Data-Driven Measures of High-Frequency Trading

Gbenga Ibikunle [a,b], Ben Moews [a,c], Dmitriy Muravyev [d,e], Khaladdin Rzayev [a,f,g*]
[a]*Edinburgh Centre for Financial Innovations, The University of Edinburgh*
[b]*RoZetta Institute, Sydney*
[c]*Centre for Statistics, The University of Edinburgh*
[d]*Department of Finance, University of Illinois at Urbana-Champaign*
[e]*Canadian Derivatives Institute*
[f]*Koç University*
[g]*Systemic Risk Centre, London School of Economics*

Abstract

High-frequency trading (HFT) accounts for a large share of equity trading volume but is not readily identifiable in public data. We introduce data-driven measures of HFT activity that distinguish between liquidity-supplying and liquidity-demanding strategies. We train machine learning models on a proprietary dataset with observed HFT activity, then apply these models to public intraday data to generate HFT measures across the entire U.S. stock universe from 2010 to 2023. Our measures significantly outperform conventional proxies, which struggle to capture HFT's temporal dynamics. Consistent with theory, our measures respond to latency arbitrage opportunities, as well as a quasi-exogenous speed bump introduction and data feed upgrade. The measures help uncover HFT's differential impact on information acquisition. Liquidity-supplying HFT improves price informativeness around earnings announcements, while liquidity-demanding HFT impedes it. This heterogeneity has important implications for market structure policy.

JEL classification: G10, G12, G14
Keywords: High-frequency trading, machine learning, liquidity, information acquisition.

1. Introduction

High-frequency trading (HFT) firms execute a large share of equity trading volume, focusing on nanoseconds and processing millions of orders through automated algorithms (e.g., surveys by Jones 2013; Menkveld 2016). Their dominance has sparked extensive research into their market impact, revealing an important distinction between strategies that take versus provide liquidity. Many HFT firms operate as market makers, leveraging their speed advantage to provide liquidity, lowering trading costs, and enhancing liquidity (e.g., Hendershott et al. 2011; Menkveld 2013; Brogaard et al. 2015). Conversely, liquidity-demanding HFTs aggressively consume liquidity, potentially increasing adverse selection costs and amplifying price volatility (e.g., Easley et al. 2011; Biais et al. 2015; Foucault et al. 2017).

Measuring HFT activity is challenging because standard market feeds do not identify it. Researchers have pursued two approaches, each with important limitations. Some studies employ private datasets, which identify HFT, most notably NASDAQ's 120-stock sample from 2008-2009, but these cover relatively few stocks over short periods.[1] Others propose proxies from public data, such as the quote-to-trade ratio (e.g., Hendershott et al. 2011) or odd-lot volume (e.g., Weller 2018). However, these proxies capture HFT and algorithmic trading jointly. They also do not distinguish between liquidity-demanding and liquidity-supplying HFT strategies (Boehmer et al. 2018; Chakrabarty et al. 2023) and as we show, mainly reflect cross-stock rather than temporal variation in HFT.

We introduce novel measures of liquidity-supplying and liquidity-demanding HFT activity (HFT_S and HFT_D). Using machine learning (ML) techniques, our method combines a proprietary dataset of directly observed HFT activity with concurrent public intraday data. Specifically, we train ensemble models to predict NASDAQ's HFT activity using 24 same-day

[1] NASDAQ's 120 stock sample from 2008-2009 that we use is the most popular, but prior studies also used proprietary data from the Investment Industry Regulatory Organization of Canada (IIROC)'s S&P/TSX 60 stocks, and the National Stock Exchange of India (NSE)'s 100-stock dataset from 2015.

measures of trading activity, liquidity, and volatility from WRDS Intraday Indicators. NASDAQ HFT remains the most widely used HFT dataset, while WRDS Intraday Indicators enable us to aggregate public intraday data to the daily level. This data-driven approach aims to capture nonlinear patterns in HFT behavior as well as relevant variable interactions. Moreover, because our approach is trained directly on HFT data, it can better distinguish between HFT and broader algorithmic trading, which is a limitation of conventional measures. Once the models are trained on a NASDAQ HFT dataset, we apply them to generate HFT measures for the entire TAQ universe of 8,314 common stocks from 2010 to 2023.

We evaluate our HFT measures against five popular HFT proxies: quote-to-trade ratio, mid-quote volatility, odd-lot volume, quoted price and depth changes, and the trade and quote message count. Importantly, quote data and these measures are not among the 24 intraday training variables. Using NASDAQ HFT data from January-June 2009 for training and July-December 2009 for out-of-sample evaluation, we find that while conventional proxies predict HFT activity individually, our measures largely subsume their predictive power in joint regressions. Moreover, standard measures capture cross-sectional but struggle to capture temporal HFT variation, while our measures reflect both dimensions. Thus, our measures capture a dimension of HFT that other measures largely miss.

We validate the measures using two natural experiments: one from years after the training sample and another occurring near the training period. The first event is NYSE Amex's introduction of speed bumps discouraging fast trading in 2017 (Khapko and Zoican 2021; Aït-Sahalia and Sağlam 2024). The speed bump introduction is linked to declines of 2.8% and 4.6% in HFT_D and HFT_S, respectively. The second is NASDAQ's 2011 data feed upgrade, which benefits HFT strategies (Ye et al. 2013). Both of our HFT measures increase in response to the event, though less than for the speed bump's negative effects. We also analyze how the HFT measures respond to latency arbitrage. Theory predicts that such arbitrage opportunities encourage liquidity demand and discourage liquidity supply (Budish et al. 2015; Foucault et al. 2017; Aquilina et al. 2022). Indeed, we find that as the number of latency arbitrage opportunities increases, HFT_D activity increases as fast traders exploit stale quotes, while HFT_S activity decreases as market makers withdraw to avoid being picked off.

Our HFT measures can be useful in a wide range of applications. We focus on one such application and examine how HFT activity affects fundamental information acquisition, a core market function. Our measures distinguish between HFT strategies, allowing us to test competing theories. Do HFTs enhance information acquisition by providing liquidity and reducing trading costs (e.g., Menkveld 2013; Stiglitz 2014; Brogaard et al. 2015; Aït-Sahalia and Sağlam 2024), or do they impair it by adversely selecting informed investors (e.g., Van Kervel and Menkveld 2019; Yang and Zhu 2020; Hirschey 2021)?

To answer this question, we study information acquisition around earnings announcements following Weller (2018). To measure information acquisition, he compares announcement returns to pre-announcement returns, with higher ratios indicating information was not discovered until publicly revealed. He finds that algorithmic trading reduces price informativeness (see also Gider et al. 2019). In contrast, we show that liquidity-supplying HFTs *enhance* information acquisition while liquidity-demanding strategies *impede* it. We also confirm this conclusion using an alternative measure of price informativeness: the future earnings response coefficient (Lundholm and Myers 2002). Our results are potentially consistent with Weller (2018), as we find that his proxies (quote-to-trade ratio and odd-lot volume) primarily capture liquidity-demanding HFT in his sample. Overall, this application highlights the advantages of our measures over existing alternatives.[3]

[3] Another approach would be to employ datasets with directly observed HFT; however, existing proprietary datasets are too small for the analysis, e.g., NASDAQ's dataset contains only several hundred earnings announcements.

Several other results further validate our approach. First, adding quote-level information to the model training only marginally improves model performance, consistent with a strong correlation between quote and trade activity. Second, we measure HFT activity as HFT volume divided by total volume in the main analysis; however, the results hold when we use unscaled HFT volume. Third, the HFT measures exhibit nonlinear relationships that are consistent with theory and highlight the value of ML methods. HFT liquidity demand responds strongly to intermarket sweep orders and decreases convexly with market depth, while HFT liquidity supply increases concavely with depth (Klein 2020; Goldstein et al. 2023). Finally, both HFT types increase around news events, with larger responses for liquidity suppliers.

Our approach assumes that the relationships between HFT activity and intraday variables in the 2009 NASDAQ dataset can be extrapolated beyond this training sample. Consistent with this assumption, prior literature finds that the results based on this widely-used HFT dataset typically hold for other U.S. exchanges (e.g., Shkilko and Sokolov 2020) including in the post-2009 period in the U.S. (e.g., Aït-Sahalia and Sağlam 2024; Brogaard et al. 2025) and internationally (e.g., Benos and Sagade 2016; Malceniece et al. 2019; Chakrabarty et al. 2025). Indeed, core HFT strategies have remained stable despite technological advances (Brogaard et al. 2014; Malceniece et al. 2019).[4] Moreover, our results for the 2011 and 2017 natural experiments demonstrate that the measures capture meaningful variation in HFT activity near and far from the training period.

This study advances the HFT literature stream in several ways. First, we develop novel measures that separate liquidity-demanding and liquidity-supplying HFT strategies, which outperform popular alternatives. We compute the measures for the entire U.S. equity market from 2010 to 2023 and plan to share them. Second, prior research shows that public HFT

[4] The features of HFT strategies developed in recent theories (Li et al. 2021a) are similar to those from a decade ago (e.g., Biais et al. 2015; Foucault et al. 2017), suggesting continuity in these core approaches. Also, many recent studies continue to rely on datasets from 2009-2012 (e.g., Boehmer et al. 2018; Goldstein et al. 2023; Nimalendran et al. 2024).

proxies combine liquidity supply and demand (Boehmer et al. 2018; Chakrabarty et al. 2023), while we separate the two and show that this distinction matters. Specifically, liquidity-supplying HFT facilitates information acquisition while liquidity-demanding strategies hurt this process. This explains why Weller (2018) finds that HFTs harm information acquisition, as his measures capture mainly liquidity-demanding trades, not the full picture. Finally, we find that conventional HFT measures struggle to capture temporal variation in HFT, while our measures are more successful.

Our work also provides an example of the successful application of machine learning in market microstructure. Recent studies show ML's effectiveness in analyzing informed trading (Bogousslavsky et al. 2023), hidden liquidity (Bartlett and O'Hara 2024), price discovery (Kwan et al. 2021), and volatility (Easley et al. 2021). We show that ML methods also effectively capture HFT, quantitatively and qualitatively outperforming the common HFT measures. Given its prevalence, we must measure HFT to understand how markets work.

2. Data and Variable Definitions

Our approach combines two primary datasets. The first is the widely used NASDAQ dataset that labels each trade transaction as executed by HFT or non-HFT for 120 stocks in 2009 (e.g., Brogaard et al. 2014). It also provides detailed trade attributes including the date and time (to the millisecond), volume, price, direction, and the counterparty type, identified as HH (both parties are HFTs), HN (an HFT demanding liquidity from a non-HFT), NH (a non-HFT demanding liquidity from an HFT), and NN (both parties are non-HFTs). NASDAQ identifies liquidity supply as all passive limit-order submissions by HFTs and liquidity demand as all aggressive order executions by HFTs (e.g., Brogaard et al. 2014). The main dependent variables are the shares of trading volume attributed to liquidity-demanding and liquidity-supplying HFTs. Specifically, $NASD\_HFT\_D_{i,t}$ ($NASD\_HFT\_S_{i,t}$) is calculated as the sum of

HH and HN (HH and NH) volume divided by the total trading volume for stock i on day t in the Nasdaq dataset.

The second dataset is the WRDS TAQ's Intraday Indicators covering the same period. We select 24 variables previously identified as associated with HFT activity. Table 1 describes the variables, which include various measures related to price, trading volume, trading costs, liquidity, and volatility. Following Bogousslavsky et al. (2023), we use pre-computed WRDS variables to enhance replicability and avoid data mining concerns. We train an ML model to predict true HFT activity in the proprietary NASDAQ dataset by variables from WRDS Intraday Indicators that aggregate TAQ data. We describe the ML model in Section 3 below.

INSERT TABLE 1 HERE

To validate our data-driven HFT measures, we obtain multiple complementary datasets. We calculate commonly used HFT proxies using quote-level data from the Millisecond TAQ database and benchmark our measures against them. We obtain intraday transaction data and corresponding bid-ask quotes from Refinitiv DataScope. Corporate event dates (specifically earnings and merger and acquisition (M&A) announcements) are from I/B/E/S and the Thomson Reuters Securities Data Company (SDC) database, respectively. Stock returns and trading volume are from the Center for Research in Security Prices (CRSP).

INSERT TABLE 2 HERE

Table 2 describes and provides summary statistics for the original NASDAQ HFT variables, our data-driven HFT measures, and other variables used in the paper's analyses. The liquidity-demanding ($NASD\_HFT\_D_{i,t}$) and liquidity-supplying HFT ($NASD\_HFT\_S_{i,t}$) HFT activities average 0.331 and 0.250, respectively. The difference is statistically significant at the 0.01 level. These two shares add up to about half of total volume, consistent with HFTs' participation. The distribution of $NASD\_HFT\_D_{i,t}$ is right-skewed, while $NASD\_HFT\_S_{i,t}$ is left-skewed. Our ML-generated HFT measures ($HFT\_D_{i,t}$ and $HFT\_S_{i,t}$) exhibit similar patterns: liquidity-demanding HFT activity is, on average, higher than liquidity-supplying activity, and the former is left-skewed while the latter is right-skewed. The bid-ask spread shows a mean of 0.142% with a wide range up to 0.885%, implying diverse liquidity conditions across the sampled stocks. Our sample includes 8,314 stocks, spanning the universe of US stocks in the TAQ database.

3. Methodology

In this section, we describe our machine learning methodology. Machine learning methods are optimized to select the best model among numerous predictors and account for their non-linearities and interactions. In our case, these methods help identify intraday variables in public data that are most related to HFT activity and aggregate these relationships semi-parametrically. We first train the models on Nasdaq HFT dataset with observed HFT trading. We then apply the trained model to compute estimated HFT activity for each stock-day with available public intraday data.

3.1. Ensemble methods for HFT prediction

We employ ensemble learning to predict HFT activity. Ensemble methods combine multiple predictive models to create a stronger collective predictor than any individual model. This approach delivers two key benefits for HFT applications. First, ensemble models reduce overfitting risk by averaging predictions across multiple weak learners. This improves out-of-sample performance when trading conditions change. Second, the component models remain simpler than complex single-model alternatives. Ensemble methods have proven particularly effective for non-linear financial prediction problems, where traditional econometric models often fail to capture complex relationships between order flow and price movements (Parker 2013; Moews et al. 2021; Cao 2022).

Three ensemble methods that build multiple simple decision rules and average their predictions are used, specifically decision trees, random forests and extremely randomized trees. Decision trees (Breiman et al. 1984) create if-then rules to classify market conditions – for example, "if bid-ask spread > 0.01 and volume < 1000, then predict price decline." Random forests (Ho 1995) combine and average over many of these trees, each trained on different data samples, which reduces prediction errors that are common in volatile HFT environments (Easley et al. 2021; Bogousslavsky et al. 2023). Extremely randomized trees (Geurts et al. 2006) further randomize tree construction. All models use mean squared error optimization and the 24 input variables from Table 2 to predict two HFT outcomes: HFT_D and HFT_S.

We design model training to handle HFT's large datasets efficiently. Each model is trained on 10,000 random stock-days, offering a sufficient sample size for tree-based models to learn reliable patterns while keeping runtimes feasible (Genuer et al. 2017). We repeat each training iteration 10 times to test how consistently models predict across different data samples, to account for changing market conditions. Monte Carlo cross-validation randomly divides data into 75% for training and 25% for testing model accuracy out-of-sample. This approach handles large HFT datasets more efficiently than k-fold methods, which require training on nearly all data multiple times and become computationally prohibitive (Hastie et al. 2009). Monte Carlo validation reduces the randomness in our performance measures (Li et al. 2010).

We optimize two key parameters: ensemble size (number of trees) and minimum samples per node split. Using grid search, we test 8 values for each parameter across 64 combinations, repeating each 10 times to measure consistency. We test ensemble sizes from 10 to 500 trees and minimum node splits from 2 to 50 samples. While more sophisticated optimization methods exist, grid search proves sufficient for our tree-based models (Probst and Boulesteix 2018). Results in Online Appendix Table OA.B.1 show that larger ensembles with finer splits consistently achieve better out-of-sample R^2 values. The top five performers all use the smallest node split threshold. Standard deviations across repeated runs confirm these configurations produce stable predictions when tested on different random samples.

We also benchmark our preferred ensemble method against standard machine learning models with varying levels of complexity, including LASSO, support vector machines, and neural networks. Of our ensemble methods, extra trees achieve the highest prediction accuracy with low variance, outperforming simpler (e.g., support vector machines) and more complex alternatives (neural networks). These results hold across different prediction setups (multi-target versus single target) and justify our model choice. Full model specifications and performance comparisons are reported in Online Appendix Subsection OA.B.2.

3.2. Comparison with common HFT proxies

In this section, we show that the proposed data-driven HFT measures outperform popular HFT measures on the NASDAQ dataset with directly observed HFT. The conventional HFT measures include the flickering quotes count ($Flick_{i,t}$), odd-lot volume ($OLV_{i,t}$), quote intensity ($QuoteInt_{i,t}$), quote-to-trade volume ratio ($QT_{i,t}$), and message count ($MG_{i,t}$). We compute these measures from the Millisecond TAQ database. Motivated by Hasbrouck (2018), $Flick_{i,t}$ measures quote volatility by first calculating the standard deviation of quote midpoints over 100ms intervals, and then averaging these deviations by stock-day. $OLV_{i,t}$ captures the daily sum of trades smaller than 100 shares (Weller 2018). $QuoteInt_{i,t}$ counts daily changes in best quotes or quote depth (Conrad et al. 2015); $QT_{i,t}$ is the ratio of quoted shares to traded shares (Hendershott et al. 2011; Weller 2018). Finally, $MG_{i,t}$ is defined as the sum of the daily number of trade and quote messages (Hendershott et al. 2011; Boehmer et al. 2018).

