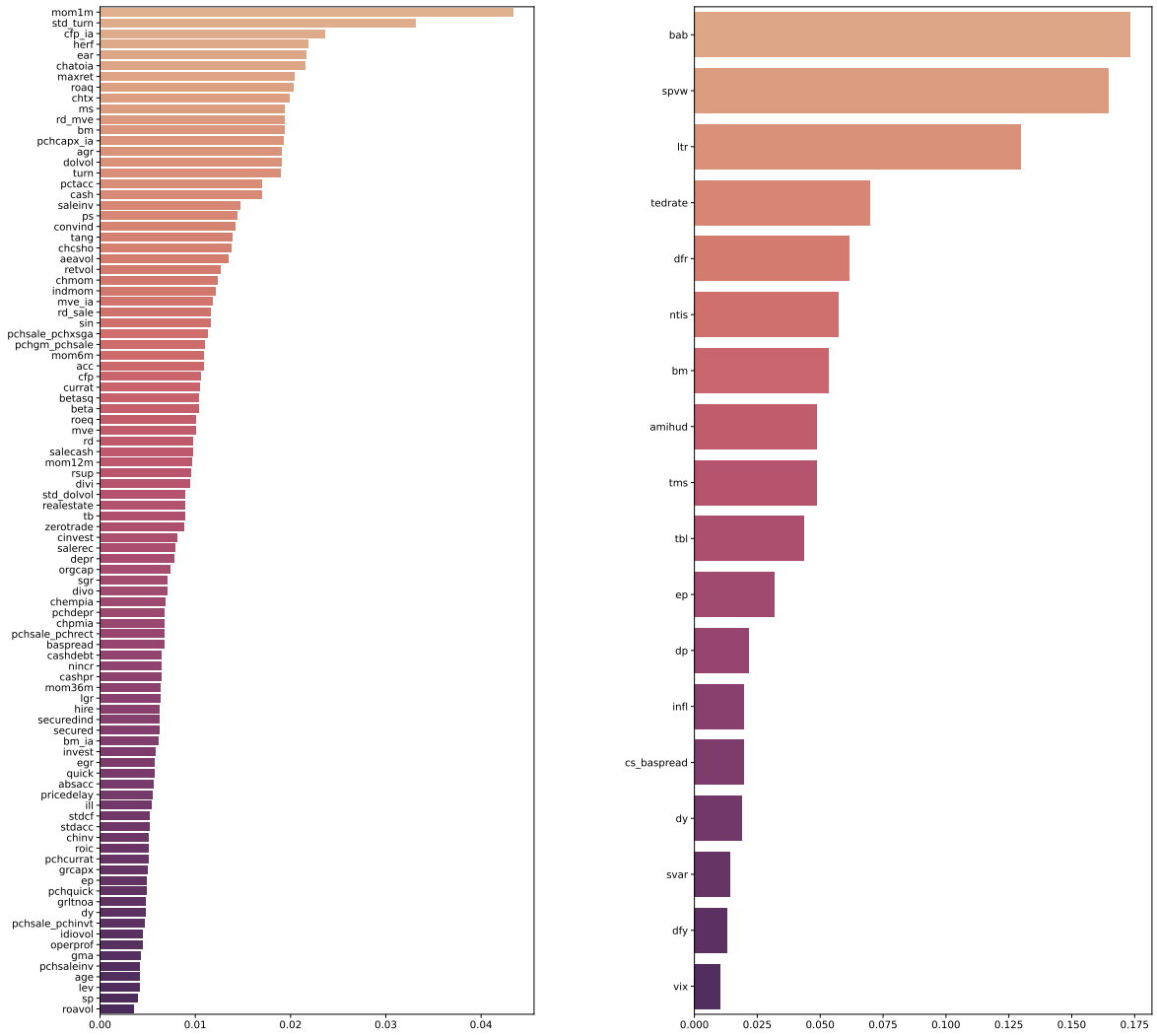
In this paper we approach the most fundamental problem in finance - optimal portfolio construction - with attention guided deep learning.

We adapt the architecture of Neural Basis Expansion Analysis for Interpretable time series forecasting (N-BEATS) of Oreshkin et al. (2019) for asset pricing and apply it to the cross-section of all US publicly listed common stocks, their 94 Green et al. (2017) firm-specific characteristics and 18 macro-economic variables.

We show the advantages of deep learning and big data for stochastic discount factor (SDF) identification. Similar to Avramov et al. (2021) we also find that deep learning on its own, without fundamentally built in economic guidance, fails to maintain the superior properties of SDF after accounting for the limits to arbitrage.

However, the asset attention we introduce in this paper, which serves as a guidance for deep learning network about stock-specific feature importance, allows to overcome limits-to-arbitrage and clears the protocol introduced by Avramov et al. (2021).

