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

MANAGEMENT SCIENCE
*Articles in Advance*, pp. 1–21
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

# Intraday Market Return Predictability Culled from the Factor Zoo

**Saketh Aleti,[a] Tim Bollerslev,[a,*] Mathias Siggaard[b]**

[a] Department of Economics, Duke University, Durham, North Carolina 27708; [b] CREATES, Aarhus University, 8210 Aarhus, Denmark
*Corresponding author
**Contact:** saketh.aleti@duke.edu, https://orcid.org/0000-0002-2880-7573 (SA); tim.bollerslev@duke.edu, https://orcid.org/0000-0003-2743-3087 (TB); masiggaard@gmail.com (MS)

**Received:** June 1, 2023
**Revised:** April 29, 2024
**Accepted:** June 11, 2024
**Published Online in Articles in Advance:** January 2, 2025

**https://doi.org/10.1287/mnsc.2023.01657**



**Abstract.** We provide strong empirical evidence for time-series predictability of the intraday return on the aggregate market portfolio by exploiting lagged high-frequency cross-sectional returns on the factor zoo. Our results rely on the use of modern machine-learning techniques to regularize the predictive regressions and help tame the signals stemming from the zoo together with techniques from financial econometrics to differentiate between continuous and theoretically nonpredictable discontinuous high-frequency price increments. Using the predictions from the model estimated for the aggregate market portfolio in the formulation of simple intraday trading strategies for a set of highly liquid ETFs results in sizeable out-of-sample Sharpe ratios and alphas after accounting for transaction costs. Further dissecting the abnormal intraday returns, we find that most of the superior performance may be traced to periods of high economic uncertainty and a few key factors related to tail risk and liquidity, pointing to slow-moving capital and the gradual incorporation of new information as the underlying mechanisms at work.

**History:** Accepted by Kay Giesecke, finance.
**Supplemental Material:** The online appendix and data files are available at https://doi.org/10.1287/mnsc.2023.01657.

**Keywords:** high-frequency data • market return predictability • factor zoo • machine learning • market timing • market frictions • slow-moving capital



## 1. Introduction

Is the return on the aggregate stock market predictable, and if so, is it possible to generate abnormal returns by trading on said predictions? These dual questions lie at the heart of the efficient market hypothesis (Fama 1970) and easily rank among the most studied questions in finance. The vast majority of the existing literature has focused on return predictability over daily or longer monthly and quarterly return horizons, utilizing the information in lagged market returns, or a relatively small number of predictor variables, including valuation ratios, dividend yields, interest rates, and volatility-type measures, reaching mixed conclusions (see, e.g., Pesaran and Timmermann 1995, Welch and Goyal 2008, and Rapach et al. 2010). Meanwhile, a smaller strand of the literature has sought to explain higher-frequency intraday return predictability patterns by various imperfections or market frictions (see, e.g., Chordia et al. 2005 and Heston et al. 2010).

Set against this background, armed with high-frequency intraday returns on an expansive set of more than 200 cross-sectional risk factors (Aleti 2023), together with estimation techniques adopted from machine learning to succinctly distill the pertinent information, we provide strong empirical evidence that the intraday return on the aggregate market portfolio is indeed predictable. Using the predictions from our main forecasting model in the formulation of a simple trading strategy for the SPY ETF results in an annualized intraday transaction-cost-adjusted Sharpe ratio of 1.37 along with an annualized Fama-French six-factor risk-adjusted alpha of 20.83%. By comparison, the intraday Sharpe ratio for the SPY over the same out-of-sample forecast period is just 0.09.[1] Consistent with the idea that the predictability exploited by our strategy may be attributed to the gradual incorporation of new information and slow-moving capital, we trace the accrual of most of the abnormal intraday returns to periods of high overall economic uncertainty and the outsized importance of a few key high-frequency factors related to liquidity and tail risk.

Our work may be seen as part of the recent literature that seeks to exploit the idea that the useful information in most high-dimensional data sets effectively lies in a lower-dimensional space.[2] In that regard, we rely on various machine-learning models as a way to uncover intraday time-series return predictability for the aggregate market portfolio. Importantly, however, rather than searching for predictive signals within the set of individual stock returns, as traditionally done in the literature, we search for predictive signals in the returns on the myriad of characteristic-sorted portfolios

established in the asset pricing literature. Hypothesizing that the predictive signals lie in this alternate subspace, we in turn rely on an array of machine-learning techniques, ranging from ridge, LASSO and elastic net to neural nets, random forests, and gradient boosted regression trees, to regularize and estimate our predictive models.

In addition, motivated by the fact that, in an arbitrage-free, continuous-time setting, price jumps are inherently unpredictable (see, e.g., Fontana et al. 2019), we deliberately filter out excessively large returns before the estimation of our prediction models, thereby allowing us to more keenly focus on the continuous, and potentially theoretically predictable, smooth high-frequency returns. To render such a decomposition meaningful, we purposely restrict our forecasting models to the active portion of the trading day for which reliable high-frequency data are available. Meanwhile, it is well established that most of the return on the market actually manifests during the overnight period (see, e.g., Hendershott et al. 2020), thus in effect leaving us with a more substantial challenge when it comes to generating economic gains from trading on the intraday return predictions only.[3]

Our basic idea of melding the literature on cross-sectional return predictability with that of time-series return predictability is akin to Dong et al. (2022), who similarly exploited a large set of asset pricing factors for uncovering market return predictability at the monthly level.[4] However, in contrast to their study, which rationalized the apparent predictability by time-varying risk premia, the intraday return predictability documented here suggests a very different economic explanation rooted in market frictions and the gradual incorporation of new information. In this regard, our analysis and empirical findings are perhaps most closely related to Huddleston et al. (2023), who similarly exploited a variety of machine-learning algorithms together with a large set of individual stock returns in their search for intraday market return predictability, and Chinco et al. (2019) who relied on LASSO for estimating intraday prediction models for individual stock returns based on the full cross-section of returns.

By comparison, our reliance on well-defined asset pricing factors as predictors, together with ETF-based trading strategies, allows us to more directly pinpoint what drives the transaction-cost-adjusted abnormal returns. In particular, relying on so-called SHAP values (Lundberg and Lee 2017) for additively decomposing the superior performance unequivocally points to factors based on various measures of turnover and liquidity as the key drivers of our superior Sharpe ratios and alphas. This, of course, is in line with the classic literature on cross-autocorrelations and the finding that the returns on larger firms tend to lead the returns on smaller firms (see, e.g., Lo and MacKinlay 1990, Chordia et al. 2007, and Hou 2007). It also echoes the past work by Chordia and Swaminathan (2002) and the finding that stocks with higher turnover tend to lead those with lower turnover and the more recent work by Caporin et al. (2017) and the finding that more actively traded stocks generally respond more quickly to important news.[5] In contrast to all of these earlier studies, however, which seek to explain lower-frequency lead-lag return relations between different groups of stocks by group-specific measures of trading activity, our prediction models for the intraday return on the market portfolio derive their explanatory power from aggregate measures of turnover, as embodied in the returns on the high-frequency systematic risk factors.

The finding that our trading strategies tend to perform the best when economic uncertainty and financial market volatility are high is consistent with market frictions being more acute in such periods, thereby limiting mispricing correction (see, e.g., Shleifer and Vishny 1997 and Gromb and Vayanos 2010).[6] Relatedly, the recent study by Avramova et al. (2023) found that the anomalous returns for individual stocks tend to concentrate in more difficult-to-arbitrage stocks and periods of high overall market volatility and low liquidity. It also echoes Farmer et al. (2023) and their findings that return predictability over longer return horizons tends to be confined to relatively short calendar time spans and pockets of predictability.

At a more technical level, our finding of intraday market return predictability also connects to a series of recent studies emphasizing possible short-lived violations of the benchmark Itô semimartingale model used for continuous-time modeling of returns and corresponding arbitrage opportunities associated with drift bursts (Christensen et al. 2022), pockets of extreme predictability (Andersen et al. 2023), or intermittent trading (Bandi et al. 2023). These studies, however, focus primarily on violations that occur at even higher intraday sampling frequencies than the 15-minute frequency on which our forecasting models and practical trading strategies are based. Similarly, the recent study by Aït-Sahalia et al. (2022) on the predictability of individual stock returns primarily emphasized the effect of market microstructure frictions and order imbalances over ultrashort millisecond time scales. That study also found that the return predictability uncovered by their LASSO-based estimation procedures essentially vanished over longer five-minute return intervals.

At a broader level, our paper is also part of the rapidly growing literature on the use of tools from machine learning to address a long list of empirical questions in finance, including the estimation and forecasting of risk premia. As a case in point, Gu et al. (2020) found that the use of machine-learning algorithms results in substantial performance gains over traditional regression-based approaches (see also Bianchi et al. 2021, Ke et al. 2021,

Fan et al. 2022, Guijarro-Ordonez et al. 2022, Leippold et al. 2022, and Jiang et al. 2023). In a similar vein, several studies have exploited machine-learning techniques to help better explain the cross-sectional variation in returns through the construction of improved factor models or the direct estimation of a more accurate stochastic discount factor (SDF) (see, e.g., Korsaye et al. 2019, Lettau and Pelger 2020, Giglio and Xiu 2021, Chen et al. 2023, Aleti and Bollerslev 2024, and Bryzgalova et al. 2024). Our paper aligns more closely with the former literature while focusing primarily on predictability induced by frictions rather than time-varying risk premia.

The remainder of the paper is organized as follows. Section 2 begins by summarizing the data underlying our empirical analyses. Section 3 discusses the formulation of our forecasting models and our implementation of the different machine-learning algorithms. Section 4 presents the resulting out-of-sample high-frequency forecasts for the market portfolio, whereas Section 5 discusses our main empirical findings pertaining to the practical implementation of the new forecasting models for trading the SPY ETF. Section 6 further shows that most of the predictability underlying the superior performance may be attributed to a few liquidity-related factors. Section 7 concludes the paper. Additional theoretical justification and empirical analyses are provided in an Online Supplemental Appendix.

## 2. Data

Our primary data set consists of high-frequency returns for the 272 portfolios constructed in Aleti (2023), spanning January 2, 1996, to December 31, 2020. The 272 portfolios comprise 218 characteristic-sorted portfolios capturing the many factors proposed in the asset pricing literature, 48 industry portfolios summarizing sector-specific movements in the stock market, and six additional factor portfolios consisting of the Fama-French five plus momentum.[7] To alleviate any concerns about market microstructure noise, we follow Aleti (2023) and sample the returns at a "coarse" 15-minute frequency for a total of 169,965 intraday return observations on each of the factors.

Table 1 reports various descriptive statistics for different subsets of the resulting intra-daily portfolio returns. The first column pertains to the Fama-French market portfolio, constructed as the value-weighted average of every common stock. It shows that the average intraday return on the market was indeed quite small, equaling just 1.07% per year. This, of course, is consistent with the above-mentioned extant literature and past work showing that most of the equity risk premium appears to have been earned during the overnight period (see, e.g., Cliff et al. 2008, Bogousslavsky 2021, and Bondarenko and Muravyev 2023).

**Table 1.** Descriptive Portfolio Statistics

| | Portfolios | | | |
|---|---|---|---|---|
| Averages (%) | Market | FF6 | Industry 48 | Factors |
| Return | 1.07 | 3.74 | 1.32 | 2.93 |
| Realized volatility | 14.42 | 8.59 | 19.62 | 8.87 |
| Relative jump measure | 2.76 | 4.76 | 5.21 | 4.69 |
| Autocorrelation | 3.56 | 6.11 | 3.28 | 4.00 |
| Correlation with market | 100.00 | 0.68 | 67.32 | −9.28 |

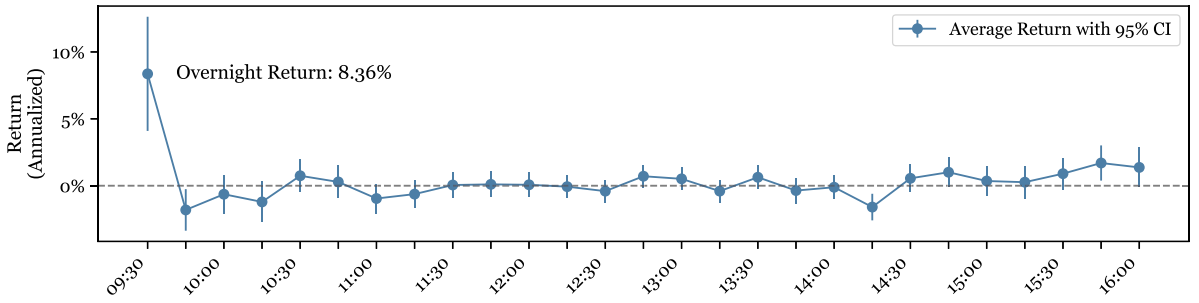
*Notes.* This table reports descriptive statistics for different subsets of the portfolios used in our empirical analyses. Market refers to the high-frequency Fama-French market portfolio. FF6 refers to the five workhorse factors from Fama and French (2015) plus the momentum factor. Industry 48 refers to the 48 industry portfolios, whereas Factors refers to the 218 characteristic-sorted portfolios. The numbers are calculated as the cross-sectional average of the time-series averages for each of the different subsets. The sample period is from 1996 to 2020. All of the measures are reported in annualized percentage form.

Further illustrating this point, Figure 1 plots the average market return earned across the day. As the figure shows, a disproportionate fraction of the market return is earned overnight, whereas the intraday returns are almost "flat" and close to zero throughout the active part of the trading day. Further underscoring the difficulty of predicting and earning a positive return during the day, the intraday market returns are also approximately serially uncorrelated, as evidenced by the average first-order autocorrelation coefficient of just 3.56% reported in Table 1.

The remaining summary statistics for the different subset portfolios, reported in Table 1, columns 2–4, tell a similar story. Not surprisingly, the 48 industry portfolios are most closely correlated with the market portfolio. Meanwhile, all of the high-frequency portfolio returns appear approximately serially uncorrelated. A cursory look at the lagged cross-autocorrelations with the return on the market portfolio also suggests little predictive information. In particular, looking at the 218 characteristic-sorted portfolios, the lagged first-, second-, and third-order cross-autocorrelations with the return on the market portfolio, corresponding to 15-, 30-, and 45-minute shifts in the factor returns, equal −0.75%, −0.43%, and −0.35%, respectively. Similar results obtain for the Fama-French workhorse factors and the 48 industry portfolios. Of course, as further explored in the Online Appendix, these averages may mask stronger correlations for a few of the individual portfolios, although the ex ante identification of such predictors remains a challenge.

