Can machine learning unlock new insights into high-frequency trading?

Gbenga Ibikunle [a,b], Ben Moews [a,c], Khaladdin Rzayev [a,d,e,*]
[a]*Edinburgh Centre for Financial Innovations, The University of Edinburgh*
[b]*RoZetta Institute, Sydney*
[c]*Centre for Statistics, The University of Edinburgh*
[d]*Koç University*
[e]*Systemic Risk Centre, London School of Economics*

Abstract

We design and train machine learning models to capture the nonlinear interactions between financial market dynamics and high-frequency trading (HFT) activity. In doing so, we introduce new metrics to identify liquidity-demanding and -supplying HFT strategies. Both types of HFT strategies increase activity in response to information events and decrease it when trading speed is restricted, with liquidity-supplying strategies demonstrating greater responsiveness. Liquidity-demanding HFT is positively linked with latency arbitrage opportunities, whereas liquidity-supplying HFT is negatively related, aligning with theoretical expectations. Our metrics have implications for understanding the information acquisition process in financial markets.

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

* Corresponding author. Emails: gbenga.ibikunle@ed.ac.uk (Gbenga Ibikunle), ben.moews@ed.ac.uk (Ben Moews), khaladdin.rzayev@ed.ac.uk (Khaladdin Rzayev).
We thank Phil Mackintosh and Heinrich Lutjens at NASDAQ for data provision. The data is made available at no cost to academics following the provision of a project description and the signing of a nondisclosure agreement. For helpful comments, we are grateful to Albert J. Menkveld.

1. Introduction

The rise of high-frequency trading (HFT) has dramatically transformed financial markets, introducing extremely high speed and complexity to the trading of financial instruments. High frequency traders (HFTs), whose trading is characterized by rapid order execution, high turnover rates, and sophisticated algorithmic strategies, are often credited with a reduction in trading costs and an enhancement of liquidity. Specifically, if endogenous liquidity-supplying HFTs trade faster than other market participants, they face lower "picking off" (adverse selection) risk and are, therefore, encouraged to provide liquidity (e.g., Hendershott et al. 2011; Brogaard et al. 2015). However, an arbitrage HFT algorithm can use market orders to demand liquidity and in the process impose adverse selection risk on other market participants (e.g., Biais et al. 2015; Foucault et al. 2017). Liquidity-demanding or consuming HFT activity may also exacerbate extreme price movements (e.g., Easley et al. 2011). Thus, the differentiation in the effects of contrasting HFT strategies highlights the conflicting nature of the evidence with respect to the effects of HFT on market quality characteristics. It also underscores the importance of distinguishing between the strategies HFTs deploy in order to accurately assess their effects on market quality and other market phenomena.

A core limitation in the existing empirical literature is its reliance on general trading activity-related metrics to proxy HFT activity. While these proxies are readily available, they vary in their ability to capture variations in liquidity demand and supply, and therefore, using them constrains the ability to investigate the role of HFTs in the evolution of key market phenomena, such as liquidity and price discovery (e.g., Boehmer et al. 2018). To overcome this, a selection of papers uses proprietary data, such as the Investment Industry Regulatory Organization of Canada (IIROC) data for constituents of the S&P/TSX 60 Index, the larger NASDAQ-provided HFT data based for a random selection of 120 NASDAQ- and NYSE-

listed stocks for 2009 and 2010 and the National Stock Exchange of India (NSE) data of 100 stocks in 2015. These proprietary datasets, while detailed, are typically limited in terms of stock and time coverage, making it difficult to conduct a detailed investigation on the wider economic effects of HFT over long periods, or to selectively investigate specific periods in time. Thus, the lack of detailed and comprehensive data presents a significant challenge to fully understanding the impact of HFT activity, especially in terms of strategy differentiation and assessment of their economic outcomes.

