Data-driven measures of high-frequency trading

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 almost half of equity trading volume, yet it is not identified in public data. We develop novel data-driven measures of HFT activity that separate strategies that supply and demand liquidity. We train machine learning models to predict HFT activity observed in a proprietary dataset using concurrent public intraday data. Once trained on the dataset, these models generate HFT measures for the entire U.S. stock universe from 2010 to 2023. Our measures outperform conventional proxies, which struggle to capture HFT's time dynamics. We further validate them using shocks to HFT activity, including latency arbitrage, exchange speed bumps, and data feed upgrades. Finally, our measures reveal how HFT affects fundamental information acquisition. Liquidity-supplying HFTs improve price discovery around earnings announcements while liquidity-demanding strategies impede it.

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

* Corresponding author. Emails: gbenga.ibikunle@ed.ac.uk (Gbenga Ibikunle), ben.moews@ed.ac.uk (Ben Moews), dmuravy2@illinois.edu (Dmitriy Muravyev), khaladdin.rzayev@ed.ac.uk (Khaladdin Rzayev).
We thank Phil Mackintosh and Heinrich Lutjens at NASDAQ for providing data. The data is made available at no cost to academics following the provision of a project description and the signing of a nondisclosure agreement.

1. Introduction

High-frequency trading (HFT) firms now execute nearly half of U.S. equity trading volume, processing millions of orders in microseconds via automated algorithms (e.g., surveys by Jones 2013; Menkveld 2016). Their dominance has sparked extensive research into their market impact, with a crucial distinction between strategies that take versus provide liquidity. Most HFTs operate as market makers, leveraging their speed advantage to provide liquidity and reduce adverse selection risk, lowering trading costs and enhancing market liquidity (e.g., Hendershott et al. 2011; Menkveld 2013; Brogaard et al. 2015). However, other HFTs aggressively consume liquidity, which can amplify adverse selection costs and exacerbate price movements (e.g., Easley et al. 2011; Biais et al. 2015; Foucault et al. 2017).

Measuring HFTs' activity is challenging because standard market feeds do not identify it. Researchers have pursued two approaches, each with important limitations. Some studies use private datasets with explicit HFT flags, 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 do not distinguish between liquidity-demanding and liquidity-supplying HFT strategies (Boehmer et al. 2018; Chakrabarty et al. 2023), and as we show below, they mainly reflect cross-stock rather than time-series variation in HFT activity.

We introduce a data-driven approach to measure both liquidity-supplying and liquidity-demanding HFT activity using machine learning (ML) techniques. We call these measures *HFT-S* and *HFT-D*. Our method combines a proprietary dataset of directly observed HFT activity with concurrent public intraday data. Specifically, we train ensemble models to predict

[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.

NASDAQ's HFT activity using 24 measures of trading activity, liquidity, and volatility from WRDS Intraday Indicators for the same stock and day. Consistent with Easley et al. (2021) and Bogousslavsky et al. (2023), the ML approach captures complex non-linear relations and variable interactions in HFT behavior. Once the models are trained out-of-sample on a relatively small NASDAQ HFT dataset, we use them to generate HFT activity measures for the entire TAQ universe covering 8,314 U.S. stocks from 2010 to 2023. Thus, this approach is limited only by TAQ coverage.

We evaluate our HFT-S and HFT-D measures against five conventional HFT proxies, including the quote-to-trade ratio, quote midpoint volatility, odd-lot volume, quoted price and depth changes, and the number of trade and quote messages. Notably, quote data and these measures were not among the 24 intraday variables used to train our measures. Using NASDAQ HFT data from January to June 2009 for training and July to December 2009 for out-of-sample validation, we find two key results. First, while conventional measures are significant predictors of HFT activity in isolation, both HFT-S and HFT-D subsume their information content in joint regressions. Second, conventional measures effectively capture cross-stock variation but struggle to explain time variation in HFT activity, whereas our measures remain significant in both dimensions. Thus, these findings are consistent with our measures offering a more comprehensive and robust framework for capturing HFT activity.

We extensively validate our HFT measures outside the original training sample. Our first key test examines a speed bump that NYSE Amex introduced in 2017 (Khapko and Zoican 2021; Aït-Sahalia and Sağlam 2024), which adds intentional delays hindering fast trading. This test is particularly valuable as it occurred well after our training period, helping validate our measures' temporal stability. The speed bump led to a substantial decline in HFT activity on both liquidity supply and demand sides. In our second natural experiment, NASDAQ's increased its data feed speed in 2011 (Ye et al. 2013), benefiting HFT strategies. Indeed, both

HFT-S and HFT-D increase but naturally less than for the speed bump shock. Further validation comes from examining latency arbitrage—when speed disparities among traders reacting to public information create profitable opportunities (e.g., Budish et al. 2015). Our measures capture theoretically predicted behaviors: a one standard deviation increase in arbitrage opportunities leads to a 1% rise in HFT-D activity (as fast traders race to pick off stale quotes) while reducing liquidity-supplying activity by 1.6% (as market makers withdraw to avoid being picked off), consistent with prior evidence (Foucault et al. 2017; Aquilina et al. 2022).

Our HFT measures cover all U.S. stocks from 2010 to 2023, enabling applications that require broad market coverage. We focus on one such application by examining how different HFT strategies affect fundamental information acquisition. Price discovery—acquiring and incorporating new fundamental information—remains a core function of financial markets. Our ability to distinguish between HFT strategies allows us to test competing theories: whether HFTs enhance information acquisition by providing liquidity and reducing costs (e.g., Menkveld 2013; Stiglitz 2014; Brogaard et al. 2015; Aït-Sahalia and Sağlam 2024), or impair it by adversely selecting informed investors (e.g., Van Kervel and Menkveld 2019; Yang and Zhu 2020; Hirschey 2021).

In this test, we follow Weller's (2018) methodology of measuring price informativeness around earnings announcements.[2] While Weller (2018) finds that algorithmic trading reduces price informativeness, we find that liquidity-supplying HFTs *enhance* information acquisition, while liquidity-demanding strategies *impede* it.[3] This divergence explains Weller's (2018) results, as his proxies—quote-to-trade ratio and odd-lot volume—primarily capture HFTs'

[2] Weller computes the "price jump ratio" as a measure of relative information acquisition by dividing the return on earnings announcement by the total return over the pre-announcement period. A high price jump ratio indicates that information was not discovered until publicly revealed, suggesting low pre-announcement information acquisition. Unlike absolute measures like cumulative abnormal returns, this ratio captures how much information enters prices early relative to potentially acquirable information.

[3] We also employ an additional measure of information acquisition, the future earnings response coefficient (e.g., Lundholm and Myers 2002), and obtain results consistent with Weller's (2018) price jump ratio.

liquidity demand in his sample. Finally, datasets with directly observed HFT trading are too small for this application; for example, the NASDAQ HFT dataset only contains several hundred earnings announcements.

We further validate our HFT measures through multiple complementary tests. First, we document theoretically consistent nonlinear relationships between our HFT measure and intraday variables. For example, HFT-D responds strongly to intermarket sweep orders (consistent with Klein 2020) and is decreasing and convex in market depth, while HFT-S exhibits an increasing, concave pattern (e.g., Goldstein et al. 2023). Second, both HFT types increase around scheduled and unscheduled news events, with liquidity suppliers showing larger responses—consistent with their role in information processing. Our measures also capture these relationships more effectively than traditional linear regressions, highlighting the advantage of ML. Finally, adding granular quote-level features only marginally improves model performance, and our results hold for alternative volume scaling approaches.

A potential limitation of our approach is its reliance on NASDAQ's 2009 HFT dataset for training. While more recent proprietary data would be ideal, several factors mitigate this concern. First, fundamental HFT strategies have remained relatively stable despite technological advances, as evidenced by consistent patterns in market-making and directional trading (Brogaard et al. 2014; Malceniece et al. 2019).[4] Second, our natural experiments demonstrate that our measures capture meaningful variation in HFT activity both near and far from the training period—particularly the introduction of the 2017 NYSE Amex speed bump. Third, HFT strategies exhibit similar patterns across venues and firms, suggesting that training on NASDAQ data generalizes to the broader market. Finally, while conventional HFT

[4] For instance, the features of HFT strategies developed in more recent theories (Li et al. 2021a) are similar to those outlined a decade ago (e.g., Biais et al. 2015; Foucault et al. 2017), suggesting continuity in these core approaches. Additionally, many recent empirical papers still rely on datasets from 2009-2012 when investigating the distinct roles of liquidity-demanding and supplying HFT strategies (e.g., Boehmer et al. 2018; Goldstein et al. 2023; Nimalendran et al. 2024), confirming the ongoing relevance of the NASDAQ HFT data.

measures face similar data vintage limitations, our approach offers the advantage of capturing both cross-sectional and time-series variation in HFT activity.

Our study advances the HFT literature in two key dimensions. First, we develop novel measures that separately capture liquidity-demanding and liquidity-supplying HFT strategies using ML techniques. While previous research has shown that public HFT proxies can reflect both types of strategies (Boehmer et al. 2018; Chakrabarty et al. 2023), our approach generates separate measures for each, enabling a more targeted analysis of HFT behavior. This distinction proves crucial for understanding HFT's market impact—for instance, our finding that Weller's (2018) documented negative relationship between HFT and information acquisition stems from liquidity-demanding strategies reconciles seemingly contradictory results in the literature. Moreover, our methodology creates an open-access HFT dataset covering the entire U.S. equity market from 2010 to 2023, allowing researchers to investigate the long-term effects of different HFT strategies and test theoretical predictions.

Our work also contributes to the growing application of ML 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). Given that HFT firms now execute nearly half of U.S. equity trading volume, developing reliable measures of their activity is crucial for understanding modern markets. Our ML approach shows that complex trading patterns can be effectively captured through public data. Moreover, by revealing how HFTs respond differently to public versus private information compared to traditional informed traders, our measures offer new insights into price discovery.

2. Data and variables

We use two primary datasets. The first is the NASDAQ-provided dataset that labels HFT and non-HFT transactions for 120 randomly selected stocks listed on NASDAQ and

NYSE in 2009. In this dataset, NASDAQ classifies transactions into those executed by HFTs and non-HFTs (e.g., Brogaard et al. 2014), and provides detailed information such as the date and time (to the millisecond), volume, price, direction, and the liquidity profile of each trade, 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). The second primary dataset is obtained from the TAQ's Intraday Indicators for the same period and contains 24 variables identified in the relevant literature as associated with HFT activity. The variables include various measures based on aspects such as price, trading volume, trading costs, liquidity, volatility, and the dynamics of retail and institutional trading. The list of these variables and their detailed descriptions are provided in Table 1.

We employ these two datasets in our ML model to generate a secondary dataset, which estimates HFT activity from publicly available TAQ data, spanning January 4th 2010 and October 18th 2023, based on training enabled by the proprietary NASDAQ dataset. The main output variables of our ML model are the fractions of trading volume attributed to liquidity-demanding ($NHFT_{i,t}^{D}$) and liquidity-supplying ($NHFT_{i,t}^{S}$) HFTs. Specifically, $NHFT_{i,t}^{D}$ ($NHFT_{i,t}^{S}$) 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*. Our ML model is presented in Section 3 below.

INSERT TABLE 1 HERE

To validate our ML-generated HFT measures, we obtain multiple complementary datasets. We calculate commonly used HFT proxies using quote-level data from the Millisecond TAQ database and benchmark ML-generated 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 collected from I/B/E/S and the Thomson Reuters Securities Data

Company (SDC) database, respectively. Stock returns and trading volume data are sourced from the Center for Research in Security Prices (CRSP).

To jointly test the empirical relevance of ML-generated HFT metrics and the association between HFT and various market quality measures, we estimate different regression models as specified in subsequent sections. The main and control variables employed in these models are also introduced within their corresponding sections. Definitions and summary statistics for these variables, along with the summary statistics for the ML-generated HFT measures, are presented in Table 2.

INSERT TABLE 2 HERE

The mean values for ML-generated liquidity-demanding ($HFT_{i,t}^{ML,D}$) and -supplying HFT ($HFT_{i,t}^{ML,S}$) activity stand at 0.316 and 0.208, respectively, and the difference is statistically significant at the 0.01 level. This indicates a predominance of demand over supply within the observed sample. The standard deviation indicates a non-negligible level of variability in HFT activity, with demand showing slightly more variability (0.112) than supply (0.101). Furthermore, the comparison of mean and median values shows that $HFT_{i,t}^{ML,S}$ is right-skewed while $HFT_{i,t}^{ML,D}$ is left-skewed. $Spread_{i,t}$ shows a mean of 0.142% with a wide range up to 0.886%, implying diverse liquidity conditions across the sampled stocks. This is to be expected since our sample includes 8,314 stocks—essentially the universe of US stocks available in the TAQ database. $Volume_{i,t}$ has a high degree of variability (mean: 2.614, max: 47.392).

