Order Flow and Cryptocurrency Returns

August 2024

Abstract

We assess the information content of order flow in the emerging cryptocurrency market. Our analysis is based on a large cross-section of cryptocurrencies and a set of disaggregated order flows denominated in 11 major currencies that reflect world order flow. Using machine learning techniques for out-of-sample prediction, we find that order flow has strong predictive ability for daily and weekly cryptocurrency returns. For example, long-short portfolios sorted on machine learning forecasts conditioning on daily order flow exhibit mean returns of up to 1.2% per day with an annualized Sharpe ratio of about 5. Order flow tends to dominate economic fundamentals for prediction, especially for weekly returns. Finally, our findings indicate that order flow has a transitory effect for daily returns (short-term reversal) and a permanent effect for weekly returns (long-term momentum).

JEL classification: C58, G11, G15.

Keywords: Order Flow, Cryptocurrencies; Machine Learning; Forecasting; Portfolio Sorts.

1 Introduction

The emergence of cryptocurrencies over the past decade has had a transformational effect in financial markets. It has facilitated the provision of decentralized financial services by performing financial transactions in a less costly, more efficient and more transparent manner compared to traditional financial intermediaries (see, e.g., Harvey, Ramachandran and Santoro, 2021). Accordingly, the growth of the cryptocurrency market has been astounding. Market capitalization has increased from $11 billion US dollars (USD) in 2014 to a peak of $2.9 trillion USD in 2021. Similarly, daily trading volume has risen from $40 million USD in 2014 to a peak of $308 billion USD in 2021.[1] As a result of this phenomenal growth, cryptocurrencies are at the centre of current public discourse, investment practice and academic research.

This paper contributes to the emerging cryptocurrency literature by assessing the informational content of order flow for explaining and predicting cryptocurrency returns. Specifically, we address three related questions: (1) can US and world order flow explain the cross-section of cryptocurrency returns?; (2) does US and world order flow convey predictive information that can generate reliable out-of-sample forecasts for the cross-section of cryptocurrency returns?; and (3) do machine learning techniques (linear and non-linear) improve the predictive power of US and world order flow relative to standard predictive approaches? Although for fiat currencies order flow has emerged as an exciting alternative to traditional economic models of exchange rate determination, for cryptocurrencies this remains an open question.[2]

[1] According to coinmarketcap.com, the market capitalization of the cryptocurrency market on January 1, 2014 was $10.8 billion USD and subsequently peaked on November 8, 2021 at $2.9 trillion USD. Daily volume was $40.2 million USD on January 1, 2014 and peaked on April 16, 2021 at $308 billion USD.

[2] For the theory and empirical evidence on the effect of order flow on fiat exchange rates, see among others, Evans and Lyons (2002, 2005), Evans (2010), Rime, Sarno and Sojli (2010), Evans and Rime (2012) and Menkhoff et al. (2016).

Order flow is a measure of the net demand for a particular cryptocurrency defined as the value of buyer-initiated orders minus the value of seller-initiated orders. In general, there are three ways in which order flow might be related to cryptocurrency returns. First, it is possible that order flow is in fact unrelated to cryptocurrency returns. In this case, it is economic fundamentals, not order flow, that may explain and predict cryptocurrency returns in the short- or the long-run. We refer to this view as the "no-predictability view" and it is essentially the null hypothesis tested in our empirical analysis.

Second, it is possible that order flow has a transitory effect on cryptocurrency returns. In this case, order flow has a temporary but not a permanent effect. The market microstructure literature associates transitory effects with changes in liquidity, price pressure and temporary preference shocks (see, e.g., Froot and Ramadorai, 2005 and Menkhoff et al., 2016). We refer to this view as the "transitory view."

Finally, third, it is possible that order flow has a permanent effect on cryptocurrency returns. The market microstructure literature associates permanent effects with asymmetric information among market participants so that trades convey information that has a persistent effect on security prices (see, e.g., Glosten and Milgrom, 1985, Kyle, 1985 and Hasbrouck, 1988, 1991). According to this view, if trades convey information about future economic fundamentals that is not currently known by all market participants, then order flow acts as the key vehicle that impounds views about economic fundamentals into cryptocurrency returns through the process of price discovery (see, e.g., Evans and Lyons, 2002, 2005 and Menkhoff et al., 2016). We refer to this view as the "permanent view."

Our empirical analysis employs daily data from a cross-section of 82 cryptocurrencies for the sample period of January 1, 2018 to June 30, 2022. We use a rich dataset of order flows denominated in 11 major currencies including the G10 currencies plus South Korea, which is a prominent centre for cryptocurrency trading. We then form a measure of world cryptocurrency order flow by aggregating the 11 international order flows. Accordingly, our

analysis assesses the predictive ability of US order flow, aggregated world order flow and disaggregated international order flows. In what follows, we discuss our empirical approach and main findings step by step.

We begin by estimating panel regressions that allow us to assess whether order flow can explain and predict cryptocurrency returns in sample. The panel regressions condition on combinations of US order flow, world order flow and the 11 individual order flows together with a set of coin-specific control variables and economic fundamentals. We estimate contemporaneous and predictive panel regressions at the daily and weekly frequency.

With respect to the in-sample evidence, our main finding is that US order flow has significant explanatory and predictive power for the cross-section of cryptocurrency returns. Notably, the contemporaneous effect is positive but the predictive effect is negative. In other words, higher US order flow leads to higher returns in the same day (according to the contemporaneous regressions) but lower returns in the following day (according to the predictive regressions). This daily pattern is consistent with the transitory view of the effect of order flow because there is an overreaction on the same day and a correction on the following day. The daily pattern is short-lived since it disappears for weekly returns for which both the contemporaneous and the predictive effects are positive. Therefore, the weekly pattern is consistent with the permanent view of the effect of order flow. Overall, the in-sample panel regression results indicate that order flow predicts a short-term reversal at the daily frequency and a long-term momentum at the weekly frequency.

For international order flows, the US, world, euro and South Korean order flows have the strongest effects in contemporaneous regressions. This result justifies the use of both aggregate and disaggregated world order flow over and above US order flow. It also motivates the inclusion of South Korea to the sample since its order flow is consistently significant. In the contemporaneous regressions, order flow explains about 1% of cryptocurrency returns for daily data and 5% for weekly data. In the predictive regressions, the US order flows has

the strongest effect for daily returns but world order flow has the strongest effect for weekly returns. Overall, the empirical evidence indicates that order flows on a global scale have strong explanatory and predictive power for daily and weekly cryptocurrency returns.

Next, we assess the out-of-sample predictive ability of order flow for cryptocurrrency returns using machine learning (ML) techniques (see, e.g., Gu, Kelly and Xiu, 2020 and Filippou, Rapach and Thimsen, 2024). These methods are ideally suited for out-of-sample forecasting because they emphasize techniques for variable selection and dimension reduction, which can accommodate a large set of predictors as well as a richer specification of functional forms. Therefore, ML forecasting allows us to condition on the 11 separate order flows and additional economic fundamentals for prediction of the one-day or one-week ahead cryptocurrency returns. We estimate a set of linear and non-linear ML models, which include the following specifications: ridge regression (RR), lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), stochastic gradient boosted regression trees (SGB), and neural nets with 1-4 hidden layers (NN1-NN4).

In terms of the out-of-sample statistical analysis, we find that machine learning forecasts consistently outperform the zero-forecast benchmark. Importantly, models conditioning on the 11 disaggregated order flows outperform models conditioning on economic fundamentals. Although the performance of the order flow models is strong for daily returns, it is in fact stronger for weekly returns. Notably, non-linear ML models perform better for daily returns but non-linear models are the best performers for weekly returns. Combined ML forecasts also perform consistently well. Taken together, these results indicate that ML models conditioning on order flow have significant out-of-sample predictive ability for cryptocurrency returns, especially at the weekly frequency.

We assess the out-of-sample economic value of conditioning on order flow using portfolio sorts. Our approach is based on generating cryptocurrency portfolios sorted on lagged values of one of the following: US order flow, world order flow and the ML return forecast of models

conditioning on all 11 international order flows. The ML approach enables us to exploit the predictive information of all 11 international order flows. Specifically, on each day, we allocate the 82 coins into quintile portfolios and examine the one-day ahead performance of the five quintile portfolios as well as the long-short (top-minus-bottom) portfolio. We then repeat the same exercise at the weekly frequency.

In assessing portfolio performance, we first demonstrate that portfolios conditioning directly on US or world order flow exhibit strong performance. For example, the long-short return spread when conditioning on daily US order flow is equal to 0.40% per day, it is highly significant (t-stat=3.94) and exhibits an annualized Sharpe ratio of 1.84. The three factor alpha, which conditions on the market factor, the size factor and the momentum factor, is also equal to 0.40% per day with a t-stat=3.78. These findings remain strong whether we have daily or weekly sorts and whether we condition on US or world order flow.

Importantly, portfolio performance improves substantially when the portfolio sorts are based on machine learning models that generate forecasts which condition on all 11 order flows. At the daily frequency, the best performing model is the non-linear ML combination, which delivers a mean long-short spread of 1.23% per day (t-stat=7.61) and an annualized Sharpe ratio of 5.05. At the weekly frequency, the best model is the linear ML combination, which delivers a mean long-short spread of 1.91% per week (t-stat=3.55) and an annualized Sharpe ratio of 2.49. Again, non-linear ML models perform better for daily returns but linear ML models perform better for weekly returns. Overall, portfolios sorted on ML forecasts conditioning on all order flows perform substantially better that just directly conditioning on US or world order flow. Moreover, models conditioning on just order flow outperform models conditioning on just economic fundamentals. Taken together, these findings indicate that there is significant additional economic value in conditioning on all 11 international order flows in the context of out-of-sample machine learning forecasts.

Motivated by Avramov, Cheng and Metzker (2023), we conduct an extensive analysis on

the profitability of our models in the presence of economic restrictions such as short-selling constraints, transaction costs and limits-to-arbitrage. To be more specific, in the cryptocurrency market, it may be difficult or even impossible to short certain coins. Therefore, to generate a realistic trading strategy that avoids shorting, we may focus on the long portfolio than the long-short portfolio spread. In light of this, we show that relative to the long-short portfolios, the long portfolios exhibit a similar mean return, slightly lower Sharpe ratio, lower turnover and hence a higher break-even transaction cost. For example, for weekly returns, the long portfolio of the linear ML combination exhibits a mean return of 3.10% per week (t-stat=2.06), an annualized Sharpe ratio of 1.47 and a break-even transaction cost of 3.87% per week. In contrast, the same model for the long-short portfolio delivers a mean spread of 1.91% per week (t-stat=3.55), an annualized Sharpe ratio of 2.49 and a break-even transaction cost of 1.16% per week. We conclude, therefore, that ML models that condition on order flow generate high economic value for long-only investors who pursue a realistic trading strategy investing in cryptocurrencies.

In summary, our analysis provides a preponderance of empirical evidence to show that order flow has strong and economically valuable out-of-sample predictive power for cryptocurrency returns. Our findings indicate that order flow has a transitory effect for daily returns (short-term reversal) and a permanent effect for weekly returns (long-term momentum). ML Models conditioning on order flow consistently outperform ML models conditioning on economic fundamentals, especially for weekly returns. The best performers are forecast combinations of non-linear ML models for daily returns and linear ML models for weekly returns. Finally, our empirical findings are robust to economic restrictions such as short-selling constraints and high transaction costs that characterize the cryptocurrency market. In short, order flow matters for explaining and predicting cryptocurrency returns.

The remainder of the paper is organized as follows. In the next section, we describe the cryptocurrency data and define order flow. In Section 3, we describe the in-sample panel

regression framework for explaining and predicting cryptocurrency returns. The framework for out-of-sample prediction based on machine learning models is examined in Section 4. In Section 5, we assess the economic value of order flow based on portfolio sorts. Finally, we conclude in Section 6. An Online Appendix provides further details on the machine learning models and additional results.

2 Data

2.1 Cryptocurrency Returns

Our empirical analysis employs data from a cross-section of 82 cryptocurrencies for the sample period of January 1, 2018 to June 30, 2022. For each cryptocurrency (crypto or coin), we collect daily US dollar (USD) prices from cryptocompare.com (henceforth CryptoCompare) at 00:00am GMT time. The CryptoCompare data aggregates prices across 150 exchanges. We exclude weekends and US holidays from the daily sample.

The CryptoCompare database contains data on more than 3000 coins. To obtain the final cross-section of 82 coins, we impose three criteria to ensure the liquidity of our crypto sample. First, we require that coins must have a market cap of greater than 1 million USD on the last day of the sample period. Second, coins must be continuously traded with a non-zero price and non-zero volume in each time period of our sample. Finally, third, we exclude all stable coins from the cross-section. As a result, we are left with a balanced panel of 82 coins.[3]

[3] Note that we set January 2018 as the beginning of the sample to allow for a reasonable number of observations across time, while maintaining a reasonably large cross section of coins. Starting earlier than 2018 severely limits the breadth of the cross-section.

We use daily prices to compute returns in the standard way:

$$r_{i,t} = \frac{P_{i,t}}{P_{i,t-1}} - 1, \tag{1}$$

where $P_{i,t}$ is the price of coin i at time t.[4]

2.2 Order Flow

Our primary variable of interest is order flow, which is defined as the log difference between buyer-initiated and seller-initiated transaction volume denominated in a particular currency. We compute order flow using data on signed volume (buy volume and sell volume) obtained from CryptoCompare. Using signed volume data from the same database as prices ensures that the main dependent and independent variables of our analysis cover the same exchanges.

A critical advantage of the CryptoCompare data is that they provide signed volume data denominated in several currencies. In our analysis, we compute the order flow denominated in each of the G10 currencies plus South Korea, which is a prominent centre for trading cryptocurrencies. We designate this group as the G11 currencies, which include the following: US dollar, Euro, British pound, Japanese yen, Swiss frank, Canadian dollar, Australian dollar, New Zealand dollar, Norwegian krone, Swedish krona and the Korean won.

