Predicting High-Frequency Industry Returns: Machine Learners Meet News Watchers

Hao Jiang*, Sophia Zhengzi Li†, and Peixuan Yuan‡

September 26, 2020

Abstract

This paper uses machine learning-based as well as fundamental-driven, news-based approaches to uncover patterns of high-frequency return predictability for sector exchange-traded funds (ETFs). A LASSO predictor that aggregates high-frequency price movements of a broad universe of individual stocks predicts ETF returns out-of-sample. The news-driven return on ETF constituent firms positively predicts ETF returns, but the component of ETF returns orthogonal to the news return negatively predicts them. These different signals contain independent information, and have different strengths, with the LASSO predictor providing continuous flows of information most powerful during trading hours and the news return offering sporadic information particularly useful during market close. A composite signal combining all three signals with Gradient Boosted Regression Trees (GBRT) has very strong power to forecast ETF returns, especially during the Covid-19 pandemic.

JEL classification: G10; G14; G40.

Keywords: Industry Return; ETF; Return Predictability; LASSO; Boosted trees; News.

*Eli Broad College of Business, Michigan State University, East Lansing, MI 48824; E-mail: jiangh@broad.msu.edu.

†Rutgers Business School, Newark, NJ 07102; E-mail: zhengzi.li@business.rutgers.edu.

‡Rutgers Business School, Newark, NJ 07102; E-mail: peixuan.yuan@rutgers.edu.

1. Introduction

A central theme in finance is return predictability. Despite the voluminous literature on the predictability of aggregate stock market and individual stock returns,[1] relative less research is devoted to the industry return predictability,[2] and even less is known about industry return predictability at high frequency. In this paper, we fill the gap by studying high-frequency return predictability for sector exchanged-traded funds (ETFs), which allow us to accurately measure high-frequency industry returns.

We apply two broad approaches to study the predictability of sector ETF returns. First, previous studies have established the gradual diffusion of information along the supply chain and across the economy (see, e.g., Hong, Torous, and Valkanov, 2007; Cohen and Frazzini, 2008; Menzly and Ozbas, 2010). This insight inspires us to use a broad set of information contained in stock markets to predict the price movements of individual sector ETFs. What distinguishes our approach is that we exploit the granular information contained in price movements of individual stocks in the entire US equity market to predict subsequent sector ETF returns. To solve the problem of high dimensionality, for each sector ETF in each short-term interval, we use LASSO regressions to select stocks exhibiting the strongest power to predict the ETF return in the subsequent period. Then we examine the out-of-sample performance of the LASSO predictor. This approach essentially combines the merits of two recent machine learning applications in return predictability by Chinco et al. (2019) and Rapach et al. (2019).

Second, we build on the literature that the stock market tends to underreact to recent corporate news announcements. For instance, Jiang et al. (2020) show that intraday individual stock returns driven by firm news tend to exhibit momentum, but those returns without accompanying firm news tend to exhibit a reversal. Therefore, we compute value-weighted

[1] See Rapach and Zhou (2013) for an extensive survey on the literature of stock return predictability.

[2] See, e.g., Boudoukh et al. (1994), Moskowitz and Grinblatt (1999), DellaVigna and Pollet (2007), Bustamante and Donangelo (2017).

high-frequency news-driven returns on an ETF's constituent firms as our second predictor, and the component of the ETF return orthogonal to the news return as our third predictor.

With the three predictors, we proceed to study sector ETF return predictability. Since corporate news tend to arrive disproportionately after market close, we start the analyses by training the LASSO predictor using overnight returns and compute news and non-news predictors with overnight news. The results indicate that overnight LASSO and news predictors have positive and significant predictive power for ETF returns after the market opens, and the overnight non-news predictor negatively predicts subsequent ETF returns. In Fama-MacBeth regressions, the slope coefficients for these predictors have large t-statistics with absolute values exceeding 5 and high adjusted R^2 above 15%. The return predictability is also economically large: when we form quintile portfolios based on the return predictors, the average return spread between Quintile 5 and Quintile 1 is sizable. For instance, the long-short portfolio based on the LASSO predictor is 0.88% per month, with an annualized Sharpe ratio of 1.21; the corresponding numbers are 1.61% per month and 2.15 for the news return and -2.89% per month and 2.56 for the non-news return.

Comparing the patterns of the return predictability over the course of the trading day, we find that the predictive power for the first half-hour from 10:00 to 10:30 is disproportionately high.[3] For the overnight LASSO predictor, approximately two thirds of the predictability during the trading day concentrates in the first half-hour; for the news predictor, less than half; for the non-news predictor, close to nine tenths. When we extend the return forecasting window to the next trading week, we find that the return predictive power for the LASSO signal and non-news signal concentrates in the first trading day, but that for the news signal persists up to the second trading day. It is important to note that the three signals contain independent information about future ETF returns because they are all significant and strong in the Fama-MacBeth regressions containing the three predictors.

Moving beyond the overnight predictors, we generate all three predictors: LASSO, news

[3]We compute overnight returns based on the last trade price before 16:00 of day $d-1$ and the last trade price before 10:00 of day d.

and non-news at half-hour frequency using intraday half-hour returns. Under this design, we still observe strong return predictive power of the three predictors. It also reveals that their relative return forecasting strength tends to be different. Specifically, the LASSO predictor shows its strength in absorbing the continuous flow of information contained in price movements of thousands of stocks traded when the market is open. The freshly updated LASSO predictor at each half-hour interval shows strong power in predicting subsequent half-an-hour ETF returns in a smooth and hump-shaped fashion: the slope coefficients in Fama-MacBeth regressions remain high and statistically significant throughout the 12 half-hour intervals from 10:00 to 16:00, and the coefficient is the highest and most significant when predicting 12:30-13:00 return. By contrast, the news predictor suffers from the fact that corporate news tends to arrive sporadically during the market open period; the resulting sparsity in news signals appears to weaken the predictive power of the news returns computed based on the past-half-hour corporate news. The slope coefficients for the news signal in Fama-MacBeth regressions fluctuate substantially across the day and are frequently statistically insignificant except for the first half-hour trading interval. On the other hand, the half-hour non-news signal maintains a strong pattern of reversal during the trading day.

In univariate portfolio sorts, the long-short portfolio based on the overnight LASSO signal is 0.88% per month, with an annualized Sharpe ratio of 1.21; the corresponding numbers are 1.61% per month and 2.15 for the overnight news signal, and -2.89% per month and 2.56 for the overnight non-news signal.

Considering the incremental return predictive power of each predictor, we devise two ways to aggregate the predictors. First, we consider a naive strategy, which rescales each predictor to be between 0 and 1 and then takes a simple average of the three scores. Second, we use Gradient Boosted Regression Trees (GBRT) to optimize the weights of each return signal based on the strength of its return predictive power. A portfolio long the top quintile and short the bottom quintile based on the simple average score generates an average return of 2.80% per month, with an annualized Sharpe ratio of 3.23. A similar long-short portfolio

based on the GBRT composite signal yields a higher average return of 3.33% per month, with a higher annualized Sharpe ratio of 4.13.

The strong evidence of ETF return predictability illustrates the great value of big data and machine learning in efficient ETF pricing.[4] The large trading profits exploiting the return predictability also shed fresh light on the stellar performance of certain hedge funds trading with very low latency. What would happen to the return predictability and trading profits during periods with large turmoil disrupting ETF pricing? To address this question, we perform a special analysis for January to July 2020, around the period of the Covid-19 pandemic. We conjecture that the unprecedented shocks to the economy and the financial market induced by the global pandemic may both distract investor attention and reduce the responsiveness of arbitrage capital to mispricing. As a result, the return predictability in high-frequency sector ETF returns could be even stronger during the Covid pandemic. Our results show that the average return to a long-short portfolio based on the LASSO predictor reaches 3.76% per month with an annualized Sharpe ratio of 4.47; the corresponding numbers are 7.37% per month and 5.19 for the news predictor, and -7.29% and 4.36 for the non-news predictor. The GBRT composite signal generates a long-short portfolio with an average monthly return of 10.87% and an annualized Sharpe ratio of 8.34. The magnitude of return predictability is more than twice that of the full sample average. These results are consistent with the idea that investor distraction and slow-moving capital during major market dislocations contribute to excessive return predictability.

[4]There is a small but rapidly growing literature that applies the techniques of machine learning in asset pricing. This literature shows the power of machine learning techniques in identifying the cross-sectional variation in individual stock returns (See, e.g., Chen et al. (2019), Chinco et al. (2019), Freyberger et al. (2020), and Gu et al. (2020)). Our paper extends this literature to high-frequency industry return predictability.

2. Empirical Methodology

2.1. Data Description

2.1.1. Intraday Return Data

Universe: Our ETF sample consists of all sector ETFs from the ETFGlobal database. We obtain monthly holding data of the ETFs from the Center for Research in Security Prices (CRSP) Mutual Fund database. Our stock sample includes all firms listed on the New York Stock Exchange (NYSE), National Association of Securities Dealers Automated Quotations (NASDAQ), and American Stock Exchange (AMEX) with share code of 10 or 11. Our intraday price and quote data for ETFs and stocks come from the NYSE trade and quote (TAQ) database. The final sample includes 226 unique sector ETFs and 10,076 unique stocks over the period of 5,178 days between January 2000 and July 2020. The average number of unique stocks per day is around 4,300.

Return data: To prepare the intraday return data, we collect minute-by-minute observations of intraday prices by applying the cleaning rules of Bollerslev, Li, and Todorov (2016), Bollerslev, Li, and Zhao (2020), and Jiang, Li, and Wang (2020) to the TAQ database. Our high-frequency data cleaning procedures are detailed in the Appendix. Based on intraday prices, we then compute intraday returns as every 30-minute return between 10:00 and 16:00 and the overnight return as the return between 16:00 on the previous trading day and 10:00 on the current trading day.[5] Since transaction prices from TAQ are raw prices without adjusting for corporate actions such as dividend payout and stock splits, we apply the daily "cumulative factor to adjust price" and "dividend cash amount" variables in the CRSP database to adjust for split and dividend.

[5]We use the price at 10:00 for overnight returns to ensure that most securities have traded at least once after the market open.

2.1.2. Intraday News Data

The high-frequency firm-level news data are from the RavenPack news database, which provides a comprehensive sample of firm-specific news stories from the Dow Jones News Wire (Recent studies using this data set include Jiang and Sun, 2015; Kelley and Tetlock, 2017; Jiang, Li, and Wang, 2020). To ensure a news story is specifically about a given firm, we rely on the "relevance score" variable provided by RavenPack, which captures how closely the news story is related to a particular company. The score ranges from 0 to 100, where a score of 0 (100) means that the entity is passively (predominantly) mentioned. We only use news stories with a relevance score of 100 in our sample. To include only news stories about firm fundamentals, we select 12 news groups of acquisitions-mergers, analyst-ratings, assets, bankruptcy, credit, credit-ratings, dividends, earnings, equity-actions, labor-issues, product-services, and revenues from a total of 29 news groups. To keep only fresh news about a company, we further exclude repeated news by requiring the "event novelty score" from RavenPack to be 100. Applying these filters does not introduce look-ahead bias because all news articles are processed by RavenPack within milliseconds of receipt and the resulting data are immediately sent to subscribers. Thus, all information is available at real time.

2.2. The LASSO Predictor

2.2.1. Model

Our first sector ETF signal is the LASSO predictor that aggregates high-frequency price movements of a broad universe of individual stocks returns. Given an input of the high-frequency returns of p individual stocks $(r^{stc}_{1,d,i}, r^{stc}_{2,d,i}, \ldots, r^{stc}_{p,d,i})$ for day d and intraday interval i, we wish to predict the jth sector ETF return $r^{etf}_{j,d,i+1}$ for day d and intraday interval $i+1$. In this paper, we focus on intraday frequency of 30 minutes and consider the overnight period as from 16:00 on day $d-1$ to 10:00 on day d. Therefore, we have $i = 1, 2, \ldots, 13$ with $i = 1$ indicating the overnight period and $i = 2$ the period between 10:00 and 10:30 and so on until $i = 13$ indicating the interval between 15:30 and 16:00. For each intraday interval,

we propose the following linear model:

$$r^{etf}_{j,d,i+1} = \beta_{j,i,0} + \sum_{s=1}^{p} \beta_{j,i,s} r^{stc}_{s,d,i} + \epsilon_{j,d,i+1}. \tag{1}$$

Note that, unlike panel regressions commonly used in asset pricing where all dependent variables share the same regression coefficients cross-sectionally, our model assumes that each ETF has its own regression coefficients to add more flexibility to modeling the sector ETFs.