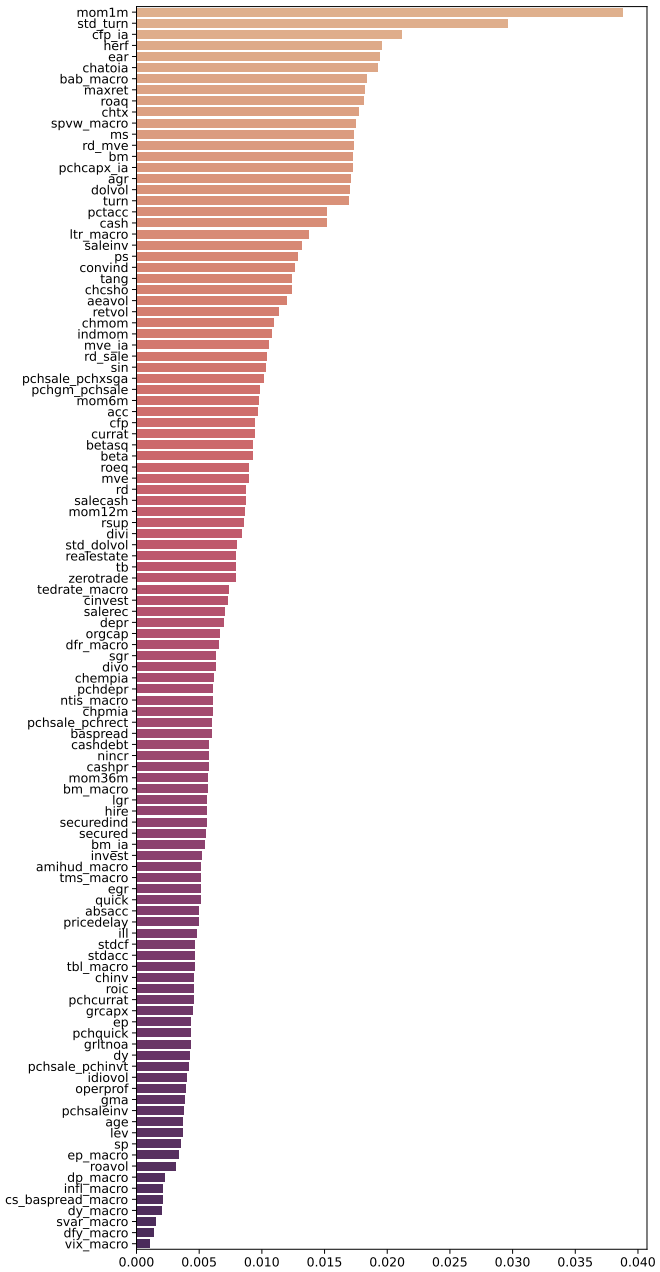
We also find an empirical support to DeMiguel et al. (2020), who argue in favor of considering multiple firm characteristics for portfolio construction to implicitly account for portfolio's trading costs. Combining all characteristics reduces trading costs as the trades required to rebalance in one underlying stock based on different characteristics often cancel out. We confirm that "paying asset specific attention" to these characteristics does the job.



(a) Firm-specific Features (b) Macro Features

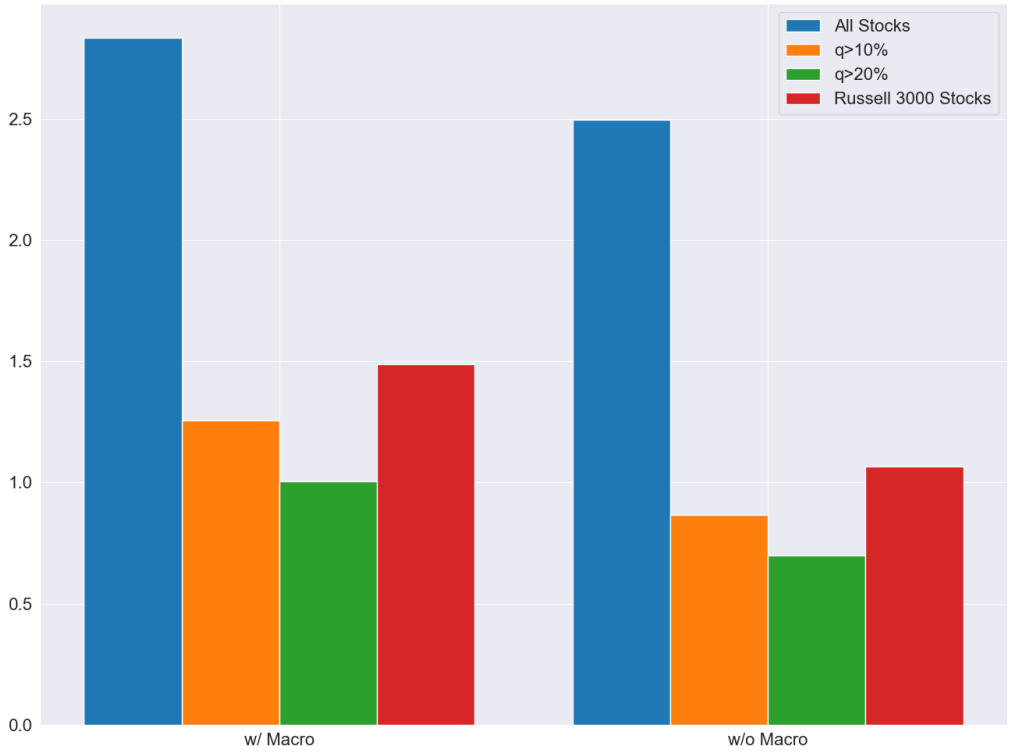
Notes: The figure presents variable importance of all variables based on median value over all test samples. The variable importance is normalized to sum to one.

Figure 8: Firm-specific and macro features importance for *Attention* model



Notes: The figure presents variable importance of all, firm-specific and macro, variables based on median value over all test samples. The variable importance is normalized to sum to one.

Figure 9: Firm-specific and macro features importance for *Attention* model.