The strong correlation between many of the factors in the zoo further complicates this. To illustrate, Figure 2 plots the full-sample high-frequency correlations between the 218 characteristic-sorted portfolios together with their lag-one and lag-two cross-autocorrelations.[8] Echoing earlier work based on more coarsely sampled daily and monthly factor returns, the figure evidences clear clusters within the factor zoo. In other words, the

**Figure 1.** (Color online) Intraday Fama-French Market Portfolio Returns




*Notes.* This figure reports the average intra-daily 15-minute returns on the Fama-French market portfolio from 1996 to 2020. The first return, labeled 09:30, corresponds to the average close-to-open return. The confidence intervals are computed using the corresponding realized volatility estimates.

relevant information in the high-dimensional factor zoo seemingly lies in a much lower dimensional space, hinting at the value of using modern machine-learning techniques in trying to identify the most useful predictors (see also the earlier related analysis in Lewellen et al. 2010). We turn next to a discussion of the specific prediction models and related estimation procedures that we rely on for doing so.

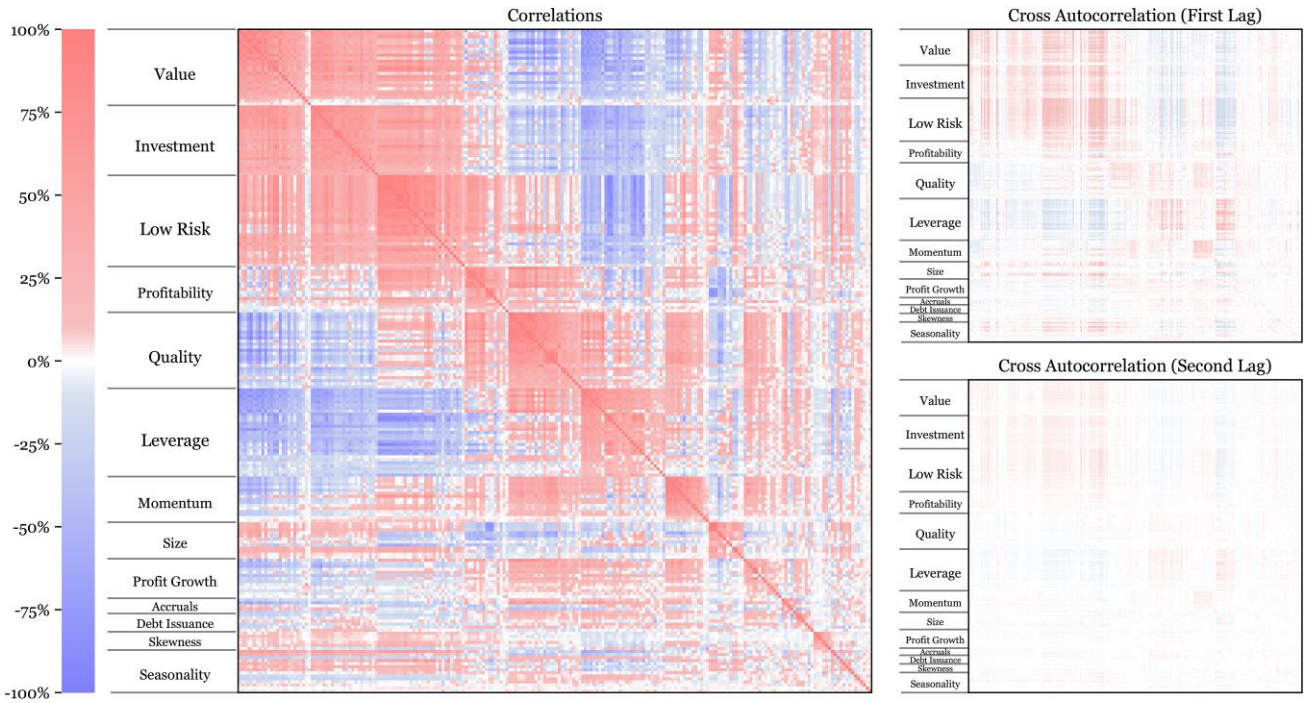
## 3. Methodology

We begin with a brief discussion of the formal setting that motivates our approach and the separation of the returns into jump and continuous components. We then describe the various forecasting specifications and machine-learning models that we rely on.

### 3.1. Continuous and Jump Price Dynamics

Let $P = (P_t)_{t\geq 0}$ denote the multivariate continuous-time log-price process representing both the market and our factor portfolios. To fix ideas, assume that the dynamics of the prices may be described by a standard jump-diffusion process, as commonly employed in the high-frequency financial econometrics literature (see, e.g.,

**Figure 2.** (Color online) High-Frequency Factor Correlations




*Notes.* The leftmost larger heatmap shows the contemporaneous full-sample pairwise cross-correlations between the 15-minute returns on each of the 218 characteristic-sorted portfolios. The upper and lower right-hand-side heatmaps show the corresponding lag-one (15-minute) and lag-two (30-minute) cross-autocorrelations, respectively.

Aït-Sahalia and Jacod 2014),

$$P_t = P_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + \sum_{s=1}^{N_t} J_s, \quad t \geq 0, \qquad (1)$$

where $\mu = (\mu_t)_{t\geq 0}$ denotes the drift, $\sigma = (\sigma)_{t\geq 0}$ is a stochastic volatility process, $W = (W_t)_{t\geq 0}$ is a standard Brownian motion, and $J = (J_s)_{s=1,\ldots,N_t}$ denotes a sequence of nonzero random variables of jump sizes with corresponding jump times $\tau = (\tau_s)_{s=1,\ldots,N_t}$. In the absence of arbitrage, the discontinuous jump component is inherently unpredictable.[9] By contrast, the drift and diffusive components, which capture the risk premia and the potential integration of slow-moving information, are both continuous and possibly jointly predictable. Hence, this naturally suggests treating the diffusive and jump components differently in the formulation of the forecasting models for future market returns.

In practice, of course, we do not observe the prices on a true continuous timescale, rendering the exact identification of the separate components impossible. The actual observed prices are also invariably contaminated by market microstructure noise stemming from bid-ask spreads, price discreteness, and other complications. As previously noted, to help ameliorate the impact of such noise, we follow the common practice in the high-frequency literature of deliberately sampling the prices at a coarse 15-minute frequency. To fix notation, let

$$r_{t,i} \equiv P_{t+i\Delta_n} - P_{t+(i-1)\Delta_n}, \qquad (2)$$

and denote the corresponding 15-minute returns, where $i \in \{1,\ldots,n\}$, $\Delta_n \equiv 1/n$, and $n = 26$ refers to the number of intraday returns per day. Armed with these returns, we rely on a standard truncation-based procedure, as further detailed in Appendix A.1, to disentangle the continuous and jump components of the factors.

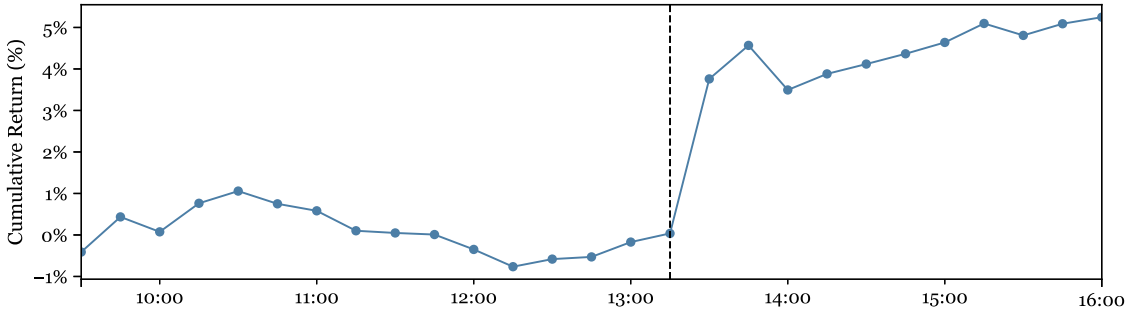
To help further intuit the idea and motivate the separate consideration of jumps in the formulation of our forecasting models, Figure 3 plots the cumulative 15-minute returns on the Fama-French market portfolio on January 3, 2001. On this day, the FOMC issued a surprise rate cut at exactly 13:15, evidently causing a sharp rise in the price, or a jump, at that exact time. As the figure further shows, there also appears to be a slight momentum in the return after the announcement. By contrast, neither the occurrence of the jump, nor its size and direction, appears to be predictable based on the earlier returns over the day. Underscoring this visual impression, the estimation of a simple AR(1) model based on the 15-minute intraday returns for the whole day results in an $R^2$ of just 0.69%. On the other hand, estimating an AR(1) model with the jump return from 13:15 to 13:30 omitted produces an $R^2$ of 8.96%. Interestingly, relying on the estimates from the former AR(1) model to predict the non-jump returns results in a negative $R^2$ of −0.19%. As such, this indirectly suggests that by including the jump return in the regression, we effectively "overfit" the unpredictable variation associated with the jump, thereby completely obscuring the smaller but potentially predictable return variation over the rest of the day. Therefore, in the results that follow, we will consider alternative forecasting specifications that exclude market jumps from the regressand.

### 3.2. Forecasting Model Specifications

Each of our specifications is based on autoregressive-type structures and the use of the lagged factor returns for predicting the future aggregate market portfolio return. Inspired by the Heterogeneous AutoRegressive (HAR) specification originally proposed by Corsi (2009) for modeling and forecasting time-varying volatility, we similarly rely on a cascade-type structure as a convenient way to impose a layer of sparsity in the dynamic dependencies.

Our baseline specification, which we will refer to as *All*, *All* in the sequel, uses the 15-minute market return $r^{mkt}_{t,i+1}$ as the regressand and all the lagged factor returns $X_{t,i} \equiv \cup_{j\in factors} X_{t,i,j}$ as the regressors. Owing to the HAR structure, each factor $j$ results in three separate regressors in the form of the lagged 15-minute, hourly, and daily portfolio returns, $X_{t,i,j} \equiv [r^j_{t,i}, r^j_{t,i|i-3}, r^j_{t-1}]$. Accordingly, our baseline specification that includes all of the

**Figure 3.** (Color online) Market Portfolio Return on January 3, 2001




*Notes.* This figure plots the cumulative 15-minute returns on the high-frequency Fama-French market portfolio on January 3, 2001. The vertical dashed line marks the time of the surprise FOMC rate cut announcement.

272 factor portfolios entails a $3\times272 = 816$ dimensional $\beta$ parameter to be estimated.

We further augment this already high-dimensional model by separating the portfolio returns into their continuous and jump components. The resulting *All, All+Jmp* specification in turn includes $X_{t,i} \cup X_{t,i}^{jmp}$ as regressors, where $X_{t,i}^{jmp} \equiv [r_{t,i}^{j,jmp}, r_{t,i|i-3}^{j,jmp}, r_{t-1}^{j,jmp}]$ is defined analogously to $X_{t,i}$ with the estimated jump returns for each of the portfolios in place of their total returns. It obviously encompasses the previous *All, All* specification but further allows for the possibility that the predictive content of the lagged jump factor returns may differ from that of the lagged continuous factor returns.

In addition, building on the idea that the true latent market jump returns should be inherently unpredictable, we consider two additional specifications in which we explicitly target only the continuous return on the market rather than the entire market return by replacing the market return regressand in the above two specifications with $r_{t,i+1}^{mkt,cts}$. If the market jump returns are indeed unpredictable, then explicitly excluding the jump component from the regressands would naturally enhance the accuracy of the resulting *Cts, All* and *Cts, All+Jmp* based predictions.[10] Additionally, as illustrated by the discussion of the single-day market returns depicted in Figure 3 and the estimated intraday AR(1) models, omitting the large jump returns from the regressands also helps prevent potential "overfitting" of a few especially large returns.

Altogether, our two most general *All, All+Jmp* and *Cts, All+Jmp* specifications both involve a total of $3\times272\times2 = 1,632$ parameters to be estimated. Even though we have a large number of high-frequency returns at our disposal for doing so, the unrestricted estimation by standard OLS of such high-dimensional models is simply not practically feasible. Instead, we rely on various machine-learning approaches to discipline the estimation and construct effective forecasts.

In addition to the factor-based model forecasts, we also consider a series of simple benchmarks in the form of AR(1), AR(4), and AR(26) models for the market portfolio return, thus incorporating the past 15-minute, one-hour, and one-day worth of market information, respectively. In parallel to the more general specifications discussed above, we consider four different versions of each of the AR models based on the decomposition of the market return into its continuous and jump components. Because none of these models exploit the information in the factor zoo, they serve as natural benchmarks for what can be learned from the market portfolio itself, implicitly speaking to a narrower form of weak form market efficiency. As a final benchmark, we also consider the sample mean return. In so doing, we deliberately rely on the ex post average return on the market to provide a more challenging, practically infeasible competitor.

### 3.3. Machine-Learning Algorithms

Following the rapidly growing financial machine-learning (ML) literature (see, e.g., Gu et al. 2020, Dong et al. 2022, Fan et al. 2022, Guijarro-Ordonez et al. 2022, Avramova et al. 2023, and Coulombe et al. 2023), along with the many additional references therein), we employ an array of ML algorithms to construct our forecasts. All of the different procedures that we use are now considered fairly standard in the finance and machine-learning literature.

More precisely, we employ the following linear procedures: OLS, Ridge, Lasso, Elastic Net (ENet), Adaptive Lasso (AdaLasso), Principal Components Regression (PCR), and Partial Least Squares (PLS). The nonlinear ML algorithms we consider are Feedforward Neural Networks (FNN), Random Forests (RF), and Gradient Boosted Regression Trees (GBRT). Following common practice, we also consider an equally weighted average of all of the above models sans OLS, thus collapsing all the forecasts into a single so-called ensemble forecast. In order to conserve space, we defer more detailed descriptions of each of the ML algorithms and their practical implementation, including the choice of hyperparameters, to the Online Appendix. For an extensive review of the same models and various applications thereof in the finance literature, we refer to the recent review by Kelly and Xiu (2023).