We address this challenge by advancing a novel methodology based on the estimation of advanced machine learning (ML) techniques. We design and train ML algorithms using the NASDAQ HFT dataset and additional input variables obtained from the non-proprietary Trade and Quote (TAQ) database. Specifically, consistent with related work in the literature (e.g., Easley et al. 2021; Bogousslavsky et al. 2023), and following a battery of robustness-inducing tests, we employ tree-based ensembles due to their consistent performance, strength in generalization, and reduced requirements for fine-tuning. By using a multi-target approach, we retain correlations between dependent variables as part of the learning process. After an empirical optimization of the model parameters on sample data, we perform repeated random sub-sampling validation, otherwise known as Monte Carlo cross-validation, due to the independence of the chosen number of iterations on the split between training and testing data. This ensures robustness against outlier results, which we further demonstrate by estimating standard deviations across experimental runs. We then deploy the trained ensembles to generate a secondary dataset using input variables from the TAQ data; this process yields a dataset with 9,440,600 stock-day observations for 8,314 US stocks over a sample period spanning January 4th 2010 and October 18th 2023. This implies that we employ an approach that significantly

moves beyond the limits of existing studies on the identification and/or detailed examination of HFT strategies.[1]

We first examine the empirical properties of the relationship between HFT and 24 input variables that capture market quality dynamics, focusing on the interactions generated during the training stage of the ML model. The findings show that the interactions between HFT measures and market quality indicators are nonlinear. For example, HFT measures exhibit an increasingly concave relationship with trading volume. This observation confirms the necessity of employing ML techniques over traditional econometric approaches in this setting, as ML models are better equipped to capture nonlinearities (e.g., Bogousslavsky et al. 2023). Furthermore, the results suggest that liquidity-demanding and -supplying HFT activities respond differently to certain market quality indicators. Specifically, our liquidity-demanding HFT measure shows greater responsiveness to increases in intermarket sweep orders (ISOs) (e.g., Klein 2020). Additionally, while liquidity-demanding HFT activities demonstrate a decreasing and convex pattern with market depth, liquidity-supplying HFT activities exhibit an increasing and concave relationship with market depth (e.g., Goldstein et al. 2023).

We next test the empirical properties of HFT measures generated by ML in the extrapolation stage. The raison d'être for the deployment of HFT strategies is to exploit speed when reacting to the emergence of information, either to update stale quotes in a bid to avoid adverse selection (liquidity-supplying) or to use market orders to snipe stale quotes (liquidity-demanding). Hence, logically, our first test exploits this well-established understanding by investigating the evolution of ML-generated liquidity-demanding and -supplying HFT activity around scheduled and unscheduled information events. We find statistically and economically significant increases in the two types of HFT activity around both scheduled and unscheduled

[1] While HFT strategies are more nuanced and diverse in practice, for the purposes of our study, we adhere to the convention established in the academic literature by categorizing them into two broad classes: liquidity-demanding and liquidity-supplying strategies (e.g., Menkveld 2013; Boehmer et al. 2018). We argue that irrespective of the mechanics of a strategy, its sum effect will be to either demand or supply liquidity.

information events, as predicted. Consistent with the literature (e.g., Brogaard et al. 2014), there are larger increases in liquidity-supplying HFT activity than in the liquidity-demanding variant, underscoring the importance of HFTs to liquidity-provision in modern financial markets.

To strengthen our arguments, we then conduct a natural experiment based on the introduction of a symmetric speed bump, which imposes exogenous limitations on the speed of both liquidity-demanding and -supplying HFTs (e.g., Khapko and Zoican 2021; Aït-Sahalia and Sağlam 2024). As expected, the declines in both liquidity-supplying and -demanding HFT activity in stocks affected by the speed bump are economically significant, with corresponding 3.9% and 1.6% reductions relative to their average trading values, following the imposition of the speed bumps.

While our findings corroborate existing literature by showing higher response rates for liquidity-supplying HFTs compared to liquidity-demanding ones, we also directly test and confirm the accuracy of our measures in capturing these distinct underlying strategies. The extant literature suggests that when latency arbitrage opportunities arise, liquidity-demanding HFTs become more aggressive (e.g., Aquilina et al. 2022) and liquidity-supplying HFTs reduce their liquidity provision (e.g., Foucault et al. 2017). We test this hypothesis and find that our results are in line with these expectations. Specifically, the relationship between liquidity-demanding (liquidity-supplying) HFT activity and volume of latency arbitrage opportunities is positive (negative); a one standard deviation increase in latency arbitrage opportunities is linked with a 1% (1.6%) rise (decrease) in liquidity-demanding (liquidity-supplying) HFT activity.