Notes: The figure plots *OOS* Sharpe Ratios obtained with macro-variables, left, from Table 1, and those without macro-variable, reported in Table 7 for *Attention* model

Figure 10: *OOS* Sharpe Ratio with and without Macro-variables

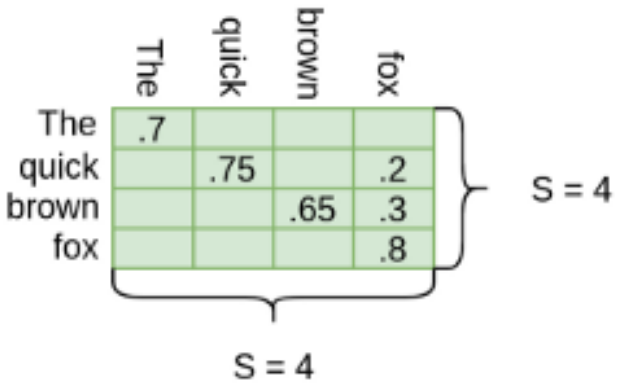
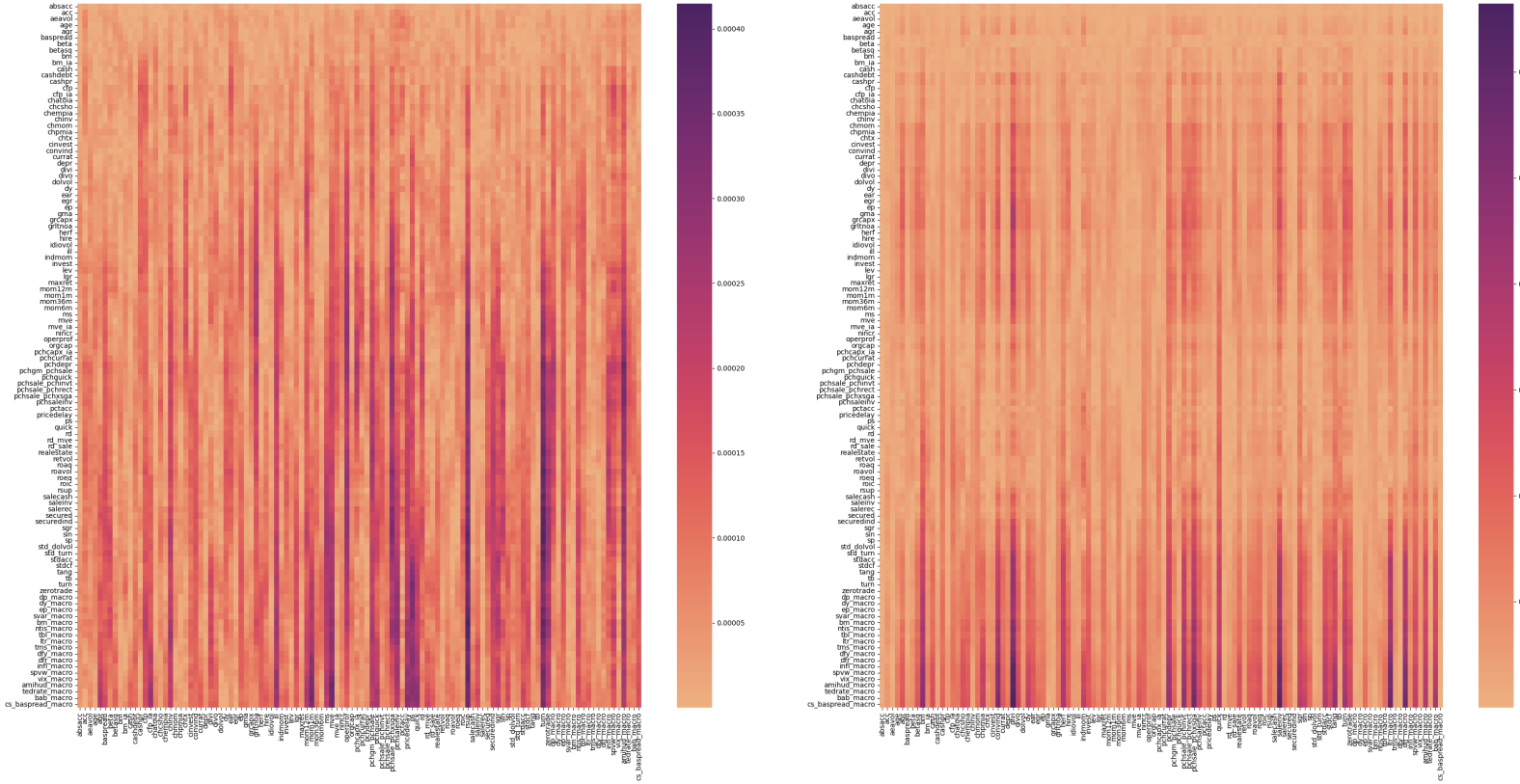