It is important for our order flow measure to be comparable across cryptocurrencies. For this reason, we standardize each series following the Menkhoff et al. (2016) approach, as follows:

$$OF_{i,t} = \frac{of_{i,t}}{\sigma(of_{i,t-30:t-1})}, \tag{2}$$

[4] The main reason for using discrete returns as opposed to log returns is that the price for some coins is less than one cent and exhibits very high volatility. Under these conditions, the approximation of returns as log price differences breaks down and log returns may exhibit values below -100%, which is not realistic.

where $OF_{i,t}$ denotes the standardized order flow of coin i at time t, $_{i,t}$ is the original log order flow, and $\sigma(of_{i,t-30:t-1})$ is the order flow volatility over the last 30 days.

In addition to individual order flows, we also define an aggregate measure across the 11 order flows denominated in different currencies. We term this the G11 or the world order flow. The world order flow is computed by aggregating all 11 buy volumes and all 11 sell volumes and then taking the log difference. We also perform the same standardization as applied to the disaggregated order flows.

To assess the contribution of each cryptocurrency to the world order flow, we conduct a variance decomposition as follows:

$$VC_i = \frac{\beta_i^2 Var(OF_i)}{\sum_{i=1}^{11} \beta_i^2 Var(OF_i)}, \tag{3}$$

where VC_i is the variance contribution of coin i to the world order flow, β_i is the estimated coefficient of each order flow when regressing world order flow on its 11 components, and $Var(OF_i)$ is the variance of each order flow.

The variance decomposition results are reported in Table 1. As expected, for daily data, the US order flow has the highest variance contribution at 39%. The second contribution is from South Korea at 24%. This motivates our decision to add South Korea to the G10 order flows. The euro is third at 10% followed by the British pound at about 6%. Overall, the world order flow is dominated by few countries since the top two contribute 63% to the total variance, whereas the top three contribute 73%. The weekly VCs are similar to the daily VCs.

2.3 Control Variables

We use a combination of coin-specific variables and economic fundamentals as controls in the empirical analysis. The coin-specific variables include total volume, market capitalization (henceforth market cap) and volatility. Total volume is the sum of a coin's 24-hour trading volume across all exchanges. The trading volume is obtained from coinmarketcap.com (henceforth CoinMarketCap). Market cap is also obtained from CoinMarketCap. Volume and market cap are expressed in logs. Volatility is defined as the log difference between a coin's high price and low price on a given day. The daily high and low prices are obtained from CryptoCompare and, therefore, are directly comparable to the end-of-day prices used to compute returns.[5]

The economic fundamentals used in our analysis include the following. The short rate is defined as the 3-month US T-bill rate. The term spread is defined as the difference between the 10-year and 3-month US treasury rates. The default spread is defined as the difference between the BAA- and AAA-rated corporate bond yields. All interest rate data are obtained from the Federal Reserve Bank of St. Louis. The TED spread is defined as the difference between the 3-month T-Bill rate and the 3-month LIBOR rate, and is obtained from the LSEG Eikon database. The S&P 500 index returns (including dividends) are from the CRSP database. The MSCI global index returns are taken from msci.com. The VIX index is obtained from CBOE.com.

In addition to the daily frequency, our analysis also uses the weekly frequency. The weekly crypto variables are defined over the interval beginning on Saturday at 00:00am GMT time and ending the following Friday at 11:59pm GMT for a total of seven days including weekends.

[5] We use two sources for cryptocurrency data because CryptoCompare has price and order flow (signed volume) data but (for our subscription) it lacks the total volume and market capitalization data that is provided by CoinMarketCap.

2.4 Summary Statistics

In Tables A2 and A3 of the Online Appendix, we report summary statistics for all coin-specific variables: returns, order flow, market cap, volume and volatility. The tables report results for all 82 coins as well as for the top 10 coins by market cap.

In brief, our main findings are as follows: (1) daily cryptocurrency returns are extremely volatile and they also display high positive skewness and immense kurtosis; (2) daily order flow displays similar extreme behaviour to returns; (3) the full cross-section of 82 coins exhibits more extreme behaviour than the top 10 coins; (4) the main source of volatility, skewness and kurtosis is the time-series, not the cross-section; and (5) the average full-sample cryptocurrency return is equal to 0.32% per day or 80.6% per year.

3 In-Sample Analysis

Our empirical approach is based on panel regressions for in-sample analysis and machine learning techniques for out-of-sample analysis. This section presents the in-sample panel regression results on the effect of order flow in explaining and predicting cryptocurrency returns. We consider US order flow, disaggregated G11 order flows and aggregate world order flow together with several control variables.

3.1 Panel Regressions with US Order Flow

3.1.1 Contemporaneous Regressions

Can US order flow explain cryptocurrency returns? We address this question by estimating panel regressions that condition on contemporaneous US order flow and the control variables.

The panel regressions are estimated using the full sample and include coin fixed effects. The results for both daily and weekly data are reported in Table 2.

Our main empirical finding is that US order flow has a positive contemporaneous relation to returns and is highly significant in all cases. Therefore, US order flow does explain cryptocurrency returns. For example, in the presence of all controls, a one standard deviation increase in US order flow is associated with a 1.6% increase in returns at the daily frequency and a 3.2% increase at the weekly frequency. These results are highly significant since the Newey and West (1987) t-statistic is equal to 25.76 for daily regressions and 9.84 for weekly regressions. It is interesting to note that the weekly coefficient is larger than the daily coefficient and this is also true for the R^2. Specifically, the R^2 is equal to 0.9% for daily regressions and 4.7% for weekly regressions.

A possible explanation for the stronger weekly results is that order flow at the daily horizon likely includes uninformed traders (otherwise known as noise traders), whose trades contribute to pricing errors. The effect of these traders on the price tends to be transitory as the price reverts back to the mean. Aggregating order flow over one week mitigates the impact of uninformed trader noise. Consequently, order flow may explain a greater variation of the contemporaneous price movements for the weekly than the daily horizon.

In terms of the control variables, in the daily regressions, the significant variables include volatility, the VIX, the S&P 500 return, the MSCI global return, the short rate, the term spread and the default spread. In contrast, the market cap, volume and the TED spread are insignificant. For the weekly regressions, the significant variables include the VIX, the S&P 500 return and the MSCI global return. In conclusion, US order flow has strong and highly significant explanatory power for the cross-section of cryptocurrency returns in the presence of crypto-specific control variables and standard economic fundamentals.

3.1.2 Predictive Regressions

Next, we assess the predictive ability of US order flow by estimating the same panel regressions but with lagged values for all independent variables. We condition on the same control variables plus lagged cryptocurrency returns. The results reported in Table 3 indicate that, across all specifications, US order flow is a highly significant predictor of both one-day ahead and one-week ahead cryptocurrency returns.

Notably, the coefficient on lagged US order flow is significantly negative for daily returns and significantly positive for weekly returns. This result is consistent with the findings of Makarov and Schoar (2020). Therefore, for daily returns, higher US order flow leads to to higher returns in the same day (according to the contemporaneous regressions) but lower returns in the following day (according to the predictive regressions). This pattern is consistent with an overreaction on the same day and a correction on the following day. The daily pattern is short-lived since it disappears for weekly returns. For the latter, an increase in US order flow leads to higher returns in the same week (according to the contemporaneous regressions) and higher returns in the following week (according to the predictive regressions). A possible explanation for this finding is that daily trading involves uninformed traders, who are contributing to an overreaction in the same day and a correction in the following day. Once data are aggregated to the weekly horizon, the actions of noise traders have less influence on price, and higher order flow predicts higher future returns.

For daily returns, the significant predictive control variables include volatility, the VIX, the SP 500 return, the MSCI global index returns and the short rate. For weekly returns, the significant predictive control variables include the lagged cryptocurrency returns, volatility, the TED spread, the S&P 500 returns, the MSCI global index returns and the short rate.

3.2 Panel Regressions with World Order Flows

In addition to panel regressions conditioning on US order flow only, we also estimate panel regressions that condition on all 11 international order flows. The panel regressions are otherwise the same as before and condition on the same control variables. We estimate two types regressions: one with just the 11 disaggregated order flows and one with the 11 order flows plus the aggregate world order flow. As previously, we use daily and weekly data and repeat for both contemporaneous and predictive regressions.[6]

The results from contemporaneous regressions are shown in Table 4. The order flows that remain significant across all specifications, for both daily and weekly returns, are the world order flow and the disaggregated order flows denominated in USD and the South Korean won. Other order flows are significant in certain specifications, most notably the euro. The world order flow exhibits the highest coefficients. Specifically, a one standard deviation increase in world order flow is associated with a 2.0% increase in returns at the daily horizon and a 2.1% increase at the weekly horizon.

These results justify the use of both disaggregated and aggregate world order flow over and above US order flow. They also motivate the inclusion of South Korea to the sample since it is consistently highly significant. In short, the empirical evidence indicates that order flows on a global scale have explanatory power for contemporaneous cryptocurrency returns.[7]

Turning to the predictive regressions, the results are somewhat different from the contemporaneous regressions. As we can see in Table 5, for daily returns, only the US order flow

[6] Note that regressions which condition on the 11 disaggregated order flows and world order flow do not suffer from collinearity due to the standardization of all order flows. Specifically, if we regress the standardized world order flow on its 11 standardized components, the R^2 is equal to 0.50.

[7] In the foreign exchange literature, there are several studies using disaggregated order flows by counterparty type (e.g., customer types). See, for example, Evans and Lyons (2005) and Menkhoff et al. (2016).

is significant in predicting one-day ahead returns. For weekly returns, the US order flow remains significant in the presence of individual order flows but becomes insignificant when including the world order flow, where the latter is significant. In short, therefore, US order flow has high predictive ability at the daily horizon but world order flow is more significant at the weekly horizon.[8]

This finding indicates that a temporal aggregation of information takes place as we move from the daily to the weekly frequency. On a daily basis, investors' trading outside of the US, represented by world order flow is not informative in predicting the next day's returns. However, when we aggregate this daily information over the span of a week, it becomes informative, more so than US order flow. In other words, aggregating order flow across time *and* across countries produces a predictor with more significance than disaggregated data in the context of in-sample regressions.

4 Out-of-Sample Analysis

We perform an out-of-sample analysis of the predictive ability of order flow for cryptocurrency returns using popular machine learning (ML) methods (see, e.g., Gu, Kelly and Xiu, 2020). Out-of-sample forecasting of asset returns is notoriously difficult (see, e.g., Welch and Goyal, 2008), which is likely exacerbated in the case of cryptos due to their very high volatility. Traditional prediction methods based on ordinary least squares (OLS) regressions are likely to underperform, when conditioning on a high number of predictors such as the 11 separate order flows and the additional control variables that we use to forecast one-day ahead cryptocurrency returns. Machine learning techniques are more suitable for out-of-

[8] For brevity, we do not report the coefficient estimates for the control variables in the panel regressions with international order flows. The results for the economic controls are largely the same as with the US order flows.

sample forecasting because they emphasize techniques for variable selection and dimension reduction, which can accommodate a large set of predictors as well as a richer specification of functional forms.

4.1 Models Considered

As our benchmark model, we use linear regression estimated with pooled OLS. The benchmark model is motivated by the predictive panel regression approach for in-sample analysis. In addition to the OLS benchmark, we consider a variety of linear and nonlinear models from the rapidly expanding ML literature on forecasting asset returns (Gu, Kelly and Xiu, 2020). For penalized linear models, we include Lasso, Ridge, and Elastic Net (EN) regressions. We employ principal component regression (PCR) as a linear dimension reduction technique. To incorporate nonlinearities and predictor interactions, we consider tree-based models such as random forest (RF) and gradient boosted regression trees (GBT) as well as feed-forward neural networks with 1 to 4 hidden layers (NN1-NN4). A comprehensive description of these models can be found in Section A1 of the Online Appendix.

As discussed in Section 2.4, cryptocurrency returns exhibit extremely heavy tails, characterized by high positive skewness and immense kurtosis. To mitigate the impact of heavy-tailed data, we adopt the robust Huber objective function, which combines the ℓ_2 loss for small errors and ℓ_1 loss for large errors. We estimate all models, except for RF, using the Huber objective function. Since RF has inherent robustness to outliers, we follow Gu, Kelly and Xiu (2020) and use the default ℓ_2 loss function. In unreported results, we find that models estimated using the Huber objective function generally outperform those estimated with the MSE objective function for cryptocurrency data. See Section A1.7 of the Online Appendix for a detailed description of the Huber objective function.

We also form forecast combinations, which are designed to combine the forecasts of multiple

models (see, e.g., Timmermann, 2006). Following Rapach, Strauss and Zhou (2010), we compute the equally-weighted average of all forecasts across a set of models at each point in time, referred to as the "mean" combination. We compute combined forecasts for three cases: (1) linear models (L-Mean), (2) nonlinear models (NL-Mean), (3) all models (All-Mean). Comparing the nonlinear forecast combination to the linear one allows us to assess the improved predictive ability gained from incorporating nonlinearities and predictor interactions. For a detailed description of forecast combinations, see Section A1.6 of the Online Appendix.

4.2 Estimation Procedure and Performance Evaluation

We generate out-of-sample forecasts for each model by dividing our sample into three disjoint sets: the training, validation, and test sets. First, we estimate the model parameters on a training sample $\mathcal{T}_1$ spanning one year (February 14, 2018 to February 14, 2019). Second, we conduct an extensive hyperparameter optimization on a one-year validation sample $\mathcal{T}_2$ (February 15, 2019 to February 14, 2020). Third, we evaluate the out-of-sample performance of each model using a one-month test sample $\mathcal{T}_3$ (Feb 18, 2020 to March 13, 2020). We keep the models fixed for one month and repeat this process by rolling forward the validation and test sets by one month while extending the training sample by one month in each iteration. We continue this process to the end fo the sample. Section A2 in the Online Appendix provides detailed information on the hyperparameter optimization setup for each model.