3. Machine learning and high-frequency trading measures

3.1. Modeling and experimental choices

Our ML methodology exploits ensemble learning, with an ensemble in supervised ML being a finite set of predictive models, often of the same type, used to generate outputs for a

desired set of dependent variables. The main reason for this approach is the ability to build a collective predictor that is stronger than its constituent parts, which are correspondingly referred to as "weak learners." This usually results in a better generalization when predicting data not previously seen by the model, meaning an improved performance for out-of-sample testing (see Bishop and Nasrabadi 2006, for a general overview). Models used in such ensembles are generally less complex when compared to similarly powerful single-model approaches. Coupled with their strong generalization performance, they have come to enjoy a broad adoption in the literature applying ML to non-linear problems in finance—and, indeed, many other research areas (Parker 2013; Moews et al. 2021; Cao 2022).

Our ensemble features decision trees (Breiman et al. 1984), one of the best-established supervised ML models, and the random forest model. Introduced by Ho (1995), the random forest model and its derivatives are one of the earliest ensemble learning methods that remain popular across research fields (Wu et al. 2008), including in financial economics (e.g., Easley et al. 2021; Bogousslavsky et al. 2023). Recent derivates are extremely randomized trees, generally abbreviated as "extra trees" (Geurts et al. 2006). In conducting our experiments, we implement the common mean squared error as the splitting criterion, meaning that for the true values of independent variables Y and corresponding predictions $\hat{Y}$ for a dataset of size n,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad (1),$$

In the case of extra trees, this translates to variance reduction as the selection criterion. Using independent (input) and dependent (output) variables as listed in Table 2, we construct each ensemble in our experiments with these 24 inputs and analyze two targets.

Other relevant model choices in our experimental setup are largely informed by computational feasibility. This concerns, most importantly, two parameters, the number of experiment repetitions to gauge consistency through an approximated standard deviation and the number of data points used per experiment. The former is set to 10 to allow for reasonable

runtimes, whereas 10,000 data points are used as a size more than sufficient for the type of model used. The degree of simplicity of the constituent models is an advantage over, for example, various deep learning approaches (Genuer et al. 2017). To make use of the full trade data available, we select samples in a uniform-random manner for each experimental iteration and split off 25% as the testing set. We employ Monte Carlo cross-validation, which is an attractive option choice, in comparison to e.g., k-fold cross-validation.[5] In each experiment repetition, multiple random splits into training and testing sets are performed as uniform-random samples from the full dataset. In doing so, the size and split percentage for these subsets can be chosen freely, with a lower variance at the cost of higher bias. The results exhibit Monte Carlo variation across multiple runs and, in the limit, the results become that of exhaustive cross-validation (e.g., Li et al. 2010).

Other parameter choices we make are less clear-cut and thus require optimization. This concerns the number of models per ensemble and the minimum number of samples for node splits. We employ a grid-based optimization approach, with 8 options each for a total of 64 experiments with different parameter combinations, and with 10 experiment repetitions each. Each experiment uses a tuple of values from $\{5, 10, 20, 40, 80, 160, 320, 640\}$ in a grid-based optimization approach. More complex alternatives for parameter optimization exist but are not warranted in this case. While a larger number of trees and a smaller number of samples per node split are often the optimal choice, this is primarily done as a precaution against challenges such as lack of generalizability for small node split values in some instances (Probst and Boulesteix 2018). Results of these experiments are provided in Table 3, in which we use the

[5] While the commonly used cross-validation approach in ML is k-fold cross-validation, which involves splitting the data into k subsamples followed by training on all except one of these samples, and swapping the subsample used as the test set each time until averages can be calculated for k iterations (Hastie et al. 2009). However, the benefit of using the entire data in the process is also the main drawback in the case of very large datasets. The computational complexity of a decision tree with the number of independent variables and tree depth being held constant is $\mathcal{O}(n \log(n))$, n denoting the number of entries in the training data. While the randomization component in extra trees alleviates some of that issue, the number of trees in an ensemble then acts as a further multiplicative factor.

arithmetic mean and standard deviation of R^2 across repeated iterations to assess the respective model's quality.

INSERT TABLE 3 HERE

Unsurprisingly, a larger number of trees with finer node splits until fewer samples per split are left generally correspond to better results for out-of-sample generalization with high accuracy. This preference is the clearest for the latter, with all five top and bottom results using the lowest and highest option, respectively. The standard deviations of the calculated R^2 values demonstrate the consistency of the model's performance with randomly sampled subsets of the data. We lock these parameter choices in subsequent experiments to these values and also retain 10 iterations per experiment going forward.

3.2. Comparison to related machine learning predictors.

While the described ensemble learning approaches are particularly suitable due to their accessibility, it is prudent to contrast the outcomes of our modeling and experimental choices with competing options. As we deal with a regression instead of a classification problem, potentially suitable machine learning models commonly used for similar prediction problems include random forests as the baseline for tree-based ensembles, support vector machines (SVM) in their regressor variation and feed-forward neural (FNN) networks with multiple hidden layers, the latter under the umbrella of "deep learning". Thus, also use SVM and FFN on a comparative basis, using best-practice parameters with a fast-enough computation Specifically, we implement SVM with a radial basis function kernel, provide the tree-based ensembles with 50 estimators and minimum node split samples, and built the artificial neural network with three hidden layers using rectified linear units and mean absolute error optimization. Employing both SVM and FNN means that we can benchmark the likely candidate models against a simpler approach, i.e., SVM, to gauge the difficulty of the prediction task, and a complex comparison model, i.e., FNN.

Our dataset for 2009 contains 29,880 stock-day data points, of which we drop 2,184, or around 7%, due to missing values in one or multiple of the independent and dependent variables. This is an acceptable loss, as the alternative of interpolation or imputation approaches for high-frequency trading information are inherently risky due to the assumptions that would need to be made. We employ scaling to avoid variability in values putting undue weight on some of observations over others. In contrast to the later results, these initial experiments apply z-score scaling, also commonly called standardization, in which, for a dataset, D,

$$z_{D_i} = \frac{D_i - \overline{D}}{\sigma(D)} \quad (2),$$

We choose this scaling method as opposed to min-max scaling, which is also known as normalization, due to the latter's sensitivity to outliers. We then test both multi-model (each model predicting one target variable) and multi-target (one model predicting points in the complete target space) frameworks if applicable, in this case for the tree-based ensembles. The former is only advisable in cases in which a multi-target approach does not perform well enough, as the interconnectivity between different dependent variables is lost.

While SVM can be used for regression, this is limited to the multi-model approach by default. Feed-forward artificial neural networks can handle both cases, but the complexity of these models would not benefit from simplifying the prediction. The result of this comparison is listed in Table 4. Specifically, Table 4 presents arithmetic mean and standard deviation estimates for R^2 values across 10 iterations for support vector regression/machines, feed-forward artificial neural networks, and random forests as well as extra trees for multi-model and multi-target setups.

INSERT TABLE 4 HERE

While the results by themselves are promising, SVM notably underperforms the alternatives, while the extra trees approach provides the highest mean performance and, aside from the artificial neural network, the lowest standard deviation estimate. Although the

universal approximation theorem concerns the predictability of arbitrary functions under minimal assumptions, this does not ensure the learnability of the necessary weights, which is a major challenge in the related literature (Zhang et al. 2017). We thus opt for extra trees both as the stronger average predictor and given the consideration that highly complex models should only be used when simpler ones do not suffice. This also allows for improved parameter optimization with reasonable resource spending.

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

The final experiment is implemented with the optimized parameters as described in the previous section, which across these multiple runs results in an average R^2 value of 0.824635, with a standard deviation of 0.005472. The application of z-score scaling is no longer necessary, as node splits in decision trees are not negatively affected by unscaled inputs. For this reason, the standard deviation is no longer directly comparable to the results in Section 3.2. A comparison to a prior non-optimized but unscaled implementation finds that, aside from an improved goodness of fit, the standard deviation is approximately halved through our optimization.

We then use this model, i.e., extra trees with multi-target, to extrapolate to all U.S. stocks obtained from the TAQ database as described in Section 2. The data covers an approximately 13-year period from January 4th 2010 to October 18th 2023, corresponding to a total of 9,440,600 non-missing stock-day observations for each of the 24 input variables listed in Table 1. All dependent variables are then predicted for the entirety of the above-mentioned data, leading to the creation of an ML-generated HFT dataset with 9,440,600 stock-day observations. These observations constitute the secondary dataset employed for subsequent analysis in subsequent sections.

3.4.Properties of machine learning-generated HFT measures

A key strength of ML over traditional predictive models lies in its ability to capture the nonlinearity between input and output variables. This aspect is crucial for our study, given the nonlinear nature of the relationship between HFT and market quality characteristics. For instance, Foucault et al. (2017) show how HFT arbitrage strategies might either enhance or impair liquidity, contingent on the nature of latency arbitrage opportunities (e.g., Rzayev et al. 2023). Consequently, ML emerges as an optimal approach to model HFT activity in financial markets given its adeptness at navigating the complex, nonlinear interdependencies inherent in market dynamics.

In this section, to determine if our ML modeling framework captures nonlinear interactions between HFT activity and its predictors, we analyze partial dependence plots. 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 these interactions.

INSERT FIGURE 1 HERE

Figure 1 demonstrates that most of our selected input variables significantly influence HFT activity predictions. Key among these are the number and value of trades, intermarket sweep orders (ISOs), and measures of market depth. The importance of trading volume and market depth for HFTs is intuitive: HFTs require counterparts for transactions, making volume a crucial factor. Similarly, market depth, indicative of liquidity and trading availability, is essential for HFT activities. However, the significance of ISOs predicting HFT activity is noteworthy. This finding aligns with the broader concerns in financial markets about ISOs. Originally intended for large institutional traders, ISOs are now believed to be increasingly exploited by HFTs to gain an advantage over slower market participants.[6] Supporting this, Li

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

et al. (2021b) find that ISO order sizes are generally smaller than those of traditional institutional traders and are often employed by fast traders. Our findings corroborate these observations, highlighting the potential use of ISOs by HFTs.

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 through partial dependence plot. As evident in Figure 2, the association between HFT activity and various input variables are indeed nonlinear. For instance, liquidity-demanding and -supplying HFT activity both demonstrate an increasing, yet concave, relationship with the total number of trades. This positive correlation with trading volume is expected, as HFTs are more active when trading volumes are high. This is consistent with Brogaard et al. (2014), who shows that HFTs favor trading in larger stocks, which tend to be more liquid.

A particularly compelling pattern emerges when examining the interplay between HFT metrics and ISOs, as well as market depth. The fraction of liquidity-demanding HFT activity exhibits a pronounced initial increase with ISOs, characterized by a concave curve, highlighting a significant initial influence of ISOs on liquidity-demanding HFT activity. Conversely, the relationship between liquidity-supplying HFT activity and ISOs is relatively flat, showing only a marginal rise in the HFT supply fraction as the dollar amount of ISOs increases, suggesting a lesser impact. This differential sensitivity of liquidity-demanding versus liquidity-supplying HFT activities to ISOs aligns with existing academic findings. Li et al. (2021b) demonstrate that HFTs often employ ISOs to target stale quotes, a tactic predominantly associated with liquidity-demanding strategies. Furthermore, Klein (2020) suggests that aggressive HFT strategies involve using ISOs upon the arrival of new information. A competing view is that the relationship between liquidity-demanding HFT activity and ISOs is reflective of the response of HFTs to institutional traders using ISOs to

avoid being front-run by HFTs. This is because, as noted by Chakravarty et al. (2012), the ISO exemption to Rule 611/Order Protection Rule of Reg NMS was adopted to allow institutional investors timely access to liquidity (at multiple price levels) needed to execute large block orders through the parallel submission of orders across multiple trading platforms.

The dynamics between market depth and both liquidity-demanding and -supplying HFT activities also present interesting insights. Liquidity-supplying HFT activity shows an increasing and concave relationship with market depth, suggesting that HFTs are more inclined to provide liquidity as the order book deepens. On the contrary, liquidity-demanding HFT activity demonstrates a decreasing and convex pattern with market depth, indicating a reduced tendency to demand liquidity in deep markets. This observation is in line with the findings of Goldstein et al. (2023), who show that HFTs tend to supply liquidity in deeper markets (where the order book is thick) and demand liquidity in shallower markets (where the order book is thin).

The findings from this section lead to two key implications. First, the nonlinear relationship between HFT activity and market quality underscores the necessity of ML models for forecasting HFT activity. Second, the distinct patterns observed in the relationship between market quality indicators and HFT strategies—varying across liquidity-demanding and -supplying activities—align well with existing debates in the literature. This alignment confirms the empirical relevance of our ML-derived HFT demand and supply metrics in capturing the nuanced strategies of HFTs. Below, we offer validating evidence on the relevance these ML-generated HFT metrics and examine their empirical significance in detail.

4. Testing the properties of ML-generated HFT.

4.1.HFT ahead of scheduled and unscheduled information events.