Our use of an ML-based framework, although increasingly popular in the general finance literature, still begs a question: when is it worth introducing a complicated new technique to the financial economics literature? We argue that it is when it can influence our

understanding of financial theory. For example, as it is in our case, when the new technique allows for the measurement and modelling of an important variable in a new, more flexible and detailed way. To demonstrate this contribution, we characterize information production and acquisition as part of the price discovery process by modelling the role of HFTs in it. Price discovery, one of the two fundamental functions of financial markets (e.g., O'Hara 2003), has been a preoccupation of researchers since Leon Walras's theory of convergence to equilibrium prices through *tâtonnement* in the 19th century (Biais et al. 2005).

Price discovery involves (i) the incorporation of existing information into asset price and (ii) the acquisition of new information (e.g., Brogaard and Pan 2022). Most of the existing studies find that HFTs enhance the efficiency of the price discovery process by increasing the speed of information incorporation into price. However, the research on HFTs' role in information acquisition is limited, perhaps due to data challenges,[2] a problem this paper addresses. The literature suggests that HFTs could either boost information acquisition by increasing liquidity and lowering transaction costs (e.g., Menkveld 2013; Brogaard et al. 2015; Aït-Sahalia and Sağlam 2024) or impair it by aggressively targeting informed institutional investors for profit (e.g., Van Kervel and Menkveld 2019; Yang and Zhu 2020; Hirschey 2021). An earlier attempt by Weller (2018) to investigate the mechanisms of these conflicting effects is constrained by the lack of granularity in the SEC's Market Information Data Analytics System (MIDAS) data the author employs in the paper, since theory suggests the effect of HFTs on information acquisition is a function of the strategies they deploy.

Our ML-generated data is not affected by this constraint since it delineates liquidity-supplying and -demanding HFT activity across an extended time frame for the entire universe of U.S. common stocks. Consistent with theory, we show that liquidity-demanding HFT

[2] Information acquisition proxies are typically low-frequency measures, meaning that datasets covering limited time spans and a narrow selection of stocks, like the NASDAQ HFT data, are not sufficiently comprehensive to meaningfully explore the relationship between HFT and information acquisition.

strategies impede the information production process in financial markets, while liquidity-supplying strategies facilitate it. Furthermore, we demonstrate that the variation in the generic algorithmic trading measures used by Weller (2018) primarily captures liquidity-demanding HFT activities, thereby explaining Weller's (2018) results on the overall negative impact of algorithmic trading on information acquisition.

Our primary target variable is the HFT volume scaled by total trading volume; hence it is the proportion of trading attributable to HFT. Hendershott et al. (2011) emphasize the importance of this normalization in empirical studies examining the role of HFTs in financial markets. However, the approach raises a potential concern that our ML algorithm may primarily capture variations in total volume rather than HFT volume specifically. To address this issue, we also employ ML to predict unscaled HFT trading volume. We then replicate all our analyses using these unscaled values, obtaining consistent results, which are reported in the Online Appendix.

As we present our results, it is important to address the key assumption of our study and its associated limitation. Our ML algorithm is trained on 2009 NASDAQ HFT data, consisting of 120 stocks and 26 proprietary HFT firms. This data reflects a single year and venue, which may introduce some constraints. Specifically, the dataset excludes HFT firms acting as brokers or engaging in lower-frequency proprietary trading (e.g., Goldman Sachs), and it only captures trades executed on NASDAQ in 2009.

Despite these limitations, we believe the key implications of our findings remain robust. First, as noted in Brogaard et al. (2014), HFT firms not included in the NASDAQ HFT dataset are unlikely to adopt strategies significantly different from those identified by NASDAQ. Second, Malceniece et al. (2019) show that HFT trading strategies are strongly correlated across venues, mitigating concerns about the data being limited to NASDAQ. Finally, although our dataset is from 2009, both theoretical and empirical studies indicate that liquidity-

demanding and supplying strategies remain fundamental to HFT, even amidst technological advancements and regulatory changes. 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.