Our analysis utilizes two statistical criteria to assess the out-of-sample predictive ability of the models, applied to a balanced panel of cryptocurrency returns, where coins are indexed by $i = 1, ..., N$ and days by $t = 1, ..., T$. Our primary performance metric is the Welch and

Goyal (2008) out-of-sample R^2 (R^2_{oos}) statistic as used by Gu, Kelly and Xiu (2020):

$$R^2_{oos} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3}(r_{i,t+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{T}_3}(r_{i,t+1})^2}, \tag{4}$$

where $r_{i,t+1}$ and $\hat{r}_{i,t+1}$ denote the realized future return and one-day ahead forecast for coin i, respectively, and $\mathcal{T}_3$ indicates that the metric is only assessed on the test sample. The R^2_{oos} statistic pools forecast errors across coins and over time into a comprehensive panel-level assessment of each model's predictive ability. It measures the reduction in mean squared error (MSE) achieved by each model relative compared to a naive forecast, which predicts zero returns for all coins. In unreported results, we explore several alternative benchmarks, including the use of historical mean returns typically employed in equity market forecasts. However, we find that the naive forecast of zero returns consistently produce the lowest R^2_{oos} values for cryptocurrencies, making it the most conservative benchmark.[9]

Outliers are prevalent in cryptocurrency returns, as evidenced by the enormous kurtosis shown in Section 2.4. Since the R^2_{oos} statistic is based on the MSE, which places extreme emphasis on large errors, outliers can disproportionately influence the evaluation of model predictive abilities. To address this issue, we incorporate a second criterion: the robust R^1_{oos} statistic from Koenker and Machado (1999), defined as:

$$R^1_{oos} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3}|r_{i,t+1} - \hat{r}_{i,t+1}|}{\sum_{(i,t)\in\mathcal{T}_3}|r_{i,t+1}|}. \tag{5}$$

The R^1_{oos} is based on the Mean Absolute Error (MAE), which is less sensitive to outliers relative to MSE. Using the MAE ensures that the ranking of the models is driven by overall performance not by a few outliers.

[9] Note that for the R^2_{oos} statistic, there is a standard way of assessing statistical significance based on the Clark and West (2006, 2007) testing procedure. However, we find that the R^2_{oos} values are significant at the 1% confidence level for nearly every model and specification, hence we omit these results for presentation purposes, but they are available upon request.

4.3 Out-of-Sample Performance Results

In Table 6, we report the results on the out-of-sample statistical performance. Our analysis compares models conditioning on three types of information sets: (1) ML models that condition on order flow (OF) only; (2) ML models that condition on economic fundamentals (EF) only (i.e., models that condition on just the control variables, not order flow); and (3) ML models that condition on both OF and EF. This grouping of models allows us to assess whether order flow is essential in generating reliable out-of-sample forecasts for cryptocurrency returns over and above economic fundamentals. Note that all models condition on lagged returns. Additionally, for the training, validation, and test sets, we standardize each variable using its mean and variance from the training set as it is standard in the ML literature. Our main findings can be summarized as follows.

First, based on our main statistical criterion, the R^2_{oos}, the OF models are consistently the best models and tend to deliver a positive value for R^2_{oos}. Specifically, the OF model is best for 11 out of the 14 ML models. Notably, for the three forecast combinations (L-Mean, NL-Mean and All-Mean), the OF model is consistently the best model.

In addition, the R^2_{oos} is positive for 7 out of the 14 ML models. Specifically, the R^2_{oos} tends to be negative for the linear models but positive for the non-linear models. The highest $R^2_{oos}$is for the Random Forest (0.61%) and the NL-Mean Combination (0.59%). Finally, these results are also confirmed by the R^1_{oos} criterion, where the results are reported in Table A4 of the Online Appendix. The general agreement of R^1_{oos} and R^2_{oos} indicates that the statistical performance evaluation is not driven by outliers.

In light of these results, we conclude that ML forecasts for cryptocurrency returns consistently outperform the benchmark, especially for non-linear specifications. Conditioning on order flow tends to deliver superior forecasts than just conditioning on economic fundamentals. In short, therefore, order flow has significant out-of-sample predictive ability for

cryptocurrency returns over and above economic fundamentals.

5 The Economic Value of Order Flow

5.1 Portfolio Sorts

In this section, we assess the out-of-sample economic value of conditioning on order flow. Our approach is based on generating portfolios by sorting the cross-section of cryptocurrencies on one of the following criteria: (1) the US order flow of each coin; (2) the world order flow of each coin; and (3) the machine learning (ML) return forecast for each coin generated by models conditioning on all 11 disaggregated international order flows and control variables. This approach provides a comprehensive assessment of the economic gains associated with the out-of-sample predictive ability of order flow.

The portfolio sorting procedure is implemented as follows. At each time t, we rank all coins using one of the three criteria listed above: US order flow, world order flow or the ML forecast based on all order flows. Based on this ranking, we allocate the 82 coins into five quintile portfolios. Allocating assets to quintile portfolios is a standard approach in the literature and ensures that there is a sufficient number of cryptos in each portfolio. Then, we compute the $t+1$ equally-weighted portfolio returns and the Newey and West (1987) t-statistic.[10] We report results for daily and weekly rebalancing. All portfolio sorts are performed out of sample. In short, the portfolio sorts allow us to assess whether order flow has predictive information for portfolio returns over the next period.[11]

[10] In computing the Newey and West (1987) t-statistic, we use the Bartlett kernel with the data-driven bandwidth parameter selected by the AR(1) model (see, e.g., Andrews, 1991).

[11] For cryptocurrencies, it is sensible to use equal weights as opposed to value weights because value weights would be dominated by the top 5 cryptos, which on a given day comprise approximately 90% of the total market cap. Consequently, using equal weights allows us to take full advantage of the breadth

We report results on the mean daily (or weekly) percent return for the five quintile portfolios (P_1 to P_5). We also report the mean return for ALL, which is the equally-weighted portfolio of all 82 cryptos in our sample that serves as benchmark for assessing performance. We primarily focus on the zero-cost investment portfolio that goes long on the top portfolio and short on the bottom portfolio ($P_5 - P_1$). For this portfolio, we report the mean return, alpha and Sharpe ratio. The alpha is estimated by regressing the portfolio return on three factors: (1) the market return defined as the value-weighted return of all 82 coins; (2) the size factor defined as the return difference between the small size portfolio (bottom 30% in market cap) and the big size portfolio (top 30% in market cap), and (3) the momentum factor based on three-week returns (see Liu, Tsyvinski and Wu, 2022).

In assessing economic restrictions (discussed in the next section), we also report the maximum drawdown (MDD), turnover (TO) and break-even transaction cost (BE-TC). The maximum drawdown is defined as the maximum cumulative loss from the portfolio's price peak to the following trough. A reasonably low MDD is indicative of the success of a trading strategy because large drawdowns often lead to fund redemptions.

Following Gu, Kelly and Xiu (2020), we define the turnover of the long-short portfolio at time t as follows:

$$TO_t = \frac{1}{2}\sum_{i\in L}\left| w_{i,t} - \frac{w_{i,t-1}(1+r_{i,t})}{\sum_i w_{i,t-1}(1+r_{i,t})}\right| + \frac{1}{2}\sum_{j\in S}\left| w_{j,t} - \frac{w_{j,t-1}(1+r_{j,t})}{\sum_j w_{j,t-1}(1+r_{j,t})}\right|, \tag{6}$$

where $i \in L$ ($j \in S$) indicates that coin i (j) belongs to the long (short) portfolio, $w_{i,t}$ ($w_{j,t}$) refers to the weight of coin i (j) at time t, and $r_{i,t}$ ($r_{j,t}$) refers to the return of coin i (j) at time t. By design, the turnover of long and short positions ranges between 0 and 1. Accordingly, the turnover of the long-short portfolio defined above ranges between 0 and 2.

of information available in our data set.

Finally, we compute the break-even transaction cost (BE-TC) as the portfolio return divided by turnover (see, e.g., Avramov, Cheng and Metzker, 2023). The BE-TC is equal to the daily/weekly proportional transaction cost required to eliminate the gains from the trading strategy. For example, if the portfolio return is equal to 1% per day and the BE-TC is equal to 1.5% per day, then the proportional transaction cost would have to be 1.5% per day to eliminate all the gains of the portfolio.

5.2 Portfolio Performance

5.2.1 Sorting on US/World Order Flow

We begin our discussion of portfolio performance by reporting results in Table 7 for separate portfolio sorts on US and world order flow for the full sample period. These portfolios are generated by conditioning on either daily (Panel A) or weekly order flow (Panel B). Note that for daily portfolio sorts, the long-short portfolio is defined as $P_1 - P_5$, whereas for weekly portfolio sorts it is defined as $P_5 - P_1$. This is because the portfolio sorts are consistent with the panel regressions: for daily data, the slope for the US and world order flow is negative and hence higher order flow predicts lower returns, whereas for weekly data higher order flow predicts higher returns.

We find that in all cases, the performance of the long-short portfolio is strong, which indicates that there is high economic value associated with order flow. For example, the daily return of $P_1 - P_5$ when conditioning on daily US order flow is equal to 0.40% with a t-stat= 3.94. The alpha is also equal to 0.40% per day and is highly significant (t-stat= 3.78) and the annualized Sharpe ratio is equal to 1.84.

The weekly results are similar. For example, for weekly US order flow, the mean of the $P_5 - P_1$ is equal to 4.06% per week but is insignificant (t-stat= 1.47). The alpha is lower

(1.83% per week) but is highly significant (t-stat= 3.37). The Sharpe ratio is equal to 0.72. The economic value of world order flow is essentially the same as US order flow: the mean of $P_5 - P_1$ is equal to 4.03% per week with a t-stat= 1.48, the alpha is equal to 1.83% per week with a t-stat= 2.95 and the Sharpe ratio is equal to 0.72. Overall, these results indicate that that there is strong economic value for portfolios sorted on US or world order flow.

5.2.2 Sorting on ML Forecasts Based on Order Flow

Next, we report results in Table 8 on portfolio sorts based on the daily forecasts generated by ML models, which condition on the 11 disaggregated order flows (OF models). This approach enables an assessment of whether there are economic gains with conditioning on all order flows as opposed to just one of them (i.e., the US or world order flow) as we did previously. Conditioning on ML forecasts necessarily truncates the sample period by the two years of data used for the training and validation sets. Hence, the ML forecasts begin on February 18, 2020. For the ML results to be comparable to the simple conditioning on US or world order flow, we also report results for the latter focusing on the same out-of-sample period beginning in February 2020. Panel A report results for US and world order flow and Panel B for the ML forecasts for the same out-of-sample period.

The performance of the ML models conditioning on all order flows is outstanding. Specifically, the ML models tend to deliver monotonically increasing returns from P_1 to P_5, significant returns close to 1% per day for the P_5 portfolio, similarly significant returns close to 1% per day for the $P_5 - P_1$ portfolio and very high annualized Sharpe ratios in the range of 1.5-5.0. Consider, for example, the best performing model by Sharpe ratio: the non-linear combination (NL-Mean) OF model. For this model, the $P_5 - P_1$ return is equal to 1.23% per day with a t-stat= 7.61, the alpha is equal to 1.28% with a t-stat= 8.31, and the annualized Sharpe ratio is equal to 5.05.

Overall, we can summarize the main results as follows: (1) the non-linear ML models perform the best and their performance is impressive by any standard; (2) the linear ML models still perform well but underperform the non-linear models; (3) the OLS benchmark performs very well; (4) the combined forecasts are among the best performers; and, finally, (5) the portfolios sorted on the daily ML models perform substantially better than portfolios sorted on US or world order flow. For example, the Sharpe ratios are equal to 1.55 for daily US order flow and 0.94 for daily world order flow but they rise up to 5.05 for the NL-Mean model. This result is especially important since it indicates that there is significant additional value in conditioning on all 11 order flows than just US or world order flow.

5.2.3 Sorting on ML Forecasts Based on Economic Fundamentals

The results discussed above are exclusively for ML models conditioning on order flow and no other variables. In Table 8, we also report daily results for ML models, which condition on economic fundamentals only (EF model) and both economic fundamentals and order flow (OF+EF model). The weekly results are reported in Table 9. For daily returns, the main result here is that for linear ML forecasts, all three models (OF, EF and OF+EF) display similar performance. However, for non-linear forecasts, the OF is consistently the best model.

For example, consider the NL-Mean model. The NL-Mean forecasts conditioning only on OF deliver a daily return of 1.23% with an annualized $SR = 5.05$. The NL-Mean forecasts conditioning only on EF deliver a daily return of 0.98% with an annualized $SR = 3.75$. Finally, the NL-Mean forecasts conditioning on OF+EF deliver a daily return of 1.04% with an annualized $SR = 4.17$. In short, for daily returns, non-linear ML models that condition only on OF perform better than non-linear models that also condition on EF.

For weekly returns, the results are overwhelmingly in favour of the OF model, which com-

pletely dominates the EF and OF+EF models. For weekly returns, there is little difference between linear and non-linear models but there is a large difference between OF and EF. For example, consider the ALL-Mean model. The ALL-Mean forecasts conditioning only on OF deliver a weekly return of 2.02% with an annualized $SR = 2.39$. The ALL-Mean forecasts conditioning only on EF deliver a weekly return of -0.17% with an annualized $SR = -0.16$. Finally, the ALL-Mean forecasts conditioning on OF+EF deliver a weekly return of 0.87% with an annualized $SR = 0.88$.