To test the empirical validity of the ML-generated HFT measures, we commence with an exploration of the dynamics of liquidity-demanding ($HFT_{i,t}^{ML,D}$) and liquidity-supplying

($HFT_{i,t}^{ML,S}$) HFT activity during both scheduled and unscheduled information events. As argued in Foucault (2016), one of the primary characteristics of HFTs is their rapid response to major information events (see also Brogaard et al. 2014). This characteristic forms the basis of latency arbitrage, a phenomenon that encapsulates the purpose of liquidity-demanding HFT activity (Aquilina et al. 2022), and liquidity-supplying market maker quote updates that typically follows the emergence of latency arbitrage opportunities (Boehmer et al. 2018; Rzayev et al. 2023). Thus, examining the behavior of the ML-generated HFT measures around information events is a logical first step in assessing the empirical relevance of $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$.

INSERT FIGURE 3 HERE

We first focus on earnings announcements as scheduled events. Panels A and B of Figure 3 show the dynamics of $HFT_{i,t}^{ML,S}$ and $HFT_{i,t}^{ML,D}$ surrounding earnings announcements, with 95% confidence intervals plotted over a 20-day window spanning ten days before and after the announcement dates. Both HFT measures display an increasing pattern starting three days before the announcement, reaching their maximum on the announcement day. To quantify the announcement effects, we contrast the average HFT activity during a three-day announcement window (days *t*, *t+1*, and *t+2*) with pre-announcement levels. The three-day period is chosen in line with previous research that investigates the short-term effects of earnings announcements (e.g., Ball and Shivakumar 2008). Our results show statistically significant increases in both HFT measures during the announcement window: $HFT_{i,t}^{ML,S}$ rises by 6.3% (from 0.208 to 0.221) and $HFT_{i,t}^{ML,D}$ increases by 2.8%.

Complementing our analysis of scheduled events, we examine HFT behavior around unscheduled M&A announcements, which may contain higher information content than earnings announcements (e.g., Bogousslavsky et al. 2023). This analysis also provides additional insights because HFTs predominantly engage in latency arbitrage strategies—rapidly processing public information—rather than exploiting private information (e.g., Budish

et al. 2015; Aquilina et al. 2022). This indicates that, unlike earnings announcements where we observe increased activity three days before the event, the unscheduled nature of M&A announcements may limit the availability of exploitable information before the event, which can make M&A announcements potentially less conducive to HFT strategies in the pre-announcement period.

INSERT FIGURE 4 HERE

Panels A and B of Figure 4 illustrate the dynamics of $HFT_{i,t}^{ML,S}$ and $HFT_{i,t}^{ML,D}$ around M&A announcements. Consistent with our expectations, $HFT_{i,t}^{ML,S}$ begins to increase just one day before the announcement—in contrast to three days for earnings announcements—while $HFT_{i,t}^{ML,D}$ rises only from the announcement day. This pattern differs from the informed trading intensity (ITI) measure of Bogousslavsky et al. (2023), which increases approximately three days before unscheduled events. While our HFT metrics may share some characteristics with ITI, their distinct behavior around unscheduled events provides important differentiation. The divergence suggests that while informed traders exploit private information before unscheduled events, HFTs primarily trade on public information after such announcements, consistent with prior literature (e.g., Rzayev and Ibikunle 2019).

We next compare the average $HFT_{i,t}^{ML,S}$ and $HFT_{i,t}^{ML,D}$ during the three-day announcement window with their pre-announcement levels. During this window, $HFT_{i,t}^{ML,S}$ increases by 3.2% and $HFT_{i,t}^{ML,D}$ rises by 1.2%, with both increases being statistically significant at the 1% level. Thus, liquidity-supplying HFT activities ($HFT_{i,t}^{ML,S}$) show a substantially larger increase than liquidity-demanding activities ($HFT_{i,t}^{ML,D}$) during information events, with the former being twice the magnitude of the latter. This pattern further validates our ML-based methodology in distinguishing between liquidity demand and supply dynamics, aligning with existing literature. Specifically, Brogaard et al. (2014) document that around

macroeconomic news announcements, liquidity-supplying HFTs' order flow increases more significantly than liquidity-demanding HFTs' flow, regardless of news sentiment. This asymmetry likely stems from non-HFTs intensifying their aggressive trading during unscheduled information events, with liquidity-supplying HFTs, who have become market makers in today's markets (e.g., Menkveld 2013), accommodating this increased demand. Cole et al. (2015) provide supporting evidence, showing that during earnings announcements, non-HFTs' aggressive trading increases more than HFTs', while liquidity-supplying HFTs consistently meet this elevated non-HFT demand throughout the announcement periods.

Beyond demonstrating the empirical validity of our ML-generated HFT metrics, the result in this section also carries significant economic implications. The larger increase in liquidity-supplying HFT activity highlights the flexible nature of HFT strategies under changing market conditions, particularly during times of heightened information flow. This demonstrates that HFTs are more than just aggressive arbitrageurs in high-information environments; they are key to preserving market liquidity (e.g., Hagströmer and Nordén 2013), particularly when non-HFT participants may intensify their trading in reaction to new information. These results are consistent with the literature that highlights HFTs' contribution to market efficiency and resilience during periods of significant information release (e.g., Brogaard et al. 2018). While the existing literature has already shown these trends primarily using the NASDAQ HFT dataset, limited to 120 stocks, or through other proprietary datasets with very short durations and limited samples, our study extends the insights by examining all U.S. listed common stocks over a broader thirteen-year timeframe using publicly available datasets.

4.2.HFT during exogenous technological changes.

We now examine how $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ respond to exogenous shocks affecting HFT activity through two natural experiments. The first experiment is a NASDAQ-

implemented technology enhancement that reduces trading data dissemination latency from 3 milliseconds to 1 millisecond (e.g., Ye et al. 2013). The second experiment is Amex's implementation of a symmetric speed bump, which imposes equal speed restrictions on both liquidity-demanding and liquidity-supplying HFT activity (e.g., Khapko and Zoican 2021; Aït-Sahalia and Sağlam 2024). The principle is straightforward: if our metrics capture HFT activity, they should show significant responses to these HFT-specific market structure changes.

On October 10, 2011, NASDAQ initiated a technological upgrade that reduces trading data dissemination latency from 3 milliseconds to 1 millisecond. This enhancement was implemented in stages: stocks with ticker symbols beginning with A and B were upgraded on October 10, while the remaining stocks were upgraded on October 17. Ye et al. (2013) employ this staggered implementation to study HFT's impact on market quality. Similarly, this phased technological enhancement provides an ideal setting to examine the causal relationship between our ML-generated HFT measures and technological changes. Given that the upgrade reduces trading latency, we expect it to be related to an increase in HFT activity. We test this hypothesis using the following stock-day regressions:

$$HFT_{i,t}^{ML,D} = \alpha_i + \beta_t + \gamma_1 Post_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (3);$$

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

where $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ are ML-generated liquidity-demanding and -supplying HFT activity, respectively. We include stock (α_i) and day (β_t) fixed effects to 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 short

estimation windows to capture the effect; specifically, we use a 10-working day window around the implementation dates.

$C_{i,t}^{k}$ 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.

Our second natural experiment is the speed bump introduced by Amex. In January 2017, the Amex filed 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 approved 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 ML-generated 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:

$$HFT_{i,t}^{ML,D} = \alpha_i + \beta_t + \gamma_1 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (5);$$

$$HFT_{i,t}^{ML,S} = \alpha_i + \beta_t + \gamma_2 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (6),$$

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 our results, we need to provide an important methodological clarification. Our HFT measures ($HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$) 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, late 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.[7] 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 5 HERE

Table 5 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.

Investigating the economic magnitudes of these shocks can provide additional validation of our measures. The Amex speed bump represents a stronger shock to HFT activity through its direct impact on trading speed. In contrast, NASDAQ's improvement in trading

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

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, but the impact is relatively modest. Our results support this distinction. Following the speed bump implementation, Amex-listed stocks experience decreases of 2.8% and 4.6% in $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$, 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 $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ for NASDAQ-listed stocks, respectively, relative to their pre-upgrade averages.

Overall, the results in this section have three implications. First, our ML-generated HFT metrics effectively capture HFT activity, validated by their response to technological 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 implementations 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 wider quoted spreads and reduced liquidity following Amex's speed bump implementation. 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 and Aït-Sahalia and Sağlam (2024) provides evidence that our liquidity-demanding and liquidity-supplying HFT metrics effectively capture supply and demand dynamics, we formally investigate this in the next section.

4.3.HFT and latency arbitrage opportunities.

Our analyses provide preliminary evidence that our measures capture the distinct characteristics of liquidity-demanding and -supplying HFT strategies. For example, we consistently observe larger changes in $HFT_{i,t}^{ML,S}$ compared to $HFT_{i,t}^{ML,D}$ around both informational events and technological changes. This pattern aligns with existing literature in two ways. First, it reflects HFTs' tendency to act as net liquidity providers during high information periods (e.g., Brogaard et al. 2014). Second, it corresponds to findings that speed bump implementation leads to wider quoted spreads due to its stronger impact on liquidity-supplying HFT activity (e.g., Aït-Sahalia and Sağlam 2024).

To further validate this insight, we turn to the concept of "latency arbitrage." 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 the majority of latency arbitrage scenarios, a significant portion of HFT activity is characterized by aggressive liquidity-taking behaviors (see also Aquilina et al. 2024). This is attributed to latency arbitrage opportunities making aggressive HFT strategies more profitable, thereby encouraging HFTs to engage more in such strategies (e.g., Baldauf and Mollner 2020). Therefore, we suggest that latency arbitrage events offer a context to distinguish between the specific characteristics of liquidity-demanding and -supplying HFT activity. In particular, in the wake of latency arbitrage opportunities, we expect an increase in liquidity-demanding HFT activity, in line with predictions by Baldauf and Mollner (2020) and Aquilina

et al. (2022). A consequence of this increase in aggressive trading and sniping activity is the increased risk of the imposition of adverse election on endogenous liquidity-supplying HFTs; hence, liquidity-supplying HFT transactions are expected to decline (e.g., Foucault et al. 2017; Menkveld and Zoican 2017).

To formally test this hypothesis, we estimate the following stock-day models:

$$HFT_{i,t}^{ML,D} = \alpha_i + \beta_t + \gamma_1 NLAO_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (7);$$

$$HFT_{i,t}^{ML,S} = \alpha_i + \beta_t + \gamma_2 NLAO_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (8),$$

where $NLAO_{i,t}$ is the number of latency arbitrage opportunities. We identify latency arbitrage opportunities following the arguments of Budish et al. (2015), which suggests examining the magnitude of changes in mid-prices to identify 'stale' quotes. Specifically, a quote at time $z-1$ is stale if the absolute difference in mid-price from time $z-1$ to z is greater than the half spread. Building upon this concept, we adopt a more conservative methodology by calculating the jump size based on the difference between the mid-price at time z and the ask and bid quotes at time $z-1$. Mathematically, if $Midprice_z > (Ask_{z-1} + TickSize)$, where $TickSize$ is set to 0.01\$, it suggests the existence of a profitable latency arbitrage opportunity. Under such circumstances, HFTs can leverage this opportunity by placing a limit buy order at $Ask_{z-1} + TickSize$ at time z. Similarly, if $Midprice_z > (Bid_{z-1} - TickSize)$, HFTs can capitalize on this arbitrage opportunity by submitting a limit sell order at $Bid_{z-1} - TickSize$ at time z. All other variables and notations are as previously defined.

We use the first-level quote data obtained from Refinitiv DataScope to identify latency arbitrage opportunities. The primary challenge in this process is the substantial volume of data required, which imposes a prohibitive computational cost for an analysis including the 8,314 stocks in our sample. Therefore, we narrow our focus to the sample of 120 firms included in the original NASDAQ HFT data and used in training our ML model. We calculate $NLAO_{i,t}$ for

these 120 firms across our entire sample period, spanning 2010 and 2023. As reported in Table 2, the average number of latency arbitrage opportunities per stock-day is 68. The standard deviation is high, at 169, and the maximum value is 1211, indicating considerable volatility in the occurrence of these opportunities across stocks and days.

INSERT TABLE 6 HERE

The results, as presented in Table 6, show a positive and statistically significant (at the 0.01 level) relationship between $HFT_{i,t}^{ML,D}$ and $NLAO_{i,t}$, whereas the relationship between $HFT_{i,t}^{ML,S}$ and $NLAO_{i,t}$ is negative and statistically significant (at the 0.05 level). The magnitude of the relationship between $HFT_{i,t}^{ML,D}/HFT_{i,t}^{ML,S}$ and $NLAO_{i,t}$ is also economically meaningful. A one-standard-deviation increase in $NLAO_{i,t}$ (169) is associated with a 1% rise in $HFT_{i,t}^{ML,D}$ and 1.6% decrease in $HFT_{i,t}^{ML,S}$.

While we refrain from claiming causality in Models (7) and (8), as it is not the primary objective of estimating them, our results indicate that the relationships between latency arbitrage and various HFT strategies are consistent with the existing body of research. The 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). The alignment of our findings with those of established theoretical and empirical studies highlights the empirical validity of $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ in capturing the liquidity-demanding and -supplying activities of HFTs.