In this test, we train the ML model using only data from January to June 2009 and then evaluate their performance out-of-sample against the other measures from July to December 2009. Specifically, we estimate the following stock-day regressions using the observable liquidity-supplying and liquidity-demanding shares in NASDAQ HFT data on HFT_D and HFT_S measures, popular HFT proxies, and stock and time fixed effects:

$$NASD\_HFT\_D_{i,t} = \alpha_i + \beta_t + \gamma_1 HFT\_D_{i,t} + \gamma_2 Flick_{i,t} + \gamma_3 OLV_{i,t} + \gamma_4 QuoteInt_{i,t} + \\ + \gamma_5 QT_{i,t} + \gamma_6 MG_{i,t} + \varepsilon_{i,t} \qquad (1)$$

$$NASD\_HFT\_S_{i,t} = \alpha_i + \beta_t + \gamma_1 HFT\_S_{i,t} + \gamma_2 Flick_{i,t} + \gamma_3 OLV_{i,t} + \gamma_4 QuoteInt_{i,t} + \\ + \gamma_5 QT_{i,t} + \gamma_6 MG_{i,t} + \varepsilon_{i,t} \qquad (2),$$

We first estimate univariate regressions for each HFT measure as independent variables and then evaluate them in a joint regression following Equations (1) and (2). We double-cluster standard errors by stock and date and standardize all dependent variables to make coefficients easier to compare.

INSERT TABLE 3 HERE

Panel A of Table 3 shows the results for liquidity-supplying HFT. Among all measures, HFT_S delivers the strongest association with liquidity-supplying HFT based on the highest coefficient, *t*-statistics, and within-R^2. Other measures are positively associated with HFT liquidity supply, but flickering quotes and odd-lot volume are not statistically significant. In a joint regression, HFT_S's coefficient magnitude and *t*-statistics decrease very little. It also dominates predictability as the other measures jointly add only 0.3% to within-R^2.

The results for liquidity-demanding HFT activity are broadly consistent with those for liquidity-supplying HFT, except that predicting liquidity demand is harder, as reflected in the lower R^2 estimates. HFT_D consistently shows the highest coefficient magnitude and *t*-statistics, along with the highest within-R^2. In the univariate regressions with fixed effects, of the conventional HFT proxies, only the quote-to-trade ratio predicts $NASD\_HFT\_D$ positively.

Panels A and B in Table 3's estimates account for stock and day fixed effects. The consistently strong and statistically significant relationships between data-driven HFT measures and actual HFT activity demonstrate the predictive power of the ML-generated proxies across both cross-sectional and time-series dimensions. In contrast, conventional HFT measures show relatively weak associations when both fixed effects are included. We hypothesize that these conventional measures predominantly capture cross-sectional but not time-series variation. To test this hypothesis, we re-estimated Equations (1) and (2) using only day fixed effects and report them in Panels C and D of Table 3.

The results in Panels C and D confirm that when controlling solely for day fixed effects, three conventional measures ($QuoteInt_{i,t}$, $QT_{i,t}$, and $MG_{i,t}$) display substantially stronger correlations with both liquidity-supplying and liquidity-demanding HFT activities. This pattern suggests that conventional HFT measures primarily capture cross-sectional variation. Notably, our HFT measures still outperform in this specification, showing much higher *t*-statistics and within-R^2. Also, our metrics subsume the information content of conventional HFT measures in joint regressions.

Overall, these findings show that the advantages of our data-driven measures over traditional HFT proxies. Our measures predict both liquidity-demanding and -supplying strategies with larger coefficients, *t*-statistics, and R^2. Furthermore, while our measures effectively capture both cross-sectional and time-series dimensions, conventional measures predominantly reflect cross-sectional variation.

3.3.Model assessment and extrapolation to U.S. stocks

Once we estimate our main model on a (relatively small) NASDAQ dataset with observed HFT, we apply this model to estimate HFT activity from observed intraday input variables. On each day and for each stock, we observe the 24 input variables listed in Table 1 and feed them into the model, whose parameters have been estimated on the training data. This is akin to first estimating regression coefficients in a linear regression (e.g., betas) and then applying them to current data (e.g., computing abnormal returns). Thus, we assume that the relationships in the training data are sufficiently general to be extrapolated to the broader market and later periods. The final sample covers 9,440,600 stock-days from January 4th 2010 to October 18th 2023.

A key strength of ML over traditional linear models lies in its ability to capture the nonlinearity between input and output variables. This aspect is important for us, given the nonlinear relationship between HFT and market characteristics. For instance, Foucault et al. (2017) show that whether HFT arbitrage strategies enhance or impair liquidity is contingent on the nature of latency arbitrage opportunities (e.g., Rzayev et al. 2023).

We analyze partial dependence plots to determine if our ML modeling framework captures nonlinear interactions between HFT activity and its predictors. We start by assessing the feature importance plot to identify key drivers of HFT activity. Next, we explore the relationships between HFT and these key drivers through partial dependence plots, focusing on the nature and shape of the interactions.

INSERT FIGURE 1 HERE

Figure 1 shows that most input variables significantly predict HFT activity. Trading volume, market depth, and intermarket sweep orders (ISOs) matter the most. Trading volume and market depth are important because HFTs need counterparts to trade with and deep markets to operate in. ISOs are designed for large institutional traders; nonetheless, HFTs exploit them to adversely select slower traders/market participants.[5] Indeed, Li et al. (2021b) show that ISO order sizes have shrunk below typical institutional sizes, and fast traders now dominate ISO usage.

INSERT FIGURE 2 HERE

Having pinpointed the key drivers of HFT activity, we further explore the shape of the relationships between these determinants and HFT activity using partial dependence plots.

[5] https://tabbforum.com/opinions/why-hfts-have-an-advantage-part-3-intermarket-sweep-orders/

Figure 2 documents the non-linear relationship between HFT activity and various input variables. For instance, liquidity-demanding and -supplying HFT activity both show an increasing and concave relationship with the total trade count. This positive correlation with trading volume is consistent with Brogaard et al. (2014), who show that HFTs favor trading in larger stocks, which tend to be more liquid.

Liquidity-demanding HFT spikes when ISO volume increases, following a concave curve that shows ISOs significantly influence these aggressive strategies. Liquidity-supplying HFT barely responds to ISOs as the relationship stays flat with only marginal increases as ISO dollar amounts rise. This differential response aligns with academic findings. Li et al. (2021b) show that HFTs use ISOs to target stale quotes, a tactic that defines liquidity-demanding strategies. Klein (2020) finds that aggressive HFT strategies deploy ISOs when new information arrives. An alternative explanation suggests HFTs respond to institutional traders who use ISOs to avoid getting front-run. Chakravarty et al. (2012) explain that regulators created the ISO exemption to Rule 611/Order Protection Rule of Reg NMS to give institutional investors timely access to liquidity at multiple price levels. This allows institutions to execute large block orders by submitting orders across multiple trading platforms simultaneously.

Market depth generates opposite effects on the two HFT measures. Liquidity-supplying HFT increases as markets deepen, following a concave curve that shows HFTs provide more liquidity when order books thicken. Liquidity-demanding HFT does the reverse – it decreases as depth increases, creating a convex pattern that shows HFTs demand less liquidity in deep markets. This makes economic sense. Goldstein et al. (2023) demonstrate that HFTs supply liquidity in deeper markets where order books are thick and demand liquidity in shallower markets where order books are thin.

These findings lead to two key implications. First, the nonlinear relationships between HFT activity and market quality show why ML models outperform simple proxies for measuring HFT activity. Linear models fail to reflect these curves and inflection points. Second, liquidity-demanding and liquidity-supplying HFT respond differently to market conditions, which matches ongoing academic debates about HFT's varied effects. This confirms our ML-derived metrics capture real HFT strategies rather than noise. We next validate these metrics and examine their empirical significance in detail.

4. Results

4.1.HFT during exogenous technological changes.

We show above that the data-driven HFT measures significantly outperform conventional measures. We now examine how HFT_D and HFT_S respond to exogenous shocks affecting HFT activity through two natural experiments: one occurring near the training sample period and another occurring years afterward. If data-driven metrics capture HFT activity, they should respond significantly to these HFT-specific market structure changes.

In the first quasi exogenous shock, Nasdaq introduces a technology upgrade that reduces trading data dissemination latency from 3 to 1 millisecond on October 10, 2011 (e.g., Ye et al. 2013). The upgrade is implemented in stages: stocks with ticker symbols beginning with A and B were upgrade on October 10, while the remaining stocks upgrade on October 17. Ye et al. (2013) employ this staggered implementation to study HFT's impact on market quality. We expect that the reduced latency encourages more HFT and test this hypothesis in the stock-day regressions:

$$HFT\_D_{i,t} = \alpha_i + \beta_t + \gamma_1 Post_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \tag{3}$$

$$HFT\_S_{i,t} = \alpha_i + \beta_t + \gamma_2 Post_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \qquad (4),$$

where $HFT\_D_{i,t}$ and $HFT\_S_{i,t}$ are our measures of liquidity-demanding and -supplying HFT activity, respectively. Stock (α_i) and day (β_t) fixed effects account for individual stock characteristics and daily variations, respectively. $Post_{i,t}$ is an indicator variable equal to 1 after October 10, 2011, for NASDAQ-listed stocks with tickers beginning with A and B, and after

October 17, 2011, for other NASDAQ-listed stocks, and 0 otherwise. We also include NYSE and Amex-listed stocks as control stocks ($Post_{i,t} = 0$ for these stocks throughout the sample period) to implement a DiD framework (e.g., Malceniece et al. 2019). The standard errors are double clustered by firm and day. Similar to Ye et al. (2013), we employ a 10-working day window around the implementation dates to zoom in on the effect. $C_{i,t}^k$ includes a range of control variables, such as volatility ($Volatility_{i,t}$), relative quoted spread ($Spread_{i,t}$), inverse price ($InvPrice_{i,t}$), and trading volume in dollars ($Volume_{i,t}$). $Volatility_{i,t}$ is calculated as the daily (*t*) standard deviation of the transactional-level returns for stock *i*. $Spread_{i,t}$ is the daily average of transaction-level bid-ask spreads. The transaction-level bid-ask spread is calculated as the difference between ask and bid prices divided by the average of ask and bid prices for each transaction. All these variables are obtained from the TAQ database.

In our second natural experiment, Amex introduces a speed bump. In January 2017, the Amex files a request with the SEC to introduce a deliberate delay in the communication between traders and the exchange. This proposed delay is designed to impact both inbound (from traders to the exchange) and outbound (from the exchange to traders) communications, establishing a total round-trip latency delay of 700 microseconds. The SEC approves this request, leading to the trading delay's activation on July 24, 2017. Given that the introduction of a speed bump increases trading latency, it is expected to reduce HFT activity. Therefore, if our data-driven HFT metrics capture the dynamics of HFT activity, we should observe a reduction in the metrics on Amex post the speed bump implementation. To formally test this hypothesis, we employ the following stock-day regression:

$$HFT\_D_{i,t} = \alpha_i + \beta_t + \gamma_1 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^k C_{i,t}^k + \varepsilon_{i,t} \qquad (5)$$

$$HFT\_S_{i,t} = \alpha_i + \beta_t + \gamma_2 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^k C_{i,t}^k + \varepsilon_{i,t} \qquad (6),$$

where $Post_{i,t}$ is an indicator variable, taking the value of 1 on July 24, 2017, when the speed bump was implemented and thereafter, and 0 before, while $Amex_{i,t}$ corresponds to 1 for NYSE Amex-listed stocks and 0 for NYSE- and NASDAQ-listed firms. Our models do not explicitly include $Post_{i,t}$ and $Amex_{i,t}$ indicator variables, as their effects are already accounted for through the inclusion of time and stock fixed effects. All other variables are as defined above. Similar to Models (3) and (4), we double-cluster standard errors by firm and day, and analyze a 10-day window around the implementation dates.

Before discussing the results from the estimation of Equations (3 – 6), we provide an important methodological clarification. Our HFT measures (HFT_D and HFT_S) are computed at the firm-day level, aggregating activity across all exchanges. This raises a potential concern: if HFTs redirect their orders from the treated exchanges (NASDAQ in Models (3) – (4) and Amex in Models (5) – (6)) to alternative venues, the impact of technological changes on overall HFT activity might be dampened. However, this concern is likely minimal because HFTs typically prefer a stock's primary listing exchange due to superior market quality. For instance, 2023 statistics show Amex leading in terms of quote quality (time at best prices), quoted depth (size at best prices), and spread tightness for its listed stocks.[6] These market quality advantages create strong incentives for HFTs to maintain their activity on the primary exchange, suggesting that technological changes should meaningfully impact HFT behavior.

INSERT TABLE 4 HERE

Table 4 reports the estimation results for Models (3) through (6). Columns (i) and (ii) present the findings for NASDAQ's latency reduction upgrade, while columns (iii) and (iv) show the results for Amex's speed bump implementation. Consistent with our predictions, the HFT measures show significantly higher activity following NASDAQ's upgrade and lower activity after Amex's speed bump implementation, relative to stocks listed on other exchanges.

[6] https://www.nyse.com/markets/nyse-american

We next explore the economic magnitudes of the observed changes. The Amex speed bump is a stronger shock to HFT activity because it is a direct speed impact. In contrast, Nasdaq's improvement in trading data dissemination is an indirect shock, as it only reduces latency for the consolidated feed while HFTs can access direct and faster feeds. As Ye et al. (2013) note, changes to consolidated feed latency affect HFT activity since HFTs utilize these feeds; however, the impact is relatively modest. Our results support this distinction. Following the speed bump introduction, Amex-listed stocks experience decreases of 2.8% and 4.6% in HFT_D and HFT_S, respectively, relative to their pre-speed bump averages. In comparison, Nasdaq's technological upgrade leads to more modest increases of 0.7% and 1.1% in HFT_D and HFT_S for NASDAQ-listed stocks, respectively, relative to their pre-upgrade averages.

These results have three main implications. First, our HFT metrics effectively capture HFT activity, validated by their response to HFT-relevant shocks and the varying response magnitudes between direct (speed bump) and indirect (trading data latency upgrade) shocks. Notably, while Nasdaq's trading data dissemination technology upgrade occurs in 2011, near the period the data we use to train our ML model (2009) is obtained, our measures also respond to the 2017 speed bump effects, suggesting the model's temporal robustness. Thus, the patterns learned by our ML model during the training stage remain applicable to later periods.

Second, in line with theoretical predictions, changes in data dissemination speed and speed bump introductions significantly affect HFT activity. Therefore, similar to colocation upgrades (e.g., Brogaard et al. 2015; Boehmer et al. 2021a), these technological changes provide exogenous shocks that can be used to examine HFT's impact on financial markets.

Third, our speed bump findings complement Aït-Sahalia and Sağlam (2024), who document that the speed bump caused wider quoted spreads and reduced liquidity. Their theoretical framework links speed changes to market-making HFT activity. We extend their analysis by showing that the speed bump affects both market-making and market-taking HFTs, with market makers experiencing stronger effects, explaining the overall negative liquidity impact in their study. Moreover, the alignment between our findings tentatively suggests that our liquidity-demanding and liquidity-supplying HFT metrics effectively capture supply and demand dynamics, we formally investigate this in the next section.

4.2.HFT and latency arbitrage opportunities.

Our analyses thus far provide evidence that our ML-generated measures capture the distinct characteristics of liquidity-demanding and -supplying HFT strategies. To further validate this insight, we examine “latency arbitrage” opportunities. Latency arbitrage involves fast traders using their superior response speeds to exploit newly available public information and execute against stale quotes before slower traders can (e.g., Budish et al. 2015; Foucault et al. 2017; Shkilko and Sokolov 2020; Aquilina et al. 2022). Aquilina et al. (2022) show that in most latency arbitrage scenarios, HFTs often aggressively take liquidity. The profitability of aggressive HFT strategies is enhanced by the emergence of latency arbitrage opportunities; hence, HFTs are encouraged to engage more in such strategies (e.g., Baldauf and Mollner 2020). Therefore, latency arbitrage events offer a context to distinguish between the specific characteristics of liquidity-demanding and -supplying HFT activity. In particular, we expect increase in liquidity-demanding HFT activity as the number of latency arbitrage opportunities increases, in line with predictions by Baldauf and Mollner (2020) and the findings of Aquilina et al. (2022). This increase in aggressive trading and sniping activity increases adverse selection risk on endogenous liquidity-supplying HFTs; hence, we expect liquidity-supplying HFT activity to decline (e.g., Foucault et al. 2017; Menkveld and Zoican 2017).

To formally test these arguments, we estimate the following stock-day models:

$$HFT\_D_{i,t} = \alpha_i + \beta_t + \gamma_1 NLAO_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \qquad (7)$$

$$HFT\_S_{i,t} = \alpha_i + \beta_t + \gamma_2 NLAO_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \qquad (8),$$

where $NLAO_{i,t}$ is the number of latency arbitrage opportunities. We identify latency arbitrage opportunities following Budish et al. (2015), who suggest examining the mid-price changes to identify "stale" quotes. Specifically, a quote at time $\tau - 1$ is stale if the absolute difference in mid-price from time $\tau - 1$ to τ exceeds the half spread. We adopt a more conservative methodology by calculating the jump size based on the difference between the mid-price at time τ and the ask and bid quotes at time $\tau - 1$. If $Midprice_{\tau} > (Ask_{\tau-1} + TickSize)$, where $TickSize$ is set to 0.01$, it suggests a profitable latency arbitrage opportunity. HFTs can exploit it by placing a limit buy order at $Ask_{\tau-1} + TickSize$ at time τ. Similarly, if $Midprice_{\tau} > (Bid_{\tau-1} - TickSize)$, HFTs can submit a limit sell order at $Bid_{\tau-1} - TickSize$ at time τ.

We identify latency arbitrage opportunities using the first-level quote data from Refinitiv DataScope. The data is enormous, which makes it computationally prohibitive to examine our full 8,314 stock sample. Therefore, we narrow the sample to the 120 firms in the original NASDAQ HFT data. We calculate $NLAO_{i,t}$ for these 120 firms across our entire sample period, from 2010 to 2023. Table 2 includes the average number of latency arbitrage opportunities per stock-day is 68. The standard deviation is 169 and the maximum value is 1211, indicating large variation in these opportunities across stocks and days.