### 3.4. Practical Implementation

Each machine-learning model is defined by a set of hyperparameters and their main parameters. The hyperparameters, such as the L1 and L2 penalty pair for the elastic net, must be chosen before estimating the model itself to obtain the main parameters.

To this end, we rely on a commonly used data-driven approach from the ML literature known as a training/validation/testing split. We begin by splitting a contiguous region of the data into three subsets: a train, valid, and test period. Keeping the splits contiguous ensures that we do not violate the time-series structure of the data. We then start by using the data from the training period to fit the parameters for a given ML algorithm over a grid of hyperparameter sets. Using each of these fitted models, we next construct pseudo out-of-sample forecasts over the validation period for each hyperparameter set. We then rank the resulting forecasts by their MSE and define the hyperparameter set corresponding to the forecast with the lowest MSE as the "optimal" hyperparameter set for the model. We then reuse the fitted parameters for the ML algorithm previously estimated with these optimal hyperparameters to construct out-of-sample forecasts over the test period. To summarize, the fitted parameters are derived from the training period, and the optimal hyperparameters are derived from the validation period, with the forecasts constructed using these hyperparameters and the

parameters from the testing period, thereby ensuring that our forecasts are genuinely out of sample.

This rolling train/valid/test sampling-splitting procedure requires its own set of tuning parameters. Following common practice in the ML literature, we rely on an expanding window, using the first 80% of the window for training and the latter 20% for validation. Next, we fix the size of our out-of-sample window, or testing period, to one calendar year, refitting our models on an annual basis.[11] Lastly, we begin our out-of-sample analysis starting in 2004, producing new forecasts until the end of 2020, for a total out-of-sample evaluation period of 17 years, or 111,280 15-minute intraday observations. Incidentally, this out-of-sample period also closely corresponds to the "late post-decimalization period" (Huddleston et al. 2023), commensurate with a substantial increase in market liquidity, thereby allowing for more meaningful high-frequency-based analyses. A series of robustness checks concerning the tuning parameters used in our estimation and other normalizations is detailed in the Online Appendix.

## 4. Forecasting the Market Portfolio

We begin our discussion of the empirical results by considering the statistical accuracy of the forecasts for the market portfolio. Following common practice in the literature, we rely on the out-of-sample $R^2$

$$R^2 = 1 - \frac{\sum_{(t,i)\in \text{OOS}} (r^{mkt}_{t,i} - \hat{r}^{mkt}_{t,i})^2}{\sum_{(t,i)\in \text{OOS}} (r^{mkt}_{t,i})^2}, \tag{3}$$

for doing so, where $\hat{r}^{mkt}_{t,i}$ denotes the forecasts produced by a specific model. We purposely report the uncentered rather than the centered $R^2$ to allow for a direct comparison with other recent studies (see, e.g., Huddleston et al. 2023). The uncentered $R^2$ is also especially adept in the high-frequency context, where the drift that drives expected returns is generally negligible in magnitude compared with that of the martingale component.

To establish a benchmark for the $R^2$ to expect with 15-minute returns, consider the predictability bound of $R^2 \leq \gamma^2 Var(r^{mkt}_{t,i})$ derived by Ross (2004), where $\gamma$ refers to the risk aversion coefficient for the maximally risk-averse CRRA investor in the economy. Using the realized variance of our market portfolio returns as an estimate of the variance results in an upper bound on the 15-minute $R^2$ of just 0.00777% for a conservative choice of $\gamma = 5$. As such, this clearly shows that we should not expect $R^2$'s around, say, 1% to 5%, as sometimes reported in the literature focused on predictability over longer monthly horizons.[12] Alternatively, suppose that the continuous-time process for the market portfolio is known and may be described by a simple model with constant drift and volatility. As shown in the Online Appendix, in this situation even perfect prediction of the drift would result in only a 15-minute $R^2 = 0.00008\%$.

Turning to the actual forecast results obtained for each of our different specifications and ML tools, many of the 15-minute out-of-sample $R^2$s reported in Table 2 are substantially larger than the benchmarks discussed immediately above. In particular, the overall highest $R^2$ of 0.212% is obtained by the Ensemble *All, All* specification. Using split regressors and/or targeting the continuous component of the market portfolio returns only seemingly result in marginally lower Ensemble-based $R^2$'s. Putting the results further into perspective of the numerical bounds discussed above also suggests that the forecasting models are likely not predicting temporal variation in the equity risk premium.[13] Instead, the predictability seems more likely to arise from short-lived violations of the basic semimartingale assumption, possibly associated with the "slow" incorporation of new information.

To more formally assess the statistical significance of the reported differences in $R^2$'s, we also calculate Diebold and Mariano (1995) (DM) tests based on the out-of-sample MSEs. To conserve space, Table 3 reports only the $t$-statistics for comparing the four benchmark models to the different Ensemble specifications. As evidenced by the negative DM $t$-statistics, all four Ensemble specifications offer statistically significant improvements over the benchmarks. Meanwhile, testing the four different Ensemble specifications against one another, none of the corresponding DM $t$-statistics are significant at conventional levels. In short, the Ensemble forecasts beat all the benchmarks, whereas

**Table 2.** Out-of-Sample $R^2$s

| Regressors | All | | All + Jmp | |
|---|---|---|---|---|
| Regressand | All | Cts | All | Cts |
| AR1 | 0.036 | 0.051 | 0.051 | 0.097 |
| AR4 | 0.024 | 0.051 | −0.120 | −0.063 |
| AR26 | −0.015 | 0.017 | −0.410 | −0.382 |
| Linear | −5.123 | −5.012 | −12.408 | −11.868 |
| Ridge | 0.172 | 0.152 | 0.126 | 0.148 |
| Lasso | 0.151 | 0.119 | 0.165 | 0.113 |
| ENet | 0.153 | 0.127 | 0.164 | 0.119 |
| AdaLasso | 0.137 | 0.137 | 0.099 | 0.107 |
| PCR | 0.172 | 0.177 | 0.053 | 0.069 |
| PLS | 0.148 | 0.106 | 0.100 | 0.122 |
| FNN | 0.111 | 0.131 | −0.044 | 0.002 |
| RF | 0.118 | 0.007 | 0.052 | 0.090 |
| GBRT | −0.161 | 0.116 | 0.128 | −0.225 |
| Ensemble | 0.212 | 0.205 | 0.199 | 0.178 |

*Notes.* This table reports the out-of-sample $R^2$s for each of our forecasts for the 15-minute return on the Fama-French market portfolio from 2004 to 2020. Each column corresponds to one of our four specifications, whereas each row corresponds to a particular machine-learning algorithm. All of the values are reported in percentages.

**Table 3.** Diebold-Mariano Tests on MSE for the Ensemble Model

| Regressors | All | | All + Jmp | |
|---|---|---|---|---|
| Regressand | All | Cts | All | Cts |
| SampleMean | −4.315 | −5.019 | −4.275 | −4.576 |
| AR1 | −2.544 | −2.417 | −2.264 | −1.204 |
| AR4 | −2.398 | −2.125 | −3.586 | −2.723 |
| AR26 | −2.409 | −2.101 | −4.550 | −4.163 |

*Notes.* This table reports the Diebold-Mariano test-statistics for the Ensemble model vs. the four different benchmark models discussed in the text for forecasting the 15-minute return on the Fama-French market portfolio. Each column corresponds to one of our four specifications, whereas each row corresponds to a particular baseline model. Negative values indicate that the baseline model performs worse than the Ensemble-based forecast.

the equivalent statistical comparisons of the four different Ensemble specifications do not point to a clear significant winner.

In the Online Appendix, we also explore the sign classification accuracy of each of the models by comparing the signs of the forecasts to the signs of the realized returns. In so doing, we find that the specifications that target only the continuous component of the market returns perform statistically significantly better than those targeting the total returns. In other words, foreshadowing the trading performance results discussed in the next section, the directional forecasting ability of our models is better for the *Cts, All* and *Cts, All+Jmp* specifications.

## 5. Trading ETF Market Proxies

As emphasized by Campbell and Thompson (2008), even small and seemingly insignificant improvements in $R^2$'s can result in large improvements in Sharpe ratios and other performance measures. Kelly et al. (2024) similarly cautioned against evaluating high-dimensional prediction models developed by techniques from machine learning based on their $R^2$'s. Hence, we now turn to a more direct economic comparison based on the transactions-cost-adjusted performance of various trading strategies that directly utilize the forecasts from our different specifications.

### 5.1. Practical Implementation

The Fama-French market portfolio is not directly tradeable. Instead, we rely on various market proxies in the form of different ETFs. To conserve space, for our main empirical analyses discussed below, we focus on the highly liquid SPY ETF designed to track the S&P 500. The Online Appendix reports additional complementary results for the QQQ (Nasdaq-100), the IVV (S&P 500), and the IWM (Russell 2000), which are also highly liquid with low transaction costs and may similarly serve as market proxies.[14]

We deliberately rely on the forecasts for the Fama-French market portfolio for defining our trading signals rather than re-estimating new forecasts for the ETF returns. Our motivation for doing so is driven by the fact that the different ETF returns are contaminated by contract-specific microstructure noise and tracking error. Consequently, high-frequency forecasts constructed directly for specific ETF returns may end up targeting economically irrelevant variation, thereby resulting in inferior trading performance (see also the discussion in Cui et al. 2024). Indeed, as we show in the Online Appendix, trading on the direct forecasts for the ETFs typically generates returns and Sharpe ratios that are markedly worse than the corresponding numbers obtained by trading on the forecasts for the Fama-French market portfolio.[15] This approach, of course, also keeps the trading results tightly linked with the statistical forecasting results discussed in the previous section and substantially reduces the overall computational burden.

### 5.2. Trading Strategies

To keep our analysis concise and straightforward, we consider a limited set of four trading strategies. As a benchmark, we also consider the trivial strategy of simply buying and holding the market.

Our first strategy seeks to time the market based solely on the signs of the forecasts. That is, the strategy responds to a positive (negative) forecast by going long (short) in the market. Let $w_{t,i}$ denote the weight for the market portfolio on day $t$ and intraday time-interval $i$. The *Sign* strategy is then simply defined by

$$[Sign] \qquad w_{t,i+1} = \begin{cases} 1 & \text{if } \hat{r}^{mkt}_{t,i+1} \geq 0, \\ -1 & \text{if } \hat{r}^{mkt}_{t,i+1} < 0, \end{cases} \tag{4}$$

where $\hat{r}^{mkt}_{t,i+1}$ denotes the forecast of the market return by any one of the forecasting models.

The *Sign* strategy effectively treats short and long positions symmetrically. However, short-selling is generally more costly to implement. Moreover, legal constraints and charters also impede many institutional investors from short-selling; see, for example, the discussion in Pontiff (1996) and Shleifer and Vishny (1997). Accordingly, our *Positive* strategy modifies the sign-based strategy by simply zeroing out the market exposure whenever the forecasts are negative:

$$[Positive] \qquad w_{t,i+1} = \begin{cases} 1 & \text{if } \hat{r}^{mkt}_{t,i+1} \geq 0, \\ 0 & \text{if } \hat{r}^{mkt}_{t,i+1} < 0. \end{cases} \tag{5}$$

Although this strategy helps circumvent the higher costs typically associated with short-selling, it still

rebalances the entire portfolio whenever the sign of $\hat{r}^{mkt}_{t,i+1}$ changes. As such, it can obviously be quite costly to implement.

Several procedures have been proposed in the literature to help mitigate trading costs. However, these procedures are typically geared toward the specific setting and strategy at hand, and they can be difficult to realistically implement in other situations. Instead, motivated by Gârleanu and Pedersen (2013, 2016) and the idea of only trading partially "towards the target," our third and fourth strategies trigger rebalance only in response to sufficiently strong predictive signals, as measured by the lagged half-spread

$$[S\text{-}Sign] \qquad w_{t,i+1} = \begin{cases} 1 & \text{if } \hat{r}^{mkt}_{t,i+1} > Spread_{t,i}/2, \\ -1 & \text{if } \hat{r}^{mkt}_{t,i+1} < Spread_{t,i}/2, \\ w_{t,i} & \text{otherwise} \end{cases} \tag{6}$$

and

$$[S\text{-}Positive] \qquad w_{t,i+1} = \begin{cases} 1 & \text{if } \hat{r}^{mkt}_{t,i+1} > Spread_{t,i}/2, \\ 0 & \text{if } \hat{r}^{mkt}_{t,i+1} < Spread_{t,i}/2, \\ w_{t,i} & \text{otherwise.} \end{cases} \tag{7}$$

As discussed in more detail in the Online Appendix, the *S-Sign* strategy represents the optimal solution to a static optimization problem, where an agent aims to maximize expected returns while facing transaction costs from paying the half-spread on each trade. This same argument holds true for the *S-Positive* strategy if we further assume that the trader is constrained to long-only positions only. Both of these strategies are also closely in line with the strategies previously employed by Huddleston et al. (2023) and Chinco et al. (2019), although in contrast to these earlier studies we deliberately choose to maintain the previously held position if the signal is not "strong" enough, rather than putting the weight to zero.

Lastly, whenever the *Positive* and *S-Positive* strategies do imply a zero weight for the market portfolio, we simply associate a zero return with that time interval. Of course, allowing for alternative cash investments like Treasury Bills would enhance the calculated returns.

### 5.3. Economic Performance Measures

Having defined the portfolio weights, we assess the trading performance by computing the resulting average intraday returns and Sharpe ratios over our out-of-sample period. That is,

$$\text{AvgRet}^{F,m,k} = \frac{1}{T_{OOS}} \sum_{t,i} w^k_{t,i} r^F_{t,i+1}, \tag{8}$$

and

$$Sharpe^{F,m,k} = \underbrace{\left( \frac{1}{T_{OOS}} \sum_{t,i} (w^k_{t,i} r^F_{t,i+1} - r^f_{t,i+1}) \right)}_{\text{AvgExcRet}^{F,m,k}} \Bigg/ \underbrace{\left( \sqrt{\frac{1}{T_{OOS}} \left( \sum_{t,i} (w^k_{t,i} r^F_{t,i+1})^2 \right)} \right)}_{\text{AvgVol}^{F,m,k}}, \tag{9}$$

where $T_{OOS}$ denotes the number of years in the out-of-sample period, $m$ refers to a particular model (specification + machine-learning algorithm), $k$ refers to a strategy, and $F$ refers to the return on the specific market proxy. We proxy for the intraday risk-free rate $r^f_{t,i+1}$ using the daily Treasury bill returns obtained from Ken French's website, proportionally distributing the daily returns over the intraday and overnight intervals. In line with the underlying forecasting models, both the above-defined average return and Sharpe ratio formulas are based exclusively on intraday returns. However, to allow for a more direct comparison with existing work, we also subsequently consider overnight-inclusive versions of both measures.