To estimate the regression coefficients, we have two considerations. The first one is economical. For each sector ETF, we do not expect all of the thousands of stocks are useful in predicting its return. We can only hope that a handful of stocks that are intrinsically connected with the ETF might be useful for prediction. The second one is statistical. Our input space is relatively large (thousands of stocks) in comparison with our sample size (thousands of observations). In addition, the return prediction problem is in a world of low signal-to-noise ratio. Under such situations, statistical machine learning theory tells us we can improve the prediction accuracy by introducing constraints to reduce the estimation variance at the cost of sacrificing a smaller amount of estimation bias. Motivated by these two considerations, we use a machine learning technique, **L**east **A**bsolute **S**hrinkage and **S**election **O**perator (LASSO), to estimate these coefficients. LASSO can sparsely estimate the coefficients $\{\beta_{j,i,s}\}$ by automatically setting many of them to be exactly zero, thus achieving both stock selection and variance reduction goals arising from our two considerations.

LASSO encourages sparse estimates of regression coefficients by introducing the L_1 penalty to the standard least squares for regressions. Let $\boldsymbol{\beta}_{j,i} = (\beta_{j,i,0}, \beta_{j,i,1}, \ldots, \beta_{j,i,p})$ be the regression coefficients of predicting the jth ETF in the ith intraday interval. Using the LASSO penalty, we will minimize the following loss function:

$$\hat{\boldsymbol{\beta}}^{LASSO}_{j,i} = \underset{\boldsymbol{\beta}_{j,i}}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{d} \left(r^{etf}_{j,d,i+1} - \beta_{j,i,0} - \sum_{s=1}^{p} \beta_{j,i,s} r^{stc}_{s,d,i} \right)^2 + \lambda_{j,i} \sum_{s=1}^{p} |\beta_{j,i,s}| \right\}, \tag{2}$$

where λ is the hyperparameter that controls the amount of shrinkage. Large values of λ

induces more zeros in the corresponding $\hat{\boldsymbol{\beta}}_{j,i}^{LASSO}$. In the extreme case of $\lambda = 0$, there will be no penalty so $\hat{\boldsymbol{\beta}}_{j,i}^{LASSO}$ corresponds to the ordinary least squares estimates. In the other extreme case of $\lambda = \infty$, the objective function will be dominated by the penalty term and thus $\hat{\boldsymbol{\beta}}_{j,i}^{LASSO} = 0$. Thus, the choice of λ is key to the performance of LASSO. In machine learning, the most common method for hyperparameter tuning is cross-validation (CV). The basic idea is to use only part of the data for fitting and a different part for validation. Then we will choose the λ that has the smallest prediction error in the validation set for generating LASSO signals. Intuitively, this process that fits the model on the training sample and then evaluates the fit on the validation sample mimics the process of in-sample training and then out-of-sample prediction. Therefore the λ value that works better in validation set tends to perform better out-of-sample than other values of λ.

2.2.2. Implementation

We implement LASSO in a rolling-fit fashion, where each fit consists of three steps: training, validation and testing. In particular, at the end of each year y, we use the lagged returns of the entire cross section of stocks that are available for the five-year period between year $y-4$ and y as predictors.

Training: We use the four year data between years $y-4$ and $y-1$ to train the model of Equation (2) on an ETF-by-ETF and interval-by-interval basis for a set of candidate hyperparameters λ covering a wide range of sparsity level.

Validation: Using estimated coefficients from training, we generate the return predictions for each ETF for the validation sample consisting of the data in year y. Then based on these predictions for year y and their corresponding true ETF returns, we compute the mean squared errors as the evaluation metric and estimate the optimal $\hat{\lambda}_{j,i}$ to be the one with the smallest mean squared errors.

Testing: Finally, we use the coefficients $\hat{\boldsymbol{\beta}}_{j,i}^{LASSO}$ estimated under $\hat{\lambda}_{j,i}$ for generating

LASSO signals for year $y+1$:

$$f_{j,d,i}^{etf,LASSO} = \hat{\beta}_{j,i,0} + \sum_{s=1}^{p} \hat{\beta}_{j,s,i} r_{s,d,i}^{stc}, \tag{3}$$

where the notation $f_{j,d,i}^{etf,LASSO}$ stands for the return forecast for the jth ETF on day d and at the end of interval i using LASSO. We repeat the above steps for each year from 2004 to 2019 and call it the LASSO predictor as our first proposed sector ETF signal.

2.3. The News-Driven Return Predictor

Our second signal is a fundamental-driven, news-based predictor that aggregates fundamental information of individual stocks for predicting sector ETF returns. The signal is constructed via a bottom-up process.

Stock-Level News-Driven Returns: First, we construct stock-level news-driven returns. Following Jiang et al. (2020), we classify high-frequency stock returns $r_{s,d,i}^{stc}$ as news-driven returns based on high-frequency market reaction to information. Specifically, if at least one firm-level news is released during the return measurement period, the news-driven return is equal to the total return of that period; the news-driven return is set to zero otherwise. For news occurring within regular trading hours, the news return is simply the 30-minute return over the same period that the news occurs. For news occurring during the weekend, holiday, or overnight, the news return is the surrounding overnight return to reflect that the first reaction to such news stories is incorporated into the stock's price only for the first trade of the following trading day. For example, the return for news events during the weekend is the return over the period of 16:00 of the surrounding Friday and 10:00 of the surrounding Monday.

ETF-Level News-Driven Returns: After obtaining the high-frequency firm-level news returns, we construct high-frequency ETF-level news returns as the value-weighted average of news returns of its constituent firms. More formally, suppose there are N constituent firms

in a given ETF j on day $d-1$, and let $r_{s,d,i}^{stc,news}$ denote the news return of constituent firm s on day d in interval i. The high-frequency news-driven ETF signal on day d and interval i is computed as:

$$f_{j,d,i}^{etf,news} = C_{j,d,i} \sum_{s=1}^{N} w_{j,d-1,s} r_{s,d,i}^{stc,news}, \tag{4}$$

where $w_{j,d-1,s}$ is the weight of constituent firm s in ETF j on day $d-1$;[6] $C_{j,d,i}$ is a normalizing constant such that the total weights on constituent firms with non-zero new return is equal to one (i.e., $C_{j,d,i} \sum_{s=1}^{N} w_{j,d-1,s} \mathbf{1}(r_{s,d,i}^{stc,news} \neq 0) = 1$). If all constituent firms have zero news return on day d in interval i, $r_{j,d,i}^{etf,news}$ is set to be zero.

2.4. The Non-News-Driven Return Predictor

Our third signal is the non-news-driven predictor that removes the fundamental news-driven part from the raw ETF return. To control for the news return in the sector ETFs, we perform cross-ETF regressions that use the ETF news-driven returns $r_{j,d,i}^{etf,news}$ in Equation (4) as independent variables. Specifically, for interval i on day d, we estimate the following regression:

$$r_{j,d,i}^{etf} = \alpha_{d,i} + \beta_{d,i} f_{j,d,i}^{etf,news} + \epsilon_{j,d,i}. \tag{5}$$

After obtaining the regression coefficient estimates, we compute the non-news-driven signal as the regression residual:

$$f_{j,d,i}^{etf,non\text{-}news} = r_{j,d,i}^{etf} - \hat{\alpha}_{d,i} - \hat{\beta}_{d,i} f_{j,d,i}^{etf,news}, \tag{6}$$

which will be called as the non-news signal.

[6]The weights of ETFs on constituent firms are based on monthly holding data from CRSP. For each trading day d, we rely on holdings from the nearest previous day to approximate the weights on day $d-1$.

2.5. Control Variables

In addition to the three proposed signals, we construct a number of control variables at ETF level following standard definitions in the literature. $ILLQ$ is the natural logarithm of the average daily ratio of the absolute ETF return to the dollar trading volume over the five-day period preceding each day (Amihud, 2002). $RVOL$ is annualized realized volatility, defined as the square root of the annualized realized variance, which is 252 times the sum of squared five-minute intraday ETF returns within each trading day. Mom for a given day t is the cumulative ETF return from day $d-252$ to day $d-21$. $Size$ is the natural logarithm of the net asset value of the ETF, estimated by the product of the closing price and the number of shares outstanding from CRSP and is updated daily.

2.6. Descriptive Statistics of Variables

Since corporate news tend to arrive disproportionately after market close, we start with the evaluation of the overnight signals (i.e., $i = 1$ in Equations (3), (4) and (6)) that are generated at the market open of each day. Table 1 reports descriptive statistics of overnight LASSO, news, and non-news signals, as well as other control variables for all ETFs in our sample. Return variables are reported in basis points (bps). Panel A shows the mean, standard deviation, and five quantiles, all of which are first computed cross-sectionally on each day and then averaged over time. The results imply that the distribution of overnight LASSO signal is close to symmetric, with a mean of -0.30 bps and a median of -0.16 bps. The distribution of the overnight news signal is slightly skewed to the right with a mean of 3.55 bps. For the overnight non-news signal, the mean and median are 0.00 bps and 0.22 bps, respectively. The cross-sectional dispersion of overnight news signals as measured by their cross-sectional standard deviation is 160.88 bps, which is larger than the dispersion of 10.33 for the LASSO signal and 52.71 bps for the non-news signal.

Panel B reports the average cross-sectional correlations among these variables. The three signals are almost pairwise uncorrelated. The correlations are in lower single digit, suggesting

the additivity of them as multiple signals for predicting ETF returns. The correlations between our signals and the control variables are also very low, imdicating the uniqueness of our signals in comparison with existing ones. On the other hand, the correlation structure among control variables is consistent with previous literature. For instance, we find that ETFs with high net asset value tend to have higher liquidity and lower volatility.

3. LASSO Predictive Regressions

Before evaluating the predictive performance of our signals, a natural question is: what are the stocks selected by LASSO to predict sector ETFs? Panel A in Fig. 1 plots the mean and 90% confidence interval for the number of firms selected by LASSO each year during our sample period.[7] There are several noteworthy patterns. First, on average, LASSO selects only 35 predictors (out of a possible 4,300 predictors) when making its first-30-minute return forecasts for each ETF. This finding shows that the stock signals for sector ETFs are indeed sparse. The use of the sparsity-encouraging techniques such as LASSO is critical to the discovery of these sparse signals. Theoretically, Donoho (2006) shows that under certain assumptions on the model matrix, if the true model is sparse, the LASSO solution identifies the correct predictors with high probability. Second, LASSO tends to select more predictors during the financial crisis. For example, LASSO uses more than 100 predictors for almost all ETFs in 2008. This is somehow consistent with the findings that correlations are much elevated during extreme downside periods (see, e.g., Ang and Chen, 2002).

The result in Panel B of Fig. 2 shows that, on average, the average proportion of constituent firms is only 3.7%.[8] This result implies that only a small proportion of constituents can affect future industry returns, and most part of the industry return predictability may come from firms other than the constituents. Therefore, the machine learning technique can be

[7]The number of predictors selected by LASSO is controlled by the hyperparameter λ chosen by validation. We use fixed 1-year data as our validation set. The first (last) validation set in our our sample period is in 2004 (2019).

[8]The proportion of constituent firms is defined by the number of overlapping firms between LASSO-selected and ETF constituent firms normalized by the total number of ETF constituents.

extremely useful to help us identify those firms.