6 Economic Restrictions

Our portfolio analysis has so far focused on the long-short portfolio because it is well suited for illustrating the economic value of conditioning on order flow. In other words, if order flow is a good predictor, it should be able to identify which portfolio performs the best and which one the worst. Then, investors would benefit from a zero-cost trading strategy that goes long on the top portfolio and short on the bottom portfolio. However, in the cryptocurrency market, it may be difficult, even impossible, to short some cryptocurrencies. When possible, additional shorting fees will apply (see, e.g., Liu, Tsyvinski and Wu, 2022). Therefore, if the objective is to generate a realistic trading strategy that avoids shorting, it may be better to focus on the long portfolio. As explained previously, the long portfolio is the top portfolio (P_5), except when conditioning on daily order flow in which case it is the bottom portfolio (P_1). For this reason, in Table 10, we report detailed results on the performance of the daily long portfolio and compare it to the corresponding daily long-short portfolio. Table 11 does the same for weekly returns. For a comprehensive assessment of performance of the most realistic trading strategies, we report the mean return, alpha, Sharpe ratio, maximum drawdown, turnover and break-even transaction cost.

Overall, we find that the performance of the long portfolio is strong and similar to the

long-short portfolio. This is because in most cases the return of the short portfolio is close to zero and hence the return to the long-short portfolio is driven primarily by the long portfolio. For example, consider the daily NL-Mean model for OF. This model delivers a long portfolio with mean return equal to 1.10% per day with a t-stat= 3.95, an alpha of 1.11% per day with a t-stat= 4.09, an annualized $SR = 2.60$, $MDD = 59\%$, $TO = 0.78$ and a BE-TC= 1.40% per day. By any standard, this is high economic value for a long-only investor in the cryptocurrency market. It also compares well with the long-short portfolio of the same model, which delivers: mean= 1.23% (t-stat= 7.61), alpha= 1.28% (t-stat= 8.31), $SR = 5.05$, $MDD = 24\%$, $TO = 1.57$ and BE-TC= 0.78%.

The full set of results reported in Tables 10 and 11 indicate that: (1) in terms of mean return and alpha the long portfolio has similar performance to the long-short portfolio; (2) in terms of Sharpe ratio and maximum drawdown, the long-short portfolio performs better; and (3) in terms of turnover and break-even transaction cost, the long portfolio performs substantially better, especially for weekly returns. In conclusion, conditioning on order flow, especially in the context of machine learning forecasts, generates high economic value for long-only investors who pursue a realistic trading strategy investing in cryptocurrencies.

7 Conclusion

The meteoric rise of cryptocurrencies as a new investment asset class has triggered an emerging literature in financial economics. A central theme of this literature is the out-of-sample prediction of cryptocurrency returns and the design of trading strategies, which are consistently profitable. The foreign exchange literature has established order flow as a prominent predictor of fiat currency returns representing the microstructure approach to exchange rate prediction as an alternative to models conditioning on standard economic fundamentals. For cryptocurrencies, however, there is a gap in the literature as there is little work on the ability

of order flow to explain and predict cryptocurrency returns. Our analysis bridges this gap by deploying an extensive dataset of order flow and cryptocurrency returns. In this manner, we contribute to the emerging cryptocurrency literature by highlighting the role of order flow for out-of-sample return prediction over and above standard economic fundamentals.

We find that order flow matters for cryptocurrencies. It matters both for explaining and for predicting the cross-section of cryptocurrency returns. The combination of machine learning techniques with order flow information provides an especially powerful predictive toolkit for this market. In addition to US order flow, we employ order flow data for all G10 currencies plus South Korea, which is prominent in this market. This enables us to perform an analysis of both aggregate world order flow and disaggregated international order flows. In the end, our main finding is that the information in all 11 international order flows is essential for cryptocurrency prediction and in most cases dominates the information in standard economic fundamentals.

Importantly, sorting directly on order flow or indirectly through machine learning forecasts allows us to generate high performing portfolios out of sample. In particular, the machine learning forecasts based on daily or weekly order flow produce long-only and long-short quintile portfolios, which exhibit a mean return of about 1% per day or 2% per week. This mean return is effectively three times higher than the equally-weighted cryptocurrency market portfolio return for the same period. In conclusion, the high economic value of conditioning on order flow together with our in-sample and out-of-sample statistical findings strongly indicate that information conveyed by order flow matters for cryptocurrency returns.

References

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 817-858.

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

Cakici, N., S. J. H. Shahzad, B. Bedowska-Sojka, and A. Zaremba (2022). Machine learning and the cross-section of cryptocurrency returns. Available at SSRN: https://ssrn.com/abstract=4295427.

Clark, T. E. and K. D. West (2006). Using out-of-sample mean squared prediction errors to test the Martingale difference hypothesis. *Journal of Econometrics* 135, 155-186.

Clark, T. E., and K. D. West (2007). Approximately normal tests for equal predictive accuracy. *Journal of Econometrics* 138, 291-311.

Diebold, F.X. and M. Shin (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting* 35, 1679–1691.

Evans, M. D. D. (2010). Order flows and the exchange rate disconnect puzzle. *Journal of International Economics* 80, 58-71.

Evans, M. D. D. and R. K. Lyons (2002). Order flow and exchange rate dynamics. *Journal of Political Economy* 110, 170-180.

Evans, M. D. D. and R. K. Lyons (2005). Meese-Rogoff redux: Micro-based exchange-rate forecasting. *American Economic Review Papers and Proceedings* 95, 405-414.

Evans, M. D. D. and D. Rime (2012). Micro approaches to foreign exchange determination, in James, J., L. Sarno and I. W. Marsh (eds.) *Handbook of Exchange Rates*. Hoboken, NJ: Wiley.

Filippou, I., D. Rapach, and C. Thimsen (2024). Cryptocurrency return predictability: A machine-learning analysis. Available at SSRN: https://ssrn.com/abstract=3914414.

Froot, K.A. and T. Ramadorai (2005). Currency returns, intrinsic value, and institutional-investor flows. *Journal of Finance* 60, 1535–1566.

Glosten, L.R. and P.R. Milgrom (1985). Bid, ask and transaction prices in a market-maker market with heterogeneously informed traders. *Journal of Financial Economics* 14, 71-100.

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

Harvey, C.R., A. Ramachandran, and J. Santoro (2021). *DeFi and the Future of Finance* Hoboken, New Jersey: Wiley.

Hasbrouck, J. (1988). Trades, quotes, inventories, and information. *Journal of Financial Economics* 22, 229–252.

Hasbrouck, J. (1991). Measuring the information content of stock trades. *Journal of Finance* 46, 179–207.

Koenker, R. and J.A.F Machado (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* 94, 1296-1310.

Kyle, A.S. (1985). Continuous auctions and insider trading. *Econometrica* 53, 1335–1355.

Liu, Y., and A. Tsyvinski (2021). Risks and returns of cryptocurrency. *Review of Financial Studies* 34, 2689–2727.

Liu, Y., A. Tsyvinski and X. Wu (2022). Common risk factors in cryptocurrency. *Journal of Finance* 77, 1133-1177.

Makarov, I. and A. Schoar (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics* 135, 293–319.

Menkhoff, L., L. Sarno, M. Schmeling, and A. Schrimpf (2016). Information flow in foreign exchange markets: Dissecting customer currency trades. *Journal of Finance* 71, 601–633.

Newey, W. K. and K. D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703-708.

Rapach, D. E., J. K. Strauss, and G. Zhou (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies* 23, 821-862.

Rime, D., L. Sarno and E. Sojli (2010). Exchange rate forecasting, order flow and macroeconomic information. *Journal of International Economics* 80, 72-88.

Timmermann, A. (2006). Forecast combinations, in Elliott, G., C. W. J. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Volume 1, Amsterdam: North-Holland.

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

Table 1: Variance Decomposition of World Order Flow

The table reports the variance decomposition of world (G11) order flow into its 11 country components. The values represent the percent variance contribution of each country to world order flow at the daily and weekly frequency. The sample period ranges from January 1, 2018 to June 30, 2022.

	Daily %	Weekly %
OF^{USD}	39.19	37.74
OF^{KRW}	23.82	31.29
OF^{EUR}	10.26	9.52
OF^{GBP}	5.59	3.77
OF^{NZD}	3.61	4.01
OF^{CHF}	3.58	1.91
OF^{AUD}	3.37	2.49
OF^{CAD}	3.23	2.01
OF^{NOK}	3.01	2.80
OF^{SEK}	2.71	2.74
OF^{JPY}	1.62	1.72
OF^{G11}	100	100

Table 2: Contemporaneous Panel Regressions with US Order Flow

The table displays panel regression results of cryptocurrency returns on contemporaneous US order flow and control variables. Order flow is in logs and standardized as described in the data section. Market cap and volume are in logs. Volatility is the log difference between a coin's high and low price. The VIX, S&P 500 and MSCI are log returns. The TED spread is the difference between the three-month Treasury bill and the three-month LIBOR. The short rate is the three-month T-bill rate. The term spread is the difference between the ten-year and three-month T-bill rates. The default spread is the difference between the AAA and BAA corporate bonds. All panel regressions use balanced panels with fixed coin effects. Columns (1)-(3) are for daily returns, where weekends and holidays have been removed. Columns (4)-(6) are for weekly returns. The t-statistics shown in parentheses are computed using Newey and West (1987) standard errors. The sample period ranges from February 14, 2018 to June 30, 2022.

	Daily			Weekly		
	(1)	(2)	(3)	(4)	(5)	(6)
constant	0.001	0.001	-0.003	0.006	-0.36	-0.42
	(0.18)	(0.64)	(-0.84)	(1.04)	(-1.31)	(-1.20)
$OF_{i,t}$	0.016***	0.016***	0.016***	0.036***	0.036***	0.031***
	(24.83)	(25.77)	(25.76)	(11.35)	(10.16)	(10.17)
$MCAP_{i,t}$		-0.001	0.004		0.016	0.023
		(-0.22)	(1.09)		(0.75)	(0.97)
$Volume_{i,t}$		0.003	0.002		-0.005	-0.009
		(0.97)	(0.65)		(-0.37)	(-0.59)
$Volatility_{i,t}$		0.209***	0.214***		0.57	0.62
		(4.82)	(4.82)		(1.39)	(1.43)
VIX_t			-0.043***			-0.11***
			(-4.37)			(-6.45)
TED $Spread_t$			0.001			0.001
			(1.45)			(1.56)
S&P 500_t			-0.546***			-4.42***
			(-6.20)			(-7.35)
$MSCI_t^{Global}$			0.740***			6.36***
			(5.58)			(6.66)
Short $Rate_t$			-0.002**			-0.011
			(-2.27)			(-1.57)
Term $Spread_t$			-0.009***			-0.051***
			(-5.39)			(-2.60)
Default $Spread_t$			0.015***			0.015
			(3.27)			(0.47)
$\bar{R}^2$ (%)	0.3	0.8	0.9	3.1	3.8	4.7
# obs.	90364	90364	90364	16728	16728	16728

Table 3: Predictive Panel Regressions with US Order Flow

The table displays results of predictive panel regressions of cryptocurrency returns on lagged US order flow and control variables. The table contains the same information as Table 2 but all variables are lagged and $r_{i,t-1}$ is the lagged return of each cryptocurrency. The t-statistics shown in brackets are computed using Newey and West (1987) standard errors. The sample period ranges from February 14, 2018 to June 30, 2022.

	Daily			Weekly		
	(1)	(2)	(3)	(4)	(5)	(6)
constant	0.001	0.001	0.008*	0.007	0.039**	-0.001
	(0.15)	(0.11)	(1.81)	(1.23)	(2.18)	(-0.02)
$OF_{i,t-1}$	-0.002***	-0.001***	-0.002***	0.007**	0.007**	0.007*
	(-3.87)	(-3.05)	(-3.12)	(2.28)	(2.38)	(1.93)
$r_{i,t-1}$		-0.012	-0.012		-0.007*	-0.006
		(-1.20)	(-1.20)		(-1.72)	(-1.33)
$MCAP_{i,t-1}$		-0.003	-0.002		-0.002	0.0004
		(-0.94)	(-0.64)		(-0.40)	(0.11)
$Volume_{i,t-1}$		-0.001	-0.002		0.001	-0.002
		(-0.23)	(-0.76)		(0.09)	(-0.39)
$Volatility_{i,t-1}$		0.035***	0.038***		0.051***	0.048***
		(2.56)	(2.71)		(3.72)	(2.88)
VIX_{t-1}			-0.051*			-0.021
			(-1.66)			(-0.25)
TED $Spread_{t-1}$			-0.001			-0.0002
			(-0.69)			(-1.00)
S&P 500_{t-1}			-0.34*			1.85***
			(-1.64)			(7.22)
$MSCI_{t-1}^{Global}$			0.55***			-2.39***
			(3.07)			(-7.40)
Short $Rate_{t-1}$			-0.004**			-0.011
			(-2.88)			(-0.78)
Term $Spread_{t-1}$			-0.002			-0.0037
			(-1.10)			(-0.16)
Default $Spread_{t-1}$			-0.001			0.043*
			(-0.31)			(1.66)
$\bar{R}^2$ (%)	0.1	0.1	0.1	0.1	0.1	0.1
# obs.	90282	90282	90282	16646	16646	16646

Table 4: Contemporaneous Panel Regressions with All Order Flows

The table displays panel regression results of contemporaneous cryptocurrency returns on disaggregated international order flows, world (G11) order flow and control variables. The table contains the same information as Table 2 but for all order flows plus the world order flow. The t-statistics shown in parentheses are computed using Newey and West (1987) standard errors. The sample period ranges from February 14, 2018 to June 30, 2022. The panel regressions also condition on the control variables reported in the previous tables but these results are not reported to save space.