To account for the transaction costs associated with the practical implementation, we begin by defining

$$Turn^{m,k}_{t,i} \equiv |w^{m,k}_{t,i} - w^{m,k}_{t,i-1}|. \tag{10}$$

We then calculate the (transaction cost) adjusted returns as

$$AdjAvgRet^{F,m,k}_{t,i} \equiv r^{F,m,k}_{t,i} - \underbrace{\text{Turn}^{m,k}_{t,i} \times \frac{\tau^F_{t,i}}{2}}_{\text{trading cost}} - \underbrace{\mathbb{1}[w^{m,k}_{t,i} < 0] \times v^F_{t,i}}_{\text{shorting cost}}, \tag{11}$$

where $\tau^F_{t,i}$ denotes the corresponding spread, and $v^F_{t,i}$ is the borrow fee for shorting. Correspondingly, our adjusted Sharpe ratio is defined by

$$AdjSharpe^{F,m,k} \equiv \text{AvgExcAdjRet}^{F,m,k} / \text{AvgVol}^{F,m,k}, \tag{12}$$

where the average excess adjusted return is computed by subtracting the risk-free rate from $\text{AdjAvgRet}^{F,m,k}_{t,i}$, analogous to Equation (9).

In the calculations reported below, we measure $\tau^F_{t,i}$ by the "Time-Weighted Quoted Spread" provided by the NYSE TAQ database. This implicitly assumes that rebalancing always incurs half the spread in transaction costs. Lower costs may likely be attained by some investors, thus further enhancing the performance of our strategies. Additionally, we rely on a very conservative fixed annualized value of 2% for the borrow rate, or shorting cost $v^F_{t,i}$.[16]

Although our trading strategies and our assessment of their performance are exclusively based on the intraday returns, the above transaction cost calculations invariably depend on the positions over the first and last 15-minute intervals of the trading day and, in turn, the treatment of the overnight period. Motivated by the finding that much of the market return has historically been earned overnight, in the calculations reported on below, we fix the weight for the overnight period for all of our strategies to be one.[17] Accordingly, the weight at the start of each day is always set to unity, or $w^{m,k}_{-1,t} \equiv 1$.

### 5.4. General Results

Motivated by the forecast results for the market portfolio discussed in Section 4, we begin our discussion of the ETF-based trading strategies with a broad overview of the performance measures achieved by the forecasts from the various Ensemble specifications to trade the SPY. In addition to the above-defined performance measures, Table 4 also reports the Fama-French six-factor alphas, computed from regressing the high-frequency returns for each of the strategies on the high-frequency factor returns.[18] For completeness, we report separate results based exclusively on the intraday returns (left part of the table) and the results that are obtained by also incorporating the overnight returns (right part of the table).

Looking at the results, the *Cts, All* specification, which uses the continuous component of the returns as the regressand and the total lagged returns as the regressors, seemingly outperforms all of the other approaches in terms of the average adjusted returns, the adjusted Sharpe ratios, and the FF6 alphas, regardless of the specific strategy used to translate the forecasts into trades. Moreover, for both the *All* and *All+Jmp* regressor specifications, using the continuous returns as the regressand results in the best performance according to all the different measures. As such, the results in Table 4 directly corroborate our earlier conjecture that targeting only the potentially predictable return variation does indeed result in more economically meaningful out-of-sample forecasts.

The table also reveals that the *S-Positive* and *S-Sign* strategies, designed to help mitigate transaction costs, generally perform better than their simpler counterparts. Further underscoring the ability of these two strategies to better manage the cost of trading, the relative performance of the same two strategies when ignoring transaction costs is markedly worse than their simpler counterparts. It is also noteworthy that allowing for short positions results in substantial economic gains, as manifest in the intraday adjusted Sharpe ratio for the *Cts, All* specification rising from 0.99 to 1.37 when shifting from the *S-Positve* to the *S-Sign* strategy

**Table 4.** Trading Performance of the Ensemble Model

| | Regressors | | | |
|---|---|---|---|---|
| | All | | All + Jmp | |
| Regressand | All | Cts | All | Cts |
| **Adjusted intraday performance** | | | | |
| Panel A1: Return (%) | | | | |
| Positive | 7.24 | 8.67 | 6.12 | 6.90 |
| Sign | 12.09 | 15.02 | 9.88 | 11.50 |
| S-Positive | 9.65 | 11.13 | 9.50 | 8.96 |
| S-Sign | 16.86 | 19.89 | 16.61 | 15.56 |
| Panel B1: Sharpe | | | | |
| Positive | 0.64 | 0.75 | 0.54 | 0.58 |
| Sign | 0.83 | 1.03 | 0.67 | 0.78 |
| S-Positive | 0.89 | 0.99 | 0.89 | 0.79 |
| S-Sign | 1.16 | 1.37 | 1.14 | 1.07 |
| Panel C1: Alpha (%) | | | | |
| Positive | 6.83 | 8.07 | 5.88 | 6.42 |
| Sign | 13.19 | 15.74 | 11.32 | 12.46 |
| S-Positive | 9.29 | 10.64 | 9.26 | 8.53 |
| S-Sign | 18.06 | 20.83 | 18.04 | 16.62 |
| Panel D1: Alpha *t*-statistic | | | | |
| Positive | 3.63 | 4.32 | 3.09 | 3.41 |
| Sign | 3.56 | 4.28 | 3.01 | 3.37 |
| S-Positive | 4.96 | 5.70 | 4.86 | 4.55 |
| S-Sign | 4.88 | 5.68 | 4.78 | 4.49 |
| **Adjusted overnight-inclusive performance** | | | | |
| Panel A2: Return (%) | | | | |
| Positive | 14.52 | 15.97 | 13.42 | 14.22 |
| Sign | 18.94 | 21.93 | 16.78 | 18.45 |
| S-Positive | 16.89 | 18.40 | 16.77 | 16.24 |
| S-Sign | 23.64 | 26.73 | 23.45 | 22.43 |
| Panel B2: Sharpe | | | | |
| Positive | 0.89 | 0.97 | 0.83 | 0.86 |
| Sign | 1.02 | 1.18 | 0.90 | 0.99 |
| S-Positive | 1.06 | 1.14 | 1.06 | 1.00 |
| S-Sign | 1.28 | 1.44 | 1.26 | 1.20 |
| Panel C2: Alpha (%) | | | | |
| Positive | 9.48 | 10.48 | 8.28 | 8.67 |
| Sign | 18.32 | 20.40 | 15.95 | 16.79 |
| S-Positive | 12.04 | 13.19 | 11.72 | 10.74 |
| S-Sign | 23.38 | 25.76 | 22.80 | 20.88 |
| Panel D2: Alpha *t*-statistic | | | | |
| Positive | 4.57 | 5.13 | 3.93 | 4.22 |
| Sign | 4.54 | 5.15 | 3.89 | 4.22 |
| S-Positive | 5.80 | 6.43 | 5.54 | 5.22 |
| S-Sign | 5.78 | 6.48 | 5.51 | 5.21 |

*Notes.* This table reports the out-of-sample, adjusted for transactions costs annualized returns, Sharpe ratios, and alphas obtained by trading the SPY ETF based on our Ensemble model forecasts. Panels D1 and D2 report the *t*-statistics for the adjusted alphas. The sample period spans 2004–2020. The intraday performance measures reported in the left panels, labeled 1, exclude the 16:00–9:30 overnight returns, whereas the overnight-inclusive performance measures reported in the right panels, labeled 2, include the overnight returns. The alphas are estimated using yearly regressions against the FF6 factors.

and the corresponding alphas more than doubling. Allowing for short positions also nearly doubles the returns, suggesting that the increase in alpha is indeed driven by genuine predictability rather than simply reducing the factor exposures through negative weights. Consistent with the idea that much of the return on the market is earned overnight, comparing the results in the left and right panels, the inclusion of the overnight returns naturally also increases all of the different performance measures.

Next, to better understand the performance of our suite of machine-learning models, Table 5 reports the intraday adjusted Sharpe ratios, along with the Sharpe ratios for our simple benchmark models. As the table shows, the benchmark models, together with traditional linear OLS-based estimation of the different specifications, all perform quite poorly under the *Positive* and *Sign* trading strategies. The use of the *S-Positive* and *S-Sign* strategies to translate the forecasts into trades improves the performance. However, the results for the benchmark models are markedly worse than the results for all the machine-learning models. The forecasts from Partial Least Squares (PLS), Feedforward Neural Networks (FFN), and Gradient Boosted Regression Trees (GRBT), in particular, when combined with the *S-Sign* strategy, all achieve Sharpe ratios in excess of 1.20. Interestingly, the Ensemble average forecasts attain the overall highest Sharpe ratios under both the *Positive* and *Sign* strategies, and very close to the highest ratios under the *S-Positive* and *S-Sign* strategies, thus effectively obviating the difficult task of trying to identify a single best model.[19] We turn next to a more detailed analysis of when these superior returns are earned.

**Table 5.** Intraday Adjusted Sharpe Ratios for (Cts, All) Specifications

| | Strategy | | | |
|---|---|---|---|---|
| Model | Positive | Sign | S-Positive | S-Sign |
| AR1 | −1.42 | −2.03 | 0.18 | 0.16 |
| AR4 | −0.18 | −0.33 | 0.44 | 0.51 |
| AR26 | −0.41 | −0.63 | 0.38 | 0.44 |
| Linear | 0.06 | −0.00 | 0.12 | 0.09 |
| Ridge | 0.49 | 0.62 | 0.61 | 0.79 |
| Lasso | 0.44 | 0.54 | 0.68 | 0.88 |
| ENet | 0.41 | 0.50 | 0.67 | 0.86 |
| AdaLasso | 0.44 | 0.52 | 0.56 | 0.68 |
| PCR | 0.47 | 0.58 | 0.75 | 0.95 |
| PLS | 0.44 | 0.54 | 0.93 | 1.21 |
| FNN | 0.77 | 1.06 | 0.90 | 1.25 |
| RF | 0.70 | 0.99 | 0.42 | 0.48 |
| GBRT | 0.66 | 0.92 | 0.99 | 1.44 |
| Ensemble | 0.79 | 1.09 | 0.98 | 1.36 |

*Notes.* The table reports the 2004–2020 out-of-sample adjusted Sharpe ratios for trading the SPY ETF based on the forecasts for each of our machine-learning algorithms under the *Cts, All* specification. All of the values are annualized.

### 5.5. Intraday and Interday Patterns

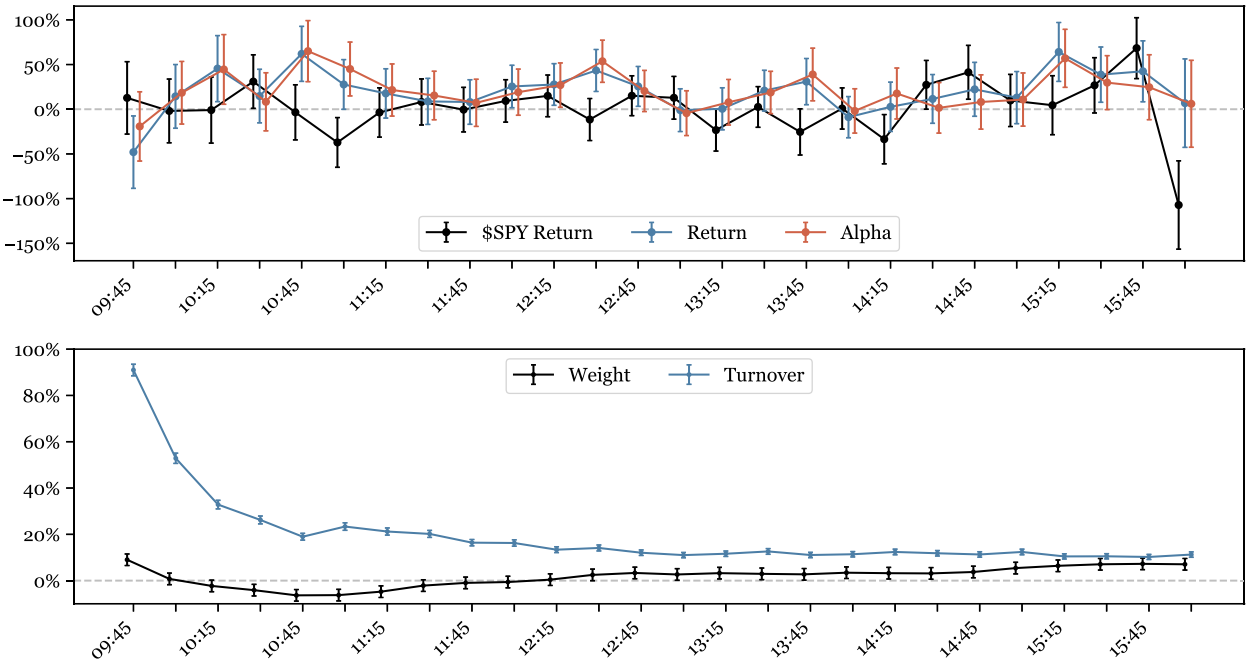
Motivated by the results in the previous section, to streamline the presentation, we will focus our more in-depth analyses on the Ensemble model and the *Cts, All* specification, along with the *S-Sign* strategy for trading the SPY. We begin by characterizing the intra-daily variation in the performance, followed by a look at the longer-run interday temporal variation and the relation to underlying economic conditions.