INSERT TABLE 5 HERE

The results from the estimation of Equations (7) and (8), as presented in Table 5, show a positive and statistically significant (at the 0.01 level) relationship between $HFT\_D_{i,t}$ and $NLAO_{i,t}$, whereas the relationship between $HFT\_S_{i,t}$ and $NLAO_{i,t}$ is negative and significant (at the 0.05 level). The relationship between $HFT\_D_{i,t}/HFT\_S_{i,t}$ and $NLAO_{i,t}$ is also economically significant. A one-standard-deviation increase in $NLAO_{i,t}$ (169) is associated with a 1% rise in HFT_D and 1.6% decrease in HFT_S. These results indicate that latency arbitrage opportunities affect various HFT strategies. Prior literature suggests that arbitrage-seeking HFTs often adopt aggressive trading strategies during latency arbitrage opportunities (e.g., Aquilina et al. 2022), and endogenous liquidity-supplying HFTs are, thus, inclined to scale back on their liquidity provision (e.g., Foucault et al. 2017). Our findings align with these arguments and validate HFT_D and HFT_S in their ability to capture the liquidity-demanding and -supplying activities of HFTs.

5. HFT's effect on information acquisition

Our data-driven HFT measures that separate liquidity supply and demand can be used in many important applications. In this section, we examine one such application focusing on price discovery, one of the fundamental functions of markets. Specifically, we show a crucial distinction between liquidity-supplying and liquidity-demanding HFT's effect on information acquisition.

Price discovery characterizes how stock prices reflect information (O'Hara 2003). This process includes both integrating existing information into asset prices and generating or acquiring new fundamental information (Brunnermeier 2005; Weller 2018; Brogaard and Pan 2022). Market microstructure researchers have extensively studied the relationship between HFT and price discovery. This growing literature primarily focuses on how existing information gets incorporated into stock prices (Menkveld 2016), often concluding that HFT enhances the speed at which existing information reaches stock prices, contributing to more efficient price discovery mechanisms.

HFTs' role in acquiring new fundamental information remains understudied. First, information acquisition happens at low frequencies rather than tick-by-tick. Second, theory implies that studying HFTs' impact on information acquisition requires separating liquidity supply and demand. Existing datasets that make this distinction, such as the Nasdaq HFT data, work well for high-frequency market quality studies. However, their limited sample periods and small stock coverage make them unsuitable for studying fundamental information acquisition. In the case of the Nasdaq data, it covers 120 firms and thus results in just 480 firm-quarter earnings announcement events per year.

Our measures offer comprehensive coverage, avoiding this problem and enabling a comprehensive analysis of how different HFT strategies influence information acquisition. HFTs can improve information acquisition by providing liquidity and thus reduce trading costs (Menkveld 2013; Brogaard et al. 2015; Aït-Sahalia and Sağlam 2024). As lower trading costs increase net profits, investors are incentivized to seek and trade on new information, and this facilitates information acquisition and dissemination. But HFTs also use aggressive strategies that weaken information acquisition. They employ order anticipation tactics, such as back-running and latency arbitrage to predict and profit from informed institutional trades (Van Kervel and Menkveld 2019; Yang and Zhu 2020; Hirschey 2021). These strategies increase trading costs for informed investors, creating a crowding-out effect that discourages information seeking and reduces overall information acquisition.

Weller (2018) studies HFT's effect on information acquisition using a novel metric "price jump ratio." This ratio divides the return at public information release by the cumulative return during the lead-up period. Bigger price jumps during announcements signal weaker information acquisition beforehand. When information gets reflected in prices only upon public release rather than gradually, it means fewer investors acquired information early. Thus, higher price jump ratios indicate lower information acquisition. Weller (2018) finds that HFT harms information acquisition.

While Weller (2018) advances our understanding of how HFTs affect information acquisition, it relies on MIDAS data that aggregates all HFT activity without separating specific trading strategies. This matters because theory suggests that different HFT strategies affect information acquisition differently. With MIDAS data, Weller (2018) shows that HFT presence reduces information acquisition, while unable to facilitate a deeper investigation.

Weller (2018) acknowledges this issue, concluding (p.2217) that future research must *"assess the precise mechanisms by which improved trading technology reduces the information content of prices."*

Responding to this call, we exploit the unique proprieties of our HFT measures to investigate how HFT affects information acquisition. In our main specification, we estimate the following regression model:

$$JUMP_{i,q} = \alpha_i + \beta_{m,q} + \gamma_1 HFT\_D_{i,q} + \gamma_2 HFT\_S_{i,q} + \textstyle\sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}, \qquad (9),$$

where $JUMP_{i,q}$ is the ratio of cumulative abnormal returns during trading days [-1, 1] relative to earnings announcements, divided by the cumulative abnormal returns during days [-21, 1]. Daily abnormal returns are calculated as the raw return minus the expected return from the market model. We calculate $HFT\_D_{i,q}$ and $HFT_S_{i,q}$by averaging the daily HFT values over the 21 trading days [-21, -1] before earnings announcements. Control variables ($C_{i,q}^{k}$) include volatility ($Volatility_{i,q}$), relative quoted spread ($Spread_{i,q}$), market value ($MValue_{i,q}$), and institutional order imbalance ($OIB20k_{i,q}$). We obtain $OIB20k_{i,q}$ directly from TAQ, capturing the price impact of trades exceeding $20,000, and compute $MValue_{i,q}$ by averaging the daily market values over the same 21-day window. The remaining control variables represent 21-day averages of their daily counterparts before earnings announcements [-21, -1]. Following Weller (2018), we include stock and month fixed effects and apply his filters.

We include both HFT_D and HFT_S in Equation (9) to examine their comparative effects on information acquisition. These metrics correlate at 0.52, so multicollinearity will not distort results. Since higher $JUMP_{i,q}$ values mean less information acquisition, we expect opposite effects from the two HFT types. HFT_D should increase $JUMP_{i,q}$ because aggressive strategies raise trading costs and discourage information seeking. HFT_S should decrease

$JUMP_{i,q}$ because liquidity-provision strategies lower trading costs and make information acquisition more profitable.

INSERT TABLE 6 HERE

The results in Table 6 show that $HFT\_D_{i,q}$ has a positive and statistically significant relationship with $JUMP_{i,q}$. An increase in a firm's $HFT\_D_{i,q}$ from the 25th percentile (0.222) to the 75th percentile (0.414) is associated with a 6.6% increase in $JUMP_{i,q}$ relative to its mean value. Conversely, $HFT\_S_{i,q}$ shows a negative and statistically significant relationship with $JUMP_{i,q}$, where an increase from the 25th percentile (0.131) to the 75th percentile (0.259) corresponds to a 3.3% decrease in $JUMP_{i,q}$ relative to its mean.

Our findings suggest that the positive relationship between common HFT measures and $JUMP_{i,q}$ shown in Weller (2018) may be driven by the measures primarily capturing liquidity-demanding HFT activity during the sample period. To examine this hypothesis, we analyze the relationship between Weller's (2018) main HFT measures and our HFT measures. Weller's (2018) measures, obtained directly from MIDAS, include cancel-to-trade ratio ($CT_{i,q}$), odd-lot rate ($OLR_{i,q}$), and trade-to-order ratio ($TO_{i,q}$). $CT_{i,q}$ is the ratio of cancelled messages to trade messages, $OLR_{i,q}$ measures the proportion of trades below 100 shares, and $TO_{i,q}$ is calculated as the ratio of executed shares to submitted shares.

INSERT TABLE 7 HERE

The results in Table 7 help reconcile our findings with Weller's (2018). $CT_{i,q}$ and $OLR_{i,q}$ are positively linked with $HFT\_D_{i,q}$, while $TO_{i,q}$ (an inverse measure of HFT) is negatively related. Conversely, the metrics display opposite relationships with $HFT\_S_{i,q}$. The directions of the relationships remain consistent in simple univariate correlation analysis. Thus, observed relationships, combined with Weller's (2018) findings of positive relationships between $CT_{i,q}/OLR_{i,q}$ and $JUMP_{i,q}$, and negative correlation between $TO_{i,q}$ and $JUMP_{i,q}$, suggest that the HFT measures in Weller (2018) predominantly capture liquidity-demanding HFT activity.

To further explore the relationship between HFT and information acquisition, we employ the future earnings response coefficient (FERC) (e.g., Lundholm and Myers 2002; Ettredge et al. 2005; Brogaard and Pan 2022) as an alternative measure. Specifically, we estimate FERC through the following model:

$$Return_{i,q} = \alpha_i + \beta_q + \sum_{n=-1}^{1}(\gamma_n Earning_{i,q+n} + \vartheta_n Earning_{i,q+n} * HFT\_D_{i,q} + \theta_n Earning_{i,q+n} * HFT\_S_{i,q}) + \rho_1 HFT\_D_{i,q} + \rho_2 HFT\_S_{i,q} + \rho_3 Return_{i,q+1} + \rho_4 Return_{i,q-1} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,q} \text{ (10),}$$

where $Return_{i,q}$ is the quarterly stock return for firm i in quarter q, and is measured as the percentage change in closing prices between quarters $q-1$ and q. The subscript n ranges from -1 to 1, capturing the temporal relationships in the model. $Earning_{i,q+n}$ denotes quarterly earnings (net income) normalized by the market value of equity at the start of quarter $q+n$. In this specification, γ_n reflects FERC; a positive value will suggest that current returns incorporate future earnings information, which indicates heightened fundamental information acquisition. We employ the same set of control variables used in the jump ratio model, averaged at the quarterly frequency.

The coefficients of interest in Model (10) are ϑ_n and θ_n, which indicate whether HFT enhances (positive coefficient) or impairs (negative coefficient) the incorporation of future earnings information into current returns. Based on our jump ratio findings, where $HFT\_D_{i,q}$ $(HFT\_S_{i,q})$ is negatively (positively) associated with information acquisition, we expect ϑ_n and θ_n to be negative and positive, respectively.

INSERT TABLE 8 HERE

Table 8 reports results that corroborate our findings from the jump ratio analysis. θ_n is positive and statistically significant at the 0.01 level, while ϑ_n is negative and also significant at the 0.01 level, indicating a positive (negative) relationship between liquidity-supplying (-demanding) HFT activity and information acquisition.

We extend our baseline results in two directions. First, we test whether existing HFT datasets that separate trading strategies can investigate HFT's role in information acquisition. This matters because if they could, it would call the need for our new measures into question. Notwithstanding, this question becomes somewhat moot since no publicly available datasets currently separate HFT strategies. We therefore employ the proprietary Nasdaq HFT dataset covering 120 stocks in 2009 in the replication of the jump ratio and FERC analyses. Table OA.C.1 shows that using the Nasdaq dataset produces no statistically significant relationship between HFT strategies and information acquisition due to limited sample size. This corroborates the relevance of our ML-generated measures. They let researchers examine how HFT affects low-frequency market outcomes that matter for real economic decisions.

The second extension addresses concerns about training our ML model on 2009 data. Researchers continue to use the Nasdaq HFT dataset because the core distinction between liquidity-demanding and liquidity-supplying strategies remains fundamental to HFT behavior (Boehmer et al. 2018; Goldstein et al. 2023; Nimalendran et al. 2024). Section 4.1 shows that our measures respond to technological shocks both near and far from the training period. We provide additional validation by examining the HFT-information acquisition relationship close to our training sample. Similar results between this restricted sample and our full sample would show that temporal distance from training data does not affect our findings. Table OA.C.2 presents results using data from January 2010 to December 2012. Both jump ratio and FERC analyses mirror our baseline results. Liquidity-demanding strategies hurt information acquisition while liquidity-supplying strategies help it.

We explore information acquisition as an important application of our novel approach to measuring HFT. While it is challenging to establish causality, our results show that our HFT measures provide valuable tools for investigating how HFT affects low-frequency economic outcomes that require large samples to study. This data-driven approach matters because econometric approaches using exogenous shocks cannot examine how different HFT strategies affect real outcomes. These shocks hit both liquidity-demanding and liquidity-supplying strategies equally, so difference-in-differences frameworks cannot separate their distinct effects. Data-driven distinction between HFT strategies becomes essential for understanding their different economic impacts. Our findings complement Weller (2018) by providing empirical evidence of specific mechanisms through which HFT affects information acquisition.

6. Extensions and further robustness analyses

In this section, we provide additional tests as a validation of the ML-generated HFT measures, and extend our baseline ML framework. First, we extend earlier exploration of the dynamics of liquidity-demanding and liquidity-supplying HFT activity around scheduled and unscheduled information announcements. Foucault (2016) and Brogaard et al. (2014) argue that HFTs rapidly respond to major information events. Hence, a detailed examination of how our ML-generated HFT measures react around these events, therefore tests their empirical validity.

INSERT FIGURE 3 HERE

Figure 3 shows how liquidity-supplying and liquidity-demanding HFT activity change around (scheduled) earnings announcements. We plot both measures over a 20-day window spanning ten days before and after announcements, with 95% confidence intervals. Both HFT types spike starting three days before and peak on announcement day. We measure this effect by comparing average HFT activity during the three-day event window (days t, t_{+1}, and t_{+2}) with pre-announcement levels. This three-day period follows previous research on short-term earnings effects (Ball and Shivakumar 2008). Both measures increase significantly during announcement windows. HFT_S jumps 6.3% (from 0.208 to 0.221) while HFT_D rises 2.8%.

Figure OA.A.1 presents the corresponding plots for HFT behavior around (unscheduled) M&A announcements, which contain higher information content than earnings announcements (Bogousslavsky et al. 2023). Our ML-generated HFT measures start increasing just one day before M&A announcements or on announcement day itself, compared to three days for earnings. This is consistent with the stream of the literature showing that HFTs primarily trade on public information by processing it rapidly (Budish et al. 2015; Aquilina et al. 2022) rather than exploiting private information as informed traders do (Bogousslavsky et al. 2023). The unscheduled nature of M&A announcements limits exploitable information beforehand. We therefore find less HFT activity before M&A announcements than before earnings announcements.

We extend our baseline ML framework by first expanding the feature space. Selecting ML input features involves competing considerations. More granular data could improve prediction precision; however, they are likely to be more expensive and challenging to access and process. More accessible datasets may sacrifice predictive power, nevertheless enable wider application and replication. Our baseline model, using daily input features derived directly from TAQ's Intraday Indicators, prioritizes accessibility, a key contribution in developing HFT measures from non-proprietary data.

These indicators lack quote-level granularity, such as message counts or quote update frequencies, which could limit ML training effectiveness. The baseline model's 82% R^2 substantially mitigates this concern by showing that our input variables capture the predominant variation in HFT activity. This suggests limited gains from incorporating more granular quote-level data. We test this empirically by adding quote-level data from the Millisecond TAQ database to evaluate potential performance improvements. The additional features, which the literature indicates are linked to HFT activity (Chakrabarty et al. 2023), include message counts, quote update frequencies, small trade volumes (under 100 shares), and high-frequency midpoint variations over 100-millisecond intervals. We calculate these measures for 2009, our training period. Using January-June 2009 data, we train a pair of models: one using only original daily features from TAQ's Intraday Indicators, another incorporating both daily indicators and granular quote features from TAQ's Millisecond database. Based on the trained models, we generate HFT measures for July-December 2009, enabling out-of-sample comparison between the two models – with and without quote information.

The analysis offers three main observations. First, adding quote-related information only marginally improves model performance, raising R^2 from 82% to 84%. Second, the corresponding pairs of ML-generated HFT measures – with and without quote information – are highly correlated. The correlation coefficients for the liquidity-supplying and -demanding HFT metrics are 0.99 and 0.96, respectively. Third, when we regress the Nasdaq HFT values on the ML measures generated with quote-level information, coefficient estimates and *t*-statistics differ only marginally from those presented in Table 3. Hence, the TAQ intraday indicator features used in the baseline ML framework sufficiently capture HFT activity. These findings are unsurprising given that our baseline feature engineering incorporates variables strongly tied to quote-level activity, such as market depth and bid-ask spreads. The levels of correlation between quote-related and trade-related features emphasize the strength of these relationships. For example, total trades and message count have a correlation coefficient of 0.90, while message count has correlation coefficients of above 0.65 with both ISO trades and market depth. Quote revision frequency correlates strongly (above 0.70) with trade frequency, ISO trades, and market depth.

Our second extension addresses HFT measure scaling. We have shown that data-driven HFT measures effectively capture both liquidity-demanding and liquidity-supplying strategies and help address important economic questions. All our tests use scaled HFT measures, where

HFT trading volume gets normalized by total trading volume. This scaling matters to account for total trading volume when examining HFTs' role (Hendershott et al. 2011). However, since our ML algorithm trains on scaled HFT values, it may capture variation in total trading volume rather than HFT trading volume. We address this by using the ML model to predict unscaled HFT trading volume using the same input variables. The key target variables become unscaled liquidity-demanding and liquidity-supplying trading volumes, calculated as the sum of HH and HN (HH and NH) volumes for stock *i* and day *t* from NASDAQ HFT data. We then replicate all tests using these unscaled values.

Our main findings remain robust when using unscaled target variables, with the complete set of results presented in the Online Appendix. We confirm that: (1) data-driven unscaled HFT measures outperform conventional HFT proxies (Table OA.D.1); (2) HFT activity responds systematically to both events in our natural experimental set up (Table OA.D.2) and scheduled and unscheduled announcements (Figures OA.D.1 and OA.D.2); (3) HFT shows distinct responses to latency arbitrage opportunities (Table OA.D.3); and (4) the two HFT types have contrasting effects on information acquisition. Liquidity-demanding strategies impair it while liquidity-supplying strategies enhance it (Tables OA.D.4 and OA.D.5).