	Daily		Weekly	
	(1)	(2)	(3)	(4)
constant	-0.005	-0.006	-0.42	-0.43
	(-1.08)	(-1.41)	(-1.21)	(-1.22)
$OF_{i,t}^{USD}$	0.013***	0.006***	0.027***	0.020***
	(21.56)	(9.66)	(7.99)	(4.72)
$OF_{i,t}^{EUR}$	0.005***	0.001	0.010***	0.006**
	(9.00)	(1.58)	(3.05)	(2.23)
$OF_{i,t}^{GBP}$	0.003**	0.001	-0.002	-0.004
	(2.05)	(0.22)	(-0.38)	(-0.85)
$OF_{i,t}^{JPY}$	0.001***	-0.001	0.0001	-0.001
	(2.97)	(-1.07)	(0.03)	(-0.74)
$OF_{i,t}^{KRW}$	0.011***	0.006***	0.023***	0.017***
	(23.94)	(12.21)	(10.17)	(6.04)
$OF_{i,t}^{CAD}$	-0.001	-0.003*	0.002	0.001
	(-0.26)	(-1.88)	(1.18)	(0.37)
$OF_{i,t}^{AUD}$	-0.001	-0.003**	0.004**	0.002
	(-0.49)	(-2.10)	(1.99)	(0.88)
$OF_{i,t}^{CHF}$	-0.001	-0.003***	0.001	-0.0004
	(-0.68)	(-2.63)	(0.67)	(-0.19)
$OF_{i,t}^{NZD}$	0.002	-0.001	0.002	-0.001
	(1.40)	(-0.11)	(0.85)	(-0.44)
$OF_{i,t}^{NOK}$	-0.001	-0.003**	-0.008	-0.010
	(-0.66)	(-2.02)	(-0.76)	(-0.99)
$OF_{i,t}^{SEK}$	0.002	0.001	-0.002	-0.004
	(1.18)	(0.05)	(-0.55)	(-1.22)
$OF_{i,t}^{G11}$		0.020***		0.021***
		(26.42)		(4.97)
$\bar{R}^2$ (%)	1.1	1.3	4.8	4.9
# obs.	90364	90364	16728	16728

Table 5: Predictive Panel Regressions with All Order Flows

The table displays results of predictive panel regressions of cryptocurrency returns on lagged disaggregated international order flows, world (G11) order flow and control variables. The table contains the same information as Table 3 but for all order flows plus the world order flow. The t-statistics shown in parentheses are computed using Newey and West (1987) standard errors. The sample period ranges from February 14, 2018 to June 30, 2022. The panel regressions also condition on the control variables reported in the previous tables but these results are not reported to save space.

	Daily		Weekly	
	(1)	(2)	(3)	(4)
constant	0.008*	0.008*	-0.002	-0.0052
	(1.78)	(1.79)	(-0.04)	(-0.09)
$OF_{i,t-1}^{USD}$	-0.002***	-0.002***	0.006*	0.001
	(-3.01)	(-3.03)	(1.95)	(0.58)
$OF_{i,t-1}^{EUR}$	0.001	0.001	-0.002	-0.005
	(0.68)	(0.71)	(-0.44)	(-0.85)
$OF_{i,t-1}^{GBP}$	0.001	0.001	0.003	0.001
	(1.38)	(1.40)	(0.68)	(0.27)
$OF_{i,t-1}^{JPY}$	0.001	0.001	0.001	0.0003
	(0.25)	(0.30)	(0.73)	(0.16)
$OF_{i,t-1}^{KRW}$	0.002	0.002	0.009	0.005
	(1.37)	(1.40)	(1.08)	(0.67)
$OF_{i,t-1}^{CAD}$	-0.001	-0.001	0.001	0.0001
	(-1.35)	(-1.33)	(0.76)	(0.04)
$OF_{i,t-1}^{AUD}$	-0.002	-0.002	-0.0002	-0.002
	(-1.23)	(-1.22)	(-0.12)	(-0.86)
$OF_{i,t-1}^{CHF}$	0.001	0.001	0.0003	-0.001
	(0.85)	(0.88)	(0.15)	(-0.41)
$OF_{i,t-1}^{NZD}$	-0.002	-0.002	0.0004	-0.001
	(-1.32)	(-1.31)	(0.12)	(-0.34)
$OF_{i,t-1}^{NOK}$	-0.001	-0.001	0.002	0.001
	(-0.77)	(-0.75)	(1.34)	(0.56)
$OF_{i,t-1}^{SEK}$	0.001	0.001	0.003	0.002
	(0.76)	(0.30)	(0.56)	(0.32)
$OF_{i,t-1}^{G11}$		-0.001		0.015***
		(-0.45)		(3.08)
$r_{i,t-1}$	-0.013	-0.013	-0.006	-0.007
	(-1.24)	(-1.23)	(-1.39)	(-1.38)
$\bar{R}^2$ (%)	0.1	0.1	0.1	0.1
# obs.	90282	90282	16646	16646

Table 6: Out-of-Sample Statistical Performance

This table reports the out-sample statistical performance using the R^2_{oos} statistic in percent based on the Mean Squared Error (MSE) of the forecasts. Daily and weekly results are reported for three models: OF only conditions on order flows; EF only conditions on economic fundamentals; and OF+EF conditions on both. All models condition on lagged returns. Bold entries indicate the highest value in each row.

	Daily R^2_{oos} (%)			Weekly R^2_{oos} (%)		
	OF	EF	OF+EF	OF	EF	OF+EF
OLS	**-0.56**	-1.24	-12.72	**1.27**	-19.48	-20.24
Ridge	**-0.56**	-11.17	-11.48	**1.24**	-20.38	-20.40
Lasso	**-0.17**	-0.17	-0.19	**1.28**	01.17	-0.05
Elastic net	-0.31	**0.21**	0.01	**1.44**	-3.65	-3.71
PCR	**-0.65**	-1.30	-0.84	**1.02**	-11.21	-10.84
Random Forest	**0.61**	-0.34	0.03	**0.69**	-0.45	0.25
SGB Model	-1.01	-0.90	**0.21**	**0.64**	-2.17	-3.33
Neural Net 1	**0.19**	0.04	0.15	**0.68**	-1.46	-1.53
Neural Net 2	0.24	**0.28**	0.20	**0.74**	-1.82	-2.43
Neural Net 3	**0.28**	-3.91	0.28	**0.78**	-1.63	-3.08
Neural Net 4	**0.25**	0.02	-1.30	**0.82**	0.30	-1.45
L-Mean	**-0.35**	-1.69	-1.71	**1.38**	-5.06	-5.10
NL-Mean	**0.59**	0.09	0.38	**0.93**	-0.01	-0.40
ALL-Mean	**0.44**	0.20	0.38	**1.36**	0.03	-0.16

Table 7: Portfolios Sorted on Order Flow - Full Sample

This table displays the performance of cryptocurrency portfolios sorted on order flow. P_1 represents the portfolio with the lowest lagged order flow and P_5 the portfolio with the highest lagged order flow. ALL is the average return across all cryptocurrencies. The portfolios are equally-weighted and rebalanced daily (Panel A) or weekly (Panel B). The alpha is based on a three-factor model, which includes the market factor, the size factor and momentum. The returns and alphas are in daily/weekly percent. SR is the annualized Sharpe ratio. The Newey and West (1987) t-statistics are shown in parenthesis. The full sample period ranges from February 14, 2018 to June 30, 2022.

Panel A: Daily Portfolios Sorted on Order Flow

	ALL	P_1	P_2	P_3	P_4	P_5	$P_1 - P_5$		
	Mean return (% daily)						Mean	alpha	SR
US OF	0.34	0.45	0.29	0.68	0.19	0.05	0.40	0.40	1.84
(t-stat)	(1.72)	(2.38)	(0.98)	(1.36)	(0.98)	(0.24)	(3.94)	(3.78)	
World OF	0.34	0.41	0.19	0.21	0.76	0.12	0.29	0.32	1.20
(t-stat)	(1.72)	(2.09)	(1.06)	(1.16)	(1.42)	(0.63)	(2.71)	(2.76)	

Panel B: Weekly Portfolios Sorted on Order Flow

	ALL	P_1	P_2	P_3	P_4	P_5	$P_5 - P_1$		
	Mean return (% weekly)						Mean	alpha	SR
US OF	1.88	0.71	1.55	1.65	1.49	4.77	4.06	1.83	0.72
(t-stat)	(1.76)	(0.73)	(1.52)	(1.60)	(1.42)	(1.64)	(1.47)	(3.37)	
World OF	1.88	1.02	1.19	1.47	1.43	5.05	4.03	1.83	0.72
(t-stat)	(1.76)	(1.07)	(1.15)	(1.53)	(1.44)	(1.73)	(1.48)	(2.95)	

Table 8: Daily Portfolios Conditioning on Order Flow - ML Test Sample

This table displays the performance of cryptocurrency portfolios that condition on daily order flow. P_1 (P_5) represents the portfolio with the lowest (highest) lagged order flow (Panel A) or lowest (highest) ML forecast conditioning on all order flows (Panel B). The portfolios are equally-weighted and rebalanced daily. Results are reported for three models: OF only conditions on order flows; EF only conditions on economic fundamentals; and OF+EF conditions on both. The alpha is based on a three-factor model, which includes the market factor, the size factor and momentum. The returns and alphas are in daily percent. SR is the annualized Sharpe ratio. The Newey and West (1987) t-statistics are shown in parenthesis. The sample period ranges from February 18, 2020 to June 30, 2022.

Panel A: Daily Portfolios Sorted on Order Flow

	$P_1 - P_5$					
	Mean	alpha	SR			
US OF	0.32	0.31	1.55			
(t-stat)	(2.63)	(2.31)				
World OF	0.21	0.23	0.94			
(t-stat)	(1.48)	(1.53)				

Panel B: Daily Portfolios Sorted on ML Forecasts Conditioning on All Order Flows

	OF Model $P_5 - P_1$			**EF Model** $P_5 - P_1$			**OF+EF Model** $P_5 - P_1$		
	Mean	alpha	SR	Mean	alpha	SR	Mean	alpha	SR
OLS	0.63	0.59	3.13	0.99	0.94	3.78	0.85	0.79	3.74
(t-stat)	(4.69)	(4.56)		(5.74)	(5.84)		(5.36)	(5.72)	
Ridge	0.60	0.55	3.00	0.83	0.77	3.31	0.76	0.69	3.59
(t-stat)	(4.51)	(4.35)		(3.60)	(5.05)		(5.20)	(5.42)	
Lasso	0.67	0.70	3.09	0.25	0.29	1.11	0.22	0.25	1.00
(t-stat)	(4.26)	(4.98)		(2.15)	(2.05)		(1.61)	(1.86)	
Elastic net	0.30	0.31	1.54	0.33	0.36	1.46	0.48	0.52	2.12
(t-stat)	(2.33)	(2.44)		(2.31)	(2.59)		(3.34)	(3.63)	
PCR	0.13	0.11	0.62	0.09	0.05	0.38	0.21	0.15	0.92
(t-stat)	(1.01)	(0.81)		(0.62)	(0.31)		(1.44)	(1.06)	
Random Forest	1.00	1.04	4.39	0.65	0.59	2.60	0.63	0.58	2.80
(t-stat)	(6.97)	(7.10)		(4.31)	(3.91)		(4.34)	(4.13)	
SGB Model	0.93	0.98	3.70	0.96	0.96	4.13	0.85	0.89	3.70
(t-stat)	(5.53)	(6.17)		(5.85)	(6.37)		(5.62)	(6.01)	
Neural Net 1	0.80	0.85	3.41	0.69	0.65	2.86	0.62	0.61	2.76
(t-stat)	(4.79)	(5.68)		(3.73)	(4.28)		(3.99)	(4.15)	
Neural Net 2	0.97	0.98	4.44	1.00	0.96	3.87	0.61	0.58	2.62
(t-stat)	(7.23)	(6.91)		(4.51)	(5.85)		(4.07)	(3.89)	
Neural Net 3	0.82	0.85	3.27	0.40	0.40	1.87	0.70	0.71	3.06
(t-stat)	(4.78)	(5.39)		(3.03)	(2.89)		(4.77)	(4.78)	
Neural Net 4	0.72	0.77	3.07	0.49	0.54	2.14	0.68	0.74	2.92
(t-stat)	(4.39)	(5.19)		(4.21)	(3.68)		(4.93)	(4.96)	
L-Mean	0.61	0.57	3.04	0.81	0.75	3.26	0.78	0.72	3.54
(t-stat)	(4.93)	(4.45)		(5.27)	(4.99)		(5.27)	(5.34)	
NL-Mean	1.23	1.28	5.05	0.98	0.92	3.75	1.04	0.95	4.17
(t-stat)	(7.61)	(8.31)		(5.72)	(5.94)		(5.92)	(6.17)	
ALL-Mean	0.97	1.00	4.46	0.97	0.91	3.83	0.88	0.81	3.85
(t-stat)	(7.14)	(7.05)		(5.90)	(6.00)		(5.77)	(5.96)	

Table 9: Weekly Portfolio Sorts Conditioning on Order Flow - ML Test Sample

This table displays the performance of portfolios that condition on weekly order flow. P_1 (P_5) represents the cryptocurrency portfolio with the lowest (highest) lagged order flow or lowest (highest) ML forecast based on all order flows. The portfolios are equally-weighted and rebalanced weekly. Results are reported for three models: OF only conditions on order flows; EF only conditions on economic fundamentals; and OF+EF conditions on both. The alpha is based on a three-factor model, which includes the market factor, the size factor and momentum. The returns and alphas are in weekly percent. SR is the annualized Sharpe ratio. The Newey and West (1987) t-statistics are shown in parenthesis. The sample period ranges from Feb 18, 2020 to June 30, 2022.