Although our primary focus is not on investigating the impacts of aggressive HFTs and latency arbitrage on financial markets, it is essential to discuss the interesting dynamics of their interplay. The rise in aggressive HFT activity, driven by latency arbitrage, contributes to the technological arms race and its associated costs (Aquilina et al. 2022). However, this process

may not be universally negative for market quality. Indeed, the presence of aggressive HFTs can enhance price efficiency. This occurs as these HFTs rapidly act on the existing information, thus enabling stock prices to more swiftly reflect current information. This dual-edged nature of latency arbitrage, where it simultaneously imposes costs due to the technological arms race while potentially improving price efficiency by quickening the information assimilation process into market prices, makes investigating the effects of HFTs in financial markets complex. It, however, underscores the importance of a balanced approach in evaluating the overall impact of HFT and latency arbitrage on market quality (e.g., Foucault et al. 2017; Rzayev et al. 2023).

4.4.Comparing ML-generated HFT Measures with Alternative Proxies.

In this section, we assess the ML-generated HFT measures against commonly used proxies from the literature. In this test, for out-of-sample validation, we train our ML model using only data from January to June 2009 to generate HFT measures. We then examine the relationships between NASDAQ's original liquidity-demanding and -supplying HFT measures from July to December 2009 with both our ML-generated measures and the following HFT proxies: flickering quotes ($Flick_{i,t}$), odd-lot volume ($OLV_{i,t}$), quote intensity ($QuoteInt_{i,t}$), quote-to-trade volume ratio ($QT_{i,t}$) and the number of messages ($MG_{i,t}$).

Motivated by Hasbrouck (2018), $Flick_{i,t}$ measures quote volatility in two steps: first calculating the standard deviation of quote midpoints over 100ms intervals, 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) and $MG_{i,t}$ is the number of messages (Hendershott et al. 2011; Boehmer et al. 2018). All five metrics are calculated using the Millisecond TAQ database.

We examine the relationship between HFT proxies and NASDAQ HFT measures using the following regression models with stock and time fixed effects:

$$NHFT_{i,t}^{D} = \alpha_i + \beta_t + \gamma_1 HFT_{i,t}^{ML,D} + \gamma_2 Flick_{i,t} + \gamma_3 OLV_{i,t} + \gamma_4 QuoteInt_{i,t} + \\ + \gamma_5 QT_{i,t} + \gamma_6 MG_{i,t} + \varepsilon_{i,t} \quad (9);$$

$$NHFT_{i,t}^{S} = \alpha_i + \beta_t + \gamma_1 HFT_{i,t}^{ML,S} + \gamma_2 Flick_{i,t} + \gamma_3 OLV_{i,t} + \gamma_4 QuoteInt_{i,t} + \\ + \gamma_5 QT_{i,t} + \gamma_6 MG_{i,t} + \varepsilon_{i,t} \quad (10),$$

where $NHFT_{i,t}^{D}$ and $NHFT_{i,t}^{S}$ are NASDAQ's liquidity-demanding and -supplying HFT measures, and $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ are our ML-generated proxies trained on data from January-June 2009. Other HFT proxies are defined as above. A key limitation of existing proxies is their inability to distinguish between liquidity-demanding and -supplying HFT strategies (Chakrabarty et al. 2023). Consequently, we include all proxies in both Models (9) and (10), estimating them both individually for each proxy and collectively, noting that including all measures simultaneously may introduce multicollinearity. The sample consists of 120 randomly selected NASDAQ- and NYSE-listed firms with available NASDAQ HFT data from July to December 2009, following the model training period of January-June 2009. We double-cluster standard errors by firm and time and standardize all dependent variables to make comparison between coefficients.

INSERT TABLE 7 HERE

Panel A of Table 7 shows the results for $NHFT_{i,t}^{S}$. Among all proxies, $HFT_{i,t}^{ML,S}$ delivers the strongest association with liquidity-supplying HFT activities from the NASDAQ HFT dataset ($NHFT_{i,t}^{S}$), demonstrated by the highest coefficient by magnitude (using standardized independent variables) and t-statistics. $HFT_{i,t}^{ML,S}$ also achieves the highest within-R^2 values. When incorporating all proxies simultaneously, the magnitude and t-statistics of $HFT_{i,t}^{ML,S}$ show only minimal decrease.

The results for liquidity-demanding HFT activity are consistent with those for liquidity-supplying HFT activities. In single-proxy regressions with stock and time fixed effects, only $QT_{i,t}$ predicts $NHFT_{i,t}^{D}$ with both correct sign and statistical significance. In contrast, $HFT_{i,t}^{ML,D}$ consistently demonstrates superior correlation with $NHFT_{i,t}^{D}$, showing the highest coefficient magnitude and t-statistics, along with the highest within-R^2 values.

Our estimation of Models (9) and (10) incorporates both stock and day fixed effects. The consistently strong correlations between ML-generated HFT measures and actual HFT values demonstrate these proxies' predictive power across both cross-sectional and time-series dimensions. In contrast, conventional HFT measures demonstrate relatively weak correlations when both fixed effects are included. One reason for this result might be that these measures predominantly capture either cross-sectional or time-series variation. To test this hypothesis, we re-estimated Models (9) and (10) using only day fixed effects.

The results in Panels C and D of Table 7 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 -demanding HFT activities. This pattern suggests that conventional HFT measures primarily capture cross-sectional variation. Notably, our ML-generated proxies maintain superior performance in this specification, too, showing much higher t-statistics and within-R^2 values. Additionally, our metrics subsume the information content of conventional HFT measures in joint regressions, where we include both sets of measures together.

Overall, our results demonstrate the superiority of our ML-generated measures over traditional HFT proxies. Our measures predict both liquidity-demanding and -supplying strategies with larger coefficients, higher t-statistics, and greater R^2 values. Furthermore, while our ML-generated measures effectively capture both cross-sectional and time-series dimensions, conventional measures predominantly reflect cross-sectional variation.

5. Application: HFT and information acquisition

The ML-based HFT measures have broad applications across various settings. In this section, we examine one such application—price discovery in financial markets; specifically, we demonstrate how the ML-based HFT measures shed new light on price formation mechanisms in modern markets.

Price discovery, a fundamental function of financial markets, is the process through which stock prices reflect information (e.g., O'Hara 2003), and it includes (i) the integration of *existing* information into asset prices and (ii) the generation or acquisition of *new* fundamental information (e.g., Brunnermeier 2005; Weller 2018; Brogaard and Pan 2022). The relationship between HFT and price discovery has been extensively examined by a fledgling stream of the market microstructure literature. The stream, which primarily concentrates on how existing information is incorporated into stock prices (for a comprehensive survey, see Menkveld 2016), largely suggests that HFT enhances the speed at which existing information is reflected in stock prices, contributing to the efficiency of price discovery mechanisms.

The role of HFTs in acquiring new information, however, remains understudied for two main reasons. First, measures of fundamental information acquisition are inherently low-frequency. Second, theoretical frameworks suggest that empirically examining HFTs' impact on information acquisition requires distinguishing between liquidity-supplying and -demanding strategies. While existing datasets that differentiate these strategies, such as NASDAQ HFT data, are valuable for analyzing high-frequency market quality metrics like liquidity, their limited sample periods and small number of stocks restrict their utility for studying low-frequency phenomena like fundamental information acquisition. For example, quarterly earnings announcements, the most frequent regular fundamental news events, yield only four information acquisition measures per stock-year. Consequently, one year of

NASDAQ HFT data covering 120 firms provides merely 480 observations. Our ML algorithm addresses this limitation by generating HFT measures for the entire TAQ universe over an extended period, enabling a more comprehensive analysis of how various HFT strategies influence information acquisition.

HFTs can improve information acquisition by increasing market liquidity and reducing trading costs through their liquidity provision function (e.g., Menkveld 2013; Brogaard et al. 2015; Aït-Sahalia and Sağlam 2024). The mechanism is intuitive: lower trading costs increase trading profitability, incentivizing investors to actively seek and capitalize on new information, thereby facilitating information acquisition and dissemination. However, HFTs may also employ aggressive strategies such as order anticipation, including back-running and latency arbitrage, to predict and profit from informed institutional investors' trades (e.g., Van Kervel and Menkveld 2019; Yang and Zhu 2020; Hirschey 2021). These strategies could increase trading costs for informed investors, potentially resulting in a crowding-out effect, which discourages them from seeking new information, thereby reducing the overall acquisition of new information.

Expanding on this discussion, Weller (2018) investigates the effect of HFTs on the information acquisition process by introducing a novel information acquisition metric known as the "price jump ratio." This ratio is calculated by dividing the return at the time of public information release by the cumulative return during the period leading up to the disclosure. The underlying concept is that a more pronounced price movement during the announcement suggests a less intense information acquisition process prior to the announcement, and implies that information predominantly becomes reflected in prices only upon public release. Thus, a higher price jump ratio means lower information acquisition. Weller (2018) concludes that ATs/HFTs have a detrimental effect on the information acquisition process.

While Weller (2018) enhances our understanding of HFTs' role in information acquisition, the study has a crucial limitation stemming from its use of MIDAS data. MIDAS aggregates HFT activity and, thus, does not differentiate between specific trading strategies. This limitation is crucial because theory suggests that HFTs' impact on information acquisition may vary fundamentally based on their trading strategies. Consequently, while MIDAS data allows Weller (2018) to document a negative relationship between HFT presence and information acquisition, it constrains the ability to investigate the underlying mechanisms driving this relationship. Weller (2018) acknowledges this limitation and offers a preliminary discussion of the strategies, ultimately emphasizing in their conclusion (p.2217) the need for future research *"to assess the precise mechanisms by which improved trading technology reduces the information content of prices."*

We respond to this call, by exploiting the unique proprieties of our ML-generated measures to investigate the role of HFTs in the information acquisition process. Specifically, we estimate the following regression model:

$$JUMP_{i,q} = \alpha_i + \beta_{m,q} + \gamma_1 HFT_{i,q}^{ML,D} + \gamma_2 HFT_{i,q}^{ML,S} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t} \quad (11),$$

where $JUMP_{i,q}$ is the ratio of cumulative abnormal returns during trading days [-1, 1] surrounding earnings announcements, divided by the cumulative abnormal returns during trading days [-21, 1] surrounding earnings announcements. Daily abnormal returns are calculated as the raw return minus the expected return, which is determined using the market model.

$HFT_{i,q}^{ML,D}$ and $HFT_{i,q}^{ML,S}$ denote our ML-generated measures of liquidity-demanding and liquidity-supplying HFT activity, respectively. We calculate these measures by averaging the daily HFT values over the 21 trading days preceding earnings announcements [-21, -1]. Our 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. We compute $MValue_{i,q}$ by averaging the daily market values over the same 21-day window, where daily market value equals closing price multiplied by shares outstanding. The remaining control variables represent 21-day averages of their daily counterparts prior to earnings announcements [-21, -1]. Following Weller (2018), we include stock and month fixed effects.

We include both $HFT_{i,q}^{ML,D}$ and $HFT_{i,q}^{ML,S}$ in our regression model to examine their comparative effects on information acquisition. The correlation coefficient between these metrics is 0.52, suggesting multicollinearity is not a concern. Given that higher $JUMP_{i,q}$ values indicate reduced information acquisition, we expect $HFT_{i,q}^{ML,D}$ to be positively associated with $JUMP_{i,q}$, as aggressive HFT strategies can increase trading costs and impede information acquisition. In contrast, we anticipate $HFT_{i,q}^{ML,S}$ to be negatively related to $JUMP_{i,q}$, since liquidity-providing HFT strategies typically reduce trading costs, making information acquisition more profitable.

INSERT TABLE 8 HERE

The estimation results of Model (11), reported in Table 8, show that $HFT_{i,q}^{ML,D}$ has a positive and statistically significant relationship with $JUMP_{i,q}$. The economic significance is notable: an increase in a firm's $HFT_{i,q}^{ML,D}$ 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 ((0.414-0.222)x0.178/0.517). Conversely, $HFT_{i,q}^{ML,S}$ demonstrate 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 value.

As Weller (2018) documents that HFT/AT reduces information acquisition, our analysis provides more nuanced insights into this relationship. Specifically, our findings

suggest that the positive association between generic 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 conjecture, we analyze the relationship between Weller's (2018) main HFT measures and our ML-generated 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 9 HERE

The results presented in Table 9 help reconcile our findings with Weller's (2018). $CT_{i,q}$ and $OLR_{i,q}$ are positively linked with $HFT_{i,q}^{ML,D}$, while $TO_{i,q}$ (an inverse measure of HFT) is negatively related. Conversely, the metrics display opposite relationships with $HFT_{i,q}^{ML,S}$. 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 activities.

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

$$Return_{i,q} = \alpha_i + \beta_q + \sum_{n=-1}^{1}(\gamma_n Earning_{i,q+n} + \vartheta_n Earning_{i,q+n} * HFT_{i,q}^{ML,D} + \theta_n Earning_{i,q+n} * HFT_{i,q}^{ML,S}) + \rho_1 HFT_{i,q}^{ML,D} + \rho_2 HFT_{i,q}^{ML,S} + \rho_3 Return_{i,q+1} + \rho_4 Return_{i,q-1} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,q} \quad (12),$$

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 our 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 encapsulates FERC, with a positive value suggesting that current returns incorporate future earnings information—an indication of heightened fundamental information acquisition in the current period. We employ the same control variables used in the jump ratio model, averaged at the quarterly frequency.