7. Conclusion

The impact of HFT on market quality has been one of the central questions for market microstructure research over the past fifteen years. However, the literature faces a key limitation in those studies either examine short-term market effects using detailed HFT data or investigate longer-term impacts using generic HFT measures that fail to differentiate between liquidity-demanding and liquidity-supplying strategies. This constraint has hampered our understanding of the mechanisms driving HFTs' effects over longer horizons.

We address this limitation by developing a data-driven approach that generates distinct measures for liquidity-demanding and liquidity-supplying HFT activity using ML techniques. By training ensembles on NASDAQ HFT data and TAQ variables, we create comprehensive HFT measures covering the entire U.S. stock universe over an extended period.

Our validation tests demonstrate that these ML-generated measures outperform traditional HFT measures and capture theoretically predicted HFT behavior. The measures respond to exogenous technological changes. Similarly, as latency arbitrage opportunities become more prevalent, liquidity-demanding HFTs increase their activity while liquidity-supplying HFTs reduce it.

We show the importance of differentiating HFT strategies by examining their role in fundamental information acquisition that requires a large sample to test. Our findings suggest that liquidity-supplying HFT activity is positively associated with information acquisition, while liquidity-demanding activity is negatively related to it. This result clarifies how different HFT strategies affect price informativeness in financial markets, highlighting the core advantages of our ML-generated proxies and empirical framework.

References

Aït-Sahalia, Y., Sağlam, M., 2024. High frequency market making: The role of speed. Journal of Econometrics 239, p.105421

Aquilina, M., Budish, E., O'neill, P., 2022. Quantifying the high-frequency trading "arms race". The Quarterly Journal of Economics 137, 493-564

Baldauf, M., Mollner, J., 2020. High-frequency trading and market performance. The Journal of Finance 75, 1495-1526

Ball, R., Shivakumar, L., 2008. How much new information is there in earnings? Journal of Accounting Research 46, 975-1016

Bartlett, R.P., O'Hara, M., 2024. Navigating the Murky World of Hidden Liquidity. Available at SSRN

Benos, E., Sagade, S., 2016. Price discovery and the cross-section of high-frequency trading. Journal of Financial Markets 30, 54-77

Biais, B., Foucault, T., Moinas, S., 2015. Equilibrium fast trading. Journal of Financial Economics 116, 292-313

Boehmer, E., Fong, K., Wu, J.J., 2021a. Algorithmic trading and market quality: International evidence. Journal of Financial and Quantitative Analysis 56, 2659-2688

Boehmer, E., Jones, C.M., Zhang, X., Zhang, X., 2021b. Tracking retail investor activity. The Journal of Finance 76, 2249-2305

Boehmer, E., Li, D., Saar, G., 2018. The competitive landscape of high-frequency trading firms. The Review of Financial Studies 31, 2227-2276

Bogousslavsky, V., Fos, V., Muravyev, D., 2023. Informed trading intensity. The Journal of Finance, Forthcoming

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. Monterey, CA: Wadsworth & Brooks. Cole Advanced Books and Software

Brogaard, J., Hagströmer, B., Nordén, L., Riordan, R., 2015. Trading fast and slow: Colocation and liquidity. The Review of Financial Studies 28, 3407-3443

Brogaard, J., Hendershott, T., Riordan, R., 2014. High-frequency trading and price discovery. The Review of Financial Studies 27, 2267-2306

Brogaard, J., Pan, J., 2022. Dark pool trading and information acquisition. The Review of Financial Studies 35, 2625-2666

Brogaard, J., Sokolov, K., Zhang, J., 2025. Strategic liquidity provision and extreme volatility spikes. Management science

Brunnermeier, M.K., 2005. Information leakage and market efficiency. The Review of Financial Studies 18, 417-457

Budish, E., Cramton, P., Shim, J., 2015. The high-frequency trading arms race: Frequent batch auctions as a market design response. The Quarterly Journal of Economics 130, 1547-1621

Cao, L., 2022. Ai in finance: challenges, techniques, and opportunities. ACM Computing Surveys (CSUR) 55, 1-38

Chakrabarty, B., Comerton-Forde, C., Pascual, R., 2023. Identifying High Frequency Trading Activity without Proprietary Data. Available at SSRN 4551238

Chakrabarty, B., Hendershott, T., Nawn, S., Pascual, R., 2025. Order exposure in high frequency markets. Journal of Financial and Quantitative Analysis Forthcoming

Chakravarty, S., Jain, P., Upson, J., Wood, R., 2012. Clean sweep: Informed trading through intermarket sweep orders. Journal of Financial and Quantitative Analysis 47, 415-435

Conrad, J., Wahal, S., Xiang, J., 2015. High-frequency quoting, trading, and the efficiency of prices. Journal of Financial Economics 116, 271-291

Easley, D., De Prado, M.L., O'Hara, M., 2011. The microstructure of the Flash Crash. Journal of Portfolio Management 37, 118-128

Easley, D., López de Prado, M., O'Hara, M., Zhang, Z., 2021. Microstructure in the machine age. The Review of Financial Studies 34, 3316-3363

Ettredge, M.L., Kwon, S.Y., Smith, D.B., Zarowin, P.A., 2005. The impact of SFAS No. 131 business segment data on the market's ability to anticipate future earnings. The Accounting Review 80, 773-804

Foucault, T., 2016. Where are the risks in high frequency trading? Financial Stability Review 20, 53-67

Foucault, T., Kozhan, R., Tham, W.W., 2017. Toxic arbitrage. The Review of Financial Studies 30, 1053-1094

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., Villa-Vialaneix, N., 2017. Random forests for big data. Big Data Research 9, 28-46

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine learning 63, 3-42

Gider, J., Schmickler, S., Westheide, C., 2019. High-frequency trading and price informativeness.

Goldstein, M., Kwan, A., Philip, R., 2023. High-frequency trading strategies. Management Science 69, 4413-4434

Hasbrouck, J., 2018. High-frequency quoting: Short-term volatility in bids and offers. Journal of Financial and Quantitative Analysis 53, 613-641

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer.

Hendershott, T., Jones, C.M., Menkveld, A.J., 2011. Does algorithmic trading improve liquidity? The Journal of Finance 66, 1-33

Hirschey, N., 2021. Do high-frequency traders anticipate buying and selling pressure? Management Science 67, 3321-3345

Ho, T.K., 1995. Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, pp. 278-282. IEEE

Jones, C.M., 2013. What do we know about high-frequency trading? Columbia Business School Research Paper

Khapko, M., Zoican, M., 2021. Do speed bumps curb low-latency investment? Evidence from a laboratory market. Journal of Financial Markets 55, 100601

Klein, O., 2020. Trading aggressiveness and market efficiency. Journal of Financial Markets 47, 100515

Kwan, A., Philip, R., Shkilko, A., 2021. The conduits of price discovery: A machine learning approach.

Lee, C.M., Radhakrishna, B., 2000. Inferring investor behavior: Evidence from TORQ data. Journal of Financial Markets 3, 83-111

Lee, C.M., Ready, M.J., 1991. Inferring trade direction from intraday data. The Journal of Finance 46, 733-746

Li, D.-C., Fang, Y.-H., Fang, Y.F., 2010. The data complexity index to construct an efficient cross-validation method. Decision Support Systems 50, 93-102

Li, S., Wang, X., Ye, M., 2021a. Who provides liquidity, and when? Journal of financial economics 141, 968-980

Li, S., Ye, M., Zheng, M., 2021b. Financial regulation, clientele segmentation, and stock exchange order types. National Bureau of Economic Research

Lundholm, R., Myers, L.A., 2002. Bringing the future forward: the effect of disclosure on the returns-earnings relation. Journal of accounting research 40, 809-839

Malceniece, L., Malcenieks, K., Putniņš, T.J., 2019. High frequency trading and comovement in financial markets. Journal of Financial Economics 134, 381-399

Menkveld, A.J., 2013. High frequency trading and the new market makers. Journal of Financial Markets 16, 712-740

Menkveld, A.J., 2016. The economics of high-frequency trading: Taking stock. Annual Review of Financial Economics 8, 1-24

Menkveld, A.J., Zoican, M.A., 2017. Need for speed? Exchange latency and liquidity. The Review of Financial Studies 30, 1188-1228

Moews, B., Davé, R., Mitra, S., Hassan, S., Cui, W., 2021. Hybrid analytic and machine-learned baryonic property insertion into galactic dark matter haloes. Monthly Notices of the Royal Astronomical Society 504, 4024-4038

Nimalendran, M., Rzayev, K., Sagade, S., 2024. High-frequency trading in the stock market and the costs of options market making. Journal of Financial Economics 159, 103900

O'Hara, M., 2003. Presidential address: Liquidity and price discovery. The Journal of Finance 58, 1335-1354

Parker, W.S., 2013. Ensemble modeling, uncertainty and robust predictions. Wiley interdisciplinary reviews: Climate change 4, 213-223

Probst, P., Boulesteix, A.-L., 2018. To tune or not to tune the number of trees in random forest. Journal of Machine Learning Research 18, 1-18

Rzayev, K., Ibikunle, G., Steffen, T., 2023. The market quality implications of speed in cross-platform trading: Evidence from Frankfurt-London microwave. Journal of Financial Markets 66, 100853

Shkilko, A., Sokolov, K., 2020. Every cloud has a silver lining: Fast trading, microwave connectivity, and trading costs. The Journal of Finance 75, 2899-2927

Stiglitz, J.E., 2014. Tapping the brakes: Are less active markets safer and better for the economy? In: Federal Reserve Bank of Atlanta 2014 Financial Markets Conference Tuning Financial Regulation for Stability and Efficiency, April

Van Kervel, V., Menkveld, A.J., 2019. High-frequency trading around large institutional orders. The Journal of Finance 74, 1091-1137

Weller, B.M., 2018. Does algorithmic trading reduce information acquisition? The Review of Financial Studies 31, 2184-2226

Yang, L., Zhu, H., 2020. Back-running: Seeking and hiding fundamental information in order flows. The Review of Financial Studies 33, 1484-1533

Ye, M., Yao, C., Gai, J., 2013. The externalities of high frequency trading. WBS Finance Group Research Paper

Zhang, Y., Lee, J., Wainwright, M., Jordan, M.I., 2017. On the learnability of fully-connected neural networks. In: Artificial Intelligence and Statistics, pp. 83-91. PMLR

Figure 1
Feature importance plot.

This figure shows the feature importance of each input variable in terms of how relevant it is to the construction of the model, meaning how much each feature contributes to the predictions made. Using the Gini impurity, importance values are calculated through the mean decrease and standard deviation in node impurity for tree-based models as the normalized total reduction of the measurement because of this feature.

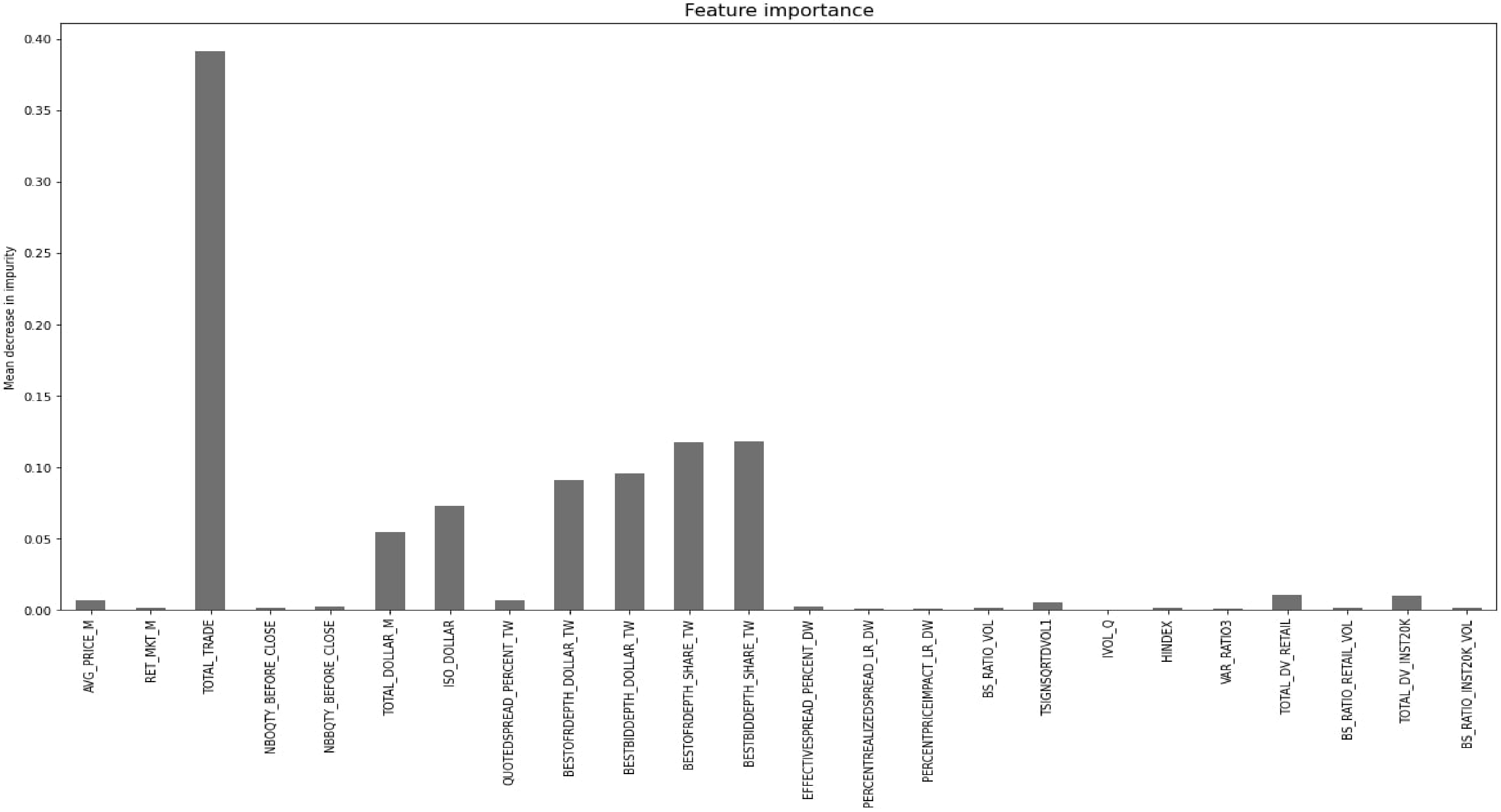


Figure 2

Partial dependence plots of ML-generated HFT proxies on selected variables.

This figure shows the marginal effect that input variables have on model predictions, and whether these relationships are nonlinear. Predictions are marginalized over the distribution of input variables resulting in a function that includes other variables and depends solely on the features of interest. This provides the average marginal effect on predictions for given values of these features.

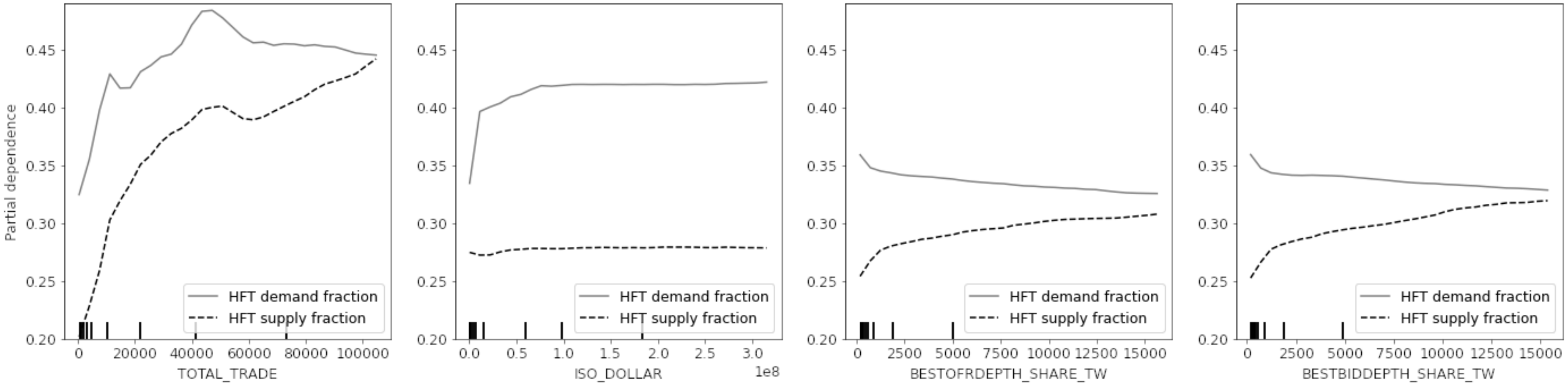
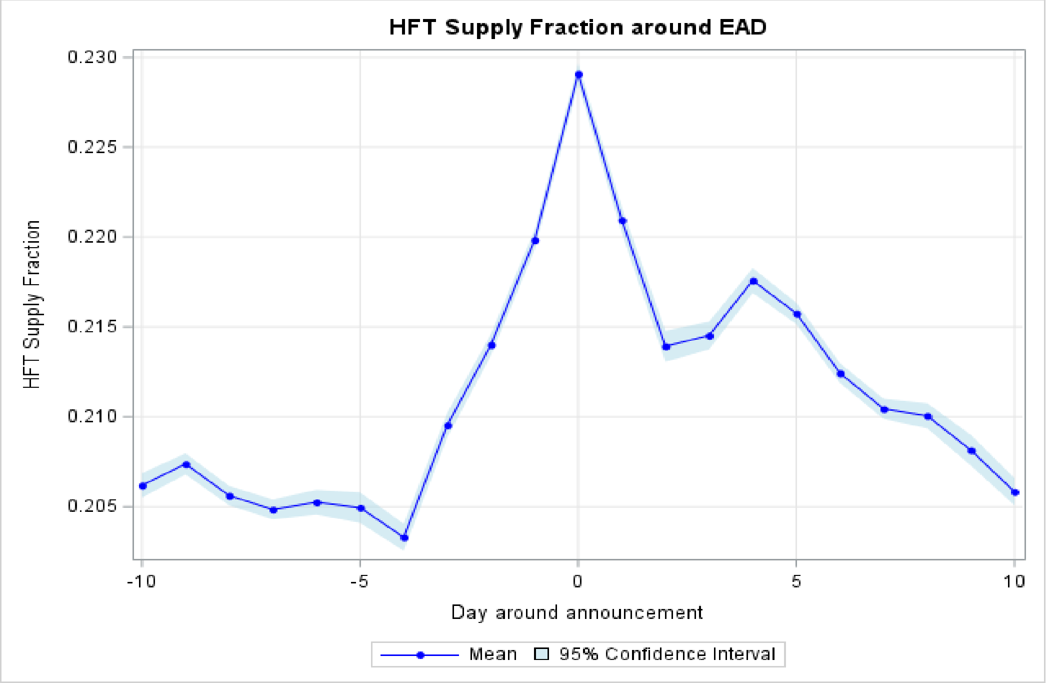


Figure 3
HFT around earnings announcements
This figure illustrates the evolution of ML-generated HFT measures with their 95% confidence interval surrounding scheduled events, specifically earnings announcements. The event window spans 10 days before and after the announcement dates, which are sourced from the I/B/E/S database. The analysis encompasses all U.S. listed common stocks, with the sample period extending from 2010 to 2023.