Panel A: Weekly Portfolios Sorted on Order Flow

	$P_5 - P_1$		
	Mean	alpha	SR
US OF	1.53	1.71	1.67
(t-stat)	(2.75)	(2.45)	
World OF	1.79	1.56	1.97
(t-stat)	(2.53)	(2.31)	

Panel B: Weekly Portfolios Sorted on ML Forecasts Conditioning on All Order Flows

	OF Model $P_5 - P_1$			**EF Model** $P_5 - P_1$			**OF+EF Model** $P_5 - P_1$		
	Mean	alpha	SR	Mean	alpha	SR	Mean	alpha	SR
OLS	1.81	1.54	2.41	-0.02	-0.28	-0.02	1.37	1.26	1.35
(t-stat)	(3.38)	(2.77)		(-0.02)	(-0.34)		(2.03)	(1.67)	
Ridge	1.86	1.59	2.48	0.08	-0.14	0.09	1.02	0.92	1.08
(t-stat)	(3.34)	(2.86)		(0.12)	(-0.20)		(1.58)	(1.30)	
Lasso	1.83	1.83	2.23	1.01	0.99	1.13	0.95	0.92	1.07
(t-stat)	(3.14)	(2.94)		(1.46)	(1.47)		(1.39)	(1.38)	
Elastic net	1.35	1.16	1.62	0.84	0.79	0.89	0.84	0.78	0.90
(t-stat)	(2.40)	(1.87)		(1.25)	(1.11)		(1.28)	(1.10)	
PCR	0.98	0.93	1.10	-0.29	-0.03	-0.24	-0.90	-0.79	-0.87
(t-stat)	(1.48)	(1.40)		(-0.30)	(-0.03)		(-1.17)	(-1.03)	
Random Forest	0.89	0.89	1.09	0.11	0.10	0.10	1.17	0.84	1.30
(t-stat)	(1.41)	(1.42)		(0.16)	(0.11)		(2.00)	(1.27)	
SGB Model	1.84	1.64	1.80	0.87	0.73	0.91	1.75	1.64	1.72
(t-stat)	(2.93)	(2.14)		(1.45)	(1.04)		(2.38)	(2.19)	
Neural Net 1	0.01	0.14	0.01	-0.54	-0.96	-0.47	0.41	0.10	0.41
(t-stat)	(0.01)	(0.21)		(-0.70)	(-1.14)		(0.54)	(0.15)	
Neural Net 2	1.01	0.93	1.27	0.35	0.05	0.35	-0.79	-0.56	-0.88
(t-stat)	(2.00)	(1.54)		(0.42)	(0.06)		(-1.09)	(-0.85)	
Neural Net 3	1.55	1.53	2.14	-1.58	-1.59	-1.40	-0.03	-0.28	-0.03
(t-stat)	(3.59)	(2.76)		(-1.94)	(-1.93)		(-0.04)	(-0.34)	
Neural Net 4	0.95	0.90	1.05	-0.70	-0.79	-0.63	-0.14	-0.38	-0.14
(t-stat)	(1.30)	(1.30)		(-0.82)	(-0.98)		(-0.23)	(-0.53)	
L-Mean	1.91	1.68	2.49	-0.19	-0.46	-0.19	0.55	0.43	0.52
(t-stat)	(3.55)	(2.93)		(-0.28)	(-0.61)		(0.79)	(0.55)	
NL-Mean	1.61	1.47	1.82	-0.69	-0.90	-0.64	1.82	1.48	1.94
(t-stat)	(2.42)	(2.20)		(-0.87)	(-1.10)		(2.88)	(2.14)	
ALL-Mean	2.02	1.73	2.39	-0.17	-0.44	-0.16	0.87	0.67	0.88
(t-stat)	(3.32)	(2.76)		(-0.22)	(-0.58)		(1.31)	(0.95)	

Table 10: Daily Long and Long-Short Portfolios Conditioning on Order Flow

This table compares the performance of the long portfolio to the long-short portfolio. The portfolios are sorted on daily order flow (Panel A) or ML forecasts from the OF model that conditions only on daily order flow (Panel B). In Panel A, the long portfolio is the bottom quintile portfolio (P_1) and the long-short portfolio is P_1-P_5. In Panel B, the long portfolio is the top quintile portfolio (P_5) and the long-short portfolio is P_5-P_1. The portfolios are equally-weighted and rebalanced daily. The alpha is based on a three-factor model, which includes the market factor, the size factor and momentum. The returns and alphas are in daily percent. SR is the annualized Sharpe ratio. MDD is the maximum drawdown. TO is the daily turnover. BE-TC is the break-even transaction cost in daily percent. The Newey and West (1987) t-statistics are shown in parenthesis. The sample period ranges from February 18, 2020 to June 30, 2022.

Panel A: Daily Portfolios Sorted on Order Flow

	Long Portfolio (P_1)						Long-Short Portfolio ($P_1 - P_5$)					
	Mean	alpha	SR	MDD	TO	BE TC	Mean	alpha	SR	MDD	TO	BE TC
US OF	0.65	0.62	1.68	0.69	0.75	0.86	0.32	0.31	1.54	0.37	1.51	0.21
(t-stat)	(2.54)	(2.47)					(2.63)	(2.31)				
World OF	0.57	0.58	1.43	0.71	0.74	0.77	0.21	0.23	0.93	0.52	1.49	0.15
(t-stat)	(2.13)	(2.25)					(1.48)	(1.53)				

Panel B: Daily Portfolios Sorted on ML Forecasts Conditioning on All Order Flows (OF Model)

	Long Portfolio (P_5)						Long-Short Portfolio ($P_5 - P_1$)					
	Mean	alpha	SR	MDD	TO	BE TC	Mean	alpha	SR	MDD	TO	BE TC
OLS	0.76	0.73	1.91	0.63	0.80	0.94	0.63	0.59	3.13	0.48	1.58	0.39
(t-stat)	(2.91)	(2.82)					(4.69)	(4.56)				
Ridge	0.75	0.72	1.88	0.64	0.80	0.93	0.60	0.55	3.00	0.43	1.58	0.37
(t-stat)	(2.83)	(2.78)					(4.51)	(4.35)				
Lasso	0.80	0.82	1.93	0.64	0.60	1.32	0.67	0.70	3.09	0.43	1.19	0.56
(t-stat)	(2.88)	(3.06)					(4.26)	(4.98)				
Elastic net	0.61	0.61	1.56	0.61	0.60	0.99	0.30	0.31	1.54	0.56	1.19	0.24
(t-stat)	(2.46)	(2.40)					(2.33)	(2.44)				
PCR	0.59	0.57	1.48	0.65	0.79	0.73	0.13	0.11	0.62	0.66	1.57	0.08
(t-stat)	(2.27)	(2.20)					(1.01)	(0.81)				
Random Forest	1.02	1.02	2.40	0.63	0.77	0.32	1.00	1.04	4.39	0.26	1.56	0.64
(t-stat)	(3.55)	(3.74)					(6.97)	(7.10)				
SGB Model	0.94	0.94	2.25	0.62	0.70	1.34	0.93	0.98	3.70	0.27	1.40	0.66
(t-stat)	(3.59)	(3.49)					(5.53)	(6.17)				
Neural Net 1	0.80	0.82	1.96	0.65	0.81	1.00	0.80	0.85	3.41	0.33	1.60	0.51
(t-stat)	(3.04)	(3.10)					(4.79)	(5.68)				
Neural Net 2	0.89	0.88	2.23	0.61	0.81	1.09	0.97	0.98	4.44	0.23	1.61	0.60
(t-stat)	(3.46)	(3.38)					(7.23)	(6.91)				
Neural Net 3	0.85	0.87	2.03	0.68	0.76	1.12	0.82	0.85	3.27	0.31	1.50	0.54
(t-stat)	(3.13)	(3.20)					(4.78)	(5.39)				
Neural Net 4	0.89	0.90	2.08	0.58	0.42	2.09	0.72	0.77	3.07	0.40	0.82	0.86
(t-stat)	(3.26)	(3.28)					(4.39)	(5.19)				
L-Mean	0.78	0.75	1.96	0.63	0.80	0.96	0.61	0.57	3.04	0.38	1.59	0.38
(t-stat)	(3.02)	(2.91)					(4.93)	(4.45)				
NL-Mean	1.10	1.11	2.60	0.59	0.78	1.40	1.23	1.28	5.05	0.24	1.57	0.78
(t-stat)	(3.95)	(4.09)					(7.61)	(8.31)				
ALL-Mean	0.96	0.98	2.33	0.62	0.80	1.19	0.97	1.00	4.46	0.31	1.60	0.61
(t-stat)	(3.54)	(3.66)					(7.14)	(7.05)				

Table 11: Weekly Long and Long-Short Portfolios Based on Order Flow

This table compares the performance of the long portfolio to the long-short portfolio. The portfolios are sorted on weekly order flow (Panel A) or ML forecasts from the OF model that conditions only on weekly order flow (Panel B). The long portfolio is the top quintile portfolio (P_5) and the long-short portfolio is P_5-P_1. The portfolios are equally-weighted and rebalanced weekly. The alpha is based on a three-factor model, which includes the market factor, the size factor and momentum. The returns and alphas are in weekly percent. SR is the annualized Sharpe ratio. MDD is the maximum drawdown. TO is the weekly turnover. BE-TC is the break-even transaction cost in weekly percent. The Newey and West (1987) t-statistics are shown in parenthesis. The sample period ranges from February 18, 2020 to June 30, 2022.

Panel A: Weekly Portfolios Sorted on Order Flow

	Long Portfolio (P_5)						Long-Short Portfolio ($P_5 - P_1$)					
	Mean	alpha	SR	MDD	TO	BE TC	Mean	alpha	SR	MDD	TO	BE TC
US OF	2.44	0.83	1.19	0.67	0.73	3.84	1.53	1.71	1.65	0.26	1.49	1.14
(t-stat)	(1.67)	(1.42)					(2.75)	(2.45)				
World OF	3.10	1.22	1.46	0.68	0.75	4.51	1.79	1.56	1.94	0.26	1.49	1.30
(t-stat)	(2.04)	(1.93)					(2.53)	(2.31)				

Panel B: Weekly Portfolios Sorted on ML Forecasts Conditioning on All Order Flows (OF Model)

	Long Portfolio (P_5)						Long-Short Portfolio ($P_5 - P_1$)					
	Mean	alpha	SR	MDD	TO	BE TC	Mean	alpha	SR	MDD	TO	BE TC
OLS	3.13	1.23	1.47	0.65	0.77	3.87	1.81	1.54	2.41	0.27	1.54	1.09
(t-stat)	(2.06)	(2.39)					(3.38)	(2.77)				
Ridge	3.14	1.24	1.48	0.66	0.77	3.88	1.86	1.59	2.48	0.30	1.54	1.12
(t-stat)	(2.05)	(2.37)					(3.34)	(2.86)				
Lasso	2.94	1.28	1.38	0.63	0.61	4.79	1.83	1.83	2.23	0.31	1.22	1.52
(t-stat)	(1.89)	(2.20)					(3.14)	(2.94)				
Elastic net	2.76	0.87	1.31	0.66	0.68	4.07	1.35	1.16	1.62	0.31	1.01	1.35
(t-stat)	(1.84)	(1.51)					(2.40)	(1.87)				
PCR	2.34	0.57	1.09	0.73	0.78	2.80	0.98	0.93	1.10	0.80	0.53	1.55
(t-stat)	(1.53)	(0.95)					(1.48)	(1.40)				
Random Forest	2.27	0.46	1.13	0.70	0.75	2.95	0.89	0.89	1.09	0.27	0.58	1.51
(t-stat)	(1.58)	(0.85)					(1.41)	(1.42)				
SGB Model	2.78	0.87	1.28	0.66	0.71	3.83	1.84	1.64	1.80	0.27	1.41	1.39
(t-stat)	(1.90)	(1.34)					(2.93)	(2.14)				
Neural Net 1	1.82	0.03	0.89	0.69	0.75	2.48	0.00	0.14	0.00	0.59	1.52	0.13
(t-stat)	(1.27)	(0.05)					(0.00)	(0.21)				
Neural Net 2	2.84	1.08	1.39	0.63	0.78	3.46	1.01	0.93	1.27	0.74	1.56	0.61
(t-stat)	(1.87)	(1.99)					(2.00)	(1.54)				
Neural Net 3	2.50	0.74	1.20	0.64	0.73	3.37	1.55	1.53	2.14	0.62	1.49	1.03
(t-stat)	(1.66)	(1.35)					(3.59)	(2.76)				
Neural Net 4	2.70	0.82	1.22	0.79	0.46	5.87	0.95	0.90	1.05	0.68	0.94	1.03
(t-stat)	(1.51)	(1.36)					(1.30)	(1.30)				
L-Mean	3.10	1.27	1.47	0.65	0.76	3.87	1.91	1.68	2.49	0.39	1.53	1.16
(t-stat)	(2.06)	(2.27)					(3.55)	(2.93)				
NL-Mean	2.81	0.92	1.35	0.64	0.74	3.71	1.61	1.47	1.82	0.31	1.50	1.06
(t-stat)	(1.94)	(1.61)					(2.42)	(2.20)				
ALL-Mean	3.14	1.29	1.48	0.63	0.75	3.94	2.02	1.73	2.39	0.32	1.53	1.22
(t-stat)	(2.10)	(2.27)					(3.32)	(2.76)				

Order Flow and Cryptocurrency Returns

Online Appendix

Table of Contents:

- Section A1 provides details on the **models** used in this paper.
- Section A2 provides details on the **setup of the hyperparameter optimization**.
- Section A3 provides details on our estimation of variable importance using **Shapley values**.
- Table **??** presents the **set of hyperparameters** used in this paper.
- Additional tables report summary statistics and the robustness of the empirical results.