The coefficients of interest in Model (12) 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 $HFT_{i,q}^{ML,D}$ $\left(HFT_{i,q}^{ML,S}\right)$ is negatively (positively) associated with information acquisition, we expect ϑ_n and θ_n to be negative and positive, respectively.

INSERT TABLE 10 HERE

Table 10 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 also extend our baseline results in two directions. First, we assess whether existing HFT datasets that differentiate trading strategies are suitable for investigating HFT's role in information acquisition. This question is crucial because a positive answer would challenge the need for ML-generated HFT measures. In a way, the relevance of this question in of itself is debatable because, as severally noted, publicly available datasets that differentiate between strategies used by HFTs do not exist at this time. Therefore, we use the time- and sample-

limited and proprietary NASDAQ HFT dataset that covers 120 stocks for 2009 to replicate our jump ratio and FERC analyses.

INSERT TABLE 11 HERE

Table 11 presents the results. Using the NASDAQ HFT dataset, we find no statistically significant relationship between HFT strategies and information acquisition, due to the limited sample size. This finding underscores the value of our ML-generated HFT measures for examining HFT's impact on low-frequency market quality metrics and, by extension, real economic outcomes that typically rely on low-frequency data.

The second extension addresses concerns about our reliance on 2009 data to train the ML model. First, it is important to note that the current literature continues to use the NASDAQ HFT dataset because the core distinction between liquidity-demanding and liquidity-supplying strategies remains fundamental to HFT behavior (e.g., Boehmer et al. 2018; Goldstein et al. 2023; Nimalendran et al. 2024). Moreover, in Section 4.2, we show that our measures respond to technological shocks both near and far from the training period. Nevertheless, we provide additional validation by examining the HFT-information acquisition relationship in the period close to our training sample. Similar results between this restricted sample and our full sample would indicate that our findings are not sensitive to the temporal distance from the training data.

INSERT TABLE 12 HERE

Table 12 presents results using data from January 2010 to December 2012. For both the jump ratio and FERC analyses, the findings mirror our baseline results. Specifically, liquidity-demanding strategies show a negative relationship with information acquisition, while liquidity-supplying strategies demonstrate a positive association.

A cautionary note regarding our results in this section is important. Our study's primary contribution is not the investigation of HFT's role in information acquisition, but rather the

development of ML techniques to identify and measure HFT strategies using publicly available data. We explore information acquisition as one important application of the ML-generated HFT measures. While the relationship between HFT and information acquisition is an understudied yet economically significant question, we do not claim to establish a causal link between HFT activity and information acquisition, acknowledging the complexity and potential endogeneity concerns. Nevertheless, our results demonstrate that ML-generated HFT measures provide valuable tools for investigating the relationship between HFT and real economic outcomes, which are inherently low-frequency in nature. This data-driven approach is particularly important because econometric approaches using exogenous shocks, such as in a DiD framework, cannot effectively examine the real effects of various strategies employed by HFTs, as these shocks symmetrically impact both liquidity-demanding and -supplying strategies. Therefore, distinguishing between different HFT strategies through data is essential for understanding their distinct economic impacts. Furthermore, our findings complement Weller (2018) by providing empirical evidence of specific mechanisms through which HFT activity affects information acquisition.

6. Extensions to the ML framework: Feature space and scaling analysis

In this section, we extend our baseline ML framework in two ways. First, we augment the feature space of our ML methodology. The selection of input features in ML involves two competing considerations. More granular data could potentially enhance prediction precision although are typically less accessible, costlier, and harder to process. Alternatively, more accessible data sources may sacrifice some predictive power while enabling wider application and replication. Our baseline model prioritizes the latter—a key contribution in developing HFT measures from non-proprietary data. Hence, our ML framework employs daily input features derived directly from TAQ's Intraday Indicators.

However, these indicators lack quote-level granularity, such as message counts or quote update frequencies, potentially constraining the ML model's training effectiveness. We believe that this concern is substantially mitigated by our ML model's R^2 of 82%, indicating that our input variables effectively capture the predominant variation in HFT activity. This suggests that the potential gains from incorporating more granular quote-level data are limited. Nevertheless, we assess this empirically by augmenting our feature set with quote-level data from the Millisecond TAQ database to evaluate potential improvements in ML algorithm performance during training. The additional quote-level features include message counts, quote update frequencies, small trade volumes (under 100 shares), and high-frequency midpoint variations over 100-millisecond intervals—metrics previously linked to HFT activity (e.g., Chakrabarty et al. 2023). We calculate these measures for 2009, our ML model training period. Using data from January to June 2009, we train two pairs of ML models: one pair using only the original daily features from TAQ's Intraday Indicators database, and another incorporating both daily indicators and the granular quote-related features from the TAQ's Millisecond database. We then generate HFT measures for July to December 2009, enabling an out-of-sample comparison between models with and without quote-related information.

We document three key findings. First, incorporating quote-related information marginally improves the ML model's performance, increasing the R^2 from 82% to 84%. Second, the HFT measures generated with and without quote-related information demonstrate remarkably high correlations. Specifically, the correlation coefficient between liquidity-supplying HFT measures with and without quote-related information is 0.99, while the corresponding correlation for liquidity-demanding measures is 0.96. Third, when we regress the HFT values from the NASDAQ HFT data on the ML-based HFT measures generated with quote-level information, the coefficient estimates and t-statistics differ only marginally from those in Table 8, where we report the correlation between the NASDAQ HFT values and the

ML-generated HFT measures based on the baseline model without quote-level information. These results suggest that the input features from the TAQ intraday indicator database sufficiently capture HFT activity, with additional quote-related information providing only minimal, even negligible, incremental value.

This finding is unsurprising as our initial input features incorporate variables strongly associated with quote-level activity, including market depth and bid-ask spreads. Indeed, analysis of correlations between quote-related input features and our original trade-related features demonstrates strong relationships. Specifically, the total number of trades exhibits a 0.90 correlation with message count, while message count shows correlations exceeding 0.65 with both ISO trades and market depth. Additionally, the frequency of quote revisions demonstrates strong correlations (exceeding 0.70) with trade frequency, ISO trades, and market depth.

Our second extension addresses the scaling of HFT measures. Thus far, we have demonstrated that the ML-generated HFT measures effectively capture both liquidity-demanding and liquidity-supplying strategies. Additionally, these measures allow us to address important economic questions that would otherwise remain unanswered. All of our tests are based on scaled HFT measures, where HFT trading volume is normalized by total trading volume. This scaling is important, as highlighted by Hendershott et al. (2011), to account for total trading volume when examining the role of HFTs in financial markets. However, it also raises a valid concern for our study. Specifically, since our ML algorithm is trained on scaled HFT values, it may capture variation in total trading volume rather than HFT trading volume. To address this, we also use the ML model to predict unscaled HFT trading volume using the same input variables. In this test, the key target variables are unscaled liquidity-demanding ($NUHFT_{i,t}^{D}$) and liquidity-supplying ($NUHFT_{i,t}^{S}$) trading volumes, calculated as the sum of HH

and HN (HH and NH) volumes for stock i and day t from the NASDAQ HFT data. We then replicate all our tests using these unscaled values.

Our main findings remain robust when using scaled target variables, with the complete set of results presented in the Online Appendix to this paper. Specifically, we confirm that: (1) HFT activity systematically responds to both scheduled and unscheduled announcements and technological changes, (2) shows distinct responses to latency arbitrage opportunities, (3) ML-generated unscaled HFT measures outperform conventional HFT proxies, and (4) demonstrates contrasting effects on information acquisition, i.e., negative for liquidity-demanding strategies but positive for liquidity-supplying ones.

7. Conclusion

The impact of HFT on market quality has been a central focus of microstructure research for the past fifteen years. However, this literature faces a key limitation: 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 -supplying strategies. This constraint has hampered our understanding of the mechanisms driving HFTs' effects over longer horizons.

We address this limitation by developing an ML approach that generates distinct measures for liquidity-demanding and -supplying HFT activity. By training ensembles on NASDAQ HFT data and TAQ variables, we create comprehensive HFT measures covering 8,314 U.S. stocks with 9,440,600 stock-day observations—spanning the entire universe of the U.S. equity market over an extended period.

Our validation tests demonstrate that these ML-generated measures capture theoretically predicted HFT behavior. The measures respond to both scheduled and unscheduled information releases and exogenous technological changes. Moreover, they

reflect strategy-specific reactions to latency arbitrage opportunities: liquidity-demanding HFTs increase their activity while liquidity-supplying HFTs reduce it. Comparative analysis confirms that our measures outperform alternative HFT proxies in correlating with actual HFT activity.

We demonstrate the importance of differentiating HFT strategies by examining their role in fundamental information acquisition, a key market quality measure that can be tested at low frequency. 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 provides clarity on how different HFT strategies affect price discovery in financial markets, highlighting the value of our data and methodology in advancing financial theory.

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

Aquilina, M., Foley, S., O'Neill, P., Ruf, T., 2024. Sharks in the dark: quantifying HFT dark pool latency arbitrage. Journal of Economic Dynamics and Control 158, 104786

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

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

Bishop, C.M., Nasrabadi, N.M., 2006. Pattern recognition and machine learning. Springer.

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., Carrion, A., Moyaert, T., Riordan, R., Shkilko, A., Sokolov, K., 2018. High frequency trading and extreme price movements. Journal of Financial Economics 128, 253-265

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

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

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

Cole, B., Daigle, J., Van Ness, B., Van Ness, R., 2015. Do high frequency traders care about earnings announcements? An analysis of trading activity before, during, and after regular trading hours. The Handbook of High Frequency Trading. San Diego: Academic Press. Elsevier Inc, 255-270

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

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

Hagströmer, B., Nordén, L., 2013. The diversity of high-frequency traders. Journal of Financial Markets 16, 741-770

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., 2019. A state-space modeling of the information content of trading volume. Journal of Financial Markets 46, 100507

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

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., 2008. Top 10 algorithms in data mining. Knowledge and information systems 14, 1-37

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 in Equation 1, 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 as a result of said feature.

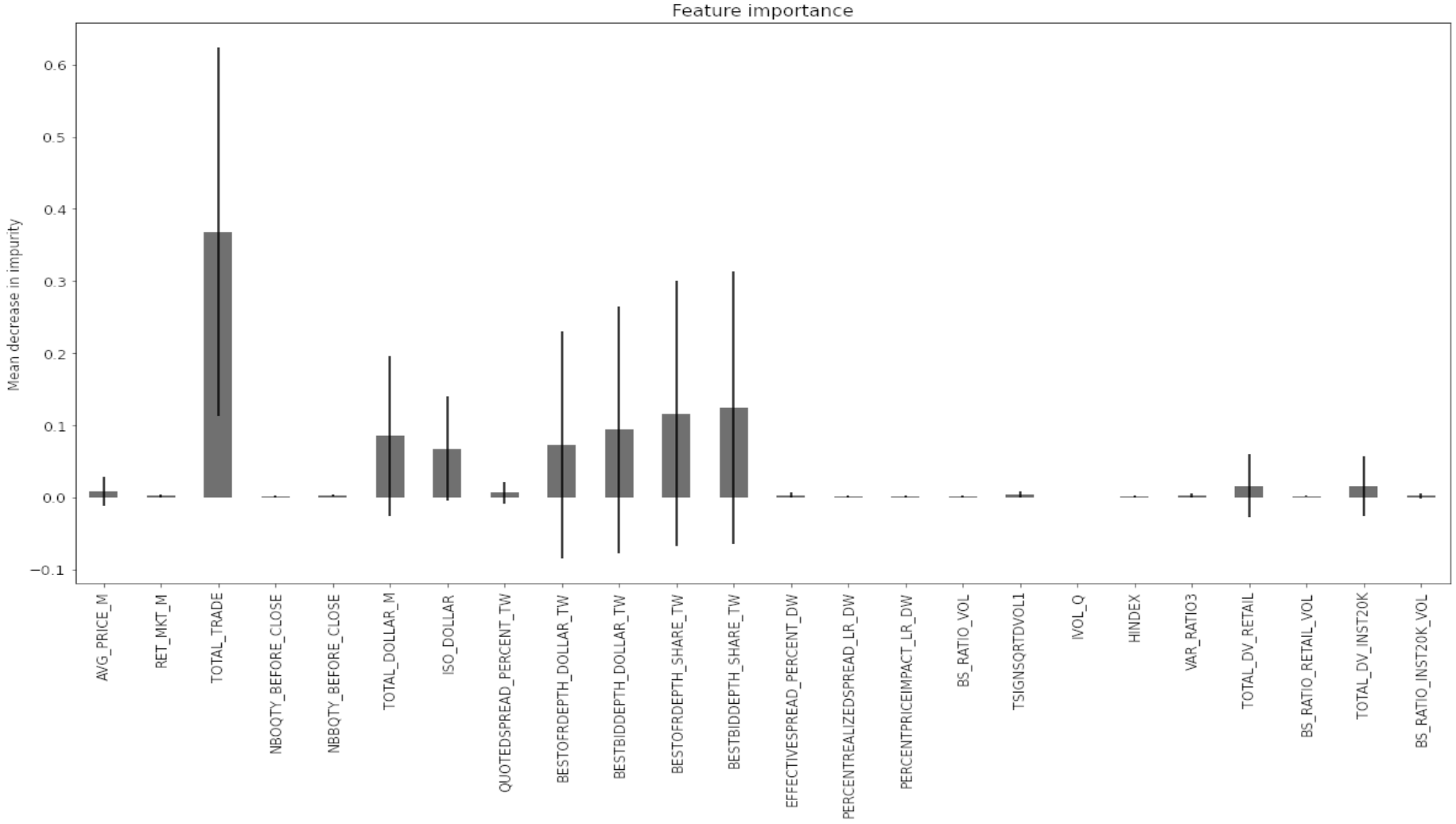


Figure 2

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.