Panel A: $HFT\_S_{i,t}$ around earning announcements.



Panel B: $HFT\_D_{i,t}$ around earning announcements.

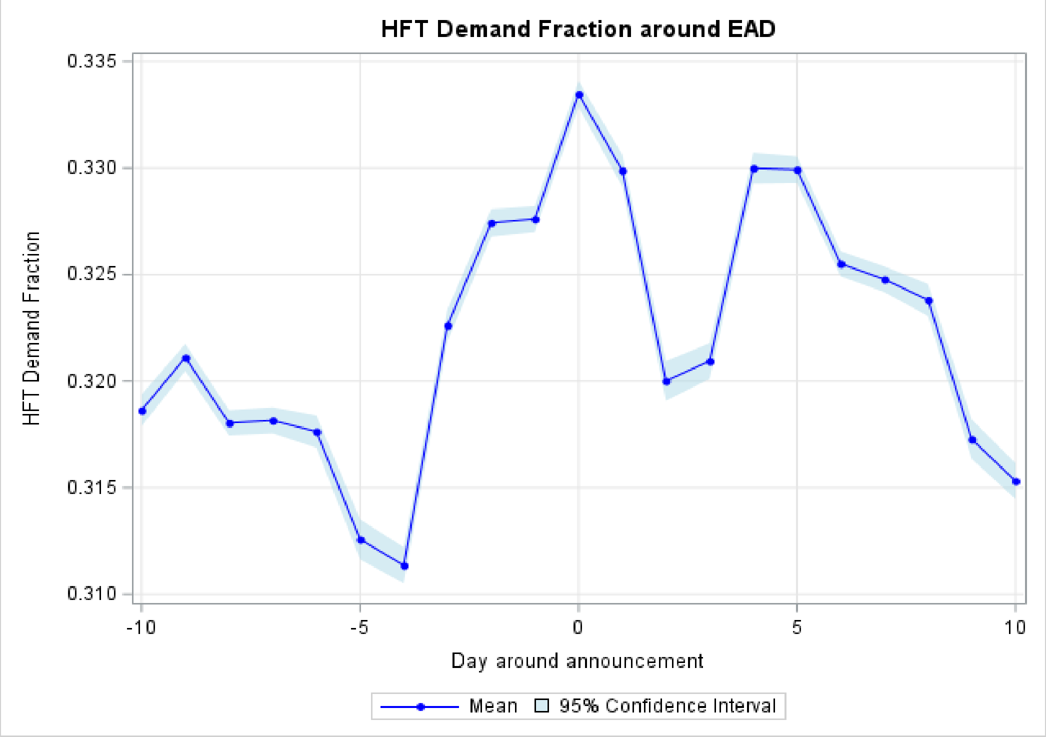


Table 1
Input and output variables in the ML model training process
This table presents the variables used to train the ML model, including their notation, descriptions, and data sources. Panel A contains output variables from NASDAQ HFT data. Panel B details input variables derived from the TAQ database, with variable labels matching the WRDS TAQ Data Manual for easy reference.

Variable	Description	Data source
Panel A: Output variables used in the ML model.		
$NASD\_HFT\_D_{i,t}$	Liquidity-demanding HFT activities for stock i in day t is computed as the daily number of shares traded by liquidity - demanding HFTs (HH and HN) divided by the total number of shares (HH, HN, NH, and NN) trading in day *t*.	NASDAQ HFT
$NASD\_HFT\_S_{i,t}$	Liquidity-supplying HFT activities for stock i in day t is computed as the daily number of shares traded by liquidity - supplying HFTs (HH and HN) divided by the total number of shares (HH, HN, NH, and NN) trading in day *t*.	NASDAQ HFT
Panel B: Input variables (features) used in the ML model.		
$AVG\_PRICE\_M_{i,t}$	Average trade price during market hours (Open to Close) for stock i in day t.	TAQ
$RET\_MKT\_M_{i,t}$	Open to close return for stock i in day t is computed as the log return of the official opening price over the official closing price.	TAQ
$TOTAL\_TRADE_{i,t}$	The total number of trades for stock i in day t.	TAQ
$NBOQTY\_BEFORE\_CLOSE_{i,t}$	The best offer size of the last quote before market close for stock i in day t.	TAQ
$NBBQTY\_BEFORE\_CLOSE_{i,t}$	The best bid size of the last quote before market close for stock i in day t.	TAQ
$TOTAL\_DOLLAR\_M_{i,t}$	The total trade value in dollars during market hours for stock i in day t.	TAQ
$ISO\_DOLLAR_{i,t}$	The sum of intermarket sweep order trade dollar value (during market hours) for stock i in day t.	TAQ
$QUOTEDSPREAD\_PERCENT\_TW_{i,t}$	The time-weighted percentage quoted spread (during market hours) for stock i in day t. The quoted spread is calculated as the difference between ask and bid prices for each transaction divided by the mid-price (the average of ask and bid prices).	TAQ
$BESTOFRDEPTH\_DOLLAR\_TW_{i,t}$	The time-weighted best offer dollar depth (during market hours) for stock i in day t is determined based on the size of the best ask price.	TAQ
		(continued)

$BESTBIDDEPTH\_DOLLAR\_TW_{i,t}$	The time-weighted best bid dollar depth (during market hours) for stock i in day t is determined as the size of the best bid price.	TAQ
$BESTOFRDEPTH\_SHARE\_TW_{i,t}$	The time-weighted best offer share depth (during market hours) for stock i in day t is determined based on the size of the best ask price.	TAQ
$BESTBIDDEPTH\_SHARE\_TW_{i,t}$	The time-weighted best bid share depth (during market hours) for stock i in day t is determined based on the size of the best bid price.	TAQ
$EFFECTIVESPREAD\_PERCENT\_DW_{i,t}$	The dollar value-weighted percentage effective spread for stock i in day t. The effective spread is calculated using the following equation: $Effective\ Spread = 2D_k(P_k - M_k)/M_k$, where k denotes transaction, D_k denotes the sign of transaction (-1 for sale and +1 for buy), P_k is the transaction price, and M_k is the prevailing mid-price for each transaction. Lee and Ready (1991) algorithm is used for trade classification.	TAQ
$PERCENTREALIZEDSPREAD\_LR\_DW_{i,t}$	The dollar value-weighted percentage realized spread for stock i in day t. The realized spread is calculated using the following equation: $Realized\ Spread = 2D_k(P_k - M_{k+5})/M_k$, where M_{k+5} is the bid-ask mid-point five minutes after the kth trade, and all other variables are as previously defined. Lee and Ready (1991) algorithm is used for trade classification.	TAQ
$PERCENTPRICEIMPACT\_LR\_DW_{i,t}$	The dollar value-weighted percentage price impact for stock i in day t. The price impact is calculated using the following equation: $Percent\ Price\ Impact = 2D_k(M_{k+5} - M_k)/M_k$, where all variables are as previously defined. Lee and Ready (1991) algorithm is used for trade classification.	TAQ
$BS\_RATIO\_VOL_{i,t}$	The absolute percentage order imbalance for stock i in day t is calculated as the absolute value of buy volume minus sell volume divided by the total trade volume. Lee and Ready (1991) algorithm is used for trade classification.	TAQ
$TSIGNSQRTDVOL1_{i,t}$	The lambda (price impact coefficient) with intercept for stock i in day t is calculated using the following equation: $Ln\frac{M_{i,s}}{M_{i,s-300}} = \alpha + \lambda * \text{SSqrtDvol} + \epsilon$, where SSqrtDvol $= Sgn(\sum_{s-300}^{s} BuyDollar - \sum_{s-300}^{s} SellDollar) \times \sqrt{\lvert\sum_{s-300}^{s} BuyDollar - \sum_{s-300}^{s} SellDollar\,\rvert}$, where $M_{i,s}$ is the mid-price for stock i at second s.	TAQ
$IVOL\_Q_{i,t}$	The quote-based intraday volatility for stock i in day t is calculated using the following equation: $Intraday\ Volatility = \frac{\sum_{s=1}^{S}(Ret_{i,s}-\overline{Ret_{i,s}})^2}{S-1}$, where $Ret_{i,s} = Ln\frac{M_{i,s}}{M_{i,s-1}}$ and $M_{i,s}$ is the mid-price for stock i at second s.	TAQ

(continued)

$HINDEX_{i,t}$	The Herfindahl index calculated across 30-minute time units for stock i in day t is calculated using the following equation: $HIndex = \frac{\sum_{s=1}^{1800}\sum_{k=1}^{N}(P_k \times SHR_k)^2}{(\sum_{s=1}^{1800}\sum_{k=1}^{N} P_k \times SHR_k)^2}$, where SHR_k is the shares of trade for transaction k.	TAQ
$VAR\_RATIO3_{i,t}$	The variance ratio for stock i in day t is calculated using the following equation: $Variance\ Ratio = \left\lvert\frac{Var(Ret_{300t})}{5 \times Var(Ret_{60t})} - 1\right\rvert$, where $Var(Ret_{300t})$ is the variance of 5-minute log returns.	TAQ
$TOTAL\_DV\_RETAIL_{i,t}$	The total dollar value of retail trades for stock i in day t. Retail trades are identified by using the methodology described in Boehmer et al. (2021b).	TAQ
$BS\_RATIO\_RETAIL\_VOL_{i,t}$	The absolute percentage order imbalance for retail trading volume for stock i in day t. Retail trades are identified by using the methodology described in Boehmer et al. (2021b).	TAQ
$TOTAL\_DV\_INST20K_{i,t}$	The total dollar value of $20,000 institutional trades for stock i in day t. $20,000 cutoff is based on Lee and Radhakrishna (2000).	TAQ
$BS\_RATIO\_INST20K\_VOL_{i,t}$	The absolute percentage order imbalance for $20,000 institutional trades' trading volume for stock i in day t. $20,000 cutoff is based on Lee and Radhakrishna (2000).	TAQ

Table 2
Regression variables and summary statistics
This table provides summary statistics and definitions of variables used in our regression analyses. Variable names in the first column are followed by their measurement units in parentheses. For variables used in multiple regressions with different frequencies (daily, quarterly, etc.), we report summary statistics corresponding to their first appearance in our analyses. All variables are winsorized at the 1st and 99th percentiles. For the original NASDAQ dataset variables ($NASD\_HFT\_D_{i,t}$ and $NASD\_HFT\_S_{i,t}$), the sample covers the year 2009 and includes 120 randomly selected NASDAQ- and NYSE-listed firms with NASDAQ HFT data. For all other variables, the sample includes all U.S.-listed common stocks from 2010 to 2023.

Variable	Definition	Mean	Std	Min	p.25	p.50	p.75	Max
$NASD\_HFT\_D_{i,t}$	Liquidity-demanding HFT activities for stock i in day t is computed as the daily number of shares traded by liquidity - demanding HFTs (HH and HN) divided by the total number of shares (HH, HN, NH, and NN) trading in day *t*.	0.331	0.160	0.013	0.202	0.342	0.453	0.662
$NASD\_HFT\_S_{i,t}$	Liquidity-supplying HFT activities for stock *i* in day t is computed as the daily number of shares traded by liquidity - supplying HFTs (HH and HN) divided by the total number of shares (HH, HN, NH, and NN) trading in day *t*.	0.250	0.169	0.010	0.110	0.206	0.375	0.636
$HFT\_D_{i,t}$	The liquidity-demanding HFT activity for stock *i* on day *t*, estimated using the ML model outlined in Section 3.	0.316	0.112	0.025	0.222	0.335	0.414	0.602
$HFT\_S_{i,t}$	The liquidity-supplying HFT activity for stock *i* on day *t*, estimated using the ML model outlined in Section 3.	0.208	0.101	0.036	0.131	0.174	0.259	0.626
$Volatility_{i,t}$ (1/00,000)	Daily volatility for stock *i* on day *t*, measured as the standard deviation of transaction-level returns.	0.008	0.018	0.000	0.000	0.001	0.007	0.123
$Spread_{i,t}$ (%)	Daily average of transaction-level spreads for stock *i* on day *t*, where each transaction-level spread is calculated as (ask price - bid price)/(0.5 × (ask price + bid price)).	0.142	0.154	0.012	0.037	0.090	0.189	0.885
$InvPrice_{i,t}$	The inverse of stock price for stock *i* on day *t*.	0.039	0.050	0.001	0.013	0.024	0.047	0.344
$Volume_{i,t}$ ($’000,000,00)	Daily trading volume in dollars for stock *i* on day *t*.	2.614	6.305	0.007	0.070	0.330	2.556	47.392
$NLAO_{i,t}$ (000)	The number of latency arbitrage opportunities for stock *i* on day *t*, identified using the methodology detailed in Section 4.2.	0.068	0.169	0.001	0.006	0.017	0.047	1.211
$Flick_{i,t}$ (0)	Quote volatility for stock *i* on day *t*, measured as the daily average of standard deviations of quote midpoints calculated over 100 ms intervals.	6.942	42.24	0.000	0.009	0.021	0.086	365.523
$OLV_{i,t}$	Daily average of trades smaller than 100 shares for stock *i* on day *t*.	3.040	12.47	0.000	0.000	0.000	1.000	80.000
$QuoteInt_{i,t}$ (000,000)	Daily count of changes in best quotes or quote depth for stock *i* on day *t*.	0.191	0.264	0.002	0.031	0.059	0.253	2.775
$QT_{i,t}$	The ratio of quoted shares to traded shares for stock *i* on day *t*.	15.82	16.23	2.19	5.88	9.51	18.71	85.70
$MG_{i,t}$ (000,000)	The total number of messages (trade and quote) for stock *i* on day *t*.	2.111	2.864	0.078	0.332	0.643	2.853	12.637

$JUMP_{i,q}$	Information acquisition proxy for stock i in quarter q, measured as the ratio of cumulative abnormal returns over [-1, 1] to cumulative abnormal returns over [-21, 1] around earnings announcements.	0.517	0.427	-0.543	0.227	0.510	0.794	1.663
$MValue_{i,q}$ ($'000,000,000)	Market value for stock i in quarter q, calculated as the average of daily market values over [-21, -1] around earnings announcements, where daily market value is closing price times shares outstanding.	0.567	1.652	0.001	0.024	0.089	0.330	12.474
$OIB20k_{i,q}$	Institutional order imbalance for stock i in quarter q, measured as the price impact of trades exceeding $20,000 over [-21, -1] around earnings announcements, obtained from TAQ.	0.351	0.183	0.050	0.200	0.333	0.494	0.763
$CT_{i,q}$	The natural logarithm of the cancel-to-trade ratio for stock i in quarter q, where the ratio is calculated as the average of daily (cancel messages/trade messages) over [-21, -1] around earnings announcements, obtained from MIDAS database.	0.507	0.540	-0.548	0.150	0.462	0.810	2.227
$OLR_{i,q}$	The natural logarithm of the odd-lot ratio for stock i in quarter q, where the ratio is calculated as the average of daily proportions of trades below 100 shares over [-21, -1] around earnings announcements, obtained from MIDAS database.	1.202	0.664	-0.430	0.777	1.288	1.735	2.212
$TO_{i,q}$	The natural logarithm of the trade-to-order ratio for stock i in quarter q, where the ratio is calculated as the average of daily (executed shares/submitted shares) over [-21, -1] around earnings announcements, obtained from MIDAS database.	-1.064	0.639	-2.972	-1.450	-1.017	-0.628	0.194

Table 3
Comparative analysis of HFT measures

This table evaluates our ML-generated HFT measures against alternative proxies using the following models:

$$NASD\_HFT\_S_{i,t} = \alpha_i + \beta_t + \gamma_1 HFT\_S_{i,t} + \gamma_2 Flick_{i,t} + \gamma_3 OLV_{i,t} + \gamma_4 QuoteInt_{i,t} + \gamma_5 QT_{i,t} + \gamma_6 MG_{i,t} + \varepsilon_{i,t}$$

$$NASD\_HFT\_D_{i,t} = \alpha_i + \beta_t + \gamma_1 HFT\_D_{i,t} + \gamma_2 Flick_{i,t} + \gamma_3 OLV_{i,t} + \gamma_4 QuoteInt_{i,t} + \gamma_5 QT_{i,t} + \gamma_6 MG_{i,t} + \varepsilon_{i,t}$$

where $NASD\_HFT\_D_{i,t}$ and $NASD\_HFT\_S_{i,t}$ are NASDAQ's liquidity-demanding and -supplying HFT measures, and $HFT\_D_{i,t}$ and $HFT\_S_{i,t}$ are our ML-generated proxies (trained on January-June 2009 data), and alternative proxies from TAQ: quote volatility ($Flick_{i,t}$, average standard deviation of quote midpoints over 100 ms intervals), $OLV_{i,t}$ ($OLV_{i,t}$, sum of sub-100 share trades), quote intensity ($QuoteInt_{i,t}$, count of quote/depth changes), quote-to-trade ratio ($QT_{i,t}$, quoted shares/traded shares), and the number of messages ($MG_{i,t}$). All dependent variables are standardized. The analysis presents results for liquidity-supplying HFT in Panels A and C, while Panels B and D focus on liquidity-demanding HFT. Panels A and B incorporate both stock and day fixed effects, whereas Panels C and D employ only day fixed effect. The sample covers July-December 2009 for 120 randomly selected NASDAQ- and NYSE-listed firms with NASDAQ HFT data. Standard errors are double-clustered by stock and day, with t-statistics in brackets. *, **, and *** indicate significance at 10%, 5%, and 1%. R^2 values are within-R^2.