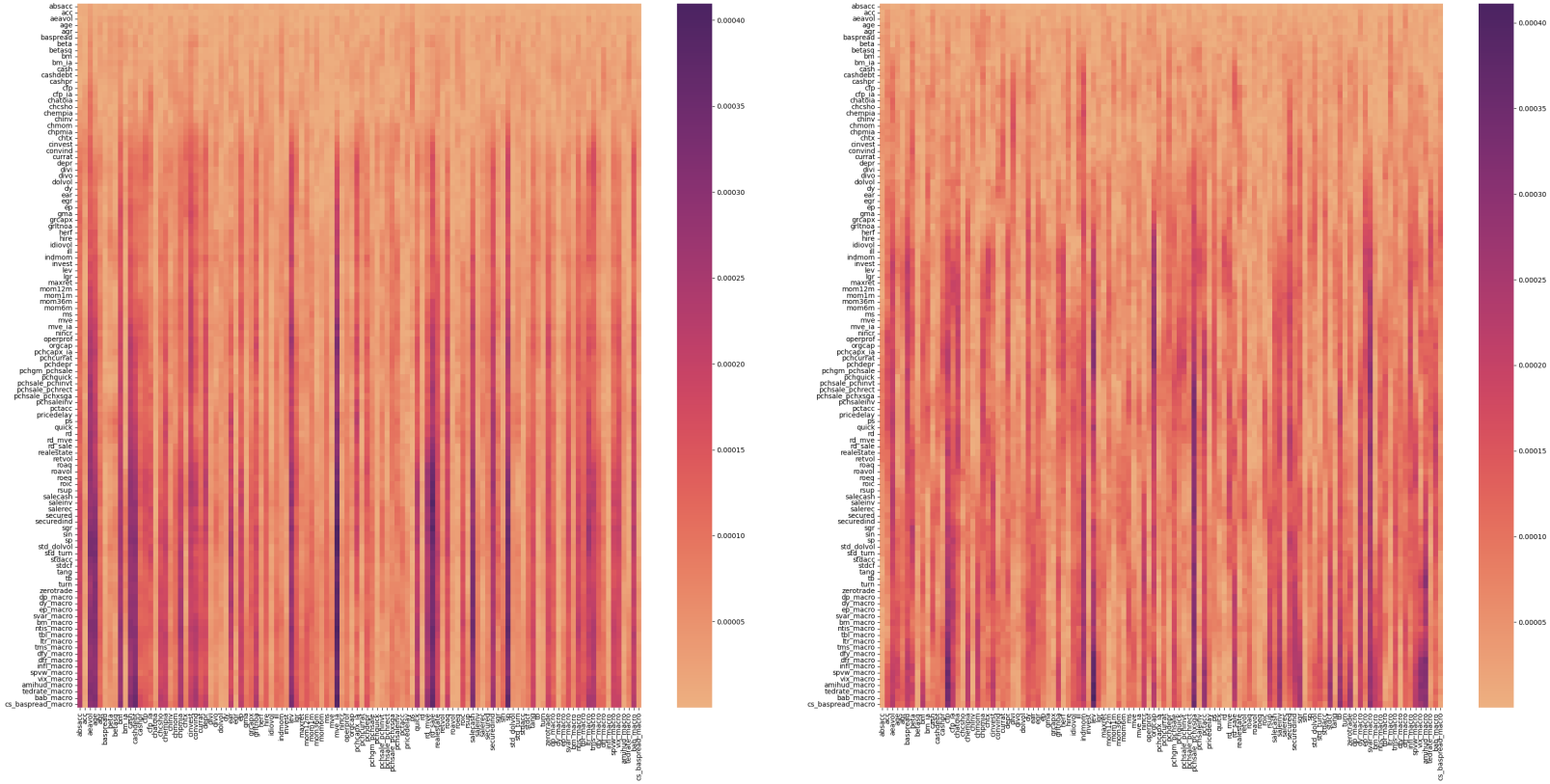
Figure 11: Embedding via *Attention* layer: NLP example.



(a) Block 1

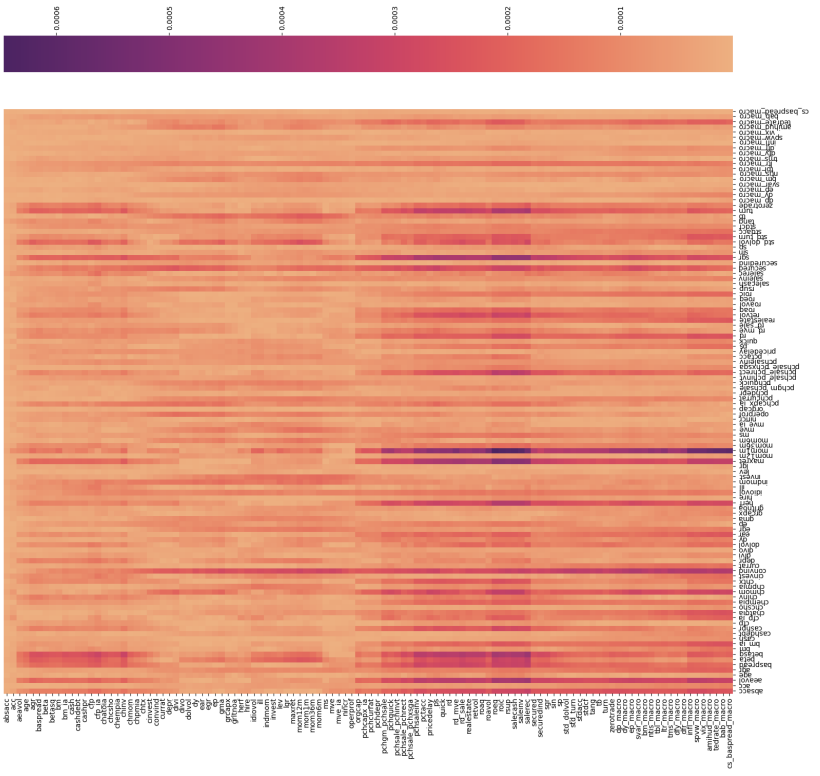
(b) Block 2

Figure 12: *Attention* feature importance by block, for *OSS* 12/2020.