The top panel in Figure 4 reports the average returns over each of the 15-minute intervals of the trading day together with the average returns and FF6 alphas earned from our trading strategy together with the corresponding 90% confidence intervals. The average returns for the SPY, not surprisingly, closely mirror the on-average close to zero and flat average intraday returns for the Fama-French market portfolio previously shown in Figure 1. The most notable systematic differences from zero seemingly occur around 14:15–14:45, probably as a result of FOMC announcements at that time, and toward the end of the trading day, likely because of increased selling pressure at that time.[20] The average intraday returns for our main trading strategy, in contrast, show some notable variation, especially around 10:45, coincident with the time at which many macroeconomic data are announced (see, e.g., Andersen et al. 2003), and around the 14:15–14:45 FOMC announcement time. The alphas display a similar pattern, suggesting that our strategy may be in part exploited a delayed reaction to economic information. We also find slightly larger returns during the middle of the day, mirroring Huddleston et al. (2023), who found an inverse-U pattern in the degree of predictability across the day, and their trading strategies correspondingly earning the highest abnormal returns in the middle of the day and performing worse than the SPY at the end of the day.

Looking at the bottom panel in Figure 4 reveals that the proportion invested in the SPY remains fairly stable across the day, with a slight dip at around 10:45. This again is most likely explained by the systematic release of macroeconomic news at that time. Meanwhile, turnover peaks at the beginning of the day and proceeds to decline throughout the rest of the day.[21] The sharp decline observed at the start of the day is readily explained by the fact that our overnight weight is fixed at 100%, thus resulting in high turnover as the strategy readjusts to the on-average much lower optimal intraday weights. More broadly, the overall patterns in both the weights and the turnover also mimic the well-documented U-shaped pattern in intraday volatility (see, e.g., Andersen and Bollerslev 1997), suggesting that there is more predictable variation in the high-frequency returns to be exploited when volatility is high.

Turning to the longer-run interday dynamic dependencies, the top panel in Figure 5 shows the cumulative

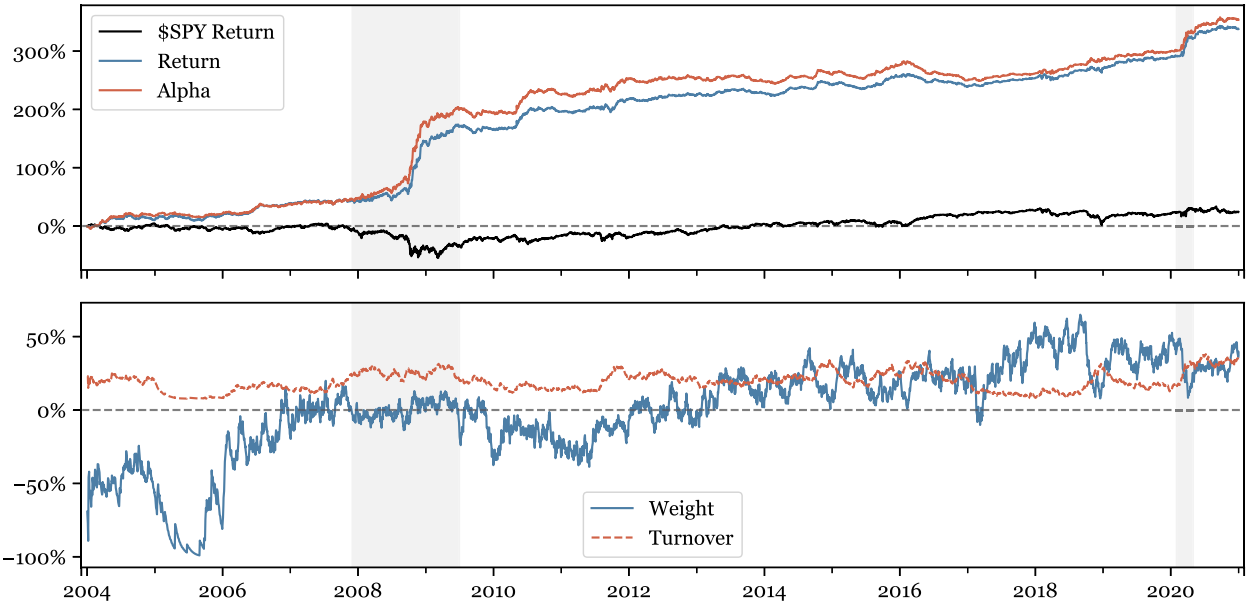
**Figure 4.** (Color online) Average Intraday Returns and Turnover




*Notes.* The first subplot shows the average annualized returns over each 15-minute interval within the day for the SPY ETF, the *(Cts, All)-(Ensemble)-(S-Sign)* strategy for trading the SPY, and the corresponding FF6 alphas. The bars refer to 90% confidence intervals. The bottom subplot shows the average portfolio turnover and weight across the day. Both plots are based on returns and portfolio weights for the 2004–2020 out-of-sample period. The markers are staggered for visibility.

daily returns over the full 2004–2020 out-of-sample period. The cumulative return on the SPY at the end of the sample is less than 50%, consistent with the on-average very low intraday return on the market and the SPY, discussed earlier. By comparison, our main strategy delivers a cumulative return of more than

**Figure 5.** (Color online) Cumulative Interday Returns and Turnover—(Cts, All)-(Ensemble)-(S-Sign)




*Notes.* The first subplot reports the cumulative intradaily returns for the SPY ETF, the *(Cts, All)-(Ensemble)-(S-Sign)* strategy for trading the SPY, and the corresponding FF6 alphas. Both the trading strategy returns and alphas are adjusted for transaction costs. The second subplot shows the average monthly weights for the SPY and turnover for our main trading strategy. The average weights and turnover are computed on a daily basis and further smoothed with a 21-trading-day EWMA. Shaded regions mark NBER-defined recessions.

300% by the end of the sample. Interestingly, it appears that the largest gains manifest during recessions, exactly when the market tends to perform poorly. However, the returns are not simply coming from exiting the market during bad times. In fact, using the long-only *S-Positive* strategy instead of the *S-Sign* strategy still delivers markedly higher returns than the SPY itself over the same recession periods. Put differently, our trading strategy does not simply exit the market during bad economic times but instead actively trades to exploit market dislocations that occur at such times. The large returns and alphas observed at the beginning of 2020 and the time of the COVID-induced recession also support this conjecture.

Further elaborating on this theme, the bottom panel in Figure 5 shows the rolling monthly weight on the SPY and the turnover of the main trading strategy. The average turnover hovers around 20% for most of the sample, suggesting a moderate but not excessive amount of portfolio rebalancing. On the other hand, the average weight tends to fluctuate quite substantially throughout the sample, again underscoring the previous point that the strategy actively seeks to find alpha via trading at all times.

### 5.6. Performance vs. Economic Conditions

To help further assess the relationship between overall economic conditions and the performance of our main strategy, we consider its relation to the monthly uncertainty measures proposed by Jurado et al. (2015). Lagging the measures by a month to avoid any look-ahead bias, we regress the total monthly intraday returns and the monthly FF6 alphas on each of the different measures. For ease of interpretation, we further normalize each of the uncertainty measures by their sample standard deviation. As a reference, we also include the same regressions involving the monthly intraday returns for the SPY. All of the regressions span the full 2004–2020 out-of-sample period for a total of 204 monthly observations.

The resulting estimates, reported in Table 6, show that the performance of the strategy is indeed strongly positively related to all three uncertainty measures. A one-standard-deviation increase in "Financial Uncertainty" or "Macro Uncertainty" increases the monthly return by 1% to 2%. The "Financial Uncertainty" and "Macro Uncertainty" measures also both do a remarkable job in terms of explaining the monthly cumulative alphas, with $R^2$s of 23.9% and 19.3%, respectively. As such, this again supports the idea that our trading strategy is able to exploit dislocations that tend to occur during highly uncertain, or bad, economic times. By contrast, all of the estimated slope coefficients for the SPY returns are insignificant and the $R^2$'s close to zero. This lack of significance for the SPY, of course, is also in line with the vast market predictability literature referenced earlier, emphasizing the difficulty of reliably predicting aggregate stock market returns; see, for example, Welch and Goyal (2008).

To provide an even more detailed look at when the strategy performs well, Table 7 reports the results from finer daily regressions against the lagged CBOE VIX index and the lagged realized volatility of the SPY. To be comparable with the VIX, the realized volatility of the SPY is computed using the summation of the squared 15-minute intraday and overnight returns, with both measures rescaled to a daily magnitude. Looking at the results reported in the second column, a 1pp increase in the daily realized volatility is associated with an impressive 17.33pp increase in the annualized intraday return for our main strategy. The results obtained for the alphas are similar, with the equivalent annualized increase equal to 18.38pp. The corresponding daily regression $R^2$'s of 2.4% and 2.3%, respectively, are also quite high. By comparison, the regression coefficient for the SPY returns, reported in the first column, is insignificant, and the $R^2$ is also effectively zero.

Looking at the second set of regressions, the results for the SPY appear reminiscent of Bollerslev et al. (2009) and the finding that the variance risk premium, defined

**Table 6.** Trading Performance vs. Monthly Uncertainty Measures

| | SPY | Return | Alpha | SPY | Return | Alpha | SPY | Return | Alpha |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | 1.63 | −7.98*** | −9.77*** | 1.47 | −7.99** | −9.61** | 0.60 | −5.06* | −5.92* |
| | (1.18) | (−2.82) | (−2.80) | (0.94) | (−2.30) | (−2.22) | (0.60) | (−1.66) | (−1.68) |
| Financial uncertainty | −0.28 | 1.81*** | 2.16*** | | | | | | |
| | (−1.00) | (3.28) | (3.14) | | | | | | |
| Macro uncertainty | | | | −0.22 | 1.55*** | 1.82** | | | |
| | | | | (−0.80) | (2.61) | (2.46) | | | |
| Real uncertainty | | | | | | | −0.08 | 1.09** | 1.25* |
| | | | | | | | (−0.44) | (1.99) | (1.95) |
| $R^2$ | 0.011 | 0.209 | 0.239 | 0.008 | 0.171 | 0.193 | 0.001 | 0.110 | 0.114 |

*Notes.* This table reports the results from regressions of the monthly returns of the SPY ETF, the *(Cts, All)-(Ensemble)-(S-Sign)* strategy for trading the SPY, and its corresponding FF6 alphas. All of the regressions are based on 204 monthly observations. The three uncertainty measures are obtained from Jurado et al. (2015) and are normalized by their sample standard deviations. Robust *t*-statistics are reported in parentheses, with *significance at the 90% level, **significance at the 95% level, and ***significance at the 99% level.

**Table 7.** Trading Performance vs. Daily Volatility Measures

| | SPY | Return | Alpha | SPY | Return | Alpha |
|---|---|---|---|---|---|---|
| Intercept | −0.00 | −0.07*** | −0.08** | −0.10** | −0.24*** | −0.22*** |
| | (−0.13) | (−2.60) | (−2.54) | (−2.30) | (−3.58) | (−3.11) |
| SPY realized volatility | 1.10 | 17.33*** | 18.38*** | −8.55 | 0.96 | 4.29 |
| | (0.26) | (4.69) | (4.50) | (−1.04) | (0.15) | (0.66) |
| VIX | | | | 15.51* | 26.33*** | 22.66** |
| | | | | (1.86) | (2.76) | (2.27) |
| $R^2$ | 0.000 | 0.024 | 0.023 | 0.003 | 0.034 | 0.030 |

*Notes.* This table reports the results from regressions of the daily returns of the SPY ETF, the *(Cts, All)-(Ensemble)-(S-Sign)* strategy for trading the SPY, and its corresponding FF6 alphas. All of the regressions are based on 4,282 daily observations. The realized variance (*RV*) of the SPY is computed using 15-minute intraday and overnight returns. The VIX index is obtained from the CBOE website and is rescaled to match the daily realized volatility estimates. Both regressors are lagged by one day. Robust *t*-statistics are reported in parentheses, with *significance at the 90% level, **significance at the 95% level, and ***significance at the 99% level.

as the difference between the two volatility measures, is able to predict the return on the market. However, the estimated coefficients are still insignificant, and they also have the "wrong" sign vis-a-vis the results in the extant literature based on the returns over longer monthly and quarterly horizons that include the overnight returns. Meanwhile, the regressions involving the intraday returns on our *(Cts, All)-(Ensemble)-(S-Sign)* trading strategy and its FF6 alphas again point to significantly better performance when the VIX is high, suggesting that implied volatility, or "investor fears" (Bollerslev and Todorov 2011), rather than realized volatility is most closely associated with its superior performance.

## 6. Feature Importance and Model Interpretability

To help clarify which of the factors drive this performance, it is instructive to compute so-called SHAP (SHapley Additive exPlanations) values (Lundberg and Lee 2017). SHAP values are now routinely used in the machine learning literature as a way to succinctly quantify variable importance in high-dimensional settings. They have also recently been used in the context of finance applications (see, e.g., Erel et al. 2021, Moehle et al. 2021, and Coulombe et al. 2023) and more general economic time-series forecasting applications (see, e.g., Borup et al. 2022). Put simply, SHAP values split a specific performance measure for a given model into the marginal contributions stemming from each of the variables included in said model; unlike certain other variable importance measures, the SHAP values are additive; the values for each regressor variable ultimately sum to the performance of the baseline model. Precise details concerning our actual calculation of the SHAP values are provided in Appendix A.2.