Panel A*:* $NASD\_HFT\_S_{i,t}$

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$HFT\_S_{i,t}$	0.104***						0.096***
	(8.52)						(7.52)
$Flick_{i,t}$		0.002*					0.001
		(1.79)					(1.00)
$OLV_{i,t}$			0.001				0.001
			(0.84)				(0.80)
$QuoteInt_{i,t}$				0.015**			-0.020***
				(2.26)			(-3.16)
$QT_{i,t}$					0.008**		0.008**
					(2.10)		(2.37)
$MG_{i,t}$						0.021***	0.032***
						(3.21)	(3.10)
Stock and Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N obs.	14,238	14,238	14,238	14,238	14,238	14,238	14,238
R^2	3%	0.1%	0%	0.4%	0.2%	0.7%	3.3%

Panel B: $NASD\_HFT\_D_{i,t}$

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$HFT\_D_{i,t}$	0.068***						0.083***
	(3.70)						(4.38)
$Flick_{i,t}$		-0.003***					-0.003***
		(-3.27)					(-3.24)
$OLV_{i,t}$			-0.000				-0.000
			(-0.31)				(-0.07)
$QuoteInt_{i,t}$				-0.002			0.030***
				(-0.42)			(2.60)
$QT_{i,t}$					0.016***		0.019***
					(3.34)		(3.91)
$MG_{i,t}$						-0.006	-0.040***
						(-1.16)	(-2.82)
Stock and Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N obs.	14,238	14,238	14,238	14,238	14,238	14,238	14,238
R^2	0.8%	0.1%	0%	0.5%	0.5%	0.3%	1.4%

Panel C: $NASD\_HFT\_S_{i,t}$

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$HFT\_S_{i,t}$	0.246***						0.239***
	(38.28)						(23.27)
$Flick_{i,t}$		-0.009***					0.001
		(-2.69)					(0.06)
$OLV_{i,t}$			-0.004				0.002*
			(-0.91)				(1.68)
$QuoteInt_{i,t}$				0.140***			-0.006
				(14.02)			(-0.62)
$QT_{i,t}$					0.084***		0.005
					(5.41)		(1.12)
$MG_{i,t}$						0.144***	0.010
						(15.97)	(0.79)
Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N obs.	14,238	14,238	14,238	14,238	14,238	14,238	14,238
R^2	74%	0.3%	0%	51%	15%	53%	74%

Panel D: $NASD\_HFT\_D_{i,t}$

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$HFT\_D_{i,t}$	0.423*** (23.23)						0.383*** (17.93)
$Flick_{i,t}$		-0.003 (-0.41)					-0.003 (-0.93)
$OLV_{i,t}$			0.000 (0.24)				0.002** (1.99)
$QuoteInt_{i,t}$				0.091*** (7.92)			0.013 (0.92)
$QT_{i,t}$					0.021** (2.04)		0.014*** (2.69)
$MG_{i,t}$						0.092*** (8.37)	0.015 (0.94)
Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N obs.	14,238	14,238	14,238	14,238	14,238	14,238	14,238
R^2	50%	0.1%	0%	23%	1.0%	24%	54%

Table 4
Impact of exchange technological changes on HFT activity
This table examines how our ML-generated HFT measures respond to two technological changes: NASDAQ's reduced data dissemination latency and Amex's speed bump implementation. We estimate the following difference-in-difference models:

$$HFT\_D_{i,t} = \alpha_i + \beta_t + \gamma_1 Post_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (4.1)$$

$$HFT\_S_{i,t} = \alpha_i + \beta_t + \gamma_2 Post_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (4.2)$$

$$HFT\_D_{i,t} = \alpha_i + \beta_t + \gamma_1 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (4.3)$$

$$HFT\_S_{i,t} = \alpha_i + \beta_t + \gamma_2 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (4.4)$$

where $HFT\_D_{i,t}$ and $HFT\_S_{i,t}$ represent the ML – generated liquidity – demanding and – supplying HFT activities for stock i on day t. α_i and β_t capture stock and day fixed effects, respectively. For the NASDAQ upgrade analysis (Models 4.1 and 4.2), $Post_{i,t}$ equals 1 after October 10, 2011, for NASDAQ-listed stocks with tickers A-B, and after October 17, 2011, for other NASDAQ stocks. NYSE and Amex stocks serve as control groups in these models. For the Amex speed bump analysis (Models 4.3 and 4.4), $Post_{i,t}$ equals 1 after July 24, 2017, and $Amex_{i,t}$ equals 1 for Amex-listed stocks. NYSE and NASDAQ stocks serve as control groups in these models. Control variables ($C_{i,t}^{k}$) include daily volatility ($Volatility_{i,t}$, standard deviation of transaction-level returns), relative quoted spread ($Spread_{i,t}$, daily average of (ask-bid)/mid-quote for each transaction), inverse price ($InvPrice_{i,t}$), and dollar trading volume ($Volume_{i,t}$). The analysis uses 10-working day windows around implementation dates. Panel A reports results for the NASDAQ upgrade and Panel B for the Amex speed bump. Standard errors are double-clustered by stock and day, with t-statistics in brackets. *, **, and *** indicate significance at 10%, 5%, and 1%. R^2 values are within-R^2.

	Panel A: NASDAQ upgrade		Panel B: Amex speed bump	
	(i) $HFT\_D_{i,t}$	(ii) $HFT\_S_{i,t}$	(iii) $HFT\_D_{i,t}$	(iv) $HFT\_S_{i,t}$
$Post_{i,t}$	0.002** (2.12)	0.002** (2.10)		
$Post_{i,t} * Amex_{i,t}$			-0.005** (-2.34)	-0.007*** (-3.31)
$Volatility_{i,t}$	0.013** (2.19)	0.000 (0.07)	0.001 (1.29)	0.001 (1.33)
$Spread_{i,t}$	-0.066*** (-12.58)	-0.024*** (-5.58)	-0.015*** (-10.96)	-0.006*** (-6.05)
$InvPrice_{i,t}$	-0.151*** (-3.08)	0.037 (0.92)	-0.026 (-1.57)	-0.023* (-1.96)
$Volume_{i,t}$	0.001 (1.30)	0.020*** (17.75)	0.001** (2.25)	0.005*** (4.24)
Stock and Day FE	Yes	Yes	Yes	Yes
N obs.	43,234	43,234	45,530	45,530
R^2	5%	11%	1.3%	3.5%

Table 5
HFT response to latency arbitrage opportunities
This table examines how our ML-generated HFT measures respond latency arbitrage opportunities using the following OLS models:

$$HFT\_D_{i,t} = \alpha_i + \beta_t + \gamma_1 NLAO_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t}$$

$$HFT\_S_{i,t} = \alpha_i + \beta_t + \gamma_2 NLAO_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t}$$

where $HFT\_D_{i,t}$ and $HFT\_S_{i,t}$ represent our liquidity-demanding and -supplying HFT activity measures for stock i and day t. α_i and β_t capture stock and day fixed effects, respectively. $NLAO_{i,t}$ is the number of latency arbitrage opportunities, identified using the methodology detailed in Section 4.2. Control variables ($C_{i,t}^{k}$) include daily volatility ($Volatility_{i,t}$, standard deviation of transaction-level returns), relative quoted spread ($Spread_{i,t}$, daily average of (ask-bid)/mid-quote for each transaction), inverse price ($InvPrice_{i,t}$), and dollar trading volume ($Volume_{i,t}$). Columns (i) and (ii) present the results for $HFT\_D_{i,t}$ and $HFT\_S_{i,t}$, respectively. The sample consists of 120 randomly selected NASDAQ- and NYSE-listed firms. Standard errors are double-clustered by stock and day, with t-statistics in brackets. *, **, and *** indicate significance at 10%, 5%, and 1%. R^2 values are within-R^2.

	(i) $HFT\_D_{i,t}$	(ii) $HFT\_S_{i,t}$
$NLAO_{i,t}$	0.018***	-0.020**
	(3.78)	(-2.02)
$Volatility_{i,t}$	-0.302***	-0.353***
	(-5.91)	(-4.50)
$Spread_{i,t}$	-0.069***	-0.033**
	(-4.66)	(-2.09)
$InvPrice_{i,t}$	-0.390***	0.428***
	(-6.04)	(7.98)
$Volume_{i,t}$	-0.002***	0.003***
	(-3.81)	(7.96)
Stock and Day FE	Yes	Yes
N obs.	246,139	246,139
R^2	17%	12%

Table 6

HFT activity and information acquisition – jump ratio

This table examines how HFT activity affects information acquisition using the following OLS model:

$$JUMP_{i,q} = \alpha_i + \beta_{m,q} + \gamma_1 HFT\_D_{i,q} + \gamma_2 HFT\_S_{i,q} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

where $JUMP_{i,q}$ measures information acquisition for stock *i* as the ratio of cumulative abnormal returns over [-1, 1] to cumulative abnormal returns over [-21, 1] around quarterly earnings announcements (q). $HFT\_D_{i,q}$ and $HFT\_S_{i,q}$ are our liquidity-demanding and liquidity-supplying HFT activities, measured as averages of daily values over [-21, -1] around earnings announcements. Models include stock (α_i) and month ($\beta_{m,q}$) fixed effects, respectively. Control variables ($C_{i,q}^{k}$) all measured as averages of daily values over [-21, -1] around earnings announcements, include volatility ($Volatility_{i,q}$), relative quoted spread ($Spread_{i,q}$), market value ($MValue_{i,q}$, price times shares outstanding), and institutional order imbalance ($OIB20k_{i,q}$, price impact of trades over $20,000 from TAQ). Columns (i) and (ii) present results from models without and with control variables, respectively. The sample includes all U.S.-listed common stocks from 2010 to 2023. Standard errors are double-clustered by stock and quarter, with t-statistics in brackets. *, **, and *** indicate significance at 10%, 5%, and 1%. R^2 values are within-R^2.

	(i) $JUMP_{i,q}$	(ii) $JUMP_{i,q}$
$HFT\_D_{i,q}$	0.208*** (5.42)	0.178*** (4.57)
$HFT\_S_{i,q}$	-0.162*** (-3.32)	-0.133*** (-2.71)
$Volatility_{i,q}$		-0.048*** (-2.87)
$Spread_{i,q}$		-0.106*** (-6.45)
$MValue_{i,q}$		-0.009*** (-3.52)
$OIB20k_{i,q}$		0.132*** (7.26)
Stock and Month FE	Yes	Yes
N obs.	49,515	49,515
R^2	0.1%	0.4%

Table 7

Comparing our HFT measures with Weller (2018) measures

This table analyzes the relationship between our ML-generated HFT measures and Weller's (2018) HFT proxies using the following OLS models:

$$CT_{i,q} = \alpha_i + \beta_{m,q} + \gamma_1 HFT\_D_{i,q} + \gamma_2 HFT\_S_{i,q} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

$$OLR_{i,q} = \alpha_i + \beta_{m,q} + \gamma_1 HFT\_D_{i,q} + \gamma_2 HFT\_S_{i,q} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

$$TO_{i,q} = \alpha_i + \beta_{m,q} + \gamma_1 HFT\_D_{i,q} + \gamma_2 HFT\_S_{i,q} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

The dependent variables are Weller's (2018) HFT proxies obtained from the MIDAS database: $CT_{i,q}$ (natural logarithm of cancel-to-trade ratio), $OLR_{i,q}$ (natural logarithm of odd-lot ratio), and $TO_{i,q}$ (natural logarithm of trade-to-order ratio), where each ratio is calculated as the average of daily values over [-21, -1] around earnings announcements. The key independent variables are $HFT\_D_{i,q}$ and $HFT\_S_{i,q}$ are liquidity-demanding and liquidity-supplying HFT activities, measured as averages of daily values over [-21, -1] around earnings announcements. Models include stock (α_i) and month ($\beta_{m,q}$) fixed effects, respectively. Control variables ($C_{i,q}^{k}$) all measured as averages of daily values over [-21, -1] around earnings announcements, include volatility ($Volatility_{i,q}$), relative quoted spread ($Spread_{i,q}$), market value ($MValue_{i,q}$, price times shares outstanding), and institutional order imbalance ($OIB20k_{i,q}$, price impact of trades over $20,000 from TAQ). The sample includes all U.S.-listed common stocks from 2012 to 2023. Standard errors are double-clustered by stock and quarter, with t-statistics in brackets. *, **, and *** indicate significance at 10%, 5%, and 1%. R^2 values are within-R^2.

	(i)	(ii)	(ii)
	$CT_{i,q}$	$OLR_{i,q}$	$TO_{i,q}$
$HFT\_D_{i,q}$	0.839***	2.714***	-1.208***
	(10.64)	(24.76)	(-15.71)
$HFT\_S_{i,q}$	-1.133***	-2.343***	1.340***
	(-12.01)	(-26.72)	(13.38)
$Volatility_{i,q}$	0.036	-0.476***	0.492***
	(0.74)	(-11.05)	(10.05)
$Spread_{i,q}$	-0.005	0.727***	-0.190***
	(-0.17)	(12.03)	(-5.76)
$MValue_{i,q}$	0.050***	0.120***	-0.071***
	(7.52)	(11.74)	(-8.63)
$OIB20k_{i,q}$	0.152***	-0.111***	0.063*
	(5.24)	(-3.22)	(1.81)
Stock and Month FEs	Yes	Yes	Yes
N obs.	43,091	43,091	43,091
R^2	2%	19%	4%

Table 8
HFT activity and information acquisition – FERC alternative measure
This table examines how HFT activity affects information acquisition using the following model:

$$Return_{i,q} = \alpha_i + \beta_q + \sum_{n=-1}^{1}(\gamma_n Earning_{i,q+n} + \vartheta_n Earning_{i,q+n} * HFT\_D_{i,q} + \theta_n Earning_{i,q+n} * HFT\_D_{i,q}) + \rho_1 HFT\_D_{i,q} + \rho_2 HFT\_S_{i,q} + \rho_3 Return_{i,q+1} + \rho_4 Return_{i,q-1} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,q}$$

where $Return_{i,q}$ is quarterly stock returns for firm i in quarter q, measured as the percentage change in closing prices between quarters $q-1$ and q. $Earning_{i,q+n}$ denotes quarterly earnings (net income) normalized by the market value of equity at the start of quarter $q+n$. The subscript n ranges from -1 to 1. $HFT\_D_{i,q}$ and $HFT\_S_{i,q}$ are our liquidity-demanding and liquidity-supplying HFT activity measures, measured as the quarterly averages of daily values. Control variables ($C_{i,q}^{k}$) all measured as quarterly averages of daily values, include volatility ($Volatility_{i,q}$), relative quoted spread ($Spread_{i,q}$), market value ($MValue_{i,q}$, price times shares outstanding), and institutional order imbalance ($OIB20k_{i,q}$, price impact of trades over $20,000 from TAQ). Columns (i) and (ii) present results from models without and with control variables, respectively. The sample includes all U.S.-listed common stocks from 2010 to 2023. Standard errors are double-clustered by stock and quarter, with t-statistics in brackets. *, **, and *** indicate significance at 10%, 5%, and 1%. R^2 values are within-R^2.