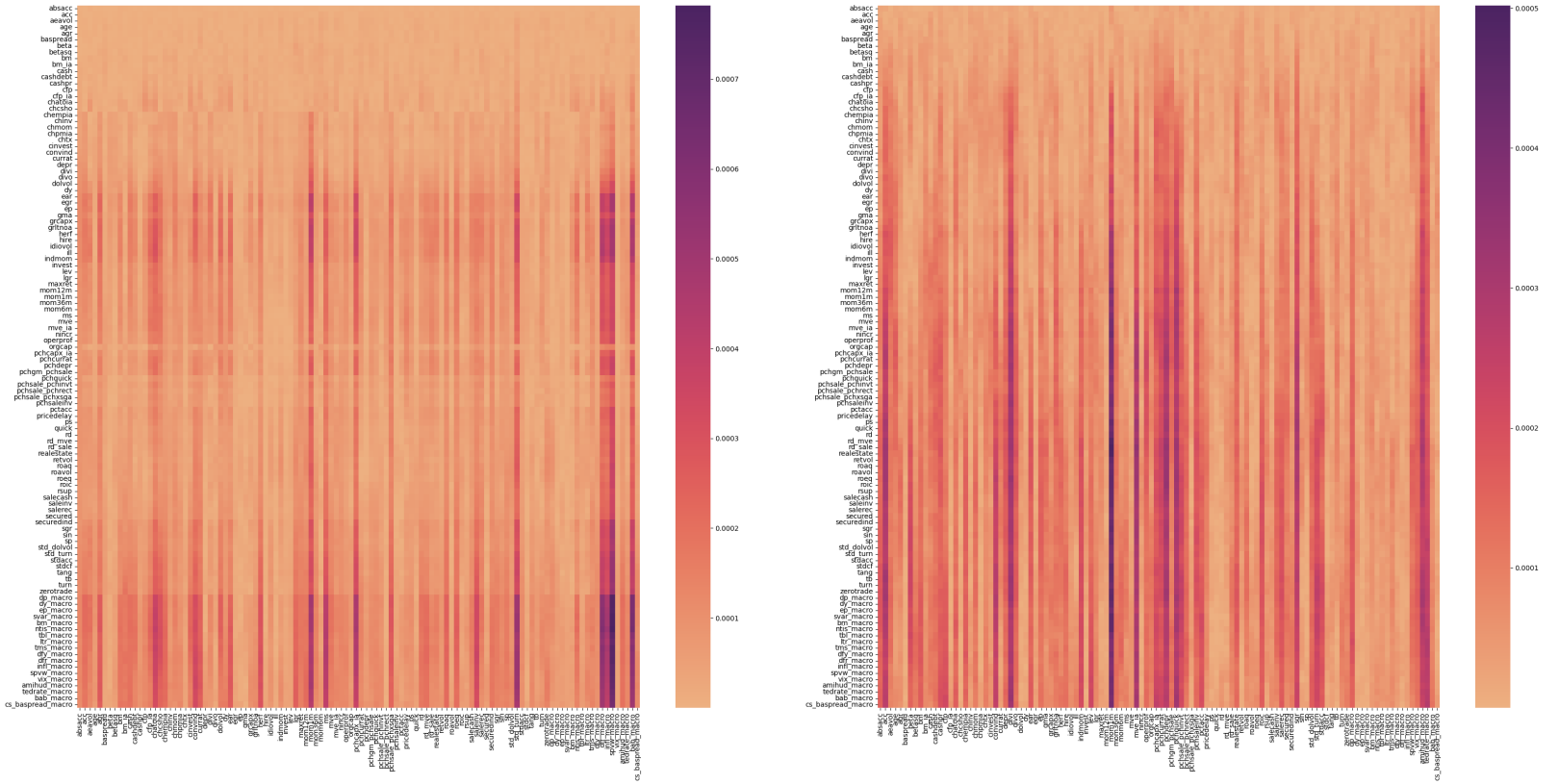
(c) Block 3 (d) Block 4

Figure 12: *Attention* feature importance by block, for *OSS* 12/2020.