To further illustrate the firms selected by LASSO, we report the LASSO-coefficient estimates in 2019 for nine sector ETFs that have high, median, and low Assets Under Management (AUM), including XLK (Technology SPDR Fund), XLI (Industrial SPDR Fund), XLV (Health care SPDR Fund), FXU (Utilities AlphaDEX Fund), XTN (Transportation SPDR Fund), FXN (Energy Alpha DEX Fund), PSCC (Consumer Staples ETF), ENFR (Infrastructure ETF), and BIZD (VanEck Income ETF). For each ETF, we only report the coefficients of the top 20 firms ranked by the absolute value of their LASSO-coefficient estimates as well as the constituent weight of each firm in parentheses.[9] If the selected firm is not an ETF constituent, the weight is reported as zero. The table clearly shows only a small proportion of constituent firms are selected by LASSO. Furthermore, we find LASSO tends to select more constituent firms for ETFs with low AUM: the total weight of LASSO-selected firms in the last row is around 30% for small ETFs, whereas the number is only around 4% for large ETFs. In unreported figures, we document that this pattern exists during our full sample period, and the total number of LASSO-selected firms is uncorrelated with the size of the ETF.

4. ETF Return Predictability

4.1. Overnight Signals

We start with the evaluation of overnight signals that are generated at the market open of each day. The market open period also tends to be the most liquid during the entire trading session and thus if our signals predict ETF returns from market open, they are relatively easier to be monetized than other intervals. To evaluate the predictive performance over different horizons, we construct different forward returns all starting from the same market open at 10:00 but end at different intervals. Let $r^{etf}_{j,d,2:i}$ be the return between the

[9] In the implementation of LASSO, we first standardize overnight stocks returns such that they have zero mean and unit standard deviation; the coefficients are multiplied by 10^4 for illustration purposes.

start of interval 2 (e.g., 10:00) to the end of interval i. That is, for $i = 2, 3, \ldots, 13$, $r^{etf}_{j,d,2:i}$ are the returns between 10:00 and 10:30, 10:00 and 11:00, and so on until between 10:00 and 16:00. We then evaluate the predictability of three overnight signals $f^{etf,LASSO}_{j,d,1}$, $f^{etf,news}_{j,d,1}$ and $f^{etf,non\text{-}news}_{j,d,1}$ in Equations (3), (4), and (6).

4.1.1. Fama-MacBeth (1973) regressions

We first use the Fama and MacBeth (1973) regressions to test the predictive power of signals. We perform the following cross-sectional regressions:

$$r^{etf}_{j,d,2:i} = \gamma_{0,d,2:i} + \gamma_{LASSO,d,2:i} f^{etf,LASSO}_{j,d,1} + \gamma_{news,d,2:i} f^{etf,news}_{j,d,1} + \gamma_{non\text{-}news,d,2:i} f^{etf,non\text{-}news}_{j,d,1} + \sum_{k=1}^{K} \gamma_{k,d,2:i} Z_{j,d-1,k} + \epsilon_{j,d,2:i}, \tag{7}$$

where the control variables $Z_{j,d-1,k}$ for ETF j are all measured at the end of day $d-1$. For a given forecasting horizon $2:i$, we obtain on each day the slope coefficients from these cross-sectional regressions. We then compute the time-series average of each slope coefficient to test if the signals are statistically significant in predicting forward returns. Our control variables include the illiquidity measure of Amihud (2002), realized volatility, ETF returns from day $d-252$ to day $d-21$ as a proxy for price momentum, and ETF size.

Table 3 reports the estimated average slope coefficients and their t-statistics for predicting different forward returns. The results clearly indicate that both the LASSO and news signals positively predict ETF returns. For the LASSO signal, the slope coefficients range from 0.107 with a t-statistic of 6.84 when predicting 30-minute ahead returns (returns between 10:00 and 10:30) to 0.189 with a t-statistic of 5.88 when predicting 210-minute ahead returns (returns between 10:00 and 15:30). For the news signal, the coefficients vary from 0.006 with a t-statistic of 6.48 when predicting 30-minute ahead returns to 0.014 with a t-statistic of 6.64 when predicting 210-minute ahead returns. The coefficients of the LASSO and news signals also tend to increase along with the forecasting horizon, indicating the predictive power of both signals could last through market close.

To put these regression coefficients into perspectives, the average cross-sectional standard deviation of overnight LASSO signal from Table 1 equals 10.33 bps. Thus, when predicting the open-to-close returns, the coefficient of 0.163 for the LASSO signal implies that a two-standard-deviation increase in the LASSO signal predicts a rise of approximately 8.5% in annual ETF returns ($2\times10.33\times10^{-4}\times0.163\times252 = 8.5\%$). Similarly, the coefficient of 0.013 for the news signal indicates that a two-standard-deviation increase in the news signal predicts a rise of around 10.5% per annum in future ETF returns ($2\times160.88\times10^{-4}\times0.013\times252 = 10.5\%$). Hence, the predictability of LASSO and news signals are both statistically and economically significant. In contrast, the non-news signal induces a reversal pattern in future ETF returns. Interestingly, this non-news reversal pattern is stronger than the overall reversal pattern induced by the raw ETF returns without adjusting for the news component (results untabulated).

Can the three overnight signals predict returns only until the same day market close or longer? To answer this question, we re-fit Equation (7) by replacing the dependent variable with interdaily returns from 10:00 on day d to 16:00 on day $d+k$ for each $k = 0, 1, ..., 4$, and present the coefficients and t-statistics in Table 4. The case with $k = 0$ (i.e., predicting the same day open-to-close return) in the first column is the same as the last column in Table 3 and is included for easy comparison. As the forecast horizon increases, the predictability of the overnight LASSO signal remains statistically significant. However, its strength decreases as is evident from the smaller t-statistics of 2.31 for $k = 4$ in comparison with the t-statistics of 5.02 for $k = 0$. A similar pattern is observed for the predictability of the overnight non-news signal. On the other hand, the overnight news signal remains a powerful predictor at longer horizon. For example, its slope coefficient increases from 0.013 when predicting one-day open-to-close return to 0.03 when predicting five-day returns, and the t-statistics are above four for all horizons, consistent with the long-lasting news momentum effect documented by Jiang et al. (2020). In summary, the Fama-MacBeth regressions support the predictability of our three signals.

4.1.2. Portfolio Sorts

We next form portfolios based on each of the three signals. At 10:00 of each trading day d, we sort ETFs into quintile portfolios based on their overnight LASSO, news, and non-news signals.[10] Then we compute equal-weighted returns on each quintile portfolio and a spread portfolio that buys ETFs in the top quintile with high signal values and sells ETFs in the bottom quintile with low signal values, with a holding period from 10:00 to 16:00 of the same trading day.

Table 5 summarizes the monthly portfolio returns in percentage, which are converted from daily returns by multiplication with a factor of 21 for ease of interpretation. The row labeled "High-Low" reports the average return spread between Quintiles 5 and 1, and the row labeled "CAPM Alpha" reports return spread adjusted by market exposure. The first two columns based on overnight LASSO and news signals show an increasing relation between signal values and future ETF returns. For instance, the average monthly returns increase from -0.09% in Quintile 1 with low overnight LASSO signal to 0.79% in Quintile 5 with high overnight LASSO signal, yielding a return spread of 0.88% per month with a t-statistic of 4.45. The quintile returns sorted by overnight news signal show a similar pattern with a high-minus-low return spread of 1.61% per month and a t-statistic of 7.34. In contrast, ETFs with high overnight non-news signal are associated with low future returns as is evident from the negative high-minus-low return spread of -2.89% per month with a t-statistic of -10.36. The rows labeled "CAPM Alpha" in Table 5 show that exposures to market factor cannot explain the return differences between High- and Low-quintiles. For example, the alpha of the spread portfolio based on the overnight LASSO signal is 0.88% per month and remains highly significant with a t-statistic of 4.44. The last row reports the annualized Sharpe ratio for each long-short portfolio. All trading strategies have a high Sharpe ratio during the full sample period. The annualized Sharpe ratios of the long-short portfolios based on overnight LASSO, news and non-news signals are 1.21, 2.15, and 2.56, respectively.

[10]The total number of ETFs per day ranges from 42 in the beginning of our sample to 224 in July 2020.

To illuminate how the long-short portfolios based on our three signals perform over time, we compute the cumulative profits for the resulting equal-weighted strategies based on an initial investment of $W_0 = \$1$. Specifically, we compound the daily returns of the spread portfolio sorted by overnight LASSO or news signals by measuring the cumulative return W_d on day d as follows:

$$W_d = W_{d-1} \times (1 + R_{high,d} - R_{low,d} + R_{rf,d}), \quad d = 1, 2, \ldots, \tag{8}$$

where $R_{high,d}$, $R_{low,d}$, and $R_{rf,d}$ are returns on Quintile 5, Quintile 1, and the risk-free rate on day d, respectively. For the spread portfolio generated by the overnight non-news signal, the cumulative return W_d is:

$$W_d = W_{d-1} \times (1 + R_{low,d} - R_{high,d} + R_{rf,d}), \quad d = 1, 2, \ldots. \tag{9}$$

Fig. 2 plots the trajectory of W_d starting from a given $W_0 = \$1$ initial investment at the start of January 2005. The portfolio value based on all three strategies continues to rise throughout the entire sample without experiencing major drawdowns. This evidence further supports the persistent predictability of all three signals.

4.2. Intraday Half-Hour Signals

Our prediction evaluation analyses have been focusing on the overnight signals computed at the market open. We find that all three signals predict the same-day intraday returns. We now evaluate the signal performance at a higher frequency of 30-minute intervals. First, we generate the signal values for every half-hour. Specifically, we follow the same procedure as in Equations (3), (4), and (6) to generate the intraday forecasts $f_{j,d,i}^{etf,LASSO}$, $f_{j,d,i}^{etf,news}$ and $f_{j,d,i}^{etf,non\text{-}news}$ for each 30-minute interval between 10:00 and 16:00 (i.e., $i = 2, \ldots, 13$).

After obtaining the intraday 30-minute signals, we run the Fama and MacBeth (1973) regressions to test for their predictive power for the subsequent 30-minute ETF returns.

Specifically, for each 30-minute interval $i = 2, 3, \ldots, 13$ on day d, we perform the following cross-sectional regressions:

$$r^{etf}_{j,d,i} = \gamma_{0,d,i} + \gamma_{LASSO,d,i} f^{etf,LASSO}_{j,d,i-1} + \gamma_{news,d,i} f^{etf,news}_{j,d,i-1} + \gamma_{non\text{-}news,d,i} f^{etf,non\text{-}news}_{j,d,i-1} + \sum_{k=1}^{K} \gamma_{k,d,i} Z_{j,d-1,k} + \epsilon_{j,d,i}. \tag{10}$$

Table 6 reports the estimated average slope coefficients and corresponding t-statistics for predicting 30-minute interval returns. The case with 10:30 in the first column is the same as the first column in Table 6 and is included for easy comparison. The results clearly show that the LASSO signals have positive and significant predictive power for every 30-minute interval, and the predictability exhibits a hump-shaped pattern over the trading day. For instance, the slope coefficients for the LASSO signal are around 0.1 near the open and close, and jump to around 0.2 in the middle of the day. The t-statistics of the LASSO signal coefficients remain above five across all 30-minute intervals.

For the news signal, Table 6 shows that only its overnight version can significantly predicts future returns. The coefficients of the news signal in other intervals are all insignificant. The news stories are less frequently released during trading hours. As shown in Panel A of Fig. 3, on average, 180 constituent firms have news after marker close, whereas less than 15 constituent firms have news within each 30-minute interval during market open. Panel B in Fig. 3 plots the intrady distribution of the number of ETFs with news. It shows that, on average, around 100 ETFs have constituents with news before 10:00 market open. However, only around 20 ETFs have news constituents within each 30-minute interval after market open. As a result, the per interval Fama-MacBeth regression might not have enough statistical power for testing the significance of the news signal. Finally, the non-news signal consistently induce short-term reversal across all 30-minute intervals.