A1 Machine Learning Models

Given our balanced panel of coin returns, we index coins as $i = 1, ..., N$ and days by $t = 1, ..., T$. In its most general form, we describe the future return of coin i at time $t+1$ as:

$$r_{i,t+1} = E_t[r_{i,t+1}] + \epsilon_{i,t+1}, \tag{A.1}$$

where

$$E_t[r_{i,t+1}] = g(z_{i,t}), \tag{A.2}$$

is the expected return at time $t+1$ of coin i and $g(\cdot)$ is a function of coin i's P-dimensional vector of characteristics $z_{i,t}$ at time t. Inspired by Gu, Kelly and Xiu (2020), we approximate $g(\cdot)$ using a variety of models with increasing complexity, with the objective of maximizing the out-of-sample predictive power for the realized return $r_{i,t+1}$.

A1.1 Linear Regression

Our benchmark model is the linear regression model, which does not incorporate nonlinearities or interactions between predictors. The linear regression model assumes $g(\cdot)$ can be approximated by a linear function of predictors $z_{i,t}$ and parameter vector θ:

$$g(z_{i,t}; \theta) = z'_{i,t}\theta. \tag{A.3}$$

The pooled OLS estimator is obtained by minimizing the standard ℓ_2 objective function:

$$\mathcal{L}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(r_{i,t+1} - g(z_{i,t}; \theta))^2.0 \tag{A.4}$$

The OLS estimator is unbiased and efficient if the number of predictors P is relatively small

compared to observations T. However, when P approaches T, the OLS estimator becomes inefficient, and begins to overfit. To mitigate this problem, we apply a variety of ML models that have been recently used in the asset pricing literature.

A1.2 Penalized Linear Regression: Ridge, Lasso, and Elastic Net

A common technique for mitigating OLS's overfitting issue is to shrink the estimated parameters by adding a penalty term to Equation (A.4). This regularization approach intentionally reduces a model's in-sample fit to improve its out-of-sample prediction. Rather than using the objective function in Equation (A.4), penalized linear models estimate θ by minimizing:

$$\mathcal{L}(\theta;\cdot) = \mathcal{L}(\theta) + \phi(\theta;\cdot), \tag{A.5}$$

where $\phi(\theta;\cdot)$ is the the widely-used elastic net (EN) penalty on θ defined as follows:

$$\phi(\theta;\lambda,\rho) = \lambda(1-\rho)\sum_{j=1}^{P}|\theta_j| + \frac{1}{2}\lambda\rho\sum_{j=1}^{P}\theta_j^2, \tag{A.6}$$

where λ is a nonnegative hyperparameter controlling the degree of shrinkage so that higher values of λ lead to greater shrinkage of the estimated parameters. When $\lambda = 0$, the model reduces to OLS. When $\lambda > 0$, EN encompasses two well-known regularization methods as special cases, determined by the nonnegative hyperparameter ρ. When $\rho = 0$, Equation (A.5) corresponds to the lasso, which imposes an ℓ_1 parameter penalization. Lasso can set a subset of θ to exactly zero, imposing sparsity on the model and thus performing variable selection. When $\rho = 1$, Equation (A.5) corresponds to ridge regression, which imposes an ℓ_2 parameter penalization. Ridge shrinks coefficient estimates toward zero, but keeps all predictors in the model. For values of ρ between 0 and 1, the elastic net combines the advantages of ridge regression and the lasso by performing both shrinkage and variable selection. A key feature

of the EN, derived from ridge regression, is its ability to handle highly correlated predictors, such as order flow variables (Diebold and Shin, 2019). We optimize the nonnegative tuning parameter, λ, using the validation sample (see Section A2 for more details).

A1.3 Principal Component Regression

Principal component regression (PCR) is a classic dimension reduction technique involving a two-step procedure. In the first step, principal components analysis (PCA) condenses P predictors into a much smaller number of K linear combinations that best capture the common variation among the predictors. Specifically, PCA linearly transforms a set of predictors into orthogonal principal components, in which the first component has the highest variance, the second component has the second highest variance, and so on, up to $K < P$ principal components.

Mathematically, the j^{th} principal component Zw_j can be solved as:

$$w_j = \arg\max_{w} Var(Zw), \quad \text{s.t. } w'w = 1, \quad Cov(Zw, Zw_l) = 0, \quad l = 1, 2, ..., j-1, \quad \text{(A.7)}$$

where Z is the $NT \times P$ matrix of standardized predictors $z_{i,t}$, and w_j is the eigenvector corresponding to the jth eigenvalue of $Var(Z)$.

In the second step, PCR uses the first K leading components as regressors in a linear regression estimated with OLS. Thus, PCR acts as a regularization technique by zeroing out coefficients associated with low variance components.

Equation (A.7) illustrates that the first K principal components are selected to optimally explain the variation in the P predictors. However, there is no guarantee that these principal components are the most effective for forecasting future returns, as $r_{i,t+1}$ is not considered in Equation (A.7). Therefore, a potential drawback of PCR is that the first K principal

components, while exhibiting high variance, may not necessarily have a significant relation with $r_{i,t+1}$. Consequently, the number of principal components used, K, is an important tuning parameter that is optimized using the validation sample as detailed in Section A2.

A1.4 Gradient Boosted Regression Trees and Random Forests

Regression trees are a nonparametric ML technique, which handles effectively nonlinearities and interactions across multiple predictors. A tree "grows" through a series of steps, with each step creating new "branches" that sort the remaining data from the previous step into bins based on one of the predictors. This sequential process partitions the predictor space into rectangular regions. The final outputs are the average values of returns $r_{i,t+1}$ within each region. Formally, the forecast of a tree with K regions and depth L, can be expressed as:

$$g(z_{i,t};\theta,K,L)=\sum_{k=1}^{K}\theta_k 1_{\left(z_{i,t}\in C_k(L)\right)}, \tag{A.8}$$

where $C_k(L)$ represents one of the K regions into which the data is partitioned, $1_{(\cdot)}$ is an indicator function denoting whether $z_{i,t}$ falls within the region $C_k(L)$, and θ_k is a constant parameter estimated as the sample average of returns within that region.

To estimate θ in Equation (A.8), we follow the algorithm of Breiman et al. (1984). At each new level, the predictor and threshold value that minimizes the forecast error in the resulting two branches is selected. This procedure, known as recursive binary splitting, is a greedy algorithm because each split is chosen based on immediate results, without considering the potential implications for future branches. The forecast error for each branch C is defined as the objective function:

$$H(\theta,C)=\frac{1}{|C|}\sum_{z_{i,t}\in C}(r_{i,t+1}-\theta)^2, \tag{A.9}$$

where $|C|$ represents the number of observations within branch C. This loss function naturally leads to the optimal choice of θ being the sample average of returns within each branch, i.e., $\theta = \frac{1}{|C|} \sum_{z_{i,t} \in C} r_{i,t+1}$. Branching continues until the maximum specified depth is reached or the number of observations in each region is below a specified count.

Trees are invariant to monotonic transformations of predictors, making them robust to outliers in the predictor data. Their high flexibility allows them to capture complex multiway interactions among predictors. However, their flexibility could also lead trees to overfit on the training sample, necessitating regularization. To address this, we employ two types of "ensemble" models that regularize by aggregating forecasts from many different trees into a single forecast.

The first type of ensemble model, random forest (RF), is a variant of a technique known as bootstrap aggregation or "bagging". This bagging process involves several steps. First, create B new samples (called bootstrap samples), each containing the same number of observations as the original dataset, by randomly drawing from the original dataset with replacement. Second, estimate a separate regression tree on each of these bootstrap samples. These trees are typically deep and complex, leading to low bias but high variance. Third, average over all B trees to reduce this variance, resulting in a more stable model.

RF improves on bagging to further decrease the correlation among trees from different bootstrap samples (Breiman, 2001). Unlike traditional bagging, where all P predictors might be considered at each split, RF reduces tree correlation by only allowing a random subset of $m < P$ predictors to be considered for each split. This procedure enhances the variance reduction of the model beyond what is achieved with traditional bagging, improving overall model stability.

The second type of ensemble model is gradient boosted regression trees (GBT), which grows regression trees sequentially (Friedman, 2001). Each tree is based on the residuals of the

previously grown trees. Unlike bagging, which favors deep trees, gradient boosting works best with shallow trees or "stumps" that have only a single split.

GBT starts by fitting the training data with a single shallow tree, which, being a weak predictor, has a large bias. Next, a second shallow tree of the same depth is estimated based on the residuals from the first tree. The forecast from this second tree is then shrunken by a learning rate $\lambda \in (0, 1)$ and added to the forecast from the first tree. Shrinkage helps the model learn gradually and prevents it from overfitting the residuals of the preceding tree. This process is repeated until the ensemble comprises B shallow trees, with the final output being the sum of the scaled forecasts from each tree.

Furthermore, we implement a regularization technique known as stochastic gradient boosting, in which a random subsample of the data is used for each tree. Similar to RF, this introduces randomness to reduce correlations between predictions, enhancing the model's variance reduction compared to the original gradient boosting algorithm. Additionally, it considerably accelerates the optimization process.

For both types of ensemble models, we optimize the number of trees (B) and the maximum depth (L) using the validation set. Additionally, for GBT, we also tune the learning rate (λ). See Section A2 for a detailed discussion.

A1.5 Neural Networks

Neural networks are powerful machine learning models that can approximate any continuous function (Hornik et al., 1989). We use feed-forward neural networks that include an "input layer" of predictors, one or more "hidden layers" that facilitate interactions between predictors and nonlinearly transform them, and an "output layer" that aggregates the hidden layers into a final forecast.

Mathematically, a feed-forward neural network can be expressed with the following general formula. Let $K^{(l)}$ denote the number of neurons in each layer, for $l = 1, ..., L$. The output of neuron k in layer l is denoted as $x^{(l)}_{(k)}$, and $x^{(l)} = (1, x^{(l)}_{(1)}, ..., x^{(l)}_{(K)})'$ is the vector of outputs for layer l, with a constant. The vector $x^{(0)}$ represents the input layer of predictors. Each neuron in hidden layer $0 < l < L$ can then be recursively expressed as:

$$x^{(l)}_{(k)} = f(x^{(l-1)'}\theta^{(l-1)}_k), \tag{A.10}$$

where $x^{(l-1)'}$ represents the transposed output vector from the previous layer $l-1$, and $\theta^{(l-1)}_k$ is the vector of weights connecting this output to neuron k in layer l. The function f denotes a nonlinear "activation function" that transforms the weighted sum of inputs. For our neural networks, we use the Rectified Linear Unit (Relu) as the activation function, defined as:

$$f(u) = \begin{cases} 0 \text{ if } u < 0 \\ u \text{ otherwise.} \end{cases} \tag{A.11}$$

This activation function is widely adopted due to its simplicity and effectiveness, particularly in improving the convergence of the neural networks during training.

The final output for the neural network is given by:

$$g(z; \theta) = x^{(L-1)'}\theta^{(L-1)}, \tag{A.12}$$

which is a linear function of the outputs from the last hidden layer, aggregating these non-linearly transformed signals into a single predictive output for the network.

We consider architectures with up to four hidden layers. Our shallowest neural network, denoted NN1, has a single hidden layer of 16 neurons. Next, NN2 has two hidden layers with 16 and 8 neurons, respectively; and NN3 has three hidden layers with 16, 8, and 4

neurons respectively. Our deepest neural network, NN4, has four hidden layers with 16, 8, 4, and 2 neurons, respectively. By comparing the performance across neural networks with different depths, we can assess the value of complexity for forecasting cryptocurrency returns.

To estimate the weight parameters, we minimize the ℓ_2 objective function given in Equation (A.4). This optimization procedure is generally computationally intensive due to the complexity and nonconvexity of neural networks, which can lead to numerous local minima. To address this challenge, we employ stochastic gradient descent (SGD). SGD operates by computing the gradient from a randomly selected small subset of the data at each iteration and simultaneously updating all model parameters. While SGD is somewhat less precise at each step compared to the standard gradient descent method, which uses the full training dataset to calculate the gradient, it offers significant computational speed benefits. To further enhance computational efficiency, we use the Adaptive Moment Estimation (Adam) algorithm. Adam optimizes the learning rate by adaptively reducing it as the gradient nears zero, thus facilitating significantly faster convergence.

Estimating neural networks involves managing a large number of weight parameters and handling severe nonlinearities, necessitating substantial regularization to ensure model stability. To achieve this, we employ several regularization techniques, similar to Gu, Kelly and Xiu (2020). First, we incorporate an ℓ_1 penalty into the loss function to encourage the model to learn a sparse representation of the predictors, reducing complexity and mitigating overfitting. Second, we implement "early stopping," which terminates parameter estimation when the validation set error begins to rise. Early stopping acts as a less computationally intensive alternative to ℓ_2 penalization while still providing shrinkage. Third, we utilize "batch normalization" to control predictor variability across different regions in the network, which stabilizes the network and accelerates the training process. Fourth, we adopt an ensemble approach by initializing neural network estimation with multiple random seeds and

averaging the forecasts from ten networks. Similar to RF, this ensemble technique reduces forecast variance, as each seed yields a different forecast. Further details on the specific hyperparameters used in these techniques are documented in Section A2.

A1.6 Forecast Combinations

We also form forecast combinations, which are designed to combine the forecasts of multiple models (see, e.g., Timmermann, 2006). Following Rapach, Strauss and Zhou (2010), we compute the equally-weighted average of all forecasts across a set of models at each point in time, referred to as the "mean" combination. Formally, if the forecast of a given model j ($j = 1, ..., J$) is denoted by $\hat{r}_{i,t+1}^{(j)}$, then the forecast combination at $t+1$ is:

$$\hat{r}_{i,t+1}^{Mean} = \frac{1}{J} \sum_{1}^{J} \hat{r}_{i,t+1}^{(j)}, \tag{A.13}$$

where J denotes the number of models combined. We compute combined forecasts for three cases: (1) linear models (L-Mean), (2) nonlinear models (NL-Mean), (3) all models (All-Mean). Comparing the nonlinear forecast combination to the linear one allows us to assess the improved predictive ability gained from incorporating nonlinearities and predictor interactions.