(e) Block 5

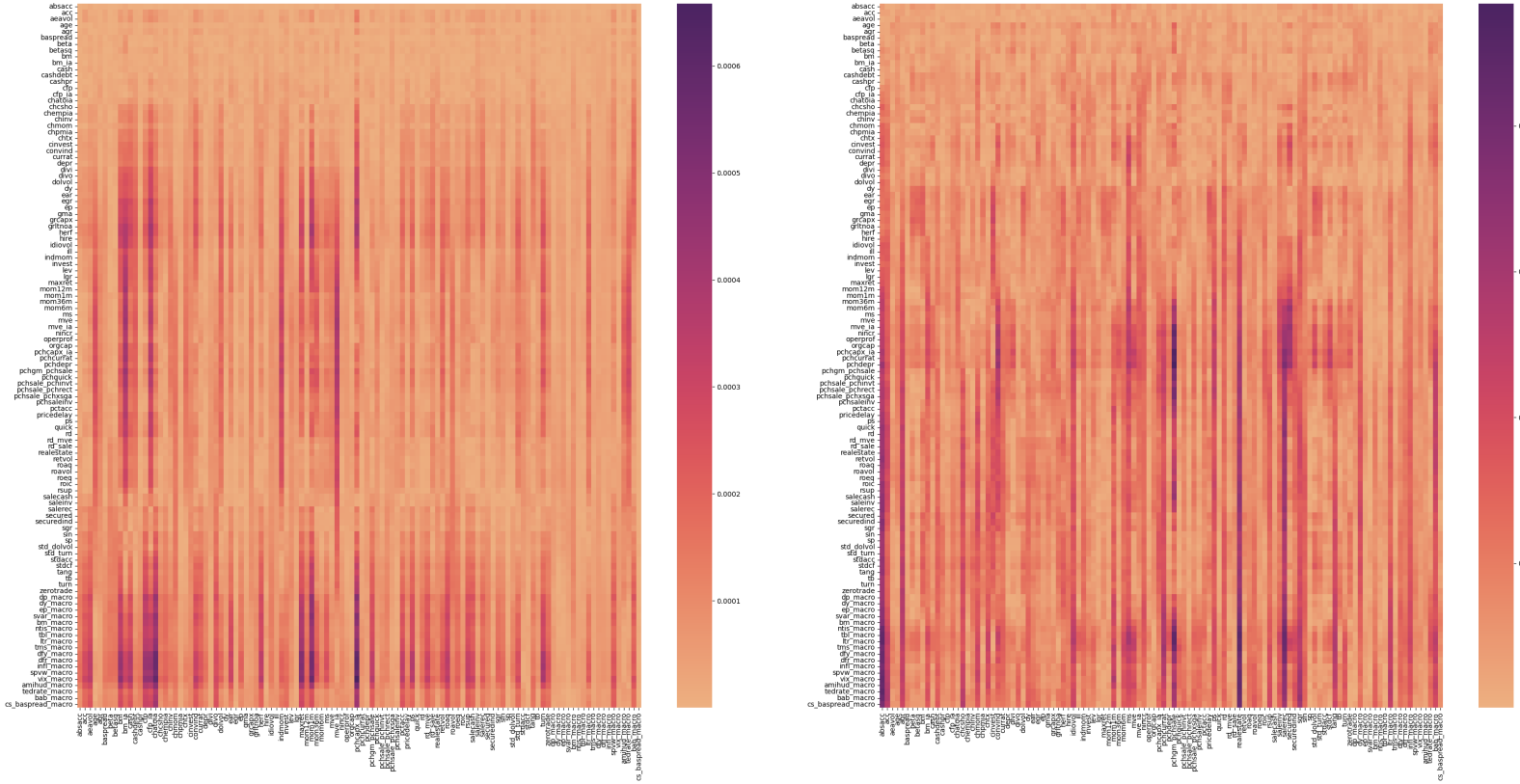
Figure 12: *Attention* feature importance by block, for *OSS* 12/2020.



(a) Block 1

(b) Block 2

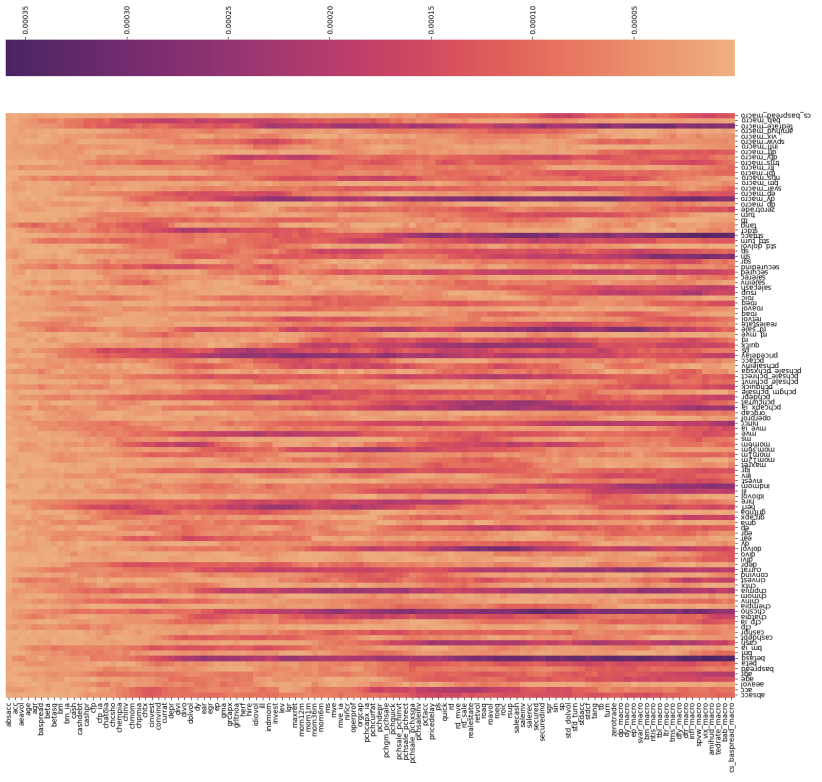
Figure 13: *Attention* feature importance by block, for *OSS* 09/2008.



(c) Block 3

(d) Block 4

Figure 13: *Attention* feature importance by block, for *OSS* 09/2008.



(e) Block 5

Figure 13: *Attention* feature importance by block, for *OSS* 09/2008.

References

Aït-sahali, Y. and Brandt, M. W. (2001). Variable selection for portfolio choice. The Journal of Finance, 56(4):1297–1351.

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. Journal of financial markets, 5(1):31–56.