5. Signal Combination

Section 4 shows that all three signals predict sector ETF returns. When there are multiple signals, the next question is then how to combine them for generating a composite signal that improves over any stand-alone signals. We investigate two approaches for signal combination below.

5.1. Simple Averaged Composite Signal

One common approach for signal combination is to simply take an "average" of them. The main motivation is that different signals protect each other from their individual errors (as long as they do not have all errors in the same direction). Since all three signals have different signal values, we first adjust their signal value to make them comparable. To do so, we cross-sectionally rank each signal period-by-period and map these ranks to the $[0, 1]$ interval.[11] As a result, we have a percentage rank score for each signal. We then take an average of those scores to obtain a new combined signal, namely $Composite^{AVG}$. By construction, the $Composite^{AVG}$ signal ignores the differences in signal strengthes.

5.2. Boosted Composite Signal

Our second signal combination approach uses the Gradient Boosted Regression Trees (GBRT), which is a powerful supervised learning method that has been widely used because it often achieves state-of-the-art prediction performance. We consider the following setup for the signal combination problem. Given an input of multiple signals, e.g., $\mathbf{f}^{etf}_{j,d,i} = (f^{etf,LASSO}_{j,d,i}, f^{etf,news}_{j,d,i}, f^{etf,non\text{-}news}_{j,d,i})$ and the forward returns, e.g., $r^{etf}_{j,d,i+1}$, GBRT uses K trees to predict the output:

$$GBRT_{j,d,i} = \sum_{k=1}^{K} \mathcal{T}_k(f^{etf}_{j,d,i}), \tag{11}$$

[11] Since the non-news signal predict future returns negatively, we first multiply the non-news signal by -1 before ranking.

where each $\mathcal{T}_k$ is a single regression tree, which partitions the space of input variables into rectangles and then fits a constant to each rectangle. GBRT then adds multiple regression trees (tree ensemble) to provide a more flexible learning method. By using the additive tree structures, GBRT can automatically capture the non-linearity and interactions that are challenging for linear models. However, such flexibility could lead to overfitting and thus poor out-of-sample performance. Regularization becomes extremely important in fitting GBRT.

Our implementation of GBRT relies on two mechanisms to reduce overfitting. First, we use cross-validation to tune two parameters: maximum depth that controls how deep each tree can grow and the learning rate that controls the relative contribution of each tree to the ensemble. The selected maximum depth is typically less than two, indicating no higher-order interactions among predictors. Second, we use early stopping to avoid overfitting the training data. Similar to the LASSO fit in Section 2.2.2, we perform a rolling fit, where each fit uses 1-year, 3-month, and 1-month data as our training, validation and testing samples, respectively. The first training and validation sample periods are from January 2005 to December 2005 and from January 2006 to March 2006, respectively. We call this composite signal $Composite^{GBRT}$.

5.3. Predictability of Composite Signals

We test the two composite signals $Composite^{AVG}$ and $Composite^{GBRT}$ by forming portfolios. Similar to the previous analysis, we focus on overnight composite signals generated at the market open. At 10:00 market open of each trading day t, we sort ETFs into quintile portfolios based on their composite signals. We compute equal-weighted returns on each quintile portfolio and a long-short portfolio that buys ETFs in the top quintile with high signal values and sell ETFs in the bottom quintile with low signal values. We hold the portfolios until market close of the same day.

Table 7 summarizes the monthly portfolio returns, which are converted from daily returns by multiplication with 21 for ease of interpretation. The row labeled "High-Low" report

average realized return of the spread portfolio. The result shows that $Composite^{AVG}$ has a realized monthly return of 2.80% which is comparable to that of the non-news signal, and a Sharpe ratio of 3.23 which is an significant improvement over any of the three individual signals. The last column shows that $Composite^{GBRT}$ further improves the performance. The equal-weighted average monthly return increases from -1.18% on Quintile 1 to 2.16% on Quintile 5, yielding a return spread of 3.34% per month with a t-statistic of 15.35 and a Sharpe ratio of 4.13. These numbers suggest that $Composite^{GBRT}$ dominates all other signals economically and statistically, illustrating the advantage of boosting tree in combining trading signals.

Fig. 4 plots the cumulative gain of the long-short portfolios based on different trading signals including three stand-alone signals and two composite signals. The y-axis presents the dollar value given $W_0 = \$1$ initial investment at the start of April 2006. The portfolio value based on all five strategies continues to rise throughout the entire sample. The $Composite^{GBRT}$ signal continues to generate superior performance during the full sample period. This result is consistent with the previous finding in Table 7 that the $Composite^{GBRT}$ signal has the highest return and Sharpe ratio.

5.4. Covid-19

The Covid-19 global pandemic wreaked havoc on the economy and financial markets. The CBOE Volatility Index (VIX) spiked up to more than 80 in March 2020, and the ETF market experienced unprecedented distress. According to BlackRock (2020),[12] trading volume in ETFs reached $5.41 trillion in March 2020, which is close to three times more than the average monthly trading volume in 2019.

What is the implication of the large turmoil in the ETF market for the return predictability we have documented? To answer this question, we take a separate look at the period from

[12]BlackRock, 2020, Lessons from COVID-19: ETFs as a Source of Stability. https://www.blackrock.com/corporate/literature/whitepaper/viewpoint-lessons-from-covid-19-etfs-as-a-source-of-stability-july-2020.pdf

January to July 2020 around the period of the Covid-19 pandemic. We conjecture that the unprecedented pandemic shock can distract investor attention and reduce the responsiveness of arbitrage capital to mispricings. As a result, the return predictability in high-frequency sector ETF returns could be even stronger during the Covid-19 pandemic.

Table 8 shows the results. The first column indicates that the average return to a long-short portfolio based on the LASSO signal reaches to 3.76% per month with an annualized Sharpe ratio of 4.47 during the pandemic; adjusting for the portfolio's market exposures slightly strengthens the results. The second and third columns show the results for the news and non-news signals: the average return on the long-short portfolio based on the news signal is 7.37% per month with a Sharpe ratio of 5.19, and the corresponding numbers for the non-news signal are −7.29% and 4.36. As shown in the fifth column, the GBRT composite signal generates a long-short portfolio with an average monthly return of 10.87% and an annualized Sharpe ratio of 8.34. The magnitude of return predictability is more than twice that of the full sample average. These results point to the notion that investor distraction and slow-moving capital during major market dislocations contribute to excessive return predictability.

6. Robustness

6.1. The GBRT Signal

In Section 2.2, we use LASSO to aggregate the information in the cross section of stock returns for ETF prediction. Alternatively, one may wonder if the GBRT is able to extract predictive information in stocks for ETF. To answer this question, we now fit forward ETF returns against the entire cross section of stock returns $\mathbf{r}^{stc}_{d,i} = (r^{stc}_{1,d,i}, r^{stc}_{1,d,i}, \ldots, r^{stc}_{p,d,i})$

$$r^{etf}_{j,d,i+1} = \sum_{k=1}^{K} \mathcal{T}_{j,i,k}(\mathbf{r}^{stc}_{d,i}), \tag{12}$$

where $\mathcal{T}_{j,i,k}$ is the k-th tree for predicting the jth ETF in interval i. As in Section 2.2.2, we perform a rolling fit on an ETF-by-ETF basis using training-validation-testing splits. Thus, we obtain the GBRT signal as:

$$f_{j,d,i}^{etf,GBRT} = \sum_{k=1}^{K} \hat{\mathcal{T}}_{j,i,k}(\mathbf{r}_{d,i}^{stc}). \tag{13}$$

First, we use Fama-MacBeth regressions to test the predictive power of the overnight GBRT signal for intraday ETF return prediction. We perform cross-sectional regressions similar to Equation (7) with a modification that replaces $f_{j,d,1}^{etf,LASSO}$ with $f_{j,d,1}^{etf,GBRT}$. Table 9 shows the regression coefficients of predicting intraday ETF returns using the GBRT signal. As can be seen, the GBRT signal positively and significantly predict ETF returns. For instance, the slope coefficients range from 0.114 with a t-statistic of 5.75 when predicting 30-minute ahead returns (returns between 10:00 and 10:30) to 0.203 with a t-statistic of 5.02 when predicting 150-minute ahead returns (returns between 10:00 and 12:30). To put these estimates into perspective, the average cross-sectional standard deviation of the GBRT signal is 11.85 bps. Thus, when predicting open-to-close return, the coefficient of 0.115 implies that a two-standard-deviation increase in the overnight GBRT signal predicts a rise of approximately 6.9% in annual returns ($2 \times 11.85 \times 10^{-4} \times 0.115 \times 252 = 6.9\%$). Therefore, the predictive power of overnight GBRT is both statistically and economically significant. Comparing Tables 3 and 9, we see that the LASSO signal is stronger than the GBRT signal when predicting open-to-close returns.

Next, we test the interdaily predictability of the overnight GBRT signal. We re-run Fama-MacBeth equation by replacing the dependent variables with cumulative returns from 10:00 on day d to 16:00 on day $d+k$ for $k = 0, 1, ..., 4$. Table 10 presents the coefficients and their t-statistics. The coefficient remains significant for day $d+1$ suggesting the predictability of overnight GBRT return lasts until the market close of the next day. Overall, GBRT's interdaily predictability appears weaker than LASSO reported in Table 4.

7. Conclusions

We have explored the high-frequency industry return predictability using a large sample of sector ETFs. We develop and compare two methodologies to generate return predictors: the machine learning-based LASSO predictor that aggregates information contained in the high-frequency stock price movements for the full cross section of the US equity market, and a simple, news-driven decomposition of ETF returns into the news and non-news ETF components.

We find that all the three predictors reliably forecast ETF returns. The LASSO and news predictors positively predict ETF returns out-of-sample; the non-news predictor negatively predicts ETF returns. These different signals contain independent information, and have different strengths, with the LASSO predictor providing continuous flows of information most powerful during trading hours and the news return offering sporadic information particularly useful during market close. We show that a composite signal combining the three predictors with boosted trees has superior power to forecast ETF returns. The strength of the return predictability more than doubles during the Covid-19 pandemic.

Appendix: High-frequency data cleaning

We begin by removing entries that satisfy at least one of the following criteria: a price less than or equal to zero; a trade size less than or equal to zero; corrected trades, i.e., trades with Correction Indicator, CORR, other than 0, 1, or 2; and an abnormal sale condition, i.e., trades for which the Sale Condition, COND, has a letter code other than @, *, E, F, @E, @F, *E, and *F. We then assign a single value to each variable for each second. If one or multiple transactions have occurred in that second, we calculate the sum of volumes, the sum of trades, and the volume-weighted average price within that second. If no transaction has occurred in that second, we enter zero for volume and trades. For the volume-weighted average price, we use the entry from the nearest previous second. Motivated by our analysis of the trading volume distribution across different exchanges over time, we purposely incorporate information from all exchanges covered by the TAQ database.