### 6.1. SHAP Values

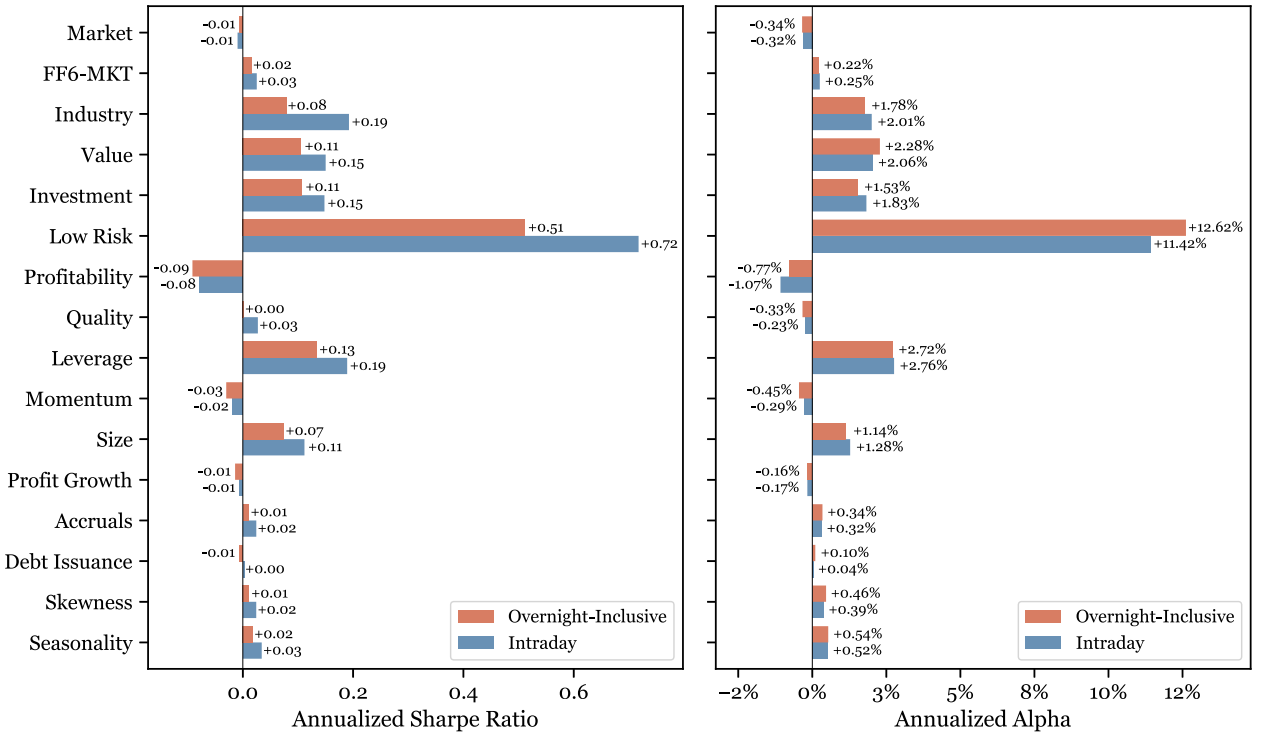
We begin with a birds-eye view of the variable importance for the different factor clusters previously illustrated in Figure 2 and their contributions to the Sharpe ratios and FF6 alphas. In addition to the 14 different factor clusters, we also include the Fama-French market portfolio, the six Fama-French portfolios, excluding the market portfolio, and the 48 industry portfolios, as additional sets of clusters. For completeness, we report the results for both the intraday exclusive and the overnight inclusive Sharpe ratios and FF6 alphas.

Looking at the results in Figure 6, quite remarkably the vast majority of the superior performance, both in terms of Sharpe ratios and alphas, may be attributed to a single cluster, namely the Low Risk cluster. Much smaller gains also arise from the Industry, Value, Investment, and Leverage clusters, whereas all of the other clusters, as well as the market and the Fama-French factors, appear to contribute very little, if anything, to the performance.

To further help identify the specific factors within the different clusters that drive the performance, Figure 7 reports the SHAP values for the top 10 factors by their intraday Sharpe ratios.[22] The names of the factors are adopted directly from Chen and Zimmermann (2022) and Jensen et al. (2023), as indicated by the "CZ" and "JKP" labeling in the figure. Consistent with the previous Figure 6, four out of the top five factors come from the Low Risk cluster, with the remaining factor stemming from the Leverage cluster. Interestingly, the SHAP values appear to be fairly evenly spread across several factors. This finding aligns with the use of models such as Ridge and PLS, which implicitly assume that the predictive signals are relatively evenly distributed among many predictor variables. By contrast, sparse models like Lasso naturally result in sparser SHAP values. The results depicted in Figure 7 pertaining to the Ensemble model naturally exhibit both features.

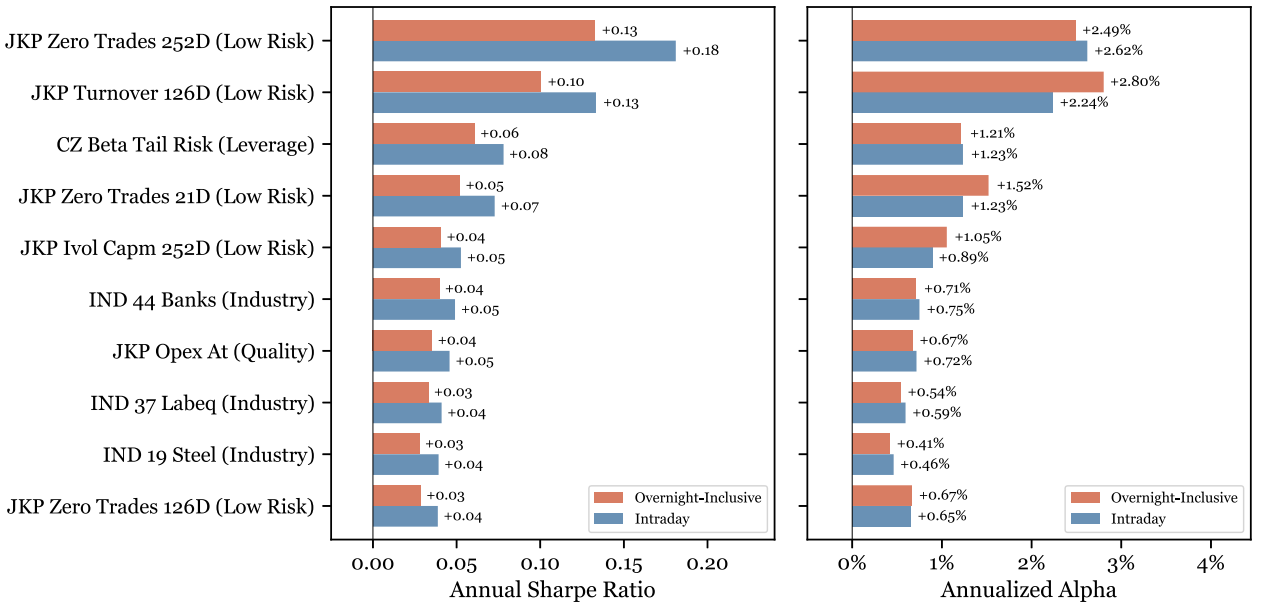
Digging further into the specific factors that matter the most, both the "JKP Zero Trades *252*D" and "JKP Zero Trades *21*D" factors are based on sorting stocks by the number of days with zero trades over the past 252 and 21 days, respectively (Liu 2006), whereas the "Turnover 126D" factor is defined by sorting stocks based on shares traded divided by shares outstanding

**Figure 6.** (Color online) Factor Cluster SHAP Contributions




*Notes.* This figure reports the SHAP values for the different factor clusters, the market portfolio (MKT), the Fama-French six portfolios except for the market (FF6-MKT), and the 48 industry portfolios (Industry), pertaining to the intraday and overnight-inclusive Sharpe ratios and FF6 alphas from trading the SPY ETF based on the *(Cts, All)-(Ensemble)-(S-Sign)* strategy. Both the Sharpe ratios and alphas are reported in annualized units.

**Figure 7.** (Color online) Individual Factor SHAP Contributions




*Notes.* This figure reports the SHAP values for each of the individual factors pertaining to the intraday and overnight-inclusive Sharpe ratios and FF6 alphas from trading the SPY ETF based on the *(Cts, All)-(Ensemble)-(S-Sign)* strategy. Both the Sharpe ratios and alphas are reported in annualized units. For brevity, the figure reports only the factors with the top 10 most positive SHAP contributions to the intraday Sharpe ratio. The specific clusters associated with each of the factors are reported in parentheses.

averaged over the previous 126 trading days (Chordia et al. 2001). The "Beta Tail Risk" factor sorts stocks based on a nuanced measure of their tail risk (Kelly and Jiang 2014), whereas the "IVol CAPM 252D" factor sorts stocks based on their rolling CAPM idiosyncratic volatility (Ali et al. 2003). Echoing the discussion in the previous section, the tail risk factor is naturally associated with time-varying economic and financial uncertainty. Meanwhile, the four Low Risk factors, which account for most of the superior performance, all speak to notions of market liquidity. It is natural to hypothesize that the corresponding long-short portfolios essentially capture the return differentials between stocks that respond more quickly to new information versus stocks that respond more slowly to news, thereby allowing for the identification of periods with more marked cross-sectional differences in the rate of news absorption and poorer overall market quality. The analysis in the next section further strengthens this argument and more directly illustrates how the Low Risk factors translate into useful trading signals.

### 6.2. Liquidity Factors and Predictability

Although each of the four factors from the Low Risk cluster identified above contribute substantially to the total trading performance, they are all designed to capture similar aggregate liquidity-type effects. Hence, not surprisingly, as previously illustrated in Figure 2, the four factors are also strongly correlated, with the full-sample realized correlations between each of the pairs lying around 90%–95%.[23] In other words, the predictive content of the factors is likely quite similar. Thus, for brevity in the analysis that follows, we will focus exclusively on the "turnover" factor, but very similar numerical results and pictures to the ones discussed below are obtained for the three other Low Risk factors.

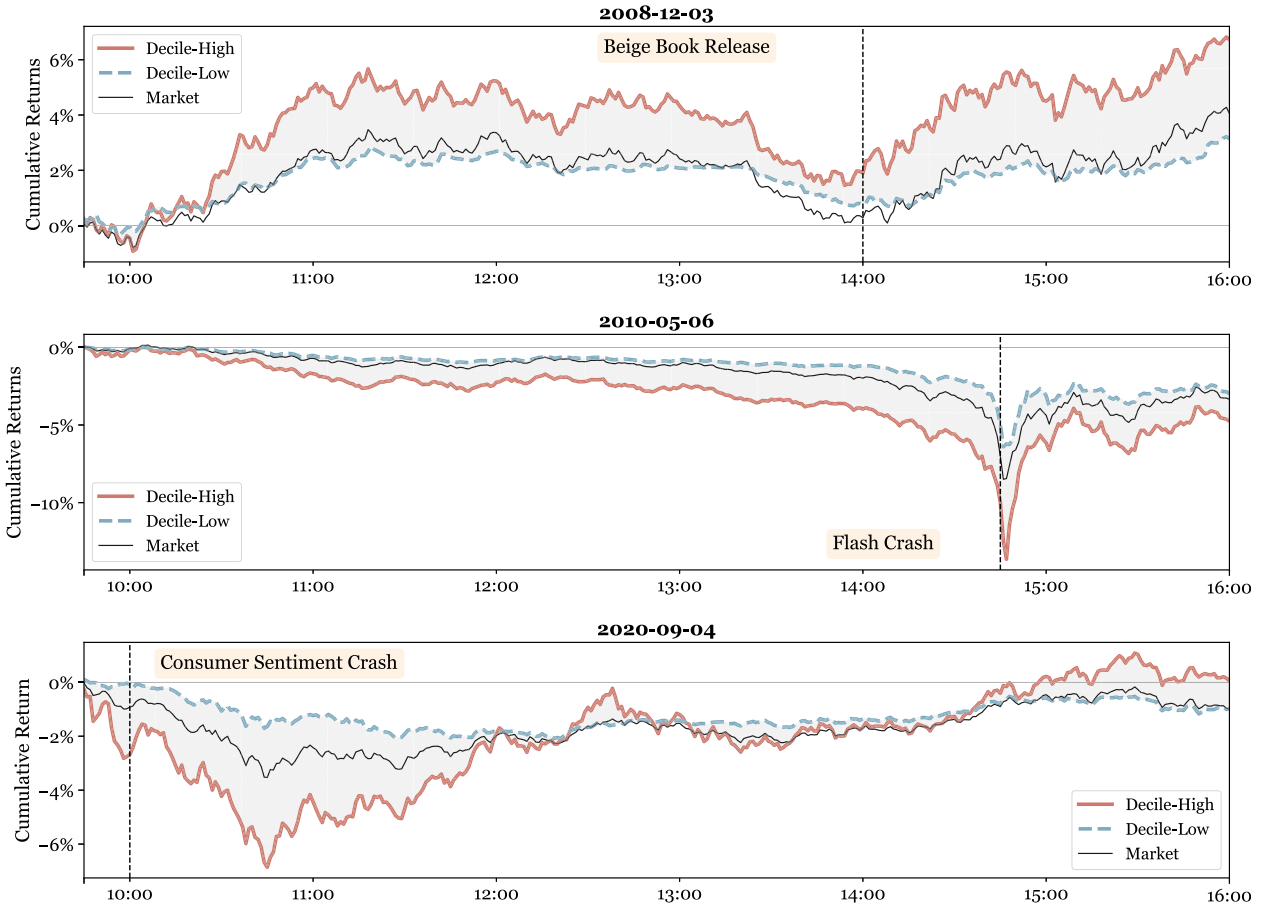
To help better illuminate which leg of the turnover factor primarily drives the performance, we begin by constructing value-weighted decile-sorted turnover portfolios at an even finer one-minute frequency, following the exact same approach used in the construction of the original "turnover" factor by Chordia et al. (2001). Figure 8 plots the resulting cumulative returns on the highest and lowest decile portfolios for three select days with various important economic news: the Beige Book release at 14:00 on March 12, 2008; the Flash Crash at around 2:45 on May 6, 2010; and the release of the Consumer Sentiment index at 10:00 on April 9, 2020. As the figure shows, the high decile portfolio, comprised of stocks that tend to trade more frequently, clearly responds more quickly to the specific news. Interestingly, the gap between the cumulative returns on the two decile portfolios, which essentially corresponds to the cumulative return on the lower-frequency "turnover" portfolio used in our forecasting models, also appears to foreshadow the subsequent direction of the market. These general features are perhaps most clearly evident in the second subplot and the day of the Flash Crash in 2010.[24] In fact, it appears as if the high decile portfolio anticipates the crash. Correspondingly, there is a steadily increasing gap between the returns on the high and low decile portfolios leading up to the crash. Our trading strategy takes advantage of this and effectively uses the high-minus-low portfolio return as a leading indicator for exiting the market in advance of the crash. Likewise, the widening negative gap observed in the bottom panel following the release of the Consumer Sentiment index signals that it is a good time to exit the market, whereas the converse widening positive gap observed in the top panel following the Beige Book release suggests that it is a good time to enter the market. Put differently, relevant new information appears to be absorbed more rapidly into the high turnover portfolio, with the high-low return differential foreseeing subsequent aggregate market moves.