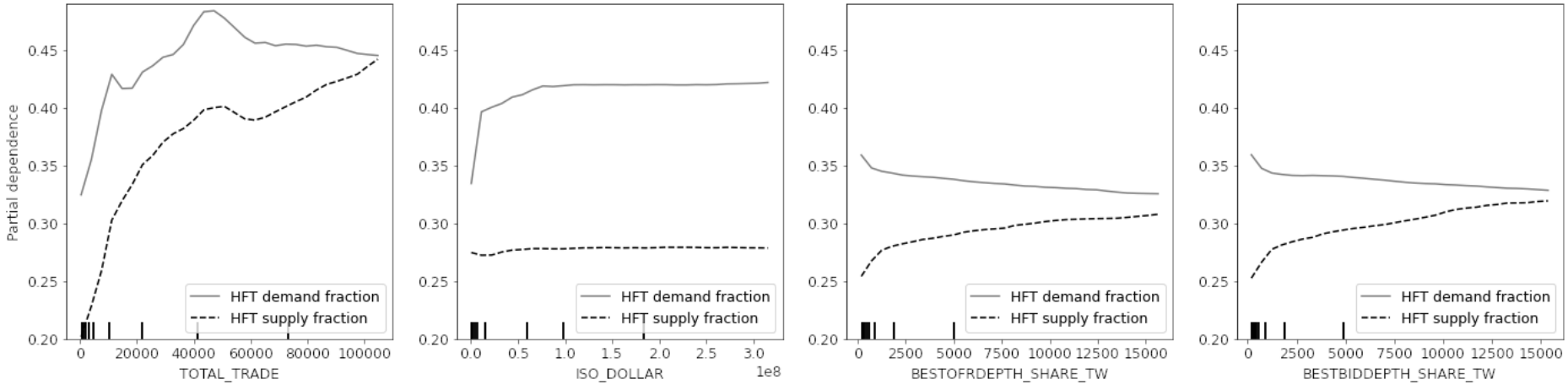
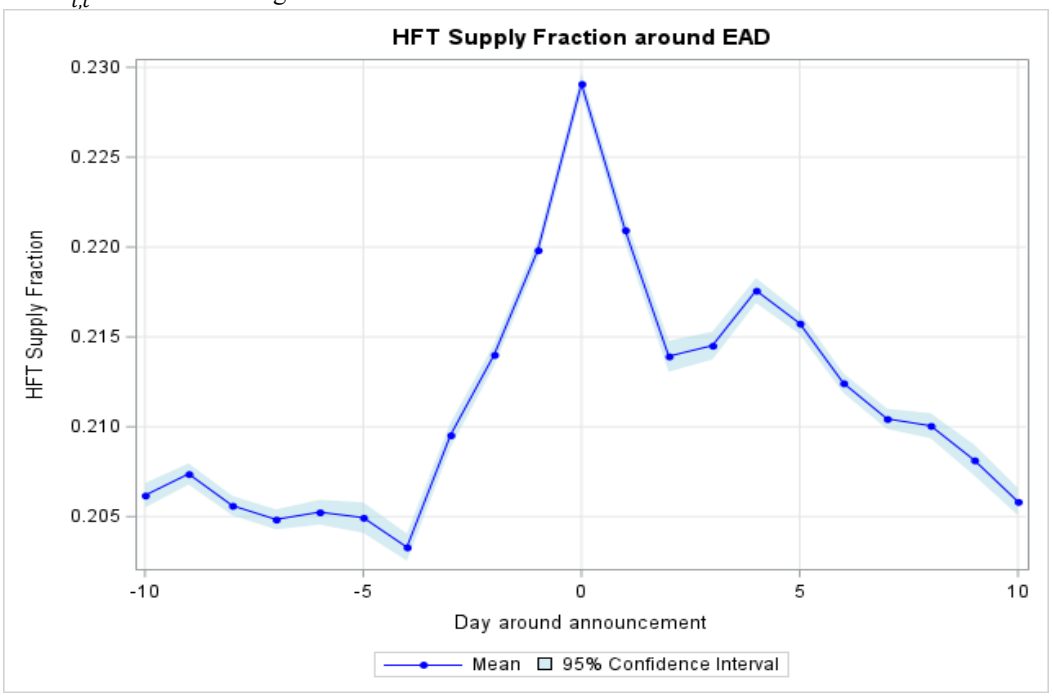


Figure 3
HFT around earnings announcements
This figure illustrates the evolution of ML-generated HFT measures with their 95% confidence interval surrounding scheduled events, specifically earnings announcements. The event window spans 10 days before and after the announcement dates, which are sourced 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: $HFT_{i,t}^{ML,S}$ around earning announcements.



Panel B: $HFT_{i,t}^{ML,D}$ around earning announcements.

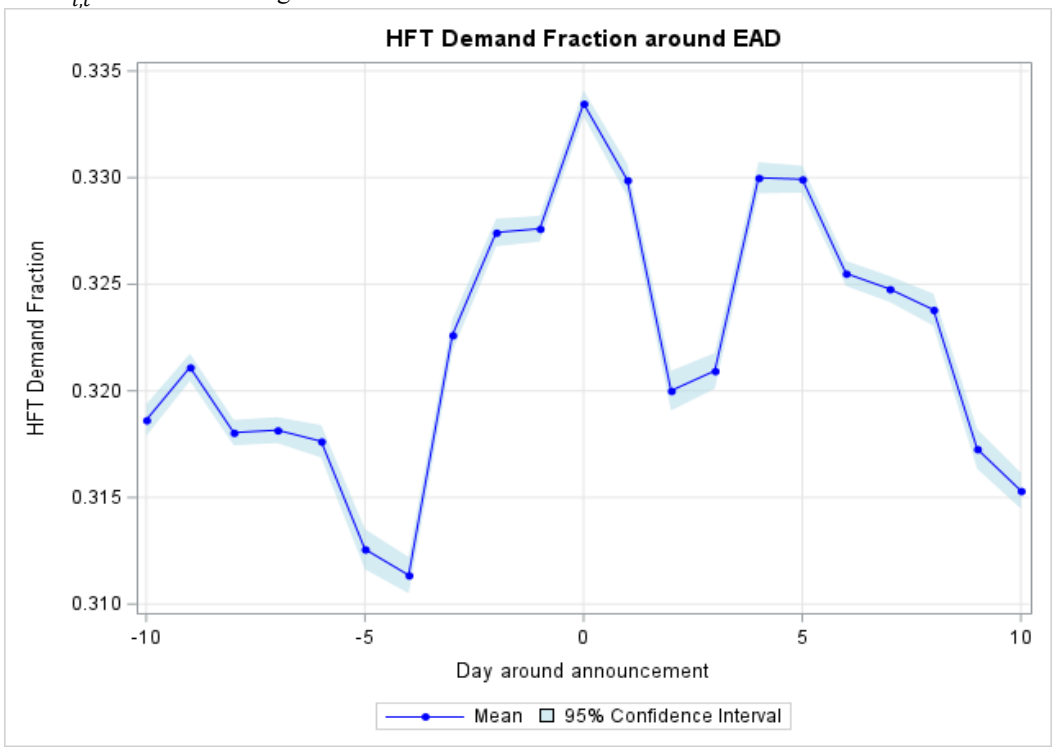
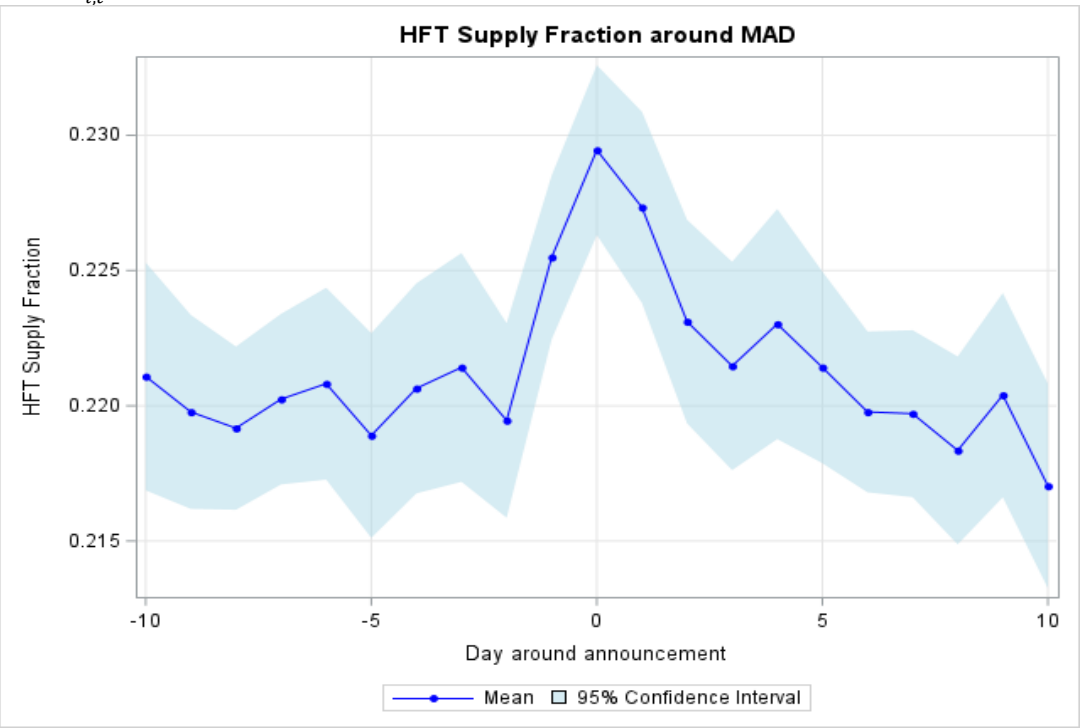


Figure 4
HFT around M&A announcements
This figure illustrates the evolution of ML-generated HFT measures with their 95% confidence interval surrounding unscheduled events, specifically mergers and acquisitions (M&A) announcements. The event window spans 10 days before and after the announcement dates, 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: $HFT_{i,t}^{ML,S}$ around M&A announcements.



Panel B: $HFT_{i,t}^{ML,D}$ around M&A announcements.

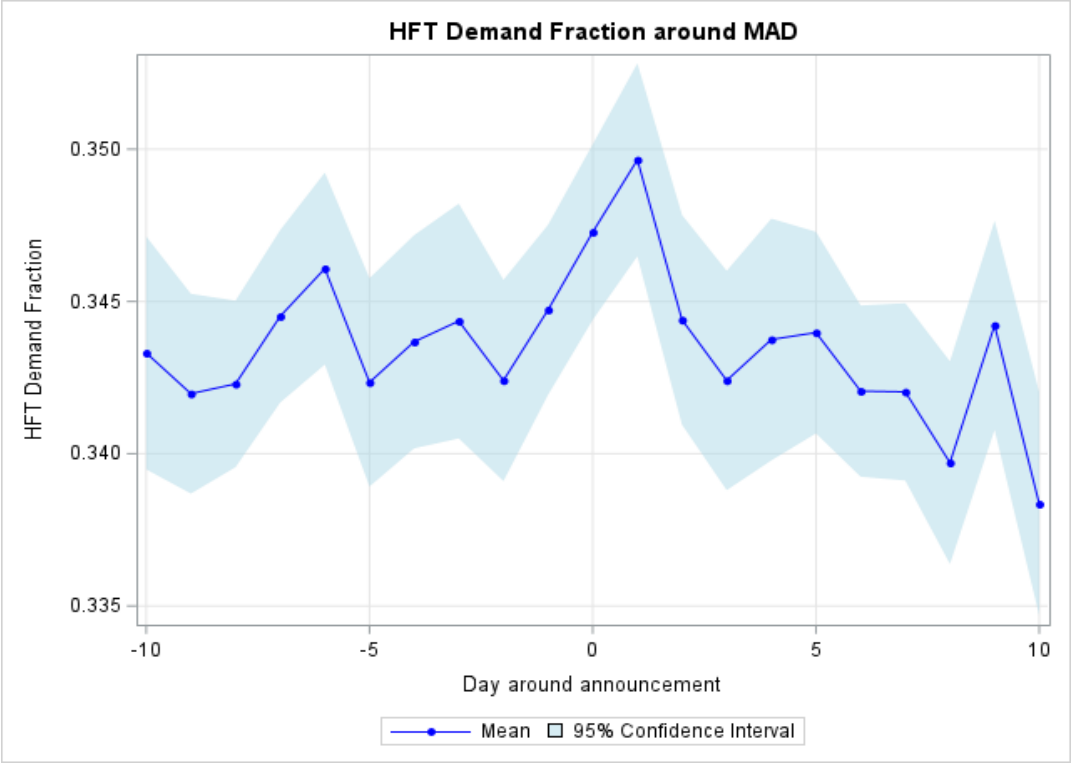


Table 1
Input and Output Variables in the ML Model Training Process
This table presents the variables used to train the ML model, including their notation, descriptions, and data sources. Panel A contains output variables from NASDAQ HFT data. Panel B details input variables derived from the TAQ database, 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.		
$NHFT_{i,t}^{D}$	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
$NHFT_{i,t}^{S}$	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 $\text{SSqrtDvol} = Sgn(\sum_{s-300}^{s} BuyDollar - \sum_{s-300}^{s} SellDollar) \times \sqrt{\lvert\sum_{s-300}^{s} BuyDollar - \sum_{s-300}^{s} SellDollar\rvert}$, 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.