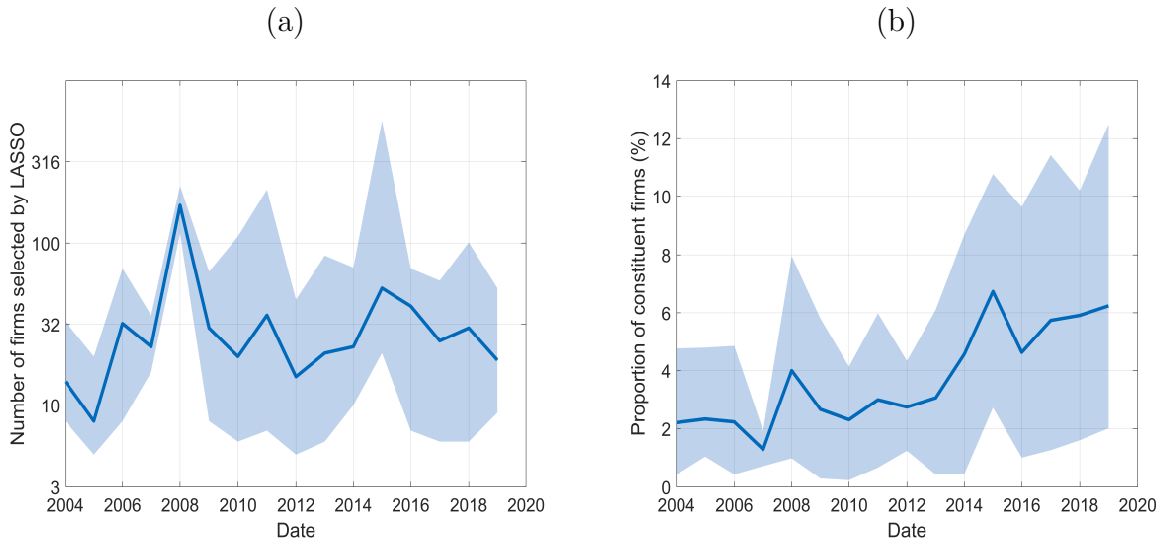

Fig. 1 Firms selected by LASSO

Panel A plots the mean and 90% confidence interval for the number of firms selected by LASSO each year during our sample period. Panel B plots the mean and 90% confidence interval for the proportion of constituent firms, defined by the number of overlapping firms between LASSO-selected and ETF constituent firms normalized by the total number of ETF constituents. The scale in Panel A is based on logarithm with base ten.

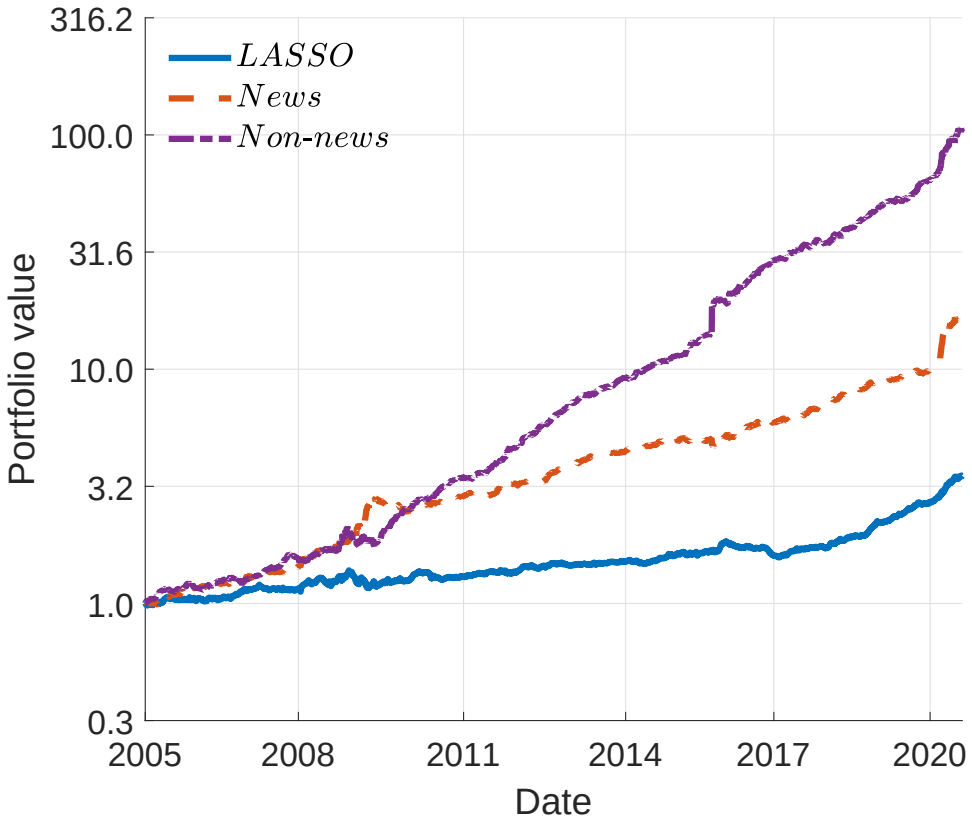

Fig. 2 Performance of strategies based on LASSO, news, and non-news signals

This figure shows cumulative gains of trading strategies based on the overnight LASSO, news, and non-news signals. *LASSO* is the overnight LASSO signal from Equation (3). *News* is the overnight news-driven signal from Equation (4). *Non-news* is the overnight non-news-driven signal from Equation (6). At the 10:00 market open of each trading day t, we sort stocks into quintile portfolios based on each signal. Then we compute equal-weighted returns on each quintile portfolio with a holding period from 10:00 to 16:00 of the same day. Let $R_{high,d}$ and $R_{low,d}$ be the returns on Quintile 5 and Quintile 1 on day d, respectively. The cumulative portfolio value based on the overnight LASSO or news signal is computed as $W_d = W_{d-1}(1 + R_{high,d} - R_{low,d} + R_{rf,d})$, where $R_{rf,d}$ is the risk-free rate on day t and the initial investment is $W_1 = \$1$. The cumulative portfolio value based on the overnight non-news signal is computed as $W_d = W_{d-1}(1 + R_{low,d} - R_{high,d} + R_{rf,d})$. Plotted is the time series of $\{W_d\}$. The scale in the figure is based on the logarithm with base ten.

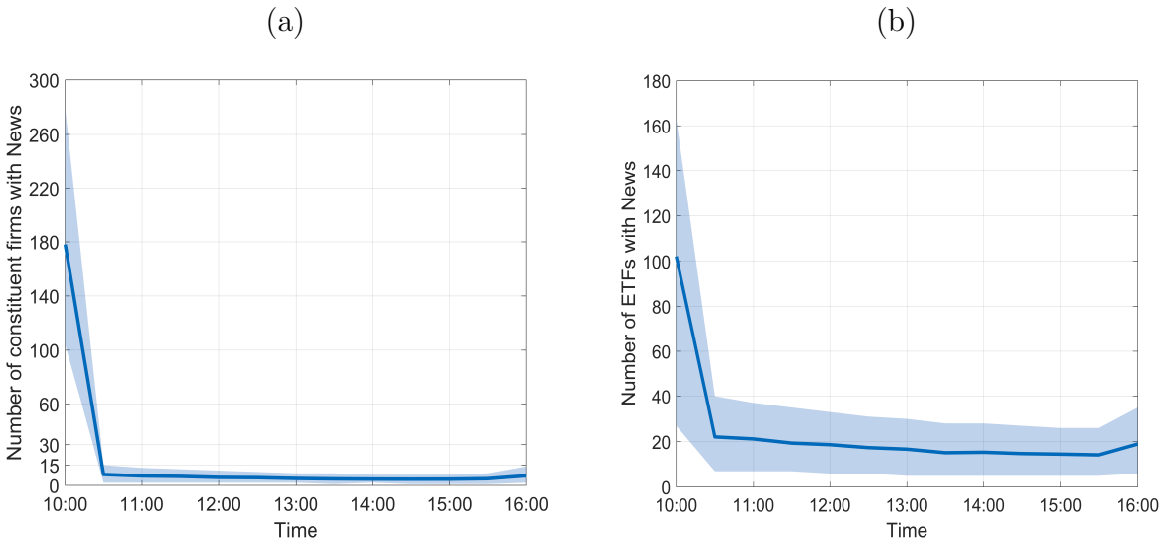

Fig. 3 Intrady news distributions

For each 30-minute interval, Panel A shows the average number of constituent firms with news and the corresponding 90% confidence interval; Panel B shows the average number of ETFs that have constituents with news and the 90% confidence interval. The x-axis displays the ending time of each 30-minute interval, e.g., 10:00 denotes the overnight interval between 16:00 the previous day and 10:00, 10:30 denotes the interval between 10:00 and 10:30, etc.

Table 1 Descriptive statistics

This table reports the descriptive statistics of our main variables. Our ETF sample consists of all sector ETFs with monthly holding data from the CRSP Mutual Fund database. Panel A reports the time-series average of the cross-sectional mean, standard deviation, and quantiles of each variable. Panel B reports the time-series average of the cross-sectional correlations of these variables. *LASSO* is the overnight LASSO signal from Equation (3). *News* is the overnight news-driven signal from Equation (4). *Non-news* is the overnight non-news-driven signal from Equation (6). *LASSO*, *News*, and *Non-news* are all reported in bps per day. *ILLQ* is the natural logarithm of the average daily ratio of the absolute ETF return to the dollar trading volume over the five-day period preceding each day. *RVOL* is annualized realized volatility, defined as the square root of the annualized realized variance, which is 252 times the sum of squared five-minute intraday ETF returns within each trading day. *Mom* for a given day d is the cumulative ETF return from day $d-252$ to day $d-21$. *Size* is the natural logarithm of the net asset value of the ETF, estimated by the product of the closing price and the number of shares outstanding from CRSP and is updated daily.

Panel A: Cross-sectional summary statistics

	Mean	Std	P1	P25	Median	P75	P99
LASSO	-0.30	10.33	-25.25	-4.47	-0.16	3.90	24.50
News	3.55	160.88	-411.20	-49.66	3.78	59.03	401.58
Non-news	0.00	52.71	-120.99	-27.22	0.22	27.57	120.89
ILLIQ	-20.30	2.42	-25.47	-21.78	-20.05	-18.60	-15.78
RVOL	22.20%	11.05%	9.24%	15.81%	19.77%	25.63%	57.59%
Mom	-0.09%	19.43%	-47.88%	-8.73%	-0.05%	8.52%	47.47%
Size	6.04	1.51	2.77	5.00	6.05	7.06	9.06

Panel B: Cross-sectional correlation

	LASSO	*News*	*Non-news*	ILLIQ	RVOL	Mom	Size
LASSO	1.000	0.017	0.033	-0.007	-0.023	0.000	0.006
News	0.017	1.000	0.000	0.001	-0.004	-0.001	0.003
Non-news	0.033	0.000	1.000	-0.004	0.005	0.001	0.007
ILLIQ	-0.007	0.001	-0.004	1.000	0.164	-0.002	-0.889
RVOL	-0.023	-0.004	0.005	0.164	1.000	0.001	-0.209
Mom	0.000	-0.001	0.001	-0.002	0.001	1.000	0.001
Size	0.006	0.003	0.007	-0.889	-0.209	0.001	1.000