Ang, A., Hodrick, R. J., Xing, Y., and Zhang, X. (2006). The cross-section of volatility and expected returns. The Journal of Finance, 61(1):259–299.

Avramov, D., Cheng, S., and Metzker, L. (2021). Machine learning versus economic restrictions: Evidence from stock return predictability. Available at SSRN 3450322.

Bianchi, D., Büchner, M., and Tamoni, A. (2021). Bond risk premiums with machine learning. The Review of Financial Studies, 34(2):1046–1089.

Brandt, M. W. (1999). Estimating portfolio and consumption choice: A conditional euler equations approach. The Journal of Finance, 54(5):1609–1645.

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

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2):123–140.

Brunnermeier, M. K. and Pedersen, L. H. (2009). Market liquidity and funding liquidity. The review of financial studies, 22(6):2201–2238.

Bryzgalova, S., Pelger, M., and Zhu, J. (2020). Forest through the trees: Building cross-sections of stock returns. Available at SSRN 3493458.

Chatigny, P., Wang, S., Patenaude, J.-M., and Oreshkin, B. N. (2021). Neural forecasting at scale.

Chen, L., Pelger, M., and Zhu, J. (2020). Deep learning in asset pricing. Available at SSRN 3350138.

Choi, D., Jiang, W., and Zhang, C. (2020). Alpha go everywhere: Machine learning and international stock returns. Available at SSRN 3489679.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al. (2020). Rethinking attention with performers. arXiv preprint arXiv:2009.14794.

Cochrane, J. H. (2011). Presidential address: Discount rates. The Journal of finance, 66(4):1047–1108.

Cong, L. W., Tang, K., Wang, J., and Zhang, Y. (2020). Alphaportfolio for investment and economically interpretable ai. SSRN, https://papers. ssrn. com/sol3/papers. cfm.

Corwin, S. A. and Schultz, P. (2012). A simple way to estimate bid-ask spreads from daily high and low prices. The Journal of Finance, 67(2):719–760.

DeMiguel, V., Martin-Utrera, A., Nogales, F. J., and Uppal, R. (2020). A transaction-cost perspective on the multitude of firm characteristics. The Review of Financial Studies, 33(5):2180–2222.

Feng, G., Polson, N., and Xu, J. (2021). Deep learning in characteristics-sorted factor models. Available at SSRN 3243683.

Frazzini, A. and Pedersen, L. H. (2014). Betting against beta. Journal of Financial Economics, 111(1):1–25.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

Green, J., Hand, J. R., and Zhang, X. F. (2017). The characteristics that provide independent information about average us monthly stock returns. The Review of Financial Studies, 30(12):4389–4436.

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

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

Gu, S., Kelly, B. T., and Xiu, D. (2019). Autoencoder asset pricing models.

Han, Y., He, A., Rapach, D., and Zhou, G. (2020). Firm characteristics and expected stock returns. Available at SSRN 3185335.

Hansen, L. P. and Jagannathan, R. (1991). Implications of security market data for models of dynamic economies. Journal of political economy, 99(2):225–262.

Hansen, L. P. and Jagannathan, R. (1997). Assessing specification errors in stochastic discount factor models. The Journal of Finance, 52(2):557–590.

Heaton, J. B., Polson, N. G., and Witte, J. H. (2017). Deep learning for finance: deep portfolios. Applied Stochastic Models in Business and Industry, 33(1):3–12.

Hou, K., Xue, C., and Zhang, L. (2020). Replicating anomalies. The Review of Financial Studies, 33(5):2019–2133.

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

Kim, S., Korajczyk, R. A., and Neuhierl, A. (2021). Arbitrage portfolios. The Review of Financial Studies, 34(6):2813–2856.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Koijen, R. S., Moskowitz, T. J., Pedersen, L. H., and Vrugt, E. B. (2018). Carry. Journal of Financial Economics, 127(2):197–225.

Kozak, S., Nagel, S., and Santosh, S. (2020). Shrinking the cross-section. Journal of Financial Economics, 135(2):271–292.

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

Markowitz, H. (1952). The utility of wealth. Journal of political Economy, 60(2):151–158.

Medhat, M. and Schmeling, M. (2022). Short-term momentum. The Review of Financial Studies, 35(3):1480–1526.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Icml.

Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y. (2019). N-beats: Neural basis expansion analysis for interpretable time series forecasting. In International Conference on Learning Representations.

Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y. (2020). Meta-learning framework with applications to zero-shot time-series forecasting. arXiv preprint arXiv:2002.02887.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

Rapach, D. E., Strauss, J. K., Tu, J., and Zhou, G. (2019). Industry return predictability: A machine learning approach. The Journal of Financial Data Science, 1(3):9–28.

Rapach, D. E. and Zhou, G. (2020). Time-series and cross-sectional stock return forecasting: New machine learning methods. Machine Learning for Asset Management: New Developments and Financial Applications, pages 1–33.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Stambaugh, R. F., Yu, J., and Yuan, Y. (2012). The short of it: Investor sentiment and anomalies. Journal of Financial Economics, 104(2):288–302.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

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

A Appendix

A.1 Stocks-Specific Covariates

Table. 1 provides a list of all stock-specifics predictors of Green et al. (2017) that we consider.