A1.7 Robust Huber Objective Function

As shown in Section 2.4, cryptocurrency returns exhibit extremely heavy tails, characterized by high positive skewness and immense kurtosis. The ℓ_2 objective function in Equation (A.4) is highly susceptible to large outliers, which can destabilize models estimated with this function. To mitigate the impact of heavy-tailed data, we adopt the robust Huber objective

function, as advocated by Gu, Kelly and Xiu (2020), defined as:

$$\mathcal{L}_H(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} H(r_{i,t+1} - g(z_{i,t};\theta), \xi), \tag{A.14}$$

where

$$H(x;\xi) = \begin{cases} x^2, & \text{if } |x| \leq \xi; \\ 2\xi|x| - \xi^2, & \text{if } |x| > \xi. \end{cases} \tag{A.15}$$

The Huber loss $H(\cdot)$ combines the ℓ_2 loss for small errors and ℓ_1 loss for large errors. The threshold parameter ξ controls the mix between ℓ_1 and ℓ_2. We estimate all models, except for RF, using the Huber objective function. We set the threshold at the 99.9% percentile of returns in the training set to more effectively handle extreme values. Since RF has inherent robustness to outliers, we follow Gu, Kelly and Xiu (2020) and use the default ℓ_2 loss function.

A2 Hyperparameter Tuning

This section outlines the hyperparameters tuned for each model, which are crucial for optimizing model performance.

For the ridge, lasso, and EN regressions, we focus on tuning λ, the regularization parameter that determines the level of shrinkage applied to the coefficients. This hyperparameter controls the trade-off between bias and variance for the model. A higher value of λ results in greater shrinkage, which generally reduces variance but increases bias, potentially leading to underfitting if set too high.

For PCR, the main hyperparameter is the number of principal components K chosen to be included in the regression model. Lower values of K can reduce variance and mitigate overfitting. However, setting K too low risks omitting predictive information.

For RF and GBT, the main hyperparameters include the number of trees and maximum depth of each tree. Increasing the number of trees generally improves the model's predictive accuracy, but can increase the risk of overfitting. Maximum depth, a stopping criterion, allows trees to capture complex multiway interactions between predictors, however deeper trees are more susceptible to overfitting. Additionally, for RF, the minimum number of samples in each leaf is another important stopping criterion; setting a lower threshold tends to produce more complex trees but increases the likelihood of overfitting. For GBT, learning rate λ controls the speed at which the model learns, with higher values leading to faster learning but at an increased risk of overfitting.

For NN1-NN4, we tune the ℓ_1 penalty, which functions similarly to the lasso. A higher ℓ_1 induces greater shrinkage, leading to a sparser model. However, setting the penalty too high can cause the model to underfit. Additionally, we tune the initial learning rate for the Adam SGD optimizer. A lower learning rate may result in slower training but can lead to more precise convergence. In contrast, a higher learning rate speeds up the training process but might yield a suboptimal solution or even cause the optimization to diverge.

Table **??** lists the hyperparameters tuned for each model considered in this paper. To select the optimal hyperparameters at a given time, we partition the sample into training, validation, and test sets as outlined in Section 4.2. For each model, we select the hyperparameters that produce the lowest MSE in the validation set. Using these chosen hyperparameters, we then re-estimate each model over the combined training and validation sets to generate the out-of-sample forecasts for the test set in the next month.

A3 Shapley Values

Shapley values, rooted in cooperative game theory, distribute payoffs among players in a game while uniquely satisfying several key axioms. For forecasting, Shapley values fairly allocate the contributions of each predictor to the model's forecasts. Specifically, the Shapley values for observation $z_{i,t}$ are the weighted average marginal contribution of each predictor to the payoff $g(z_{i,t}) - E[g(Z)]$ over all subsets of predictors, where $g(z_{i,t})$ is the output of a model for coin i at time t, Z is the $NT \times P$ matrix of stacked predictors, and $E[g(Z)]$ is the mean model output across all observations, i.e., the unconditional expectation of function $g(\cdot)$.

Following Štrumbelj and Kononenko (2014), the Shapley value for observation $z_{i,t}$ and predictor j can be mathematically expressed as the following weighted sum:

$$\phi_j(z_{i,t}) = \sum_{Q \subseteq S \setminus \{j\}} \underbrace{\frac{|Q|!(|S| - |Q| - 1)!}{|S|!}}_{\text{weight}} \underbrace{(\Delta_{Q \cup \{j\}}(z_{i,t}) - \Delta_Q(z_{i,t}))}_{\text{marginal contribution}} \tag{A.16}$$

where Q is a subset of predictors excluding predictor j, $|Q|$ denotes the number of elements in Q, and S represents the complete set of P predictors. Δ_Q denotes the model forecast conditional on the predictors in Q. The marginal contribution, $\Delta_{Q \cup \{j\}}(z_{i,t}) - \Delta_Q(z_{i,t})$, measures the change in the model forecast when predictor j is included compared to when it is excluded.

Shapley values possess the "local accuracy" (or "efficiency") property, expressed as:

$$\sum_{j=1}^{P} \phi_j(z_{i,t}) = g(z_{i,t}) - E[g(Z)], \tag{A.17}$$

which leads to the linear interpretation:

$$g(z_{i,t}) = E[g(Z)] + \phi_1(x_t) + ... + \phi_p(x_t). \tag{A.18}$$

This property allows us to interpret any forecast from a model linearly. Specifically, the sum of the Shapley values for all predictors equals the difference between the model's forecast for a specific observation and the mean forecast across all observations.

Calculating Shapley values can be extremely computationally intensive, especially when there are many predictors. This is because the process entails re-estimating the model on all possible subsets of predictors, with and without predictor j, as indicated in Equation (A.16), which leads to computations that increase exponentially with the number of predictors. Thus, approximations are necessary to make the computation feasible.

We estimate the Shapley values using the state-of-the-art SHapley Additive exPlanations (SHAP) framework developed by Lundberg and Lee (2017). This framework introduces efficient algorithms for different types of models that approximate the Shapley values without the need to evaluate every possible subset of predictors. For linear models such as OLS, ridge, lasso, EN, and PCR, we use Linear SHAP, which leverages the model coefficients to efficiently calculate the contributions of each predictor. For tree-based models such as RF and GBT, we use Tree SHAP, which exploits the structure of regression trees to significantly reduce computation time. For the neural network models, we employ Kernel SHAP, which is a model-agnostic technique that uses a weighted linear regression approach to approximate Shapley values.

References

Breiman, L. (2001). Random forests. *Machine Learning* 45, 5-32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). Cart. *Classification and Regression Trees* Wadsworth and Brooks/Cole Monterey, CA, USA.

Diebold, F.X. and M. Shin (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting* 35, 1679-1691.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics* 29, 1189-1232.

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

Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural networks* 2, 359-366.

Lundberg, S.M. and S. Lee (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30, 4768-4777.

Rapach, D. E., J. K. Strauss, and G. Zhou (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies* 23, 821-862.

Štrumbelj, E. and I. Kononenko (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* 41, 647-665.

Timmermann, A. (2006). Forecast combinations, in Elliott, G., C. W. J. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Volume 1, Amsterdam: North-Holland.

Table A1: Hyperparameters

This table describes the set of hyperparameters that are tuned for each individual model considered in the paper.

Ridge	$\lambda \in (10^{-5}, 10^{-1})$
	$\rho = 0$
Lasso	$\lambda \in (10^{-5}, 10^{-1})$
	$\rho = 1$
EN	$\lambda \in (10^{-5}, 10^{-1})$
	$\rho = \{0.25, 0.5, 0.75\}$
PCR	$K = 1 \sim P$
RF	#Trees = 50∼600
	#Features in each split = floor($\sqrt{P}$)
	Tree depth = 5∼30
	Minimum leaf samples = {0.01, 0.05, 0.1}
GBT	#Trees = 100∼600
	Tree depth = 1∼10
	Learning rate = {0.0001, 0.001, 0.01}
	Sample rate = 0.5
NN1-NN4	Activation Functions = Relu
	Adam Params. = default
	Learning Rate $\in \{0.001, 0.01\}$
	Batch size = 16
	Epochs = 100
	$\ell^1 = (10^{-5}, 10^{-3})$
	Patience = 5

Table A2: Summary Statistics

The table presents summary statistics for daily returns, order flow, market capitalization, volume and volatility. Order flow is measured in millions of USD. Market capitalization and volume are measured in billions of USD. Returns and volatility are measured in daily percent terms. Statistics across time represent the time-series average across coins. Statistics across coins represent the cross-sectional average across time. Statistics across time and coins are computed using the entire dataset across both time and coins. The sample period ranges from January 1, 2018 to June 30, 2022.

		Across Time and Coins					Across Time			Across Coins		
		Mean	Median	StD	Skew	Kurt	S.D	Skew	Kurt.	StD	Skew	Kurt
Top 10 Coins	Return	0.23	-0.07	8.41	10.78	449.13	8.03	3.03	67.84	3.37	0.51	0.40
	OF	0.44	178.69	20.90	-9.08	515.87	14.60	1.37	198.08	7.86	0.22	2.71
	MCAP	60.20	7.54	0.16	4.35	20.22	58.90	1.52	2.50	113.00	2.32	3.85
	Volume	5.40	1.44	11.17	5.78	94.55	5.08	3.85	31.51	8.08	1.93	2.37
	Volatility	8.04	6.10	7.57	6.82	125.24	7.10	4.16	36.69	3.80	0.75	0.32
All Coins	Return	0.32	-0.05	30.82	253.06	72257	12.83	3.16	61.71	7.61	1.65	12.75
	OF	0.06	0.01	7.86	-8.38	4433.50	2.50	2.27	254.35	3.20	0.82	51.32
	MCAP	7.70	0.11	60.28	12.95	190.24	7.60	2.44	8.82	44.20	7.87	63.29
	Volume	0.74	0.07	4.33	15.62	607.89	0.86	9.89	177.92	3.30	6.78	48.39
	Volatility	9.96	7.44	10.63	15.90	889.72	9.40	4.88	50.06	8.20	3.21	15.78

Table A3: Mean Returns Across Time

The table reports the mean daily (Panel A) and annualized (Panel B) returns across each year of the sample period. These are reported for the following cross-section: (1) all 82 coins, (2) the top 10 coins by market capitalization, and (3) quintile portfolios sorted on contemporaneous returns. The sample period ranges from January 1, 2018 to June 30, 2022.

Panel A: Mean Daily Returns

		2018	2019	2020	2021	2022	2018-2022
All Coins		0.06	0.10	0.64	0.93	-0.56	0.32
Top 10		-0.34	0.15	0.60	0.98	-0.73	0.23
Sort by Returns	1	-0.92	-0.28	0.17	0.36	-1.05	-0.01
	2	-0.71	-0.13	0.43	0.64	-0.81	0.08
	3	-0.54	0.03	0.57	0.81	-0.66	0.14
	4	-0.36	0.20	0.80	1.04	-0.46	0.27
	5	2.92	0.69	1.22	1.80	0.17	1.17

Panel B: Mean Annualized Returns

		2018	2019	2020	2021	2022	2018-2022
All Coins		15.1	25.2	161.3	234.4	-141.1	80.6
Top 10		-85.7	37.8	151.2	247.0	-184.0	58.0
Sort by Returns	1	-231.8	-70.6	42.8	90.7	-264.6	-2.5
	2	-178.9	-32.8	108.4	161.3	-204.1	20.2
	3	-136.1	7.6	143.6	204.1	-166.3	35.3
	4	-90.7	50.4	201.6	262.1	-115.9	68.0
	5	735.8	173.9	307.4	453.6	42.8	294.8

Table A4: Out-of-Sample Statistical Performance Based on R^1_{oos}

This table reports out-sample statistical performance using the R^1_{oos} statistic in percent based on the Mean Absolute Error (MAE) of the forecasts. Daily and weekly results are reported for three models: OF only conditions on order flows; EF only conditions on economic fundamentals; and OF+EF conditions on both. All models condition on lagged returns. Bold entries indicate the highest value in each row.

	Daily R^1_{oos} (%)			Weekly R^1_{oos} (%)		
	OF	EF	OF+EF	OF	EF	OF+EF
OLS	-1.38	**-1.16**	-11.83	**-2.31**	-22.16	-22.89
Ridge	**-1.34**	-11.26	-11.48	**-2.31**	-22.33	-22.70
Lasso	-1.05	**-1.01**	-1.05	**-2.22**	-2.87	-3.37
Elastic net	-1.06	**-0.91**	-1.11	**-2.25**	-5.78	-5.79
PCR	**-1.45**	-3.16	-2.60	**-2.28**	-14.69	-14.40
Random Forest	**0.38**	-1.30	-0.76	**-1.79**	-3.89	-2.79
SGB Model	**0.49**	-0.45	-0.36	**-0.49**	-3.78	-5.42
Neural Net 1	**0.02**	-0.49	-0.45	**-2.19**	-3.83	-4.41
Neural Net 2	**-0.17**	-0.21	-0.32	**-2.08**	-3.50	-4.61
Neural Net 3	**-0.12**	-3.34	-0.25	**-2.16**	-4.18	-4.83
Neural Net 4	**-0.20**	-0.32	-1.35	**-2.24**	-2.56	-4.48
L-Mean	**-1.16**	-3.67	-3.58	**-2.18**	-9.15	-9.34
NL-Mean	**0.23**	-0.38	-0.17	**-1.65**	-2.60	-3.24
ALL-Mean	**-0.12**	-0.95	-0.77	**-1.71**	-3.54	-3.95