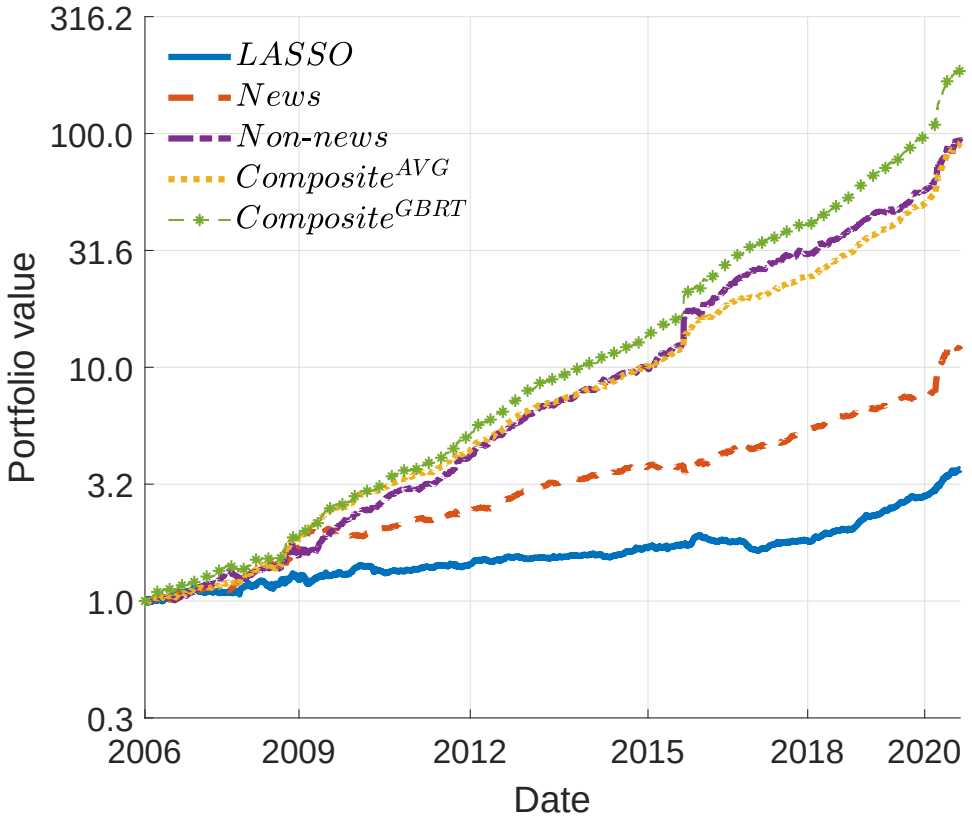

Fig. 4 Performance of strategies based on composite signals

This figure shows cumulative gains of trading strategies based on the overnight LASSO, news, and non-news signals, along with two composite signals formed by combining all three signals. *LASSO* is the overnight LASSO signal from Equation (3). *News* is the overnight news-driven signal from Equation (4). *Non-news* is the overnight non-news-driven signal from Equation (6). The first composite signal $Composite^{AVG}$ is constructed by simply averaging the three signals. The second composite signal $Composite^{GBRT}$ combines the three signals using Gradient Boosting Regression Trees (GBRT). At the 10:00 market open of each trading day t, we sort stocks into quintile portfolios based on each overnight signal. Then we compute equal-weighted returns on each quintile portfolio with a holding period from 10:00 to 16:00 of the same day. Let $R_{high,d}$ and $R_{low,d}$ be the returns on Quintile 5 and Quintile 1 on day d, respectively. The cumulative portfolio value based on each overnight signal except the non-news signal is computed as $W_d = W_{d-1}(1 + R_{high,d} - R_{low,d} + R_{rf,d})$, where $R_{rf,d}$ is the risk-free rate on day t and the initial investment is $W_1 = \$1$. The cumulative portfolio value based on the overnight non-news signal is computed as $W_d = W_{d-1}(1 + R_{low,d} - R_{high,d} + R_{rf,d})$. Plotted is the time series of $\{W_d\}$. The scale in the figure is based on the logarithm with base ten.

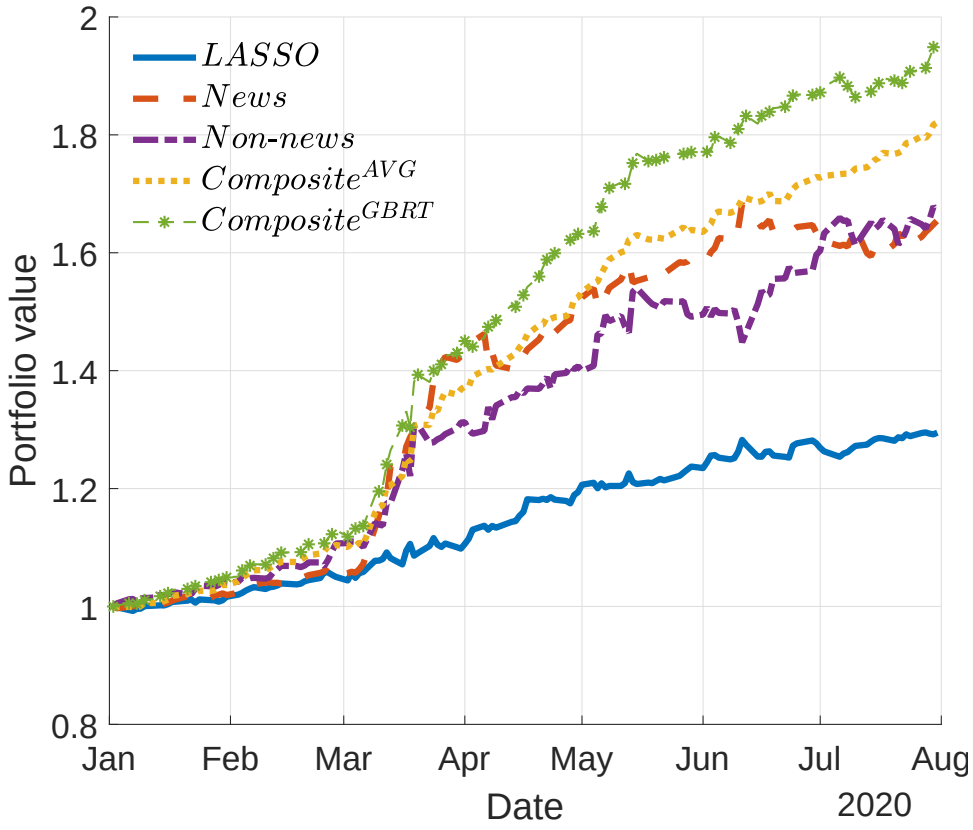

Fig. 5 Performance of strategies in 2020

This figure shows cumulative gains between January and July 2020 of trading strategies based on the overnight LASSO, news, and non-news signals, along with two composite signals formed by combining all three signals. *LASSO* is the overnight LASSO signal from Equation (3). *News* is the overnight news-driven signal from Equation (4). *Non-news* is the overnight non-news-driven signal from Equation (6). The first composite signal $Composite^{AVG}$ is constructed by simply averaging the three signals. The second composite signal $Composite^{GBRT}$ combines the three signals using Gradient Boosting Regression Trees (GBRT). At the 10:00 market open of each trading day t, we sort stocks into quintile portfolios based on each overnight signal. Then we compute equal-weighted returns on each quintile portfolio with a holding period from 10:00 to 16:00 of the same day. Let $R_{high,d}$ and $R_{low,d}$ be the returns on Quintile 5 and Quintile 1 on day d, respectively. The cumulative portfolio value based on each overnight signal except the non-news signal is computed as $W_d = W_{d-1}(1 + R_{high,d} - R_{low,d} + R_{rf,d})$, where $R_{rf,d}$ is the risk-free rate on day t and the initial investment is $W_1 = \$1$. The cumulative portfolio value based on the overnight non-news signal is computed as $W_d = W_{d-1}(1 + R_{low,d} - R_{high,d} + R_{rf,d})$. Plotted is the time series of $\{W_d\}$. The scale in the figure is based on the logarithm with base ten.

Table 2 Firms selected by LASSO

This table reports the top 20 LASSO regression coefficients in 2019 for nine ETFs with high, medium, and low AUMs, including XLK (Technology SPDR Fund), XLI (Industrial SPDR Fund), XLV (Health care SPDR Fund), FXU (Utilities AlphaDEX Fund), XTN (Transportation SPDR Fund), FXN (Energy Alpha DEX Fund), PSCC (Consumer Staples ETF), ENFR (Infrastructure ETF), and BIZD (VanEck Income ETF). For each ETF, we report coefficients of the top 20 firms ranked by the absolute value of their coefficients. The ETF constituent weights of selected firms are reported in parentheses. If a selected firm is not an ETF constituent, its weight is reported as zero. Positive constituent weights are in **bold**. The bottom row reports the total constituent weights of the top 20 firms.

Top 20 firms	High			Medium			Low		
	XLK	XLI	XLV	FXU	XTN	FXN	PSCC	ENFR	BIZD
1	4.11	2.65	4.27	0.61	3.37	3.77	11.79	7.69	8.02
	(0.00)	(0.00)	(0.00)	(**1.58**)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
2	1.82	0.36	1.87	0.32	2.18	1.44	11.52	6.33	5.68
	(0.00)	(**4.79**)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
3	1.57	0.35	1.73	0.23	1.53	1.42	7.70	3.77	2.32
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(**5.12**)	(0.00)	(**10.45**)	(0.00)
4	1.57	0.31	1.57	0.21	1.46	1.38	4.48	3.10	2.10
	(0.00)	(0.00)	(0.00)	(0.00)	(**4.25**)	(0.00)	(0.00)	(0.00)	(**18.61**)
5	1.47	0.28	1.38	0.20	1.39	1.30	3.31	1.68	1.11
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
6	1.30	0.26	1.25	0.15	1.38	1.28	3.19	1.62	1.09
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(**8.17**)	(0.00)
7	1.24	0.22	1.23	0.15	1.35	1.24	3.01	1.18	1.05
	(**2.14**)	(0.00)	(0.00)	(0.00)	(0.00)	(**0.17**)	(0.00)	(0.63)	(0.00)
8	1.21	0.20	1.18	0.13	1.19	1.21	2.71	1.00	1.03
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(**4.19**)	(0.00)	(0.00)
9	1.20	0.19	1.17	0.12	1.18	1.12	2.63	0.98	0.79
	(0.00)	(0.00)	(**3.54**)	(0.00)	(0.00)	(0.00)	(0.00)	(**2.75**)	(0.00)
10	1.15	0.17	1.17	0.12	1.16	1.11	2.62	0.71	0.78
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(**0.89**)	(0.00)
11	1.14	0.17	1.16	0.09	1.16	1.07	2.52	0.55	0.76
	(0.00)	(0.00)	(0.00)	(0.00)	(**3.32**)	(0.00)	(0.00)	(**6.57**)	(**0.98**)
12	1.12	0.16	1.15	0.09	1.13	1.05	2.28	0.40	0.75
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(**10.76**)	(0.00)	(0.00)
13	1.11	0.15	1.13	0.08	1.13	1.02	2.12	0.38	0.74
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(**1.34**)	(0.00)	(**1.32**)	(0.00)
14	1.11	0.14	1.12	0.07	1.11	1.01	2.09	0.37	0.71
	(0.00)	(0.00)	(0.00)	(**4.85**)	(0.00)	(0.00)	(**4.53**)	(0.00)	(0.00)
15	1.06	0.12	1.11	0.04	1.08	0.99	1.83	0.30	0.70
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(**5.56**)	(0.00)
16	1.06	0.12	1.05	0.03	1.08	0.99	1.75	0.29	0.69
	(0.00)	(**2.42**)	(0.00)	(0.00)	(0.00)	(0.00)	(**4.55**)	(0.00)	(**3.72**)
17	1.04	0.11	1.04	0.03	1.03	0.98	1.68	0.26	0.68
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
18	1.03	0.11	1.03	0.03	1.00	0.96	1.58	0.25	0.59
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(**4.47**)	(0.00)	(**0.88**)
19	1.02	0.11	1.01	0.02	0.99	0.92	1.53	0.24	0.50
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
20	1.02	0.09	1.00	0.02	0.97	0.90	1.49	0.20	0.50
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Total	(2.14)	(7.21)	(3.54)	(6.43)	(7.57)	(6.63)	(28.51)	(36.31)	(24.18)

Table 3 Intraday prediction

This table reports the estimated regression coefficients and Newey-West t-statistics (in parentheses) from Fama-MachBeth cross-sectional regressions predicting short-term ETF returns from 10:00 to different ending time of the day. Our ETF sample consists of all sector ETFs with monthly holding data from the CRSP Mutual Fund database. $LASSO$ is the overnight LASSO signal from Equation (3). $News$ is the overnight news-driven signal from Equation (4). $Non\text{-}news$ is the overnight non-news-driven signal from Equation (6). $ILLQ$ is the natural logarithm of the average daily ratio of the absolute ETF return to the dollar trading volume over the five-day period preceding each day. $RVOL$ is annualized realized volatility, defined as the square root of the annualized realized variance, which is 252 times the sum of squared five-minute intraday ETF returns within each trading day. Mom for a given day d is the cumulative ETF return from day $d-252$ to day $d-21$. $Size$ is the natural logarithm of the net asset value of the ETF, estimated by the product of the closing price and the number of shares outstanding from CRSP and is updated daily.