No.	Acronym	Description	No.	Acronym	Description
1	absacc	Absolute Accruals	48	mom36m	36-month momentum
2	acc	Working Capital accruals	49	mom6m	6-month momentum
3	aeavol	Abnormal earnings announcement volume	50	ms	Mohanram financial statement score
4	age	# years since first Compustat coverage	51	mve	Size
5	agr	Asset growth	52	mve_ia	Industry-adjusted size
6	baspread	Bid-ask spread	53	nincr	Number of earnings increases
7	beta	Beta	54	operprof	Operating profitability
8	betasq	Beta squared	55	orgcap	Organizational capital
9	bm	Book-to-market	56	pchcapx_ia	Industry adjusted % change in capital expenditures
10	bm_ia	Industry-adjusted book to market	57	pchcurrat	% change in current ratio
11	cash	Cash holdings	58	pchdepr	% change in depreciation
12	cashdebt	Cash flow to debt	59	pchgm_pchsale	% change in gross margin less % change in sales
13	cashpr	Cash productivity	60	pchquick	% change in quick ratio
14	cfp	Cash flow to price ratio	61	pchsale_pchinvt	% change in sales less % change in inventory
15	cfp_ia	Industry-adjusted cash flow to price ratio	62	pchsale_pchrect	% change in sales less % change in A/R
16	chatoia	Industry-adjusted change in asset turnover	63	pchsale_pchxsga	% change in sales less % chagne in SG&A
17	chcsho	Change in shares outstanding	64	pchsaleinv	% change in sales-to-inventory
18	chempia	Industry-adjusted change in employees	65	pctacc	Percent accruals
19	chinv	Change in inventory	66	pricedelay	Price delay
20	chmom	Change in 6-month momentum	67	ps	Piotroski financial statements score
21	chpmia	Industry-adjusted change in profit margin	68	quick	Quick ratio
22	chtx	Change in tax expense	69	rd	R&D increase
23	cinvest	Corporate investment	70	rd_mve	R&D to market capitalization
24	convind	Convertible debt indicator	71	rd_sale	R&D to sales
25	currat	Current ratio	72	realestate	Real estate holdings
26	depr	Deprecation/PP&E	73	retvol	Return volatility
27	divi	Dividend initiation	74	roaq	Return on assets
28	divo	Dividend omission	75	roavol	Earnings volatility
29	dolvol	Dollar trading volume	76	roeq	Return on equity
30	dy	Dividend to price	77	roic	Return on invested capital
31	ear	Earnings announcement return	78	rsup	Revenue surprise
32	egr	Growth in common shareholder equity	79	salecash	Sales to cash
33	ep	Earnings to price	80	saleinv	Sales to inventory
34	gma	Gross profitability	81	salerec	Sales to receivables
35	grcapx	Growth in capital expenditures	82	secured	Secured debt
36	grltnoa	Growth in long-term net operating assets	83	securedind	Secured debt indicator
37	herf	Industry sales concentration	84	sgr	Sales growth
38	hire	Employee growth rate	85	sin	Sin stocks
39	idiovol	Idiosyncratic return volatility	86	sp	Sales to price
40	ill	Amihud Illiquidity	87	std_dolvol	Volatility of liquidity (dollar trading volume)
41	indmom	Industry momentum	88	std_turn	Volatility of liquidity (share turnover)
42	invest	Capital expenditures and inventory	89	stdacc	Accrual volatility
43	lev	Leverage	90	stdcf	Cash flow volatility
44	lgr	Growth in long-term debt	91	tang	Debt capacity/firm tangibility
45	maxret	Maximum daily return	92	tb	Tax income to book income
46	mom12m	12-month momentum	93	turn	Share turnover
47	mom1m	1-month momentum	94	zerotrade	Zero trading days

Table 1: Description of the 94 stock-based predictors of Green et al. (2017).

A.2 Macroeconomic Variables

Table. 2 provides a list of all macro-economic variables we consider.

No.	Acronym	Description	No.	Acronym	Description
1	dp	Dividend Price Ratio	10	dfy	Default Yield Spread
2	dy	Dividend Yield	11	dfr	Default Return Spread
3	ep	Earnings Price Ratio	12	infl	Consumer Price Index
4	svar	Stock Variance	13	spvw	S&P 500 Index Returns
5	bm	Book-to-Market Ratio	14	vix	VIX Index
6	ntis	Net Equity Expansion	15	amihud	Aggregate Amihud Illiquidity
7	tbl	Treasury-bill Rates	16	tedrate	TED Rate
8	ltr	Long Term Rate of Returns	17	bab	BAB Returns
9	tms	Term Spread	18	cs_baspread	Aggregate Bid-Ask Spread Corwin and Schultz (2012)

Table 2: Description of the 18 macro-economic features.

B Training setup details

The set of hyper-parameters (HPs) used for the experimental section were fixed based on Oreshkin et al. (2019, 2020); Chatigny et al. (2021) and are detailed in Table. 3 below. All hyperparameters are fixed for the whole duration of the backtest. We find that changing the below specified configurations does not have significant impact on the output. All experiments were produced using PyTorch 1.9.0 Paszke et al. (2017). Each model was retrained after each forecast. In total 192 models were trained independently to produce weights for all assets. We repeated the procedure 10 times, i.e. 10 ensembles, for a total of 1920 models.

Hyper-parameter	**Value**
History size (months)	180
Iterations	10,000
Learning rate	0.001
Batch size	1024
Width	512
Blocks	5
Block-layers	4
Attention-head	2
Attention-head-width	64
Attention-head-feature-redraw-interval	1000
Activation function	[`Identity`]
Ensemble size	10

Table 3: Hyper parameters' specifications.