Our study relates to three streams of the literature. First, it is closely related to Boehmer et al. (2018) and Chakrabarty et al. (2023), who evaluate the efficacy of HFT proxies in identifying HFT activity, finding that these proxies can capture both liquidity-demanding and -supplying strategies, though with varying degrees of sensitivity to each. Distinctly, our ML approach utilizes publicly accessible TAQ data to construct HFT metrics that specifically discern liquidity-demanding separately from liquidity-supplying HFT activity. By using latency arbitrage opportunities and information acquisition as testing grounds, we demonstrate that the metrics reflect the unique characteristics of the two broad HFT strategies. Thus, we generate a novel and more granular HFT dataset, which future researchers can utilize to investigate the long-term effects of various HFT strategies on financial markets and the broader economy, and contribute to bridging the gap between theoretical predictions and empirical tests in the HFT literature.

Second, our research enhances the stream of the financial economics literature exploring ML's role in analyzing the microstructure of financial markets. Easley et al. (2021) provide an overview of recent studies and discuss the potential applications of ML in this field. Kwan et al. (2021) investigate the relevance of using reinforcement learning to understand the price discovery process. Bogousslavsky et al. (2023) stand out as particularly relevant to our

work; they suggest employing ML to predict informed trading activity and demonstrate its effectiveness in detecting such activity in financial markets. We contribute to this stream of the literature by generating two ML-computed HFT measures, each capturing a specific HFT strategy. Importantly, our analysis reveals a behavioral pattern of HFTs, which is different to that of traditional informed traders around unscheduled events. Unlike informed traders who typically acquire private information and intensify their trading before such events (e.g., Bogousslavsky et al. 2023), HFTs predominantly utilize public information, leading to an intensification in their trading following unscheduled information events.

Third, our study contributes to the literature on the market quality effects of HFT by providing new insights into two important aspects: the relationship between HFT and the information production process in financial markets, and the impact of a recent market intervention, asymmetric speed bumps, on liquidity. Our ML-generated HFT measures enable us to offer a more nuanced understanding of the relationship between HFT and information acquisition, which is particularly important given it shows the implications of HFT for end-users such as institutional investors (see Van Kervel and Menkveld 2019), and in light of the role of algorithmic trading-sourced price informativeness in corporate investment decisions (see Aliyev et al. 2021; Ye et al. 2023). Moreover, our results on the impact of speed bumps on different HFT strategies suggest that the documented deterioration in market quality around symmetric speed bumps (e.g., Aït-Sahalia and Sağlam 2024) stems from a more pronounced decline in liquidity-supplying HFT activities relative to liquidity-demanding ones, due to the implementation of speed bumps.

2. Data and variables

We use two primary datasets in this study. 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 database 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.[3]

We employ these two datasets in our ML model to generate a secondary dataset, which estimates HFT activity from publicly available TAQ data, spanning from January 4th 2010 to 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 ($HFT_{i,t}^{D}$) and liquidity-supplying ($HFT_{i,t}^{S}$) HFTs. Specifically, $HFT_{i,t}^{D}$ ($HFT_{i,t}^{S}$) 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

We also employ various supplementary datasets to conduct a series of tests of the relevance of the ML-generated HFT data. Intraday transaction details and the associated

[3] Selecting input variables involves trade-offs. On one hand, the input variables should be widely accessible to researchers to allow for the generation of HFT measures over an extended time period and for a comprehensive sample of instruments. On the other hand, incorporating more expensive and significantly less accessible, but more granular data could potentially increase the accuracy of HFT measure predictions. We prioritize the first consideration, as one of the key objectives of our research is to provide HFT metrics for a broad sample, enabling researchers to explore new research questions. Therefore, we rely on the TAQ dataset, which is commonly used in market microstructure studies and offers extensive data coverage for all U.S.-listed stocks. Additionally, as discussed in Section 3.4, our ML model's R^2 exceeds 82% during the testing stage, indicating that our input variables effectively capture the preponderance variation in HFT measures. This suggests that the benefits of employing more granular (e.g., quote) data is limited.

prevailing ask and bid prices are retrieved from Refinitiv DataScope, and earnings and merger and acquisition (M&A) announcement dates are sourced from I/B/E/S and the Thomson Reuters Securities Data Company (SDC) database, respectively. Furthermore, stock return and volume data are acquired from the Center for Research in Security Prices (CRSP).