Variable	Definition	Mean	Std	Min	p.25	p.50	p.75	Max
$HFT_{i,t}^{ML,D}$	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
$HFT_{i,t}^{ML,S}$	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 4.3.	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
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

Table 4
Machine Learning comparison
The table lists the arithmetic mean and standard deviation for R^2 values across 10 iterations for 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.
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

Table 5
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:

$$HFT_{i,t}^{ML,D} = \alpha_i + \beta_t + \gamma_1 Post_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (5.1)$$

$$HFT_{i,t}^{ML,S} = \alpha_i + \beta_t + \gamma_2 Post_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (5.2)$$

$$HFT_{i,t}^{ML,D} = \alpha_i + \beta_t + \gamma_1 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (5.3)$$

$$HFT_{i,t}^{ML,S} = \alpha_i + \beta_t + \gamma_2 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (5.4)$$

where $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ 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 5.1 and 5.2), $Post_{i,t}$ 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 5.3 and 5.4), $Post_{i,t}$ 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) $HFT_{i,t}^{ML,D}$	(ii) $HFT_{i,t}^{ML,S}$	(iii) $HFT_{i,t}^{ML,D}$	(iv) $HFT_{i,t}^{ML,S}$
$Post_{i,t}$	0.002**	0.002**		
	(2.12)	(2.10)		
$Post_{i,t} * Amex_{i,t}$			-0.005**	-0.007***
			(-2.34)	(-3.31)
$Volatility_{i,t}$	0.013**	0.000	0.001	0.001
	(2.19)	(0.07)	(1.29)	(1.33)
$Spread_{i,t}$	-0.066***	-0.024***	-0.015***	-0.006***
	(-12.58)	(-5.58)	(-10.96)	(-6.05)
$InvPrice_{i,t}$	-0.151***	0.037	-0.026	-0.023*
	(-3.08)	(0.92)	(-1.57)	(-1.96)
$Volume_{i,t}$	0.001	0.020***	0.001**	0.005***
	(1.30)	(17.75)	(2.25)	(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 6
HFT Response to Latency Arbitrage Opportunities
This table examines how our ML-generated HFT measures respond latency arbitrage opportunities using the following OLS models:

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

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

where $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ represent the ML – generated 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 4.3. 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 $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$, 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) $HFT_{i,t}^{ML,D}$	(ii) $HFT_{i,t}^{ML,S}$
$NLAO_{i,t}$	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 7
Comparative Analysis of HFT Measures
This table evaluates our ML-generated HFT measures against alternative proxies using the following models:

$$NHFT_{i,t}^{S} = \alpha_i + \beta_t + \gamma_1 HFT_{i,t}^{ML,S} + \gamma_2 Flick_{i,t} + \gamma_3 OLV_{i,t} + \gamma_4 QuoteInt_{i,t} + \gamma_5 QT_{i,t} + \gamma_6 MG_{i,t} + \varepsilon_{i,t}$$

$$NHFT_{i,t}^{D} = \alpha_i + \beta_t + \gamma_1 HFT_{i,t}^{ML,D} + \gamma_2 Flick_{i,t} + \gamma_3 OLV_{i,t} + \gamma_4 QuoteInt_{i,t} + \gamma_5 QT_{i,t} + \gamma_6 MG_{i,t} + \varepsilon_{i,t}$$

where $NHFT_{i,t}^{D}$ and $NHFT_{i,t}^{S}$ are NASDAQ's liquidity-demanding and -supplying HFT measures, and $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ 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: $NHFT_{i,t}^{S}$							
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$HFT_{i,t}^{ML,S}$	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: $NHFT^{D}_{i,t}$							
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$HFT^{ML,D}_{i,t}$	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: $NHFT^{S}_{i,t}$							
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$HFT^{ML,S}_{i,t}$	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: $NHFT_{i,t}^{D}$							
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$HFT_{i,t}^{ML,D}$	0.423***						0.383***
	(23.23)						(17.93)
$Flick_{i,t}$		-0.003					-0.003
		(-0.41)					(-0.93)
$OLV_{i,t}$			0.000				0.002**
			(0.24)				(1.99)
$QuoteInt_{i,t}$				0.091***			0.013
				(7.92)			(0.92)
$QT_{i,t}$					0.021**		0.014***
					(2.04)		(2.69)
$MG_{i,t}$						0.092***	0.015
						(8.37)	(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 8
HFT Activity and Information Acquisition – Jump ratio
This table examines how HFT activity affects information acquisition using the following OLS model:

$$JUMP_{i,q} = \alpha_i + \beta_{m,q} + \gamma_1 HFT_{i,q}^{ML,D} + \gamma_2 HFT_{i,q}^{ML,S} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

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). $HFT_{i,q}^{ML,D}$ and $HFT_{i,q}^{ML,S}$ are ML-generated 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.

	$JUMP_{i,q}$
$HFT_{i,q}^{ML,D}$	0.178***
	(4.57)
$HFT_{i,q}^{ML,S}$	-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
N obs.	49,515
R^2	0.4%

Table 9
Comparing ML-Generated 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:

$$CT_{i,q} = \alpha_i + \beta_{m,q} + \gamma_1 HFT_{i,q}^{ML,D} + \gamma_2 HFT_{i,q}^{ML,S} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

$$OLR_{i,q} = \alpha_i + \beta_{m,q} + \gamma_1 HFT_{i,q}^{ML,D} + \gamma_2 HFT_{i,q}^{ML,S} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

$$TO_{i,q} = \alpha_i + \beta_{m,q} + \gamma_1 HFT_{i,q}^{ML,D} + \gamma_2 HFT_{i,q}^{ML,S} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

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 $HFT_{i,q}^{ML,D}$ and $HFT_{i,q}^{ML,S}$ are ML-generated 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) $CT_{i,q}$	(ii) $OLR_{i,q}$	(ii) $TO_{i,q}$
$HFT_{i,q}^{ML,D}$	0.839***	2.714***	-1.208***
	(10.64)	(24.76)	(-15.71)
$HFT_{i,q}^{ML,S}$	-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 10
HFT Activity and Information Acquisition - FERC
This table examines how HFT activity affects information acquisition using the following model:

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

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. $HFT_{i,q}^{ML,D}$ and $HFT_{i,q}^{ML,S}$ are ML-generated 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.

	$Return_{i,q}$
$Earning_{i,q+1} * HFT_{i,q}^{ML,D}$	-2.018***
	(4.56)
$Earning_{i,q+1} * HFT_{i,q}^{ML,S}$	2.676***
	(5.25)
$HFT_{i,q}^{ML,D}$	-0.059
	(-1.59)
$HFT_{i,q}^{ML,S}$	0.010
	(0.08)
$Earning_{i,q+1}$	0.573**
	(9.67)
All controls as defined in the model	Yes
Stock and Quarter FE	Yes
N obs.	157,343
R^2	4%

Table 11
HFT Activity and Information Acquisition Using NASDAQ HFT Data
This table replicates the analyses from Tables 9 and 11 using NASDAQ's original HFT measures instead of our ML-generated proxies. $NHFT_{i,t}^{D}$ and $NHFT_{i,t}^{S}$ 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) $Return_{i,q}$
$NHFT_{i,q}^{D}$	0.997	
	(0.52)	
$NHFT_{i,q}^{S}$	-0.903	
	(-0.56)	
$Earning_{i,q+1} * NHFT_{i,q}^{D}$		0.521
		(0.06)
$Earning_{i,q+1} * NHFT_{i,q}^{S}$		-3.246
		(-0.59)
Controls	As in Table 9	As in Table 11
Stock and Month FEs	Yes	Yes
N obs.	466	401
R^2	0.7%	40%

Table 12
HFT Activity and Information Acquisition: Analysis of 2010-2012 Period
This table replicates the analyses from Tables 9 and 11 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) $Return_{i,q}$
$HFT_{i,q}^{ML,D}$	0.114*** (2.59)	
$HFT_{i,q}^{ML,S}$	-0.101** (-2.10)	
$Earning_{i,q+1} * HFT_{i,q}^{ML,D}$		-3.982*** (-3.41)
$Earning_{i,q+1} * HFT_{i,q}^{ML,S}$		3.666*** (2.81)
Controls	As in Table 9	As in Table 11
Stock and Month FEs	Yes	Yes
N obs.	9,915	30,048
R^2	0.4%	5%

Online Appendix for **"Can machine learning unlock new insights into high-frequency trading?"**

March 2025

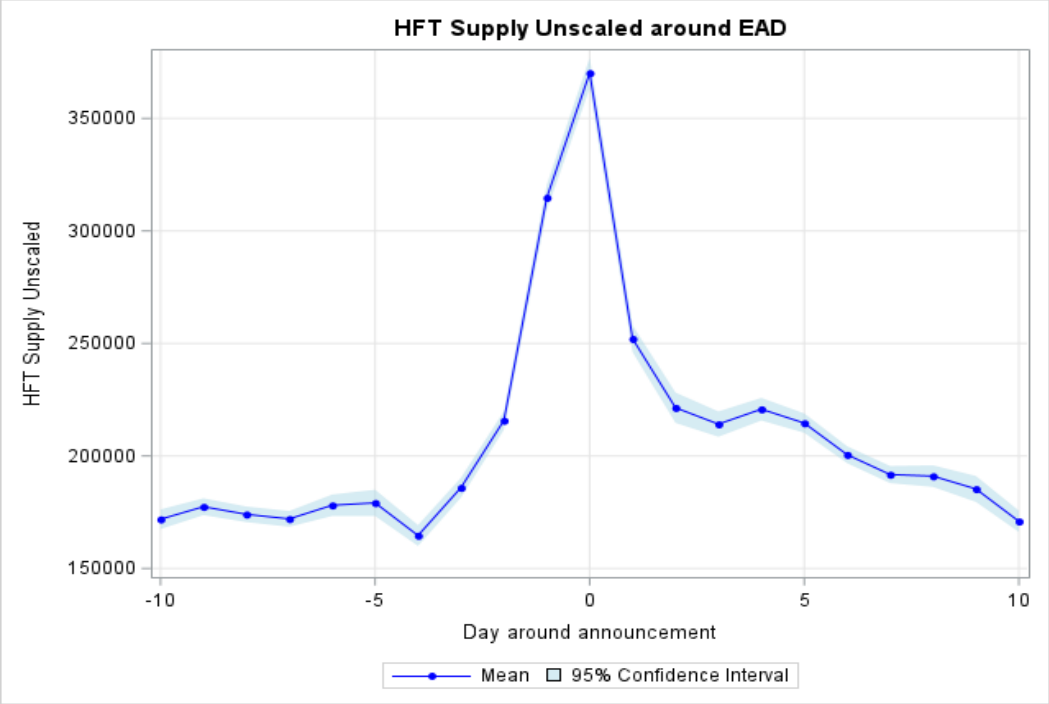
Introduction

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

- Figure OA.1: Replication of Figure 3 using unscaled HFT measures.
- Figure OA.2: Replication of Figure 4 using unscaled HFT measures.
- Table OA.1: Replication of Table 5 using unscaled HFT measures.
- Table OA.2: Replication of Table 6 using unscaled HFT measures.
- Table OA.3: Replication of Table 7 using unscaled HFT measures
- Table OA.4: Replication of Table 8 using unscaled HFT measures
- Table OA.5: Replication of Table 10 using unscaled HFT measures

Figure OA.1
HFT around earnings announcements
This figure illustrates the evolution of ML-generated unscaled HFT measures ($UHFT_{i,t}^{ML,S}$ and $UHFT_{i,t}^{ML,D}$) 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: $UHFT_{i,t}^{ML,S}$ around earning announcements.