As discussed earlier, the basic premise that high-turnover portfolios tend to lead low-turnover portfolios is hardly new to the literature. Chordia and Swaminathan (2002), for example, have previously shown that portfolios comprised of high-volume stocks appear to lead those of low-volume stocks, in turn inducing significant cross-autocorrelations in the returns over short horizons. Existing more theoretically oriented work, including Chan (1993) and Bernhardt and Mahani (2007), has also demonstrated how these differences in the absorption rates of market-wide information, or speeds of adjustments, may naturally arise from informational and trading frictions, with Mech (1993) positing that differential costs of trading provide the most likely explanation for the observed lead-lag relationships. However, we still find a nontrivial and highly statistically significant amount of alpha even after accounting for transaction costs from trading on the innovations of the high turnover stocks through our turnover-sorted factor portfolio.

To further develop the empirical support for this thesis, we consider two additional simple trading strategies based on either the long or the short leg of the high-minus-low turnover portfolio. Importantly, to eliminate irrelevant common variation, we reconstruct these high and low turnover portfolios as the 15-minute return on each of the underlying legs minus the 15-minute return on the market portfolio, following the same approach for market-neutralization as in Dong et al. (2022). We then use the lagged 15-minute returns on the resulting High-Minus-Mkt and Mkt-Minus-Low portfolios to define a simple trading strategy where we go long/short based on whether this variable is above/below a one-standard-deviation threshold.[25]

Consistent with the idea that the superior performance attributed to the turnover factor stems primarily

**Figure 8.** (Color online) Turnover Portfolio Returns




*Notes.* This figure reports the cumulative intradaily returns on the high and low decile turnover portfolios, together with the cumulative returns on the Fama-French market portfolio, for three select days. The return differentials between the high and low decile portfolios are shaded in gray. The (approximate) times of the labeled news events are indicated by dashed vertical lines.

from more frequently traded stocks that exhibit a faster speed of adjustment, this simple High-Minus-Mkt strategy attains an intraday Sharpe ratio of 0.50 from trading the SPY ETF. By comparison, the Sharpe ratio for the Mkt-Minus-Low strategy equals only −0.13. The transaction-cost adjusted intraday FF6 alphas for the same two strategies equal 4.90% ($t$ = 2.03) and −0.44% ($t$ = −0.17), respectively. In short, the economically relevant predictive content from our High-Minus-Low turnover factor lies in the "high" leg. Meanwhile, the intraday Sharpe ratio and alpha for the simple High-Minus-Mkt strategy are still markedly lower than the 1.37 Sharpe ratio and 20.83% alpha obtained for the *(Cts, All)-(Ensemble)-(S-Sign)* strategy reported in Table 4. Thus, even though this analysis points to some performance from merely exploiting the long leg of the turnover factor, much of the superior performance of our main strategy still stems from the use of modern machine-learning techniques in conjunction with the full factor zoo. Of course, the above analysis and discussion are also inherently ex post in nature, relying on the estimation results for the *(Cts, All)* Ensemble model together with the SHAP values to actually identify the most important factors and the turnover factor in particular.

## 7. Conclusion

We document significant intraday market return predictability from the use of high-frequency intraday factors combined with modern machine-learning techniques. Utilizing the ensemble forecasts from our prediction models for the market portfolio in the construction of simple trading strategies for a series of liquid ETFs, including the SPY results in large economic gains. The trading models perform particularly well during periods of increased economic uncertainty and heightened financial market volatility, generally thought to be associated with greater market dislocations. Consistent with that idea, we further trace most of the superior performance to a few liquidity-related factors that effectively capture differences in the rate of news absorption across stocks.

The factors that form the foundation for our empirical analyses have chiefly been proposed to explain

cross-sectional, as opposed to time-series, variation in equity returns. As such, it is possible that techniques similar to the ones developed here may give rise to even better prediction models and trading strategies when applied to individual stocks. Similarly, although our use of an ensemble approach involving a simple average of an array of different machine-learning-based forecasts delivers superior performance, it is possible that even better results may be obtained through the use of dynamic forecast combinations (see, e.g., Diebold and Shin 2019). Further gains may also be available through the use of more advanced dynamic trading strategies designed to more effectively control transaction costs compared with the simple S-Sign strategy that we primarily rely on here. We leave further work on all of these issues for future research.

## Acknowledgments

The authors thank the editor (Kay Giesecke) and three anonymous referees for their helpful and constructive comments. The authors also thank Peter R. Hansen, Andrew Patton, and participants at various seminars and conferences for their useful comments and suggestions.

## Appendix A

### A.1. Jump Detection

Our jump identification scheme relies on the now commonly used truncation procedure originally proposed by Mancini (2001). Intuitively, the procedure classifies a high-frequency return as a jump if the return is deemed "too large" to be compatible with the realization from a normal distribution with the locally estimated variance. More formally, we identify the return $r_{t,i}$ on day $t$ and intraday time interval $i$ as a jump if

$$|r_{t,i}| > \alpha \hat{\sigma}_{t,i} \Delta_n^{\varpi}, \tag{A.1}$$

where $\alpha$ and $\varpi$ are tuning parameters, $\hat{\sigma}_{t,i}$ denotes an estimate of the spot volatility at the time, and $n$ refers to the number of high-frequency observations per day. Following the literature, we set $\alpha$ and $\varpi$ to 3.5 and 0.49, respectively (Bollerslev and Todorov 2011). Unlike the existing literature, however, which generally estimates the spot volatility using the ex post volatility for the full day $t$, we instead rely on rolling ex ante volatility forecasts to ensure that our jump detection scheme is implementable in real time while still incorporating the most recent information.

To do so, we employ a simple exponentially weighted moving average (EWMA) of the rolling product of adjacent absolute returns,

$$\hat{\sigma}^2_{t,i} \equiv \frac{\pi}{2} \cdot n \cdot \text{EWMA}(|r_{t,i-1} \cdot r_{t,i-2}|, \text{halflife} = n) \cdot \tau_{t,i}, \tag{A.2}$$

$$\tau_{t,i} = \left( \sum_{s=1}^{k} r^2_{t-s,i} \right) \Big/ \left( \frac{1}{n} \sum_i \sum_{s=1}^{k} r^2_{t-s,i} \right), \tag{A.3}$$

where $\tau_{t,i}$ is a time-of-day adjustment factor. This construction is directly motivated by the bi-power variation measure traditionally used in the literature, $BV_t \equiv \frac{\pi}{2} \cdot \frac{n}{n-1} \cdot \sum_i |r_{t,i} \cdot r_{t,i-1}|$ (Barndorff-Nielsen and Shephard 2004). However, unlike the bi-power variation, which uses the ex post returns over the full day $t$, our EWMA-based estimator avoids look-ahead-bias by using only intraday realized returns preceding time $(t, i)$. Additional details, including comparisons with other procedures, are provided in the Online Appendix.

### A.2. Shapley Values Computation

The following algorithm provides a step-by-step guide to our computation of SHAP values.

1. Let $\pi_p$ represent a permutation of all the regressors possibly included in the model. We will write $\pi_p(j)$ to refer to the $j$'th element of the permutation of $\{1, \ldots, J\}$ defined by $\pi_p$. Correspondingly, the permuted set of regressors may be written as $\{X^{\pi_p(1)}_{t,i}, \ldots, X^{\pi_p(J)}_{t,i}\}_{t,i}$, where $X^j_{t,i}$ refers to the value of the $j$'th regressor on day $t$ and intraday time interval $i$.

2. Using this sequence of regressors, the predictions $\{\hat{r}^{\pi_p,0}_{t,i}, \hat{r}^{\pi_p,1}_{t,i}, \ldots, \hat{r}^{\pi_p,J}_{t,i}\}_{t,i}$ are based on the following expression:

$$\hat{r}^{\pi_p,j}_{t,i} = \sum_{j'=1}^{j} \hat{\beta}_{\pi_p(j')} X^{\pi_p(j')}_{t,i} + \sum_{j''=j+1}^{J} \hat{\beta}_{\pi_p(j'')} \overline{X}^{\pi_p(j'')},$$

where $\overline{X}^j$ refers to the sample average of the $j$'th regressor, and $\hat{\beta}_j$ refers to the estimated coefficient for that same regressor.

3. Produce the first set of return predictions $\{\hat{r}^{\pi_p,0}_{t,i}\}_{t,i}$ by replacing all of the regressors with their sample averages. Then, incrementally replace each regressor's mean with its actual values to produce the subsequent set of return predictions, where the order by which the regressors are replaced is determined by the permutation $\pi_p$. The resulting sequence of sets of predictions, $\{\hat{r}^{\pi_p,j}_{t,i}\}$, incrementally incorporates the information from each regressor.

4. Using each set of predictions $\{\hat{r}^{\pi_p,j}_{t,i}\}_{t,i}$ for $j \in \{1, \ldots, J\}$, compute the returns from the strategy and the performance measure of interest, say, $\{\xi^0_{\pi_p}, \ldots, \xi^J_{\pi_p}\}$.

5. Compute the marginal contribution of regressor $\pi_p(j)$ as the marginal effect of adding that regressor to the model that includes only regressors $\{\pi_p(1), \ldots, \pi_p(j-1)\}$ and replaces the remaining regressors with their sample averages as $\xi^j_{\pi_p} - \xi^{j-1}_{\pi_p}$.

6. Repeat all of the steps above for each permutation of the $J$ regressors. The SHAP value for the $j$'th regressor is then defined as $\frac{1}{J!} \sum_p \xi^{\pi_p^{-1}(j)}_{\pi_p} - \xi^{\pi_p^{-1}(j)-1}_{\pi_p}$, where $\pi_p^{-1}$ inverts the permutation $\pi_p$.

The additive nature of SHAP values, whereby the marginal contributions for each of the regressors $\pi_p(j)$ defined by $\xi^j_{\pi_p} - \xi^{j-1}_{\pi_p}$ adds up to the performance of the full model, $\xi^J_{\pi_p}$, makes them particularly attractive and easy to interpret.[26] In particular, as discussed in the main text, the additive nature of SHAP values allows us to aggregate our different performance measures by individual factors and factor clusters.

At the same time, the computation of SHAP values for large-dimensional models, including some of the ones used here, can be quite challenging from a computational perspective. Hence, rather than relying on the full set of $J!$ possible permutations, we instead randomly sample a smaller set of permutations and then take the sample average of the resulting permutation-specific estimates. This approach, dubbed "Monte Carlo Permutation Sampling" by Rozemberczki et al. (2022), was originally proposed by Castro et al.

(2017), and it is now fairly standard in the machine-learning literature. For our main results, we use a total of 400 such samples.[27]

We also further employ antithetical sampling to help reduce the standard error of the estimates. This is a common technique in Monte Carlo simulations, where samples are deliberately chosen to induce a negative correlation between draws. Following Lomeli et al. (2019), the specific Monte Carlo sampling scheme that we rely on, consists of drawing a random permutation and then reversing that permutation for the subsequent trial. In other words, the "even" Monte Carlo trial iterations consist of random permutations, whereas the "odd" trials simply use the reversed permutation underlying the preceding even trial, thereby substantially reducing the standard errors of the resulting SHAP estimates (see also the discussion in Mitchell et al. 2022).

## Endnotes

[1] The Sharpe ratio for a traditional buy-and-hold strategy for the SPY that also includes the overnight return is 0.49.

[2] This idea is closely related to the so-called "Manifold Hypothesis" in the machine learning literature. It is, of course, also directly related to the ubiquitous use of factor models in asset pricing as a way to summarize systematic risks in terms of a few common factors.

[3] Lou et al. (2019) attributed the large difference between average intraday and overnight returns to a "tug of war" between differently informed groups of investors; see also Bogousslavsky (2021) and Akbas et al. (2022). Relatedly, Gao et al. (2018) and Li et al. (2020) found that the overnight return helped forecast intraday momentum-type effects.

[4] By comparison, Engelberg et al. (2023) found little evidence that cross-sectional firm-level characteristics have out-of-sample time-series predictability for monthly market returns.

[5] In a similar vein, Huang et al. (2022) documented that the lead-lag relationship observed empirically between economically linked firms appears to be driven mainly by "slow-moving" news.

[6] The theoretical models of Gârleanu and Pedersen (2013, 2016) also further explain how limited risk-bearing capacity and transaction cost may naturally curtail mispricing correction. Similar arguments based on limits to arbitrage have also been invoked to explain a number of cross-sectional pricing anomalies, including, for example, the idiosyncratic volatility puzzle (Stambaugh et al. 2015) and the differential pricing of up- and downside beta risks (Bollerslev et al. 2022).

[7] The 218 characteristic-based factors collectively capture all the factors studied in Jensen et al. (2023) and Chen and Zimmermann (2022).

[8] The factors are ordered based on the classifications in Jensen et al. (2023) and Aleti (2023).

[9] This holds even without a semimartingale restriction, as shown by Fontana et al. (2019).

[10] These specifications can also be viewed as weighted regressions, where the weight on the regressand is an indicator variable pertaining to whether the market return is flagged as continuous.

[11] For the nonlinear FNN, RF, and GBRT models, we refit only every five years to reduce the computational burden.

[12] By comparison, evaluating the same bound with monthly returns results in 4.72%, which incidentally is also closely in line with the highest out-of-sample $R^2$'s obtained by Gu et al. (2020) through the use of ML tools for uncovering monthly market return predictability.

[13] Generating a Ross-bound of 0.2%, commensurate with the $R^2$ obtained for the *All, All* Ensemble specification, would require an implausible large risk aversion coefficient of $\gamma = 25.3$.

[14] As a point of reference, the average time-weighted percentage spreads for the SPY, QQQ, IVV, and IWM over the 2004–2020 period were 0.0067%, 0.0167%, 0.0156%, and 0.008%, respectively.

[15] It would be interesting to further explore how far these findings extend to other liquid ETFs, including different sector ETFs, for which the returns are not as strongly correlated with the return on the latent Fama-French market portfolio.

[16] As a reference, InteractiveBrokers (https://www.interactivebrokers.com/en/pricing/short-sale-cost.php) charges a 0.25% fee for shorting highly liquid securities so that our use of 2% is eight times higher than the cost faced by many investors.

[17] The additional analysis in the Online Appendix shows that the intraday performance measures discussed below are essentially unaltered when closing out the market position at the end of the day or when maintaining the weight defined by the intraday strategy for the 15:45–16:00 time interval during the overnight period. It is, of course, possible that by actively trading the overnight returns even better strategies may be obtained. This, however, would also entail the development of very different forecasting models from the high-frequency models underlying our current intraday forecasts, and we will not pursue this any further here.