To jointly test the empirical relevance of our ML-generated HFT metrics and the association between HFT and various market quality measures, we estimate different regression models as specified in their respective sections below. 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 activities ($HFT_{i,t}^{ML,S}$) 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 variability in HFT activity, with demand showing slightly more variability (0.112) than supply (0.101). Furthermore, the comparison of mean and median values for both $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ indicates a right-skewed distribution, suggesting the presence of observations with very high HFT activity. $Spread_{i,t}$ shows a mean of 0.142% with a wide range up to 0.886%, implying diverse liquidity conditions across the sampled stocks. $Volume_{i,t}$ has a high degree of variability (mean: 2.614, max: 47.391).

3. Machine learning and high-frequency trading measures

3.1.Extremely randomized trees as a robust ensemble model.

Our ML methodology exploits ensemble learning, specifically decision trees and random forests due to their robustness and strong performance. An ensemble in supervised ML is 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, or "weak learners." This generally results in a better generalization using 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). Ensemble learning algorithms/models are generally less complex when compared to similarly powerful single-model approaches. Coupled with their strong generalization performance, which allows for accessible tunings, 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).

As one of the best-established supervised ML models, decision trees are hierarchical models built analogous to a flowchart of decision subsets. With independent variables being fed into the tree's root, nodes are split along chosen variables until the end nodes, or leaves, allow for the prediction of values or labels for regression and classification problems, respectively (Breiman et al. 1984). This first iteration of decision trees uses the Gini impurity, which can be written, for a number of possible labels L with respective probabilities p_i with $i \in \{1, 2, \dots, L\}$, as

$$Gini(p) = \sum_{i=1}^{L} p_i(1 - p_i) = 1 - \sum_{i=1}^{L} p_i^2 \quad (1)$$

While the above is a generalization of the Shannon entropy, other versions use the information gain as the mutual information in probability and information theory (see, for example, Quinlan 1986, as the first instance).

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).

These tree-based ensembles make use of bootstrap aggregation, better known as "bagging" in the computer science literature, which generates subsets from the respective data using random draws with replacement. Each tree in the ensemble is then grown with a randomly chosen subset of the data, as well as a subset of the independent variables, to improve stability and accuracy while reducing variance and the risk of overfitting. For classification problems, the majority vote of the predictions is taken, while the arithmetic mean of predictions is calculated for regression problems. These types of models are also commonly used in related research, for example in recent work by Easley et al. (2021) as well as Bogousslavsky et al. (2023).

One of the above-mentioned more recent derivates are extremely randomized trees, generally abbreviated as "extra trees" and introduced by Geurts et al. (2006), which are based on the same random subspace approach as random forests but omit the bagging method. Instead, each weak learner is trained using the whole training data, with nodes being split not through an optimized choice of independent variables but a random selection of the latter. In practice, this is generally done in a mixed approach by selecting an optimal variable from a randomly chosen subset. The resulting models share a similar performance with the older random forest approach, but are computationally less intensive (Biau and Scornet 2016).

For 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,

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

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 are optimized as described in the Section 3.3.

3.2.Comparison to related machine learning predictors.

Before running a full optimization framework on the likely suitable method for the problem at hand, it is prudent to compare contending approaches. As we deal with a regression instead of a classification problem, the suite of potentially suitable machine learning models commonly used for similar prediction problems includes random forests as the similarly performing but computationally slightly more intensive baseline for tree-based ensembles – support vector machines in their regressor variation and feed-forward neural networks with multiple hidden layers.