Forecasting horizon ending time	10:30	11:00	11:30	12:00	12:30	13:00	13:30	14:00	14:30	15:00	15:30	16:00
Intercept	0.174	2.878	3.338	3.451	5.026	5.260	5.288	5.904	4.784	3.571	6.279	4.720
	(0.13)	(1.89)	(2.01)	(1.98)	(2.68)	(2.71)	(2.70)	(2.88)	(2.28)	(1.34)	(2.77)	(1.92)
$LASSO$	0.107	0.120	0.122	0.132	0.145	0.140	0.155	0.158	0.170	0.189	0.189	0.163
	(6.84)	(6.05)	(5.49)	(5.46)	(5.79)	(5.41)	(5.74)	(5.68)	(5.86)	(6.03)	(5.88)	(5.02)
$News$	0.006	0.008	0.008	0.010	0.011	0.011	0.012	0.012	0.012	0.013	0.014	0.013
	(6.48)	(6.44)	(5.62)	(6.49)	(6.68)	(6.32)	(6.33)	(6.13)	(6.26)	(6.46)	(6.64)	(6.30)
$Non\text{-}news$	-0.080	-0.089	-0.089	-0.088	-0.087	-0.087	-0.086	-0.088	-0.087	-0.090	-0.089	-0.091
	(-24.62)	(-21.09)	(-18.77)	(-17.32)	(-16.25)	(-15.88)	(-14.65)	(-14.51)	(-14.47)	(-14.18)	(-13.82)	(-13.78)
ILLIQ	-0.025	0.204	0.195	0.150	0.267	0.327	0.268	0.282	0.204	0.091	0.241	0.166
	(-0.24)	(1.80)	(1.59)	(1.11)	(1.85)	(2.13)	(1.73)	(1.77)	(1.27)	(0.45)	(1.38)	(0.89)
RVOL	-1.724	-1.873	-5.129	-6.788	-8.073	-6.842	-9.209	-11.003	-12.292	-11.324	-11.182	-8.831
	(-1.01)	(-0.82)	(-1.94)	(-2.32)	(-2.56)	(-2.13)	(-2.75)	(-3.16)	(-3.48)	(-3.10)	(-2.94)	(-2.23)
Mom	1.228	1.327	1.666	2.035	1.381	1.846	1.301	1.304	1.445	1.935	1.770	1.751
	(2.24)	(2.00)	(2.30)	(2.56)	(1.63)	(2.09)	(1.42)	(1.38)	(1.51)	(1.91)	(1.76)	(1.73)
Size	-0.051	0.204	0.170	0.056	0.190	0.362	0.217	0.159	0.082	-0.055	0.072	0.017
	(-0.29)	(1.02)	(0.79)	(0.22)	(0.71)	(1.27)	(0.74)	(0.53)	(0.26)	(-0.15)	(0.21)	(0.05)
Adj-R^2(%)	15.93	17.32	17.64	17.52	17.74	17.73	17.73	17.90	18.03	18.27	18.30	18.46

Table 4 Interdaily prediction

This table reports the estimated regression coefficients and Newey-West t-statistics (in parentheses) from Fama-MachBeth cross-sectional regressions predicting cumulative ETF returns from 10:00 on day d to 16:00 on day $d+k$ for each $k = 0, 1, ..., 4$. Our ETF sample consists of all sector ETFs with monthly holding data from the CRSP Mutual Fund database. $LASSO$ is the overnight LASSO signal from Equation (3). $News$ is the overnight news-driven signal from Equation (4). $Non\text{-}news$ is the overnight non-news-driven signal from Equation (6). $ILLQ$ is the natural logarithm of the average daily ratio of the absolute ETF return to the dollar trading volume over the five-day period preceding each day. $RVOL$ is annualized realized volatility, defined as the square root of the annualized realized variance, which is 252 times the sum of squared five-minute intraday ETF returns within each trading day. Mom for a given day d is the cumulative ETF return from day $d-252$ to day $d-21$. $Size$ is the natural logarithm of the net asset value of the ETF, estimated by the product of the closing price and the number of shares outstanding from CRSP and is updated daily.

Forecasting horizon ending time	16:00 d	16:00 $d+1$	16:00 $d+2$	16:00 $d+3$	16:00 $d+4$
Intercept	4.720	11.215	20.727	28.045	38.354
	(1.92)	(2.54)	(3.16)	(3.28)	(3.64)
$LASSO$	0.163	0.147	0.161	0.190	0.275
	(5.02)	(2.42)	(2.00)	(1.90)	(2.31)
$News$	0.013	0.022	0.027	0.030	0.030
	(6.30)	(6.29)	(5.53)	(5.22)	(4.50)
$Non\text{-}news$	-0.091	-0.067	-0.074	-0.083	-0.079
	(-13.78)	(-5.77)	(-4.86)	(-4.53)	(-3.69)
ILLIQ	0.166	0.397	0.911	1.326	1.984
	(0.89)	(1.14)	(1.71)	(1.90)	(2.37)
RVOL	-8.831	-20.494	-28.212	-35.200	-41.588
	(-2.23)	(-2.80)	(-2.63)	(-2.46)	(-2.41)
Mom	1.751	2.076	2.661	1.457	1.346
	(1.73)	(1.07)	(0.97)	(0.51)	(0.39)
Size	0.017	0.246	0.956	1.839	2.827
	(0.05)	(0.34)	(0.86)	(1.25)	(1.60)
Adj-R^2(%)	18.46	18.43	18.18	17.94	17.91

Table 5 Performance of quintile portfolios sorted on the basis of different signals

This table reports the performance of quintile portfolios formed on the basis of different overnight signals. At 10:00 of each trading day d, we sort stocks into quintile portfolios based on their overnight LASSO, news, and non-news signals. Then we compute equal-weighted returns on each quintile portfolio and a spread portfolio that buys stocks in the top quintile with high signal values and sells stocks in the bottom quintile with low signal values, with a holding period from 10:00 on day d until market close of the same day. Our ETF sample consists of all sector ETFs with monthly holding data from the CRSP Mutual Fund database. *LASSO* is the overnight LASSO signal from Equation (3). *News* is the overnight news-driven signal from Equation (4). *Non-news* is the overnight non-news-driven signal from Equation (6). We multiply daily returns by 21 to get monthly returns in percentage. The row labeled "High-Low" reports the average return spread between Quintiles 5 and 1, and the row labeled "CAPM Alpha" reports return spread adjusted by market exposure. The row labeled "Annual Sharpe Ratio" reports the annualized Sharpe ratio for each long-short portfolio.

	LASSO	*News*	*Non-news*
1 (Low)	-0.09	-0.42	1.92
2	0.34	0.13	0.80
3	0.46	0.66	0.37
4	0.52	0.79	-0.13
5 (High)	0.79	1.20	-0.97
High-Low	0.88	1.61	-2.89
	(4.45)	(7.34)	(-10.36)
CAPM Alpha	0.88	1.61	-2.87
	(4.44)	(7.35)	(-10.21)
Annual Sharpe Ratio	1.21	2.15	2.56

Table 6 Predicting next half-hour returns

This table reports the estimated regression coefficients and Newey-West t-statistics (in parentheses) from Fama-MachBeth cross-sectional regressions predicting every intraday half-hour ETF return in interval i (i.e., $i = 2, ..., 13$) by signals calculated using inputs from interval $i-1$. Our ETF sample consists of all sector ETFs with monthly holding data from the CRSP Mutual Fund database. $LASSO$ is the LASSO signal from Equation (3). $News$ is the news-driven signal from Equation (4). $Non\text{-}news$ is the non-news-driven signal from Equation (6). $ILLQ$ is the natural logarithm of the average daily ratio of the absolute ETF return to the dollar trading volume over the five-day period preceding each day. $RVOL$ is annualized realized volatility, defined as the square root of the annualized realized variance, which is 252 times the sum of squared five-minute intraday ETF returns within each trading day. Mom for a given day d is the cumulative ETF return from day $d-252$ to day $d-21$. $Size$ is the natural logarithm of the net asset value of the ETF, estimated by the product of the closing price and the number of shares outstanding from CRSP and is updated daily.

30-minute interval ending time	10:30	11:00	11:30	12:00	12:30	13:00	13:30	14:00	14:30	15:00	15:30	16:00
Intercept	0.174	0.727	1.888	0.305	0.433	0.778	-0.286	-0.449	-0.885	-1.537	1.110	-1.528
	(0.13)	(0.69)	(2.04)	(0.38)	(0.56)	(1.02)	(-0.39)	(-0.57)	(-0.68)	(-1.12)	(1.24)	(-1.51)
$LASSO$	0.107	0.182	0.221	0.146	0.215	0.229	0.204	0.194	0.143	0.157	0.149	0.101
	(6.84)	(10.31)	(13.70)	(9.57)	(13.63)	(14.26)	(11.48)	(11.04)	(7.64)	(11.69)	(7.04)	(5.80)
$News$	0.006	0.105	-0.049	0.125	0.011	-0.001	0.047	-0.081	0.089	0.121	-0.011	0.835
	(6.48)	(1.32)	(-0.57)	(1.19)	(0.39)	(-0.11)	(0.90)	(-1.57)	(0.84)	(0.90)	(-0.73)	(1.10)
$Non\text{-}news$	-0.080	-0.092	-0.115	-0.107	-0.126	-0.142	-0.152	-0.159	-0.184	-0.165	-0.178	-0.215
	(-24.62)	(-21.94)	(-28.35)	(-22.56)	(-29.49)	(-34.29)	(-31.38)	(-33.08)	(-31.18)	(-36.51)	(-31.42)	(-39.14)
ILLIQ	-0.025	0.001	0.067	-0.028	0.013	0.099	-0.043	-0.037	-0.091	-0.118	-0.007	-0.118
	(-0.24)	(0.01)	(1.02)	(-0.51)	(0.24)	(1.88)	(-0.79)	(-0.68)	(-0.90)	(-1.22)	(-0.10)	(-1.63)
RVOL	-1.724	-0.272	-1.744	-1.296	-0.357	-0.203	-1.396	-1.773	-2.429	1.397	-0.422	0.566
	(-1.01)	(-0.19)	(-1.51)	(-1.18)	(-0.36)	(-0.21)	(-1.40)	(-1.95)	(-1.63)	(1.56)	(-0.44)	(0.52)
Mom	1.228	0.415	0.048	0.309	-0.391	0.637	-0.561	-0.325	1.686	0.259	-0.121	0.027
	(2.24)	(0.95)	(0.12)	(0.86)	(-1.28)	(1.93)	(-1.77)	(-1.10)	(2.70)	(0.92)	(-0.37)	(0.08)
Size	-0.051	-0.179	-0.021	-0.081	-0.018	0.196	-0.078	-0.042	-0.068	-0.088	-0.104	-0.138
	(-0.29)	(-1.45)	(-0.18)	(-0.81)	(-0.20)	(2.16)	(-0.83)	(-0.46)	(-0.37)	(-0.71)	(-0.90)	(-1.08)
Adj-R^2(%)	15.93	14.35	13.91	14.79	13.54	13.75	14.02	14.07	14.29	15.04	14.77	14.99

Table 7 Performance of quintile portfolios sorted on the basis of composite signals

This table reports the performance of quintile portfolios formed on the basis of different overnight signals. At 10:00 of each trading day d, we sort stocks into quintile portfolios based on their overnight LASSO, news, and non-news signals, along with two composite signals formed by combining all three signals. Then we compute equal-weighted returns on each quintile portfolio and a spread portfolio that buys stocks in the top quintile with high signal values and sells stocks in the bottom quintile with low signal values, with a holding period from 10:00 on day d until market close of the same day. Our ETF sample consists of all sector ETFs with monthly holding data from the CRSP Mutual Fund database. *LASSO* is the overnight LASSO signal from Equation (3). *News* is the overnight news-driven signal from Equation (4). *Non-news* is the overnight non-news-driven signal from Equation (6). The first composite signal $Composite^{AVG}$ is constructed by simply averaging the three signals. The second composite signal $Composite^{GBRT}$ combines the three signals using Gradient Boosting Regression Trees (GBRT). We multiply daily returns by 21 to get monthly returns in percentage. The row labeled "High-Low" reports the average return spread between Quintiles 5 and 1, and the row labeled "CAPM Alpha" reports return spread adjusted by market exposure. The row labeled "Annual Sharpe Ratio" reports the annualized Sharpe ratio for each long-short portfolio.