[18] To allow the factor loadings to vary over time, we estimate the regressions on a rolling annual basis. Of course, because all the strategies invest only in the market portfolio, fixing the factor loadings for the market factor to be equal to the high-frequency market weights, all of the alphas would be identically equal to zero. Instead, the reported alphas should be interpreted as traditional measures of risk-adjusted performance "as if" the strategies were investments into some unknown assets or actively traded funds.

[19] Further along these lines, the Online Appendix reports SHAP values for each of the individual forecasting models and their contributions to both the Sharpe ratio and the alpha for the *(Cts, All)-(Ensmeble)-S-Sign* model; SHAP values are defined more precisely in Section 6 below. The patterns observed for these contributions generally resemble the patterns observed for the individual forecasts themselves. Interestingly, both the linear and the nonlinear models appear important in terms of driving the superior performance of the Ensemble average.

[20] The intraday pattern for the Fama-French market portfolio, previously depicted in Figure 1 averaged over the full 1996–2020 sample period, also evidences a negative, albeit not significantly so, average return over the last 15 minutes of the active portion of the trading day when restricting the sample to the 2004–2020 out-of-sample period.

[21] Recall that the turnover defined in (10) is computed as the absolute difference in the weights. Hence, a strategy that switches between being fully invested in the SPY and cash over each successive 15-minute interval would have a turnover of 100%.

[22] Complementary analysis in the Online Appendix further reveals that the SHAP values are generally concentrated in a few factors within each cluster.

[23] By comparison, the average 15-minute return correlations between each of the Low Risk factors and the tail risk factor from the Leverage cluster is only around 66%.

[24] The Flash Crash has, of course, been studied extensively by market regulators and elsewhere in the academic literature; see, for example, Andersen and Bondarenko (2014) and the many additional references therein.

[25] Intuitively, the strategy trades only if the forecast is sufficiently large/small. The standard deviation that we use for the threshold is

computed on a rolling basis with a three-month window, corresponding to $26 * 21 * 3 = 1,638$ high-frequency observations.

[26] For brevity, we do not report the SHAP value associated with the null model, $\xi^0_{\pi_p}$, in the various figures and tables of our paper. Note that when studying intraday trading performance, this value ends up approximately corresponding to the performance of simply holding cash, because the ML model forecasts when the regressors are zero are similarly close to zero. Thus, the SHAP value for the null model will be close to zero anyway, whether we are studying Sharpe ratios or alphas.

[27] In terms of computational time, each Monte Carlo trial takes about 17 hours, so this still amounts to 6,800 hours of processing.

## References

Aït-Sahalia Y, Fan J, Xue L, Zhou Y (2022) How and when are high-frequency stock returns predictable? Working paper, Princeton University, Princeton, NJ.

Aït-Sahalia Y, Jacod J (2014) *High-Frequency Financial Econometrics* (Princeton University Press, Princeton, NJ).

Akbas F, Boehmer E, Jiang C, Koch PD (2022) Overnight returns, daytime reversals, and future stock returns. *J. Financial Econom.* 145(3):850–875.

Aleti S (2023) The high-frequency factor zoo. Working paper, Duke University, Durham, NC.

Aleti S, Bollerslev T (2024) News and asset pricing: A high-frequency anatomy of the SDF. *Rev. Financial Stud.* Forthcoming.

Ali A, Hwang LS, Trombley MA (2003) Arbitrage risk and the book-to-market anomaly. *J. Financial Econom.* 69(2):355–373.

Andersen TG, Bollerslev T (1997) Intraday periodicity and volatility persistence in financial markets. *J. Empir. Finance* 4(2):115–158.

Andersen TG, Bollerslev T, Diebold FX, Vega C (2003) Micro effects of macro announcements: Real-time price discovery in foreign exchange. *Amer. Econom. Rev.* 93(1):38–62.

Andersen TG, Bondarenko O (2014) Vpin and the flash crash. *J. Financial Marketing* 17:1–46.

Andersen TG, Li Y, Todorov V, Zhou B (2023) Volatility measurement with pockets of extreme return persistence. *J. Econometrics* 237(2):105048.

Avramova D, Cheng S, Metzker L (2023) Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Sci.* 69(5):2587–2619.

Bandi FM, Kolokolov A, Pirino D, Renò R (2023) Discontinuous trading in continuous-time econometrics. Working paper, John Hopkins University, Baltimore.

Barndorff-Nielsen OE, Shephard N (2004) Power and bipower variation with stochastic volatility and jumps. *J. Financial Econom.* 2(1):1–37.

Bernhardt D, Mahani RS (2007) Asymmetric information and stock return cross-autocorrelations. *Econom. Lett.* 96(1):14–22.

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

Bogousslavsky V (2021) The cross-section of intraday and overnight returns. *J. Financial Econom.* 141(1):172–194.

Bollerslev T, Patton AJ, Quaedvlieg R (2022) Realized semibetas: Disentangling 'good' and 'bad' downside risks. *J. Financial Econom.* 144(1):227–246.

Bollerslev T, Tauchen G, Zhou H (2009) Expected stock returns and variance risk premia. *Rev. Financial Stud.* 22(11):4463–4492.

Bollerslev T, Todorov V (2011) Tails, fears, and risk premia. *J. Finance* 66(6):2165–2211.

Bondarenko O, Muravyev D (2023) Market return around the clock: A puzzle. *J. Financial Quant. Anal.* 58(3):939–967.

Borup D, Coulombe PG, Rapach D, Schütte ECM, Schwenk-Nebbe S (2022) The anatomy of out-of-sample forecasting accuracy. Working paper, Aarhus University, Aarhus, Denmark.

Bryzgalova S, Pelger M, Zhu J (2024) Forest through the trees: Building cross-section of stock returns. *J. Finance.* Forthcoming.

Campbell JY, Thompson SB (2008) Predicting excess stock returns out of sample: Can anything beat the historical average? *Rev. Financial Stud.* 21(4):1509–1531.

Caporin M, Kolokolov A, Renò R (2017) Systemic co-jumps. *J. Financial Econom.* 126(3):563–591.

Castro J, Gómez D, Molina E, Tejada J (2017) Improving polynomial estimation of the shapley value by stratified random sampling with optimum allocation. *Comput. Oper. Res.* 82:180–188.

Chan K (1993) Imperfect information and cross-autocorrelation among stock prices. *J. Finance* 48(4):1211–1230.

Chen AY, Zimmermann T (2022) Open source cross-sectional asset pricing. *Crit. Finance Rev.* 27(2):207–264.

Chen L, Pelger M, Zhu J (2023) Deep learning in asset pricing. *Management Sci.* 70(2):714–750.

Chinco A, Clark-Joseph AD, Ye M (2019) Sparse signals in the cross-section of returns. *J. Finance* 74(1):449–492.

Chordia T, Roll R, Subrahmanyam A (2005) Evidence on the speed of convergence to market efficiency. *J. Financial Econom.* 76(1):271–292.

Chordia T, Sarkar A, Subrahmanyam A (2007) The microstructure of cross-autocorrelations. Working paper, Federal Reserve Bank of New York, New York.

Chordia T, Subrahmanyam A, Anshuman VR (2001) Trading activity and expected stock returns. *J. Financial Econom.* 59(1):3–32.

Chordia T, Swaminathan B (2002) Trading volume and cross-autocorrelations in stock returns. *J. Finance* 55(2):913–935.

Christensen K, Oomen RC, Renò R (2022) The drift burst hypothesis. *J. Econometrics* 227(2):461–497.

Cliff M, Cooper MJ, Gulen H (2008) Return differences between trading and non-trading hours: Like night and day. Working paper, Virginia Tech University, Blacksburg.

Corsi F (2009) A simple approximate long-memory model of realized volatility. *J. Financial Econom.* 7(2):174–196.

Coulombe PG, Rapach D, Schütte ECM, Schwenk-Nebbe S (2023) The anatomy of machine learning-based portfolio performance. Working paper, University of Montreal, Montreal.

Cui W, Hu J, Wang J (2024) Nonparametric estimation for high-frequency data incorporating trading information. *J. Econometrics* 240:105690.

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. *J. Bus. Econom. Statist.* 13(3):253–265.

Diebold FX, Shin M (2019) Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *Internat. J. Forecast.* 35(4):1679–1691.

Dong X, Li Y, Rapach DE, Zhou G (2022) Anomalies and the expected market return. *J. Finance* 77(1):639–681.

Engelberg J, McLean RD, Pontiff J, Ringgenberg MC (2023) Do cross-sectional predictors contain systematic information? *J. Financ. Quant. Anal.* 58(3):1172–1201.

Erel I, Stern LH, Tan C, Weisbach MS (2021) Selecting directors using machine learning. *Rev. Financial Stud.* 34(7):3226–3264.

Fama EF (1970) Efficient capital markets: A review of theory and empirical work. *J. Finance* 25(2):383–417.

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

Fan J, Ke ZT, Liao Y, Neuhierl A (2022) Structural deep learning in conditional asset pricing. Working paper, Princeton University, Princeton, NJ.

Farmer LE, Schmidt L, Timmermann A (2023) Pockets of predictability. *J. Finance* 78(3):1279–1341.

Fontana C, Pelger M, Platen E (2019) On the existence of sure profits via flash strategies. *J. Appl. Probab.* 56(2):384–397.

Gao L, Han Y, Li SZ, Zhou G (2018) Market intraday momentum. *J. Financial Econom.* 129(1):394–414.

Gârleanu N, Pedersen LH (2013) Dynamic trading with predictably returns and transaction costs. *J. Finance* 68(6):2309–2340.

Gârleanu N, Pedersen LH (2016) Dynamic portfolios choice with frictions. *J. Econom. Theory* 165:487–516.

Giglio S, Xiu D (2021) Asset pricing with omitted factors. *J. Political Econom.* 129(7):1947–1990.

Gromb D, Vayanos D (2010) Limits of arbitrage. *Ann. Rev. Finance* 2:251–275.

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

Guijarro-Ordonez J, Pelger M, Zanotti G (2022) Deep learning statistical arbitrage. Working paper, Stanford University, Stanford, CA.

Hendershott T, Livdan D, Rösch D (2020) Asset pricing: A tale of night and day. *J. Financial Econom.* 138(3):635–662.

Heston SL, Korajczyk RA, Sadka R (2010) Intraday patterns in the cross-section of stock returns. *J. Finance* 65(4):1369–1407.

Hou K (2007) Industry information diffusion and the lead-lag effect in stock returns. *Rev. Financial Stud.* 20(4):1113–1138.

Huang S, Lee CM, Song Y, Xiang H (2022) A frog in every pan: Information discreteness and the lead-lag returns puzzle. *J. Financial Econom.* 145(2):83–102.

Huddleston D, Liu F, Stentoft L (2023) Intraday market predictability: A machine learning approach. *J. Financial Econom.* 21(2): 485–527.

Jensen TI, Kelly B, Pedersen LH (2023) Is there a replication crisis in finance? *J. Finance* 78(5):2465–2518.

Jiang J, Kelly BT, Xiu D (2023) (Re-)imag(in)ing price trends. *J. Finance* 78(6):3193–3249.

Jurado K, Ludvigson SC, Ng S (2015) Measuring uncertainty. *Amer. Econom. Rev.* 105(3):1177–1216.

Ke Z, Kelly BT, Xiu D (2021) Predicting returns with text data. Working paper, Yale University, New Haven, CT.

Kelly B, Jiang H (2014) Tail risk and asset prices. *Rev. Financial Stud.* 27(10):2841–2871.

Kelly B, Malamud S, Zhou K (2024) The virtue of complexity in return prediction. *J. Finance* 79(1):459–503.

Kelly BT, Xiu D (2023) Financial machine learning. Working paper, Yale University, New Haven, CT.

Korsaye SA, Quaini A, Trojani F (2019) Smart sdfs. Working paper, University of Geneva, Geneva.

Leippold M, Wang Q, Zhou W (2022) Machine-learning in the Chinese stock market. *J. Financial Econom.* 145(2):64–82.

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

Lewellen J, Nagel S, Shanken J (2010) A skeptical appraisal of asset pricing tests. *J. Financial Econom.* 96(1):175–194.

Li H, Diks C, Pachenko V (2020) Predicting intraday return patterns based on overnight returns for the US stock market. Working paper, University of Amsterdam, Amsterdam.

Liu W (2006) A liquidity-augmented capital asset pricing model. *J. Financial Econom.* 82(3):631–671.

Lo AW, MacKinlay AC (1990) When are contrarian profits due to stock market overreaction? *Rev. Financial Stud.* 3(2):175–205.

Lomeli M, Rowland M, Gretton A, Ghahramani Z (2019) Antithetic and Monte Carlo kernel estimators for partial rankings. *Statist. Comput.* 29:1127–1147.

Lou D, Polk C, Skouras S (2019) A tug of war: Overnight vs. intraday expected returns. *J. Financial Econom.* 134(1):192–213.

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* 30:1–10.

Mancini C (2001) Disentangling the jumps of the diffusion in a geometric jumping Brownian motion. *Giornale dell'Istituto Italiano degli Attuari* LXIV:19–47.

Mech TS (1993) Portfolio return autocorrelation. *J. Financial Econom.* 34(3):307–344.

Mitchell R, Cooper J, Frank E, Holmes G (2022) Sampling permutations for shapley value estimation. *J. Mach. Learn. Res.* 23(43): 1–46.

Moehle N, Boyd S, Ang A (2021) Portfolio performance attribution via shapley value. Working paper, BlackRock, New York.

Pesaran MH, Timmermann A (1995) Predictability of stock returns: Robustness and economic significance. *J. Finance* 50(4): 1201–1228.

Pontiff J (1996) Costly arbitrage: Evidence from closed-end funds. *Quart. J. Econom.* 111(4):1135–1151.

Rapach DE, Strauss JK, Zhou G (2010) Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Rev. Financ. Stud.* 23(2):821–862.

Ross SA (2004) *Neoclassical Finance* (Princeton University Press, Princeton, NJ).

Rozemberczki B, Watson L, Bayer P, Yang HT, Kiss O, Nilsson S, Sarkar R (2022) The shapley value in machine learning. Working paper, University of Edinburgh, Edinburgh, UK.

Shleifer A, Vishny RW (1997) The limits of arbitrage. *J. Finance* 52(1):35–55.

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

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