For these initial experiments to gauge performances, we use best-practice parameters with a fast-enough computation while making use of established developments in the ML literature regarding these models (Bishop and Nasrabadi 2006). We implement the support vector regressor 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.

Our dataset for 2009 contains 29,880 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 are inherently risky due to the assumptions made. Another concern for some of the models are differences in ranges covered by numerical variables, as this can put undue weight on some variables over others. For this reason, scaling is commonly employed as part of the data preprocessing. In contrast to the later results, these initial experiments apply z-score scaling, also commonly called standardization, in which, for a dataset D,

$$z_{D_i} = \frac{D_i - \bar{D}}{\sigma(D)} \quad (3),$$

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) setups when they apply, 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 support vector machines can be used for regression, their functionality requires the multi-model approach. Feed-forward artificial neural networks can handle both, but the complexity of these models would not benefit from simplifying the prediction. The result of this comparison is listed in Table 3. Specifically, Table 3 presents arithmetic mean and standard deviation estimates for R^2 values across 10 iterations for support vector regression/machines, feed-forward artificial neural networks, random forests for multi-model and multi-target setups, and extremely randomized trees for multi-model and multi-target setups.

INSERT TABLE 3 HERE

Here, our model selection covers both the commonly-found twin set of successful tree-based regression approaches (see Section 3.1) as well as a simple baseline model. The latter is represented by a support vector machine as a functionally close cousin to methods such as regularized least-squares and logistic regressions. In doing so, we can benchmark the likely candidate models against a simpler approach to gauge the difficulty of the prediction task. Conversely, a feed-forward neural network, which comes with its own related drawbacks, is a complex comparison model to our ensemble, as explained below.

While the results by themselves are promising, the support vector regressor 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 a more complete parameter optimization with reasonable computational resources.

3.3. Grid-based model tuning and Monte Carlo cross-validation.

Some choices in our experimental setup are the result of computational feasibility. This concerns two parameters, the number of experiment repetitions to gauge consistency through an approximated standard deviation and the number of data points used per experiment. Here, too, the degree of simplicity of the constituent models is an advantage over, for example, various deep learning approaches (Genuer et al. 2017). 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.

To make use of the full dataset available, we sample this number in a uniform-random manner for each experimental iteration and split off 25% as the testing set. Cross-validation, or rotation estimation, are common alternative names for repeated out-of-sample testing to assess the generalization performance of a given method. In ML, the commonly used variation is k-fold cross-validation, 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).

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.

In this case, Monte Carlo cross-validation, which is also more descriptively known as repeated random sub-sampling validation, is the more sensible choice. In each experiment

repetition, multiple random splits into training and testing sets are performed, in our case 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 (see, for example, Li et al. 2010).

Other parameter choices are less clear-cut and thus require optimization. This concerns, in our case, 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 4, in which we use the arithmetic mean and standard deviation of R^2 across repeated iterations to assess the respective model's quality.

INSERT TABLE 4 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.

3.4.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 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.

3.5.Properties of ML – 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 (see also 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 modelling 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.[4] Supporting this, Li 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 analysis corroborates 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

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

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 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 relevance of the ML-generated HFT data, we commence with an exploration of the dynamics of liquidity-demanding ($HFT_{i,t}^{ML,D}$) and liquidity-supplying ($HFT_{i,t}^{ML,S}$) during both scheduled and unscheduled information events. As discussed 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 (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 both $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$, which are expected to capture liquidity-demanding and -supplying HFT activity respectively.

INSERT FIGURE 3 HERE

We first focus on earnings announcements as scheduled events. Panels A and B of Figure 3 show the trajectories of $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ over a 20-day period that extends from ten days before to ten days following the earnings announcement dates. Both measures show an interesting trend, beginning to rise two days before the earnings announcement dates and peaking on the day immediately following the announcements (*t+1*). The spike on the day after the announcement is unsurprising, as many companies choose to release their earnings statements after the regular market hours (e.g., DeHaan et al. 2015), a time when HFTs, given their penchant for ending the day flat, are not traditionally very active, as noted by Bhattacharya et al. (2020) and Chakrabarty et al. (2022).