Panel B: $UHFT_{i,t}^{ML,D}$ around earning announcements.

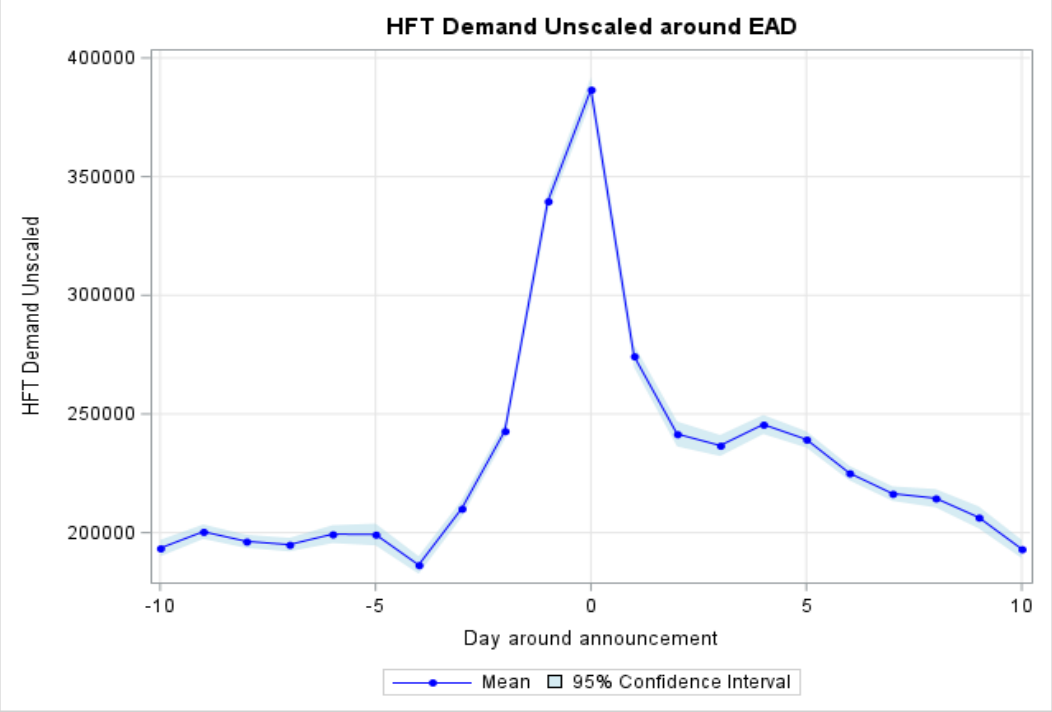
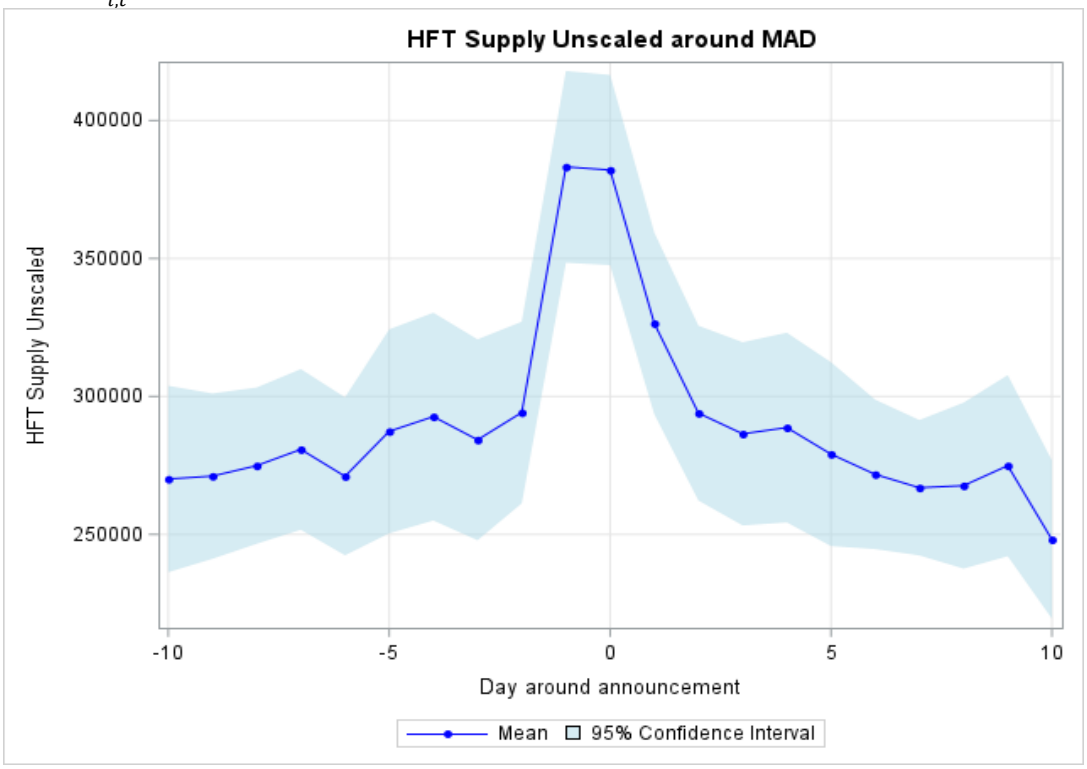


Figure OA.2
HFT around M&A announcements
This figure illustrates the evolution of ML-generated unscaled HFT measures ($UHFT_{i,t}^{ML,S}$ and $UHFT_{i,t}^{ML,D}$) with their 95% confidence interval surrounding unscheduled events, specifically mergers and acquisitions (M&A) announcements. The event window 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: $UHFT_{i,t}^{ML,S}$ around M&A announcements.



Panel B: $UHFT_{i,t}^{ML,D}$ around M&A announcements.

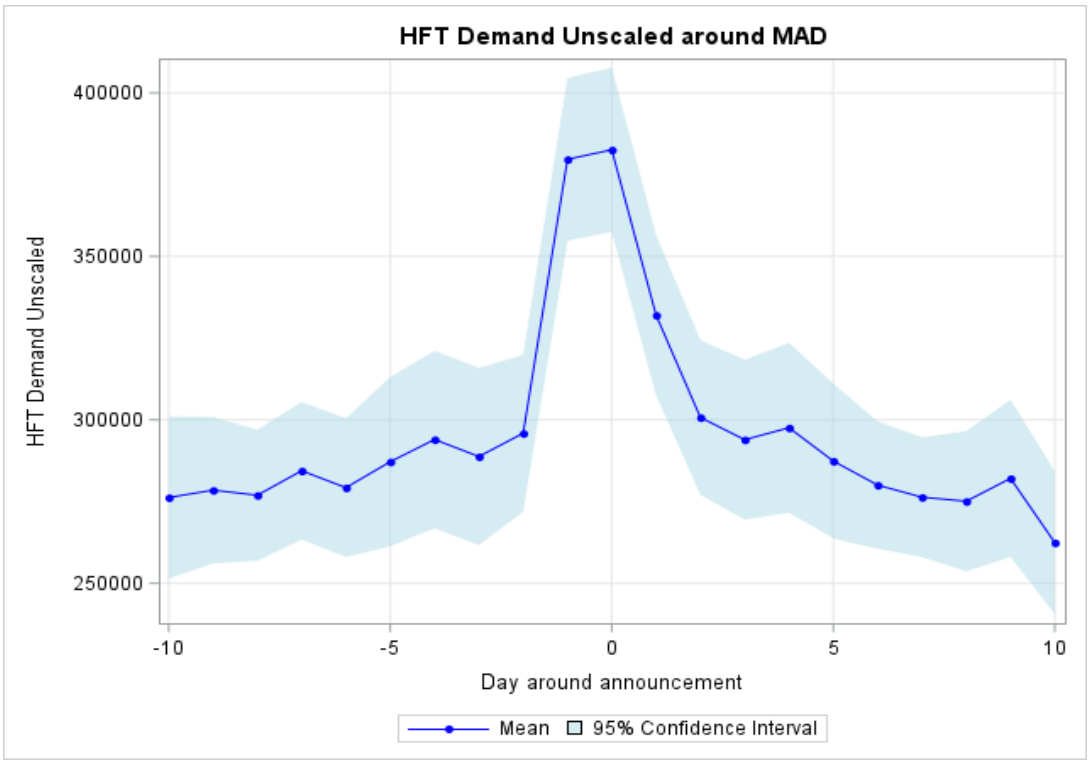


Table OA.1
Impact of Exchange Technological Changes on HFT Activity
This table examines how our ML-generated unscaled HFT measures respond to two technological changes: NASDAQ's reduced data dissemination latency and Amex's speed bump implementation. We estimate the following difference-in-difference models:

$$UHFT_{i,t}^{ML,D} = \alpha_i + \beta_t + \gamma_1 Post_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (5.1)$$

$$UHFT_{i,t}^{ML,S} = \alpha_i + \beta_t + \gamma_2 Post_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (5.2)$$

$$UHFT_{i,t}^{ML,D} = \alpha_i + \beta_t + \gamma_1 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (5.3)$$

$$UHFT_{i,t}^{ML,S} = \alpha_i + \beta_t + \gamma_2 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (5.4)$$

where $UHFT_{i,t}^{ML,D}$ and $UHFT_{i,t}^{ML,S}$ 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 5.1 and 5.2), $Post_{i,t}$ 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 5.3 and 5.4), $Post_{i,t}$ 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) $UHFT_{i,t}^{ML,D}$	(ii) $UHFT_{i,t}^{ML,S}$	(iii) $UHFT_{i,t}^{ML,D}$	(iv) $UHFT_{i,t}^{ML,S}$
$Post_{i,t}$	1.055***	1.347**		
	(2.79)	(2.27)		
$Post_{i,t} * Amex_{i,t}$			-0.977**	-0.696**
			(-2.27)	(-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.2
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:

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

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

where $UHFT_{i,t}^{ML,D}$ and Urepresent 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 4.3. 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 $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$, 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) $UHFT_{i,t}^{ML,D}$	(ii) $UHFT_{i,t}^{ML,S}$
$NLAO_{i,t}$	66.266***	-150.518**
	(3.21)	(-2.04)
Controls	Yes	Yes
Stock and Day FE	Yes	Yes
N obs.	246,139	246,139
R^2	39%	38%

Table OA.3
Comparative Analysis of HFT Measures
This table evaluates our ML-generated unscaled HFT measures against alternative proxies using the following models:

$$NUHFT_{i,t}^{D} = \alpha_i + \beta_t + \gamma_1 UHFT_{i,t}^{ML,D} + \gamma_2 Flick_{i,t} + \gamma_3 OLV_{i,t} + \gamma_4 QuoteInt_{i,t} + \gamma_5 QT_{i,t} + \gamma_6 MG_{i,t} + \varepsilon_{i,t}$$
$$NUHFT_{i,t}^{S} = \alpha_i + \beta_t + \gamma_1 UHFT_{i,t}^{ML,S} + \gamma_2 Flick_{i,t} + \gamma_3 OLV_{i,t} + \gamma_4 QuoteInt_{i,t} + \gamma_5 QT_{i,t} + \gamma_6 MG_{i,t} + \varepsilon_{i,t}$$

where $NUHFT_{i,t}^{D}$ and $NUHFT_{i,t}^{S}$ are NASDAQ's unscaled liquidity-demanding and -supplying HFT measures, and $UHFT_{i,t}^{ML,D}$ and $UHFT_{i,t}^{ML,S}$ 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: $NUHFT_{i,t}^{S}$							
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$UHFT_{i,t}^{ML,S}$	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: $NUHFT_{i,t}^{D}$							
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$UHFT_{i,t}^{ML,D}$	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: $NUHFT_{i,t}^{S}$							
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$UHFT_{i,t}^{ML,S}$	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: $NUHFT_{i,t}^{D}$							
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
$UHFT_{i,t}^{ML,D}$	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.4

HFT Activity and Information Acquisition—Jump ratio

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

$$JUMP_{i,q} = \alpha_i + \beta_{m,q} + \gamma_1 UHFT_{i,q}^{ML,D} + \gamma_2 UHFT_{i,q}^{ML,S} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

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). $UHFT_{i,q}^{ML,D}$ and $UHFT_{i,q}^{ML,S}$ 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.

	$JUMP_{i,q}$
$UHFT_{i,q}^{ML,D}$	0.042***
	(9.82)
$UHFT_{i,q}^{ML,S}$	-0.022***
	(-5.99)
Controls	Yes
Stock and Month FE	Yes
N obs.	49,515
R^2	1%

Table OA.5

HFT Activity and Information Acquisition—FERC

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

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

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. $UHFT_{i,q}^{ML,D}$ and $UHFT_{i,q}^{ML,S}$ 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.

	$Return_{i,q}$
$Earning_{i,q+1} * HFT_{i,q}^{ML,D}$	-0.003***
	(6.56)
$Earning_{i,q+1} * HFT_{i,q}^{ML,S}$	0.003***
	(7.26)
Controls	Yes
Stock and Quarter FE	Yes
N obs.	157,343
R^2	4%