	$LASSO$	$News$	$Non\text{-}news$	$Composite^{AVG}$	$Composite^{BRT}$
1 (Low)	-0.09	-0.42	1.92	-0.95	-1.18
2	0.34	0.13	0.80	0.07	-0.03
3	0.46	0.66	0.37	0.37	0.25
4	0.52	0.79	-0.13	0.80	0.67
5 (High)	0.79	1.20	-0.97	1.85	2.16
High-Low	0.88	1.61	-2.89	2.80	3.34
	(4.45)	(7.34)	(-10.36)	(13.83)	(15.35)
CAPM Alpha	0.88	1.61	-2.87	2.80	3.33
	(4.44)	(7.35)	(-10.21)	(13.78)	(15.30)
Annual Sharpe Ratio	1.21	2.15	2.56	3.23	4.13

Table 8 Performance of quintile portfolios sorted on the basis of composite signals in 2020

This table reports the performance of quintile portfolios formed on the basis of different overnight signals during the sample period from January to July 2020. At 10:00 of each trading day d, we sort stocks into quintile portfolios based on their overnight LASSO, news, and non-news signals, along with two composite signals formed by combining all three signals. Then we compute equal-weighted returns on each quintile portfolio and a spread portfolio that buys stocks in the top quintile with high signal values and sells stocks in the bottom quintile with low signal values, with a holding period from 10:00 on day d until market close of the same day. Our ETF sample consists of all sector ETFs with monthly holding data from the CRSP Mutual Fund database. *LASSO* is the overnight LASSO signal from Equation (3). *News* is the overnight news-driven signal from Equation (4). *Non-news* is the overnight non-news-driven signal from Equation (6). The first composite signal $Composite^{AVG}$ is constructed by simply averaging the three signals. The second composite signal $Composite^{GBRT}$ combines the three signals using Gradient Boosting Regression Trees (GBRT). We multiply daily returns by 21 to get monthly returns in percentage. The row labeled "High-Low" reports the average return spread between Quintiles 5 and 1, and the row labeled "CAPM Alpha" reports return spread adjusted by market exposure. The row labeled "Annual Sharpe Ratio" reports the annualized Sharpe ratio for each long-short portfolio.

	$LASSO$	$News$	$Non\text{-}news$	$Composite^{AVG}$	$Composite^{GBRT}$
1 (Low)	-0.30	-1.93	5.19	-3.17	-4.24
2	1.75	-0.12	1.07	0.15	0.00
3	-0.07	2.06	0.69	0.90	0.98
4	0.56	1.73	0.54	2.13	1.62
5 (High)	3.47	5.45	-2.11	5.48	6.63
High-Low	3.76	7.37	-7.29	8.65	10.87
	(4.51)	(3.04)	(-3.93)	(4.77)	(5.59)
CAPM Alpha	3.90	7.17	-7.24	8.65	10.92
	(4.79)	(3.00)	(-3.80)	(4.79)	(5.62)
Annual Sharpe Ratio	4.47	5.19	4.36	7.31	8.34

Table 9 Intraday prediction – GBRT signal

This table reports the estimated regression coefficients and Newey-West t-statistics (in parentheses) from Fama-MachBeth cross-sectional regressions predicting short-term ETF returns from 10:00 to different ending time of the day. Our ETF sample consists of all sector ETFs with monthly holding data from the CRSP Mutual Fund database. $GBRT$ is the overnight GBRT signal from Equation (13). $News$ is the overnight news-driven signal from Equation (4). $Non\text{-}news$ is the overnight non-news-driven signal from Equation (6). $ILLQ$ is the natural logarithm of the average daily ratio of the absolute ETF return to the dollar trading volume over the five-day period preceding each day. $RVOL$ is annualized realized volatility, defined as the square root of the annualized realized variance, which is 252 times the sum of squared five-minute intraday ETF returns within each trading day. Mom for a given day d is the cumulative ETF return from day $d-252$ to day $d-21$. $Size$ is the natural logarithm of the net asset value of the ETF, estimated by the product of the closing price and the number of shares outstanding from CRSP and is updated daily.

30-minute interval ending time	10:30	11:00	11:30	12:00	12:30	13:00	13:30	14:00	14:30	15:00	15:30	16:00
Intercept	0.171	2.345	2.861	2.710	3.636	4.315	3.998	4.302	3.531	2.487	5.209	3.772
	(0.81)	(1.75)	(1.98)	(1.74)	(2.17)	(2.40)	(2.14)	(2.21)	(1.82)	(0.97)	(2.48)	(1.74)
$GBRT$	0.114	0.156	0.170	0.158	0.203	0.197	0.193	0.178	0.162	0.151	0.151	0.115
	(5.75)	(5.12)	(4.78)	(4.30)	(5.02)	(4.87)	(4.52)	(4.11)	(3.64)	(3.28)	(3.36)	(3.42)
$News$	0.007	0.009	0.009	0.011	0.012	0.012	0.013	0.013	0.014	0.015	0.015	0.015
	(7.78)	(7.38)	(7.00)	(7.54)	(7.69)	(7.64)	(7.63)	(7.40)	(7.53)	(7.84)	(8.00)	(7.71)
$Non\text{-}news$	-0.077	-0.086	-0.089	-0.086	-0.085	-0.085	-0.083	-0.084	-0.082	-0.083	-0.083	-0.088
	(-25.20)	(-20.90)	(-18.89)	(-16.89)	(-16.05)	(-15.63)	(-14.69)	(-14.46)	(-13.76)	(-13.33)	(-13.18)	(-13.70)
ILLIQ	0.017	0.087	0.075	0.036	0.114	0.207	0.160	0.173	0.151	0.042	0.184	0.086
	(0.20)	(0.91)	(0.73)	(0.31)	(0.93)	(1.53)	(1.12)	(1.23)	(1.05)	(0.22)	(1.21)	(0.55)
RVOL	-2.074	-2.391	-5.406	-6.871	-7.828	-7.829	-9.228	-10.714	-11.753	-11.124	-11.022	-9.009
	(-1.32)	(-1.12)	(-2.21)	(-2.57)	(-2.71)	(-2.61)	(-3.01)	(-3.38)	(-3.63)	(-3.28)	(-3.20)	(-2.52)
Mom	0.303	0.186	-0.056	0.212	-0.479	0.060	-0.425	-0.402	-0.092	0.137	0.079	0.343
	(0.61)	(0.30)	(-0.08)	(0.29)	(-0.62)	(0.07)	(-0.51)	(-0.47)	(-0.10)	(0.14)	(0.09)	(0.35)
Size	-0.038	-0.066	-0.117	-0.168	-0.064	0.170	0.093	0.075	0.101	-0.022	0.093	-0.046
	(-0.26)	(-0.39)	(-0.62)	(-0.80)	(-0.28)	(0.69)	(0.36)	(0.29)	(0.37)	(-0.07)	(0.32)	(-0.15)
Adj-R^2(%)	15.54	16.90	17.10	16.85	17.04	16.98	17.15	17.36	17.47	17.62	17.70	17.88

Table 10 Interdaily prediction – GBRT signal

This table reports the estimated regression coefficients and Newey-West t-statistics (in parentheses) from Fama-MachBeth cross-sectional regressions predicting cumulative ETF returns from 10:00 on day d to 16:00 on day $d+k$ for each $k = 0, 1, ..., 4$. Our ETF sample consists of all sector ETFs with monthly holding data from the CRSP Mutual Fund database. $GBRT$ is the overnight GBRT signal from Equation (13). $News$ is the overnight news-driven signal from Equation (4). $Non\text{-}news$ is the overnight non-news-driven signal from Equation (6). $ILLQ$ is the natural logarithm of the average daily ratio of the absolute ETF return to the dollar trading volume over the five-day period preceding each day. $RVOL$ is annualized realized volatility, defined as the square root of the annualized realized variance, which is 252 times the sum of squared five-minute intraday ETF returns within each trading day. Mom for a given day d is the cumulative ETF return from day $d-252$ to day $d-21$. $Size$ is the natural logarithm of the net asset value of the ETF, estimated by the product of the closing price and the number of shares outstanding from CRSP and is updated daily.

Forecasting horizon ending time	16:00 d	16:00 $d+1$	16:00 $d+2$	16:00 $d+3$	16:00 $d+4$
Intercept	3.772	8.679	15.801	20.596	26.584
	(1.74)	(2.09)	(2.53)	(2.50)	(2.64)
$GBRT$	0.115	0.099	0.112	0.177	0.169
	(3.42)	(2.16)	(1.56)	(0.92)	(0.34)
$News$	0.015	0.022	0.025	0.027	0.027
	(7.71)	(6.86)	(5.83)	(5.11)	(4.50)
$Non\text{-}news$	-0.088	-0.070	-0.081	-0.080	-0.076
	(-13.70)	(-6.26)	(-5.50)	(-4.57)	(-3.74)
ILLIQ	0.086	0.264	0.705	0.973	1.394
	(0.55)	(0.85)	(1.47)	(1.54)	(1.83)
RVOL	-9.009	-15.787	-20.604	-22.469	-26.366
	(-2.52)	(-2.26)	(-2.01)	(-1.66)	(-1.61)
Mom	0.343	-1.311	-1.183	-2.410	-2.524
	(0.35)	(-0.77)	(-0.51)	(-0.92)	(-0.80)
Size	-0.046	0.209	1.041	1.610	2.459
	(-0.15)	(0.33)	(1.04)	(1.21)	(1.53)
Adj-R^2(%)	17.88	17.53	17.16	16.89	16.76

References

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets 5, 31–56.

Ang, A., Chen, J., 2002. Asymmetric correlations of equity portfolios. Journal of Financial Economics 63, 443–494.

Bollerslev, T., Li, S. Z., Todorov, V., 2016. Roughing up beta: continuous vs. discontinuous betas, and the cross-section of expected stock returns. Journal of Financial Economics 120, 464–490.

Bollerslev, T., Li, S. Z., Zhao, B., 2020. Good volatility, bad volatility, and the cross section of stock returns. Journal of Financial and Quantitative Analysis 55, 1–31.

Boudoukh, J., Richardson, M., Whitelaw, R. F., 1994. Industry returns and the fisher effect. the Journal of Finance 49, 1595–1615.

Bustamante, M. C., Donangelo, A., 2017. Product market competition and industry returns. The Review of Financial Studies 30, 4216–4266.

Chen, L., Pelger, M., Zhu, J., 2019. Deep learning in asset pricing. Available at SSRN 3350138 .

Chinco, A., Clark-Joseph, A., Ye, M., 2019. Sparse signals in the cross-section of returns. Journal of Finance 74, 449–492.

Cohen, L., Frazzini, A., 2008. Economic links and predictable returns. Journal of Finance 63, 1977–2011.

DellaVigna, S., Pollet, J. M., 2007. Demographics and industry returns. American Economic Review 97, 1667–1702.

Donoho, D. L., 2006. For most large underdetermined systems of equations, the minimal L_1-norm near-solution approximates the sparsest near-solution. Communications on Pure and Applied Mathematics 59, 907–934.

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

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

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

Hong, H., Torous, W., Valkanov, R., 2007. Do industries lead stock markets? Journal of Financial Economics 83, 367–396.

Jiang, H., Li, S., Wang, H., 2020. Pervasive underreaction: Evidence from high-frequency data. Journal of Financial Economics, forthcoming .

Jiang, H., Sun, Z., 2015. News and corporate bond liquidity. Unpublished working paper, Michigan State University and University of California, East Lansing and Irvine.

Kelley, E. K., Tetlock, P. C., 2017. Retail short selling and stock prices. Review of Financial Studies 30, 801–834.

Menzly, L., Ozbas, O., 2010. Market segmentation and cross-predictability of returns. Journal of Finance 65, 1555–1580.

Moskowitz, T. J., Grinblatt, M., 1999. Do industries explain momentum? The Journal of finance 54, 1249–1290.

Rapach, D., Zhou, G., 2013. Forecasting stock returns. Handbook of Economic Forecasting 2A, 328–383.

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