	(i) $Return_{i,q}$	(ii) $Return_{i,q}$
$Earning_{i,q+1} * HFT\_D_{i,q}$	-2.035***	-2.018***
	(4.59)	(4.56)
$Earning_{i,q+1} * HFT\_S_{i,q}$	2.690***	2.676***
	(5.30)	(5.25)
$HFT\_D_{i,q}$	-0.060	-0.059
	(-1.58)	(-1.59)
$HFT\_S_{i,q}$	0.011	0.010
	(0.19)	(0.08)
$Earning_{i,q+1}$	0.575***	0.573***
	(9.38)	(9.67)
Additional Controls	No	Yes
Stock and Quarter FE	Yes	Yes
N obs.	157,343	157,343
R^2	4%	4%

Online Appendix for **"Data-Driven Measures of High-Frequency Trading"**

July 2025

Introduction

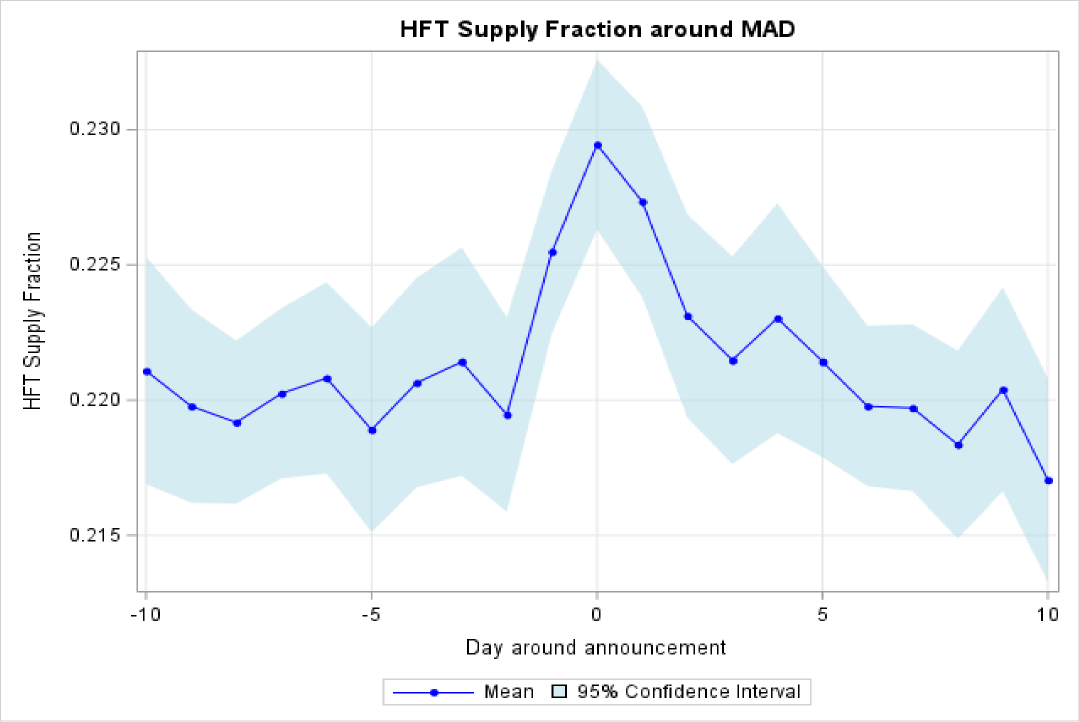
This online appendix provides supplementary results to the findings presented in Ibikunle et al. (2025). The content is as follows:

- Online Appendix A. HFT Activity Around M&A Announcements.
- Online Appendix B. Model Optimization and Machine Learning Comparisons
- Online Appendix C. Additional Tests on HFT and Information Acquisition
- Online Appendix D. Using Unscaled HFT Measures

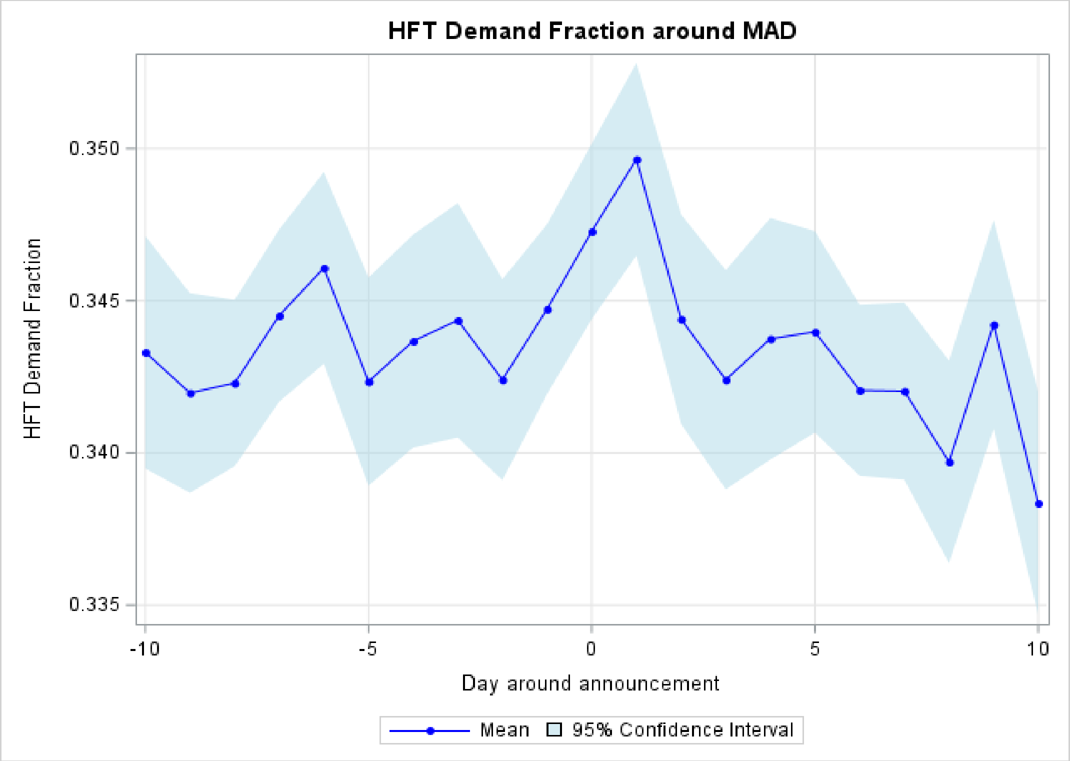
Online Appendix A. HFT activity around M&A announcements

Figure OA.1
HFT around M&A announcements
This figure illustrates the evolution of ML-generated HFT measures with their 95% confidence interval surrounding unscheduled events, specifically mergers and acquisitions (M&A) announcements. The event window spans 10 days before and after the announcement dates, which are sourced from the Thomson Reuters Securities Data Company (SDC) database. The analysis encompasses all U.S. listed common stocks, with the sample period extending from 2010 to 2023.

Panel A: $HFT\_S_{i,t}$ around M&A announcements.



Panel B: $HFT\_D_{i,t}$ around M&A announcements.



Online Appendix B. Model optimization and machine learning comparisons

Table OA.B.1

Parameter optimization results

The table lists the arithmetic mean and standard deviation for R^2 values across 10 iterations for different parameter combinations regarding the number of samples requires to split a tree node and the number of trees determining the ensemble size. Results are ranked by the Mean column.

Rank	Mean	Std.	Split samples	Ensemble size
1	0.814442	0.008260	5	640
2	0.813941	0.008360	5	320
3	0.813713	0.008455	5	160
4	0.812587	0.008609	5	80
5	0.810152	0.008016	5	40
...	...	...	...	...
60	0.659040	0.027015	640	160
61	0.658566	0.022346	640	80
62	0.657760	0.022598	640	320
63	0.655796	0.023405	640	10
64	0.654791	0.027320	640	5

OA.B.2. Comparison to simple machine learning methods

We compare our ensemble methods against standard single-model alternatives to validate our approach. Since we predict continuous HFT outcomes rather than classify discrete categories, we test four benchmark models: LASSO (which is a linear regression that shrinks coefficients to zero), Support Vector Machines (which capture non-linear patterns), and neural networks with three hidden layers (which can learn complex relationships but require more data and computation time). We configure each model with standard parameters: LASSO uses alpha=0.1 and tolerance=0.0001, SVM employs radial basis kernels, and neural networks use rectified linear activation with mean absolute error optimization. This allows us to benchmark whether tree ensembles truly outperform simpler methods (LASSO, SVM) and more complex alternatives (deep learning) for HFT prediction.

Our dataset spans 29,880 stock-days. We drop 2,184 observations (7%) with missing dependent or independent variables. We standardize each variable using z-score scaling to prevent variables with larger ranges from dominating predictions and to convert predictors to comparable units. We choose z-score over min-max scaling because HFT data contains extreme outliers that would distort min-max normalization. these initial experiments apply z-score scaling, also commonly called standardization, in which, for a dataset, D,

$$z_{D_i} = \frac{D_i - \bar{D}}{\sigma(D)} \qquad \text{(OA.1),}$$

We test two prediction approaches: multi-model (separate models for each target) versus multi-target (one model predicting both outcomes simultaneously). Multi-target models can capture relationships between our two dependent variables, potentially improving accuracy. However, LASSO and SVM require separate models by design, while neural networks can handle both approaches.

Online Appendix Table OA.B.2 reports mean R² values and standard deviations across 10 iterations for all methods, comparing single-target versus multi-target performance where applicable. Extra trees deliver the highest mean R^2 performance with low standard deviation, outperforming both simpler methods (LASSO, SVM) and complex neural networks. While neural networks can theoretically approximate any function, they struggle to learn optimal parameters from our HFT dataset—a common challenge when financial data contains high noise relative to signal strength (Zhang et al. 2017).

Our optimized extra trees model achieves an average R^2 of 0.825 with standard deviation of 0.005 across multiple runs. We no longer apply z-score scaling since tree-based models handle unscaled inputs effectively through their splitting mechanism. Optimization cuts prediction variance in half compared to our baseline model while improving accuracy. Therefore, we select extra trees as our primary method because they achieve superior prediction accuracy while remaining interpretable and computationally efficient.

Table OA.B.2
Machine learning comparison
The table lists the arithmetic mean and standard deviation for R^2 values across 10 iterations for least absolute shrinkage and selection operator (LASSO), support vector regression (SVR), feed-forward artificial neural networks (ANN), random forests for multi-model (RF-MM) and multi-target (RF) setups, and extremely randomized trees for multi-model (ET-MM) and multi-target (ET) setups. Results are inversely ranked by the Mean column.

Method	Mean	Std.
LASSO	0.625	0.013
SVR	0.684	0.058
ANN	0.783	0.0229
RF-MM	0.784	0.055
RF	0.790	0.043
ET-MM	0.804	0.036
ET	0.805	0.035

Online Appendix C. Additional tests on HFT and information acquisition

Table OA.C.1

HFT activity and information acquisition using Nasdaq HFT data

This table replicates the analyses from Tables 6 and 8 using NASDAQ's original HFT measures instead of our ML-generated proxies. $NASD\_HFT\_D_{i,q}$ and $NASD\_HFT\_S_{i,q}$ are NASDAQ's liquidity-demanding and -supplying HFT measures. The sample consists of 120 randomly selected stocks for which NASDAQ provided HFT data in 2009. All other specifications, including variable definitions, measurement periods, control variables, and fixed effects, remain identical to those in Tables 9 and 11.

	(i) $JUMP_{i,q}$	(ii) $Return_{i,q}$
$NASD\_HFT\_D_{i,q}$	0.997 (0.52)	
$NASD\_HFT\_S_{i,q}$	-0.903 (-0.56)	
$Earning_{i,q+1} * NASD\_HFT\_D_{i,q}$		0.521 (0.06)
$Earning_{i,q+1} * NASD\_HFT\_S_{i,q}$		-3.246 (-0.59)
Controls	As in Table 6	As in Table 8
Stock and Month FEs	Yes	Yes
N obs.	466	401
R^2	0.7%	40%

Table OA.C.2
HFT activity and information acquisition: analysis of 2010-2012 period

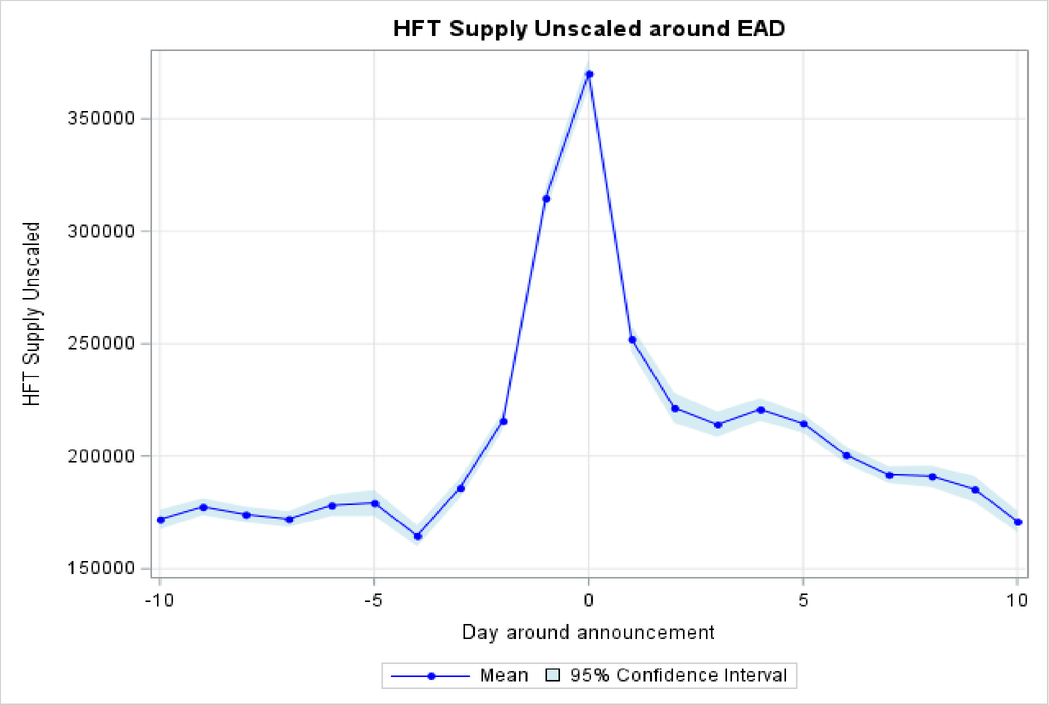
This table replicates the analyses from Tables 6 and 8 using data from 2010 to 2012, a period immediately following our ML model's training sample (2009). All other specifications, including variable definitions, measurement periods, control variables, and fixed effects, remain identical to those in Tables 9 and 11.

	(i) $JUMP_{i,q}$	(ii) $Return_{i,q}$
$HFT\_D_{i,q}$	0.114*** (2.59)	
$HFT\_S_{i,q}$	-0.101** (-2.10)	
$Earning_{i,q+1} * HFT\_D_{i,q}$		-3.982*** (-3.41)
$Earning_{i,q+1} * HFT\_S_{i,q}$		3.666*** (2.81)
Controls	As in Table 6	As in Table 8
Stock and Month FEs	Yes	Yes
N obs.	9,915	30,048
R^2	0.4%	5%

Online Appendix D. Using unscaled HFT measures

Figure OA.D.1
HFT around earnings announcements
This figure illustrates the evolution of ML-generated unscaled HFT measures ($U\_HFT\_S_{i,t}$ and $U\_HFT\_D_{i,t}$) with their 95% confidence interval surrounding scheduled events, specifically earnings announcements. The event window spans 10 days before and after the announcement dates, which are sourced from the I/B/E/S database. The analysis encompasses all U.S. listed common stocks, with the sample period extending from 2010 to 2023.

Panel A*:* $U\_HFT\_S_{i,t}$ around earning announcements.



Panel B: $U\_HFT\_D_{i,t}$ around earning announcements.

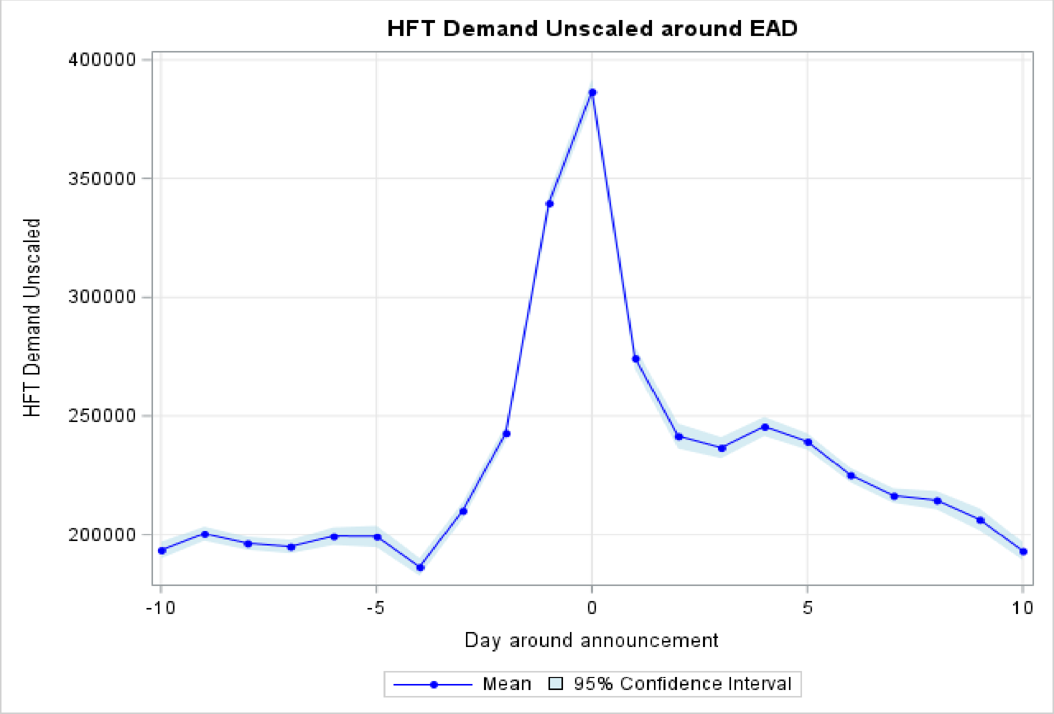
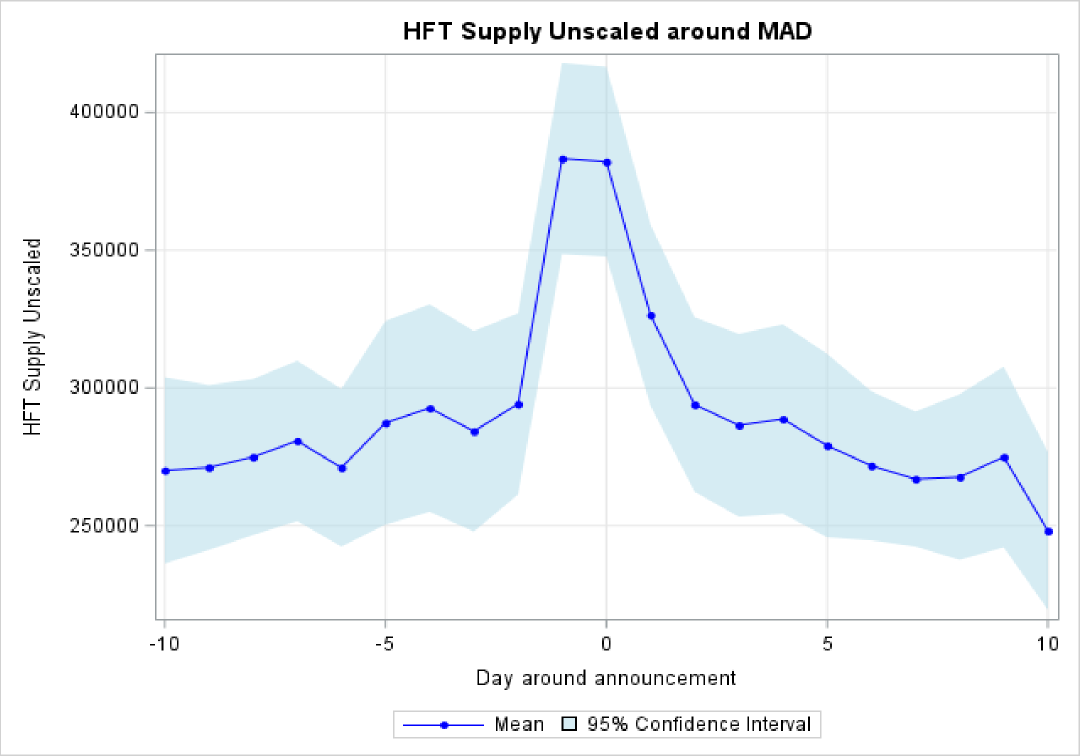


Figure OA.D.2
HFT around M&A announcements
This figure illustrates the evolution of ML-generated unscaled HFT measures ($U\_HFT\_S_{i,t}$ and $U\_HFT\_D_{i,t}$) with their 95% confidence interval surrounding unscheduled events, specifically mergers and acquisitions (M&A) announcements. The event window spans 10 days before and after the announcement dates, which are sourced from the Thomson Reuters Securities Data Company (SDC) database. The analysis encompasses all U.S. listed common stocks, with the sample period extending from 2010 to 2023.