The figure also presents the average values before and after the announcement date, showing that the average values for both HFT metrics are higher in the post-announcement period. To test the statistical significance of the increases in HFT measures during the announcement period, we compare the average $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ over a three-day announcement window (*t*, *t+1*, and *t+2*) with the averages on other days within our 20-day window. 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 a 3.11% increase in $HFT_{i,t}^{ML,D}$ during the three-day announcement window (from 0.321 to 0.331), which is statistically significant at the 0.01 level. The increase is even more substantial for $HFT_{i,t}^{ML,S}$ at 6.79% (from 0.209 to 0.223), also statistically significant at the 0.01 level.

Earnings announcements, as scheduled events, may have lower information content (e.g., Bogousslavsky et al. 2023). Thus, we extend our analysis to the dynamics of $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ during unscheduled events, with a focus on mergers and acquisitions (M&A) announcements. Exploring HFT dynamics in the context of M&A announcements is interesting and important task because HFTs are primarily recognized not for seeking private information but for their capacity to swiftly process public information for profit, a strategy termed latency arbitrage (e.g., Budish et al. 2015; Aquilina et al. 2022). This indicates that if our ML-generated HFT metrics accurately represent HFT behavior, we should not expect to see a pre-announcement rise in their values before M&A announcements. This is because M&A announcements, unlike scheduled earnings announcements, are typically unexpected, and therefore, there is minimal to no public information available beforehand that would be of interest to HFTs.

INSERT FIGURE 4 HERE

Panels A and B of Figure 4 demonstrate the trends of $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ around M&A announcements. Consistent with our expectations, there is no pre-announcement uptick in $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$; instead, their values start to increase on the day of the announcement and peak the days after. This evolution of our HFT metrics, crucially, draw a sharp contrasts with that of the informed trading intensity (ITI) introduced by Bogousslavsky et al. (2023), which rises ahead of unscheduled events. Hence, while it could be argued that the HFT metrics are characteristically similar to the ITI, this departure in evolution around unscheduled public information events is an important point of differentiation. It suggests that our HFT metrics diverge from ITI because while informed traders exploit private information before unscheduled events, HFTs, consistent with the literature (e.g., Rzayev and Ibikunle 2019), predominantly trade based on public information after such events.

We compare the average $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ during the three-day announcement window to their averages on the remaining days within the 20-day period for M&A announcements. There is a 1.2% increase in $HFT_{i,t}^{ML,D}$ and a 3.2% rise in $HFT_{i,t}^{ML,S}$ during the three-day window, with both increases statistically significant at the 0.01 level.

One of the noteworthy results in this section is the larger increase in liquidity-supplying HFT activities ($HFT_{i,t}^{ML,S}$) compared to liquidity-demanding HFT activities ($HFT_{i,t}^{ML,D}$), during information events, with the former being nearly double the latter. This observation may also corroborate the effectiveness of our ML-based methodology in distinguishing between liquidity demand and supply dynamics. For instance, Brogaard et al. (2014) examine HFT order flows around macroeconomic news announcements and find that liquidity-supplying HFTs' order flow increases more than that of liquidity-demanding HFTs in response to both negative and positive news. This may be attributed to non-HFTs ramping up their aggressive trading in response to the unscheduled information events, this demand for liquidity is then subsequently accommodated by endogenous liquidity-supplying HFTs, who have emerged as a crucial market making force in financial markets (e.g., Menkveld 2013). Supporting this conjecture, Cole et al. (2015) find that during earnings announcements, the rise in aggressive trading by non-HFTs outpaces that of HFTs, with liquidity-supplying HFTs continuing to meet the demand of these aggressive non-HFT traders throughout these periods.

Beyond demonstrating the empirical validity of our ML-generated HFT metrics, the findings in this section also carry significant economic implications. The higher increase in liquidity-supplying HFT activities 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 over a two-year period, or through other proprietary datasets with very short durations and limited samples, our study extends these insights by examining all U.S. listed common stocks over a broad thirteen-year timeframe using publicly available datasets.

4.2.HFT and speed bumps.