Panel A: $U\_HFT\_S_{i,t}$ around M&A announcements.



Panel B: $U\_HFT\_D_{i,t}$ around M&A announcements.

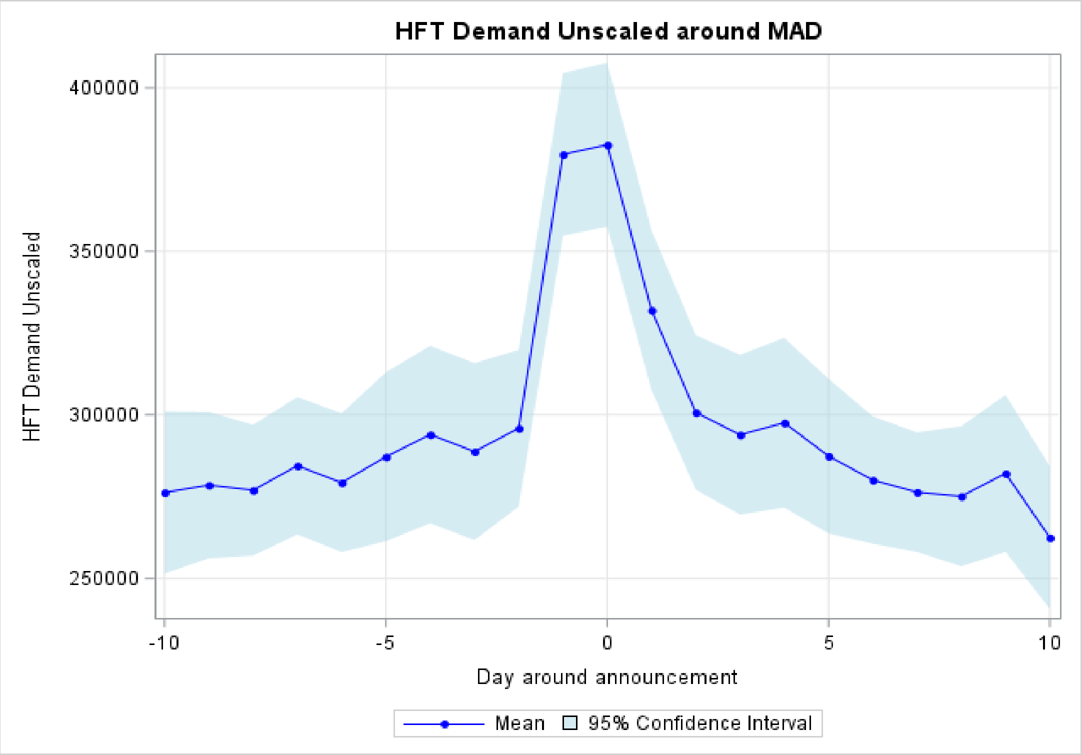


Table OA.D.1
Comparative analysis of HFT measures
This table evaluates our ML-generated unscaled HFT measures against alternative proxies using the following models:

$$NASD\_U\_HFT\_D_{i,t} = \alpha_i + \beta_t + \gamma_1 U\_HFT\_D_{i,t} + \gamma_2 Flick_{i,t} + \gamma_3 OLV_{i,t} + \gamma_4 QuoteInt_{i,t} + \gamma_5 QT_{i,t} + \gamma_6 MG_{i,t} + \varepsilon_{i,t}$$
$$NASD\_U\_HFT\_S_{i,t} = \alpha_i + \beta_t + \gamma_1 U\_HFT\_S_{i,t} + \gamma_2 Flick_{i,t} + \gamma_3 OLV_{i,t} + \gamma_4 QuoteInt_{i,t} + \gamma_5 QT_{i,t} + \gamma_6 MG_{i,t} + \varepsilon_{i,t}$$

where $NASD\_U\_HFT\_D_{i,t}$ and $NASD\_U\_HFT\_S_{i,t}$ are NASDAQ's unscaled liquidity-demanding and -supplying HFT measures, and $U\_HFT\_D_{i,t}$ and $U_HFT_S_{i,t}$are our ML-generated unscaled HFT proxies, trained on January-June 2009 data) and alternative proxies from TAQ: quote volatility ($Flick_{i,t}$, average standard deviation of quote midpoints over 100 ms intervals), $OLV_{i,t}$ ($OLV_{i,t}$, sum of sub-100 share trades), quote intensity ($QuoteInt_{i,t}$, count of quote/depth changes), quote-to-trade ratio ($QT_{i,t}$, quoted shares/traded shares), and the number of messages ($MG_{i,t}$).. All dependent variables are standardized. The analysis presents results for liquidity-supplying HFT in Panels A and C, while Panels B and D focus on liquidity-demanding HFT. Panels A and B incorporate both stock and day fixed effects, whereas Panels C and D employ only day fixed effect. The sample covers July-December 2009 for 120 randomly selected NASDAQ- and NYSE-listed firms with NASDAQ HFT data. Standard errors are double-clustered by stock and day, with t-statistics in brackets. *, **, and *** indicate significance at 10%, 5%, and 1%. R^2 values are within-R^2.

Panel A: $NASD\_U\_HFT\_S_{i,t}$

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$U\_HFT\_S_{i,t}$	1.349***						1.163***
	(10.93)						(9.14)
$Flick_{i,t}$		0.002					-0.002
		(0.78)					(-0.87)
$OLV_{i,t}$			0.012*				0.005***
			(1.80)				(2.59)
$QuoteInt_{i,t}$				0.793***			-0.073
				(5.89)			(-0.67)
$QT_{i,t}$					-0.253***		-0.133**
					(-3.16)		(-3.92)
$MG_{i,t}$						0.940***	0.440***
						(5.88)	(2.88)
Stock and Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N obs.	14,238	14,238	14,238	14,238	14,238	14,238	14,238
R^2	68%	0.1%	0.5%	24%	4.5%	27%	72%

Panel B: $NASD\_U\_HFT\_D_{i,t}$

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$U\_HFT\_D_{i,t}$	1.045***						0.849***
	(11.47)						(9.75)
$Flick_{i,t}$		0.002					-0.005***
		(0.09)					(-3.14)
$OLV_{i,t}$			0.008				0.004**
			(1.31)				(2.45)
$QuoteInt_{i,t}$				0.672***			0.090
				(6.12)			(1.15)
$QT_{i,t}$					-0.179***		-0.100***
					(-3.27)		(-4.67)
$MG_{i,t}$						0.788***	0.246**
						(6.08)	(2.07)
Stock and Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N obs.	14,238	14,238	14,238	14,238	14,238	14,238	14,238
R^2	64%	0%	0.5%	31%	4%	33%	69%

Panel C: $NASD\_U\_HFT\_S_{i,t}$

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$U\_HFT\_S_{i,t}$	1.552***						1.534***
	(31.38)						(23.86)
$Flick_{i,t}$		-0.085***					-0.001
		(-3.57)					(-0.49)
$OLV_{i,t}$			-0.052				-0.006
			(-1.21)				(-0.79)
$QuoteInt_{i,t}$				1.439***			0.082
				(6.04)			(0.50)
$QT_{i,t}$					1.054***		0.022
					(3.47)		(0.71)
$MG_{i,t}$						1.458***	-0.070
						(6.01)	(-0.39)
Stock and Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N obs.	14,238	14,238	14,238	14,238	14,238	14,238	14,238
R^2	90%	0.3%	0.1%	60%	25%	61%	95%

Panel D: $NASD\_U\_HFT\_D_{i,t}$

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$U\_HFT\_D_{i,t}$	1.167***						1.146***
	(27.10)						(16.68)
$Flick_{i,t}$		-0.067***					-0.002
		(-3.82)					(-1.62)
$OLV_{i,t}$			-0.038				-0.005
			(-1.06)				(-0.28)
$QuoteInt_{i,t}$				1.130***			0.245
				(7.68)			(1.55)
$QT_{i,t}$					0.710***		0.022
					(3.33)		(0.95)
$MG_{i,t}$						1.144***	-0.226
						(7.63)	(-1.32)
Stock and Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N obs.	14,238	14,238	14,238	14,238	14,238	14,238	14,238
R^2	92%	0.4%	0.1%	70%	22%	71%	94%

Table OA.D.2

Impact of exchange technological changes on HFT activity

This table examines how our ML-generated unscaled HFT measures respond to two technological changes: NASDAQ's reduced data dissemination latency and Amex's speed bump implementation. We estimate the following difference-in-difference models:

$$U\_HFT\_D_{i,t} = \alpha_i + \beta_t + \gamma_1 Post_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad \text{(OA.6.1)}$$

$$U\_HFT\_S_{i,t} = \alpha_i + \beta_t + \gamma_2 Post_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad \text{(OA.6.2)}$$

$$U\_HFT\_D_{i,t} = \alpha_i + \beta_t + \gamma_1 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad \text{(OA.6.3)}$$

$$U\_HFT\_S_{i,t} = \alpha_i + \beta_t + \gamma_2 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad \text{(OA.6.4)}$$

where $U\_HFT\_D_{i,t}$ and $U\_HFT\_S_{i,t}$ represent the ML – generated unscaled liquidity – demanding and – supplying HFT activities for stock i on day t. α_i and β_t capture stock and day fixed effects, respectively. For the NASDAQ upgrade analysis (Models OA.6.1 and OA.6.2), $Post_{i,t}$ equals 1 after October 10, 2011, for NASDAQ-listed stocks with tickers A-B, and after October 17, 2011, for other NASDAQ stocks. NYSE and Amex stocks serve as control groups in these models. For the Amex speed bump analysis (Models OA.6.3 and OA.6.4), $Post_{i,t}$ equals 1 after July 24, 2017, and $Amex_{i,t}$ equals 1 for Amex-listed stocks. NYSE and NASDAQ stocks serve as control groups in these models. Control variables ($C_{i,t}^{k}$) include daily volatility ($Volatility_{i,t}$, standard deviation of transaction-level returns), relative quoted spread ($Spread_{i,t}$, daily average of (ask-bid)/(0.5×(ask+bid) for each transaction), inverse price ($InvPrice_{i,t}$), and dollar trading volume ($Volume_{i,t}$). The analysis uses 10-working day windows around implementation dates. Panel A reports results for the NASDAQ upgrade and Panel B for the Amex speed bump. Standard errors are double-clustered by stock and day, with t-statistics in brackets. *, **, and *** indicate significance at 10%, 5%, and 1%. R^2 values are within-R^2.

	Panel A: NASDAQ upgrade		Panel B: Amex speed bump	
	(i) $U\_HFT\_D_{i,t}$	(ii) $U\_HFT\_S_{i,t}$	(iii) $U\_HFT\_D_{i,t}$	(iv) $U\_HFT\_S_{i,t}$
$Post_{i,t}$	1.055*** (2.79)	1.347** (2.27)		
$Post_{i,t} * Amex_{i,t}$			-0.977** (-2.27)	-0.696** (-1.98)
Controls	Yes	Yes	Yes	Yes
Stock and Day FE	Yes	Yes	Yes	Yes
N obs.	43,234	43,234	45,530	45,530
R^2	29%	18%	59%	49%

Table OA.D.3
HFT response to latency arbitrage opportunities
This table examines how our ML-generated unscaled HFT measures respond latency arbitrage opportunities using the following OLS models:

$$U\_HFT\_D_{i,t} = \alpha_i + \beta_t + \gamma_1 NLAO_{i,t} + \textstyle\sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t}$$
$$U\_HFT\_S_{i,t} = \alpha_i + \beta_t + \gamma_2 NLAO_{i,t} + \textstyle\sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t}$$

where $U\_HFT\_D_{i,t}$ and $U\_HFT\_S_{i,t}$ represent the ML – generated unscaled liquidity – demanding and – supplying HFT activities for stock i and day t. α_i and β_t capture stock and day fixed effects, respectively. $NLAO_{i,t}$ is the number of latency arbitrage opportunities, identified using the methodology detailed in Section 4.2. Control variables ($C_{i,t}^{k}$) include daily volatility ($Volatility_{i,t}$, standard deviation of transaction-level returns), relative quoted spread ($Spread_{i,t}$, daily average of (ask-bid)/(0.5×(ask+bid) for each transaction), inverse price ($InvPrice_{i,t}$), and dollar trading volume ($Volume_{i,t}$). Columns (i) and (ii) present the results for $U\_HFT\_D_{i,t}$ and $U\_HFT\_S_{i,t}$, respectively. The sample consists of 120 randomly selected NASDAQ- and NYSE-listed firms. Standard errors are double-clustered by stock and day, with t-statistics in brackets. *, **, and *** indicate significance at 10%, 5%, and 1%. R^2 values are within-R^2.

	(i) $U\_HFT\_D_{i,t}$	(ii) $U\_HFT\_S_{i,t}$
$NLAO_{i,t}$	66.266*** (3.21)	-150.518** (-2.04)
Controls	Yes	Yes
Stock and Day FE	Yes	Yes
N obs.	246,139	246,139
R^2	39%	38%

Table OA.D.4

HFT activity and information acquisition − jump ratio

This table examines how HFT activity affects information acquisition using the following OLS model:

$$JUMP_{i,q} = \alpha_i + \beta_{m,q} + \gamma_1 U\_HFT\_D_{i,q} + \gamma_2 U\_HFT\_S_{i,q} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

where $JUMP_{i,q}$ measures information acquisition for stock i as the ratio of cumulative abnormal returns over [-1, 1] to cumulative abnormal returns over [-21, 1] around quarterly earnings announcements (q). $U\_HFT\_D_{i,q}$ and $U\_HFT\_S_{i,q}$ are ML-generated unscaled liquidity-demanding and liquidity-supplying HFT activities, measured as averages of daily values over [-21, -1] around earnings announcements. Models include stock (α_i) and month ($\beta_{m,q}$) fixed effects, respectively. Control variables ($C_{i,q}^{k}$) all measured as averages of daily values over [-21, -1] around earnings announcements, include volatility ($Volatility_{i,q}$), relative quoted spread ($Spread_{i,q}$), market value ($MValue_{i,q}$, price times shares outstanding), and institutional order imbalance ($OIB20k_{i,q}$, price impact of trades over $20,000 from TAQ). The sample includes all U.S.-listed common stocks from 2010 to 2023. Standard errors are double-clustered by stock and quarter, with t-statistics in brackets. *, **, and *** indicate significance at 10%, 5%, and 1%. R^2 values are within-R^2.

	$JUMP_{i,q}$
$U\_HFT\_D_{i,q}$	0.042***
	(9.82)
$U\_HFT\_S_{i,q}$	-0.022***
	(-5.99)
Controls	Yes
Stock and Month FE	Yes
N obs.	49,515
R^2	1%

Table OA.D.5

HFT activity and information acquisition − FERC

This table examines how HFT activity affects information acquisition using the following model:

$$Return_{i,q} = \alpha_i + \beta_q + \sum_{n=-1}^{1}(\gamma_n Earning_{i,q+n} + \vartheta_n Earning_{i,q+n} * U\_HFT\_D_{i,q} + \theta_n Earning_{i,q+n} * U\_HFT\_S_{i,q}) + \rho_1 U\_HFT\_D_{i,q} + \rho_2 U\_HFT\_S_{i,q} + \rho_3 Return_{i,q+1} + \rho_4 Return_{i,q-1} + \sum_{k=1}^{4}\delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,q}$$

where $Return_{i,q}$ is quarterly stock returns for firm i in quarter q, measured as the percentage change in closing prices between quarters $q-1$ and q. $Earning_{i,q+n}$ denotes quarterly earnings (net income) normalized by the market value of equity at the start of quarter $q+n$. The subscript n ranges from -1 to 1. $U\_HFT\_D_{i,q}$ and $U\_HFT\_S_{i,q}$ are ML-generated unscaled liquidity-demanding and liquidity-supplying HFT activities, measured as the quarterly averages of daily values. Control variables ($C_{i,q}^{k}$) all measured as quarterly averages of daily values, include volatility ($Volatility_{i,q}$), relative quoted spread ($Spread_{i,q}$), market value ($MValue_{i,q}$, price times shares outstanding), and institutional order imbalance ($OIB20k_{i,q}$, price impact of trades over \$20,000 from TAQ). The sample includes all U.S.-listed common stocks from 2010 to 2023. Standard errors are double-clustered by stock and quarter, with t-statistics in brackets. *, **, and *** indicate significance at 10%, 5%, and 1%. R^2 values are within-R^2.

	$Return_{i,q}$
$Earning_{i,q+1} * U\_HFT\_D_{i,q}$	-0.003***
	(6.56)
$Earning_{i,q+1} * U\_HFT\_S_{i,q}$	0.003***
	(7.26)
Controls	Yes
Stock and Quarter FE	Yes
N obs.	157,343
R^2	4%