Next, we examine the behavior of $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ in response to exogenous shocks impacting HFT activity. To investigate this, we use a natural experiment, the implementation of a symmetric speed bump, which imposes exogenous limitations on the speed of both liquidity-demanding and liquidity-supplying HFT operations (e.g., Khapko and Zoican 2021; Aït-Sahalia and Sağlam 2024). The principle is straightforward: if our metrics indeed reflect HFT activity, they should respond to shocks specific to HFTs.

In January 2017, the NYSE American exchange (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 accurately 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:

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

$$HFT_{i,t}^{ML,S} = \alpha_i + \beta_t + \gamma_2 Post_{i,t} * Amex_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (5)$$

where $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ correspond to ML-generated liquidity-demanding and -supplying HFT activity, respectively. We incorporate stock-specific fixed effects (α_i) and day fixed effects (β_t) to account for individual stock characteristics and daily variations, respectively. $Post_{i,t}$ 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. The standard errors are double clustered by firm and day.

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

A few points of clarification are necessary with respect to the estimating of Models (4) and (5). First, our model does 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. Second, our HFT measures ($HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$) are calculated at the firm level and on a daily basis. Consequently, we do not estimate HFT activity exclusive to the NYSE Amex exchange. This raises the concern that if HFTs trading NYSE Amex-listed stocks choose to reroute their orders from NYSE Amex to other exchanges, it could potentially mitigate the

impact of the speed bump on their overall activity. In this scenario, our HFT measures may not fully respond to the implementation of the speed bump.

Although the possibility of HFTs rerouting orders to alternative exchanges is a legitimate concern, its impact on our findings is expected to be minimal. This expectation stems from the fact that, despite the ability of HFTs to trade NYSE Amex-listed stocks on various platforms, NYSE Amex is often the preferred venue due to its superior market quality for these securities. For example, the end of 2023 statistics indicate that NYSE Amex leads in terms of time spent quoting at the best prices, offering the largest quoted size at the best prices, and featuring the tightest quoted spreads for its listed stocks.[5] Hence, HFTs are incentivized to maintain their trading of NYSE Amex-listed stocks on the NYSE Amex exchange to capitalize on more favourable pricing and market conditions. Consequently, the implementation of the speed bump is expected to influence HFT activity.

INSERT TABLE 5 HERE

The findings reported in Table 5, showing the estimation results for two windows: (i) the 30 days before and after the speed bump's implementation and (ii) the 60 days before and after the speed bump's implementation, are consistent with this expectation. The coefficients of the interaction terms (γ_1 and γ_2) are negative and statistically significant at the 1% and 5% levels for both $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$. These findings hold across both estimation windows. The results indicate a decline in $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ for NYSE Amex-listed companies following the speed bump introduction, compared to those listed on NYSE and NASDAQ. Moreover, the magnitude of the effect is economically substantial. Specifically, post-speed bump, $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ for the NYSE Amex-listed stocks decrease by about 1.6% and 3.9%, respectively, relative to their average values.

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

These results have three implications. First, they validate our ML-generated HFT metrics as effective tools for capturing HFT activities in financial markets. Second, in line with theoretical predictions, speed bumps are linked to a decrease in HFT activity. Therefore, similar to the effects of colocation upgrades (e.g., Brogaard et al. 2015; Boehmer et al. 2021a), speed bumps provide an exogenous shock that can be used to examine the impact of HFT on financial markets. Third, our findings offer additional insights into the findings of Aït-Sahalia and Sağlam (2024), who show that the implementation of the NYSE Amex speed bump leads to wider quoted spreads and reduced liquidity. The study also develops a theoretical framework to understand how changes in speed affect market-making HFT activities. Our analysis expands on their work by demonstrating that the speed bump impacts not only market-making but also market-taking HFTs. However, because market makers experience a more significant effect, the overall influence on liquidity is detrimental. Furthermore, the consistency between our study and Aït-Sahalia and Sağlam (2024) provides early evidence of the capability of our liquidity-demanding and liquidity-supplying HFT metrics to capture the nuances of demand and supply dynamics, which we formally test in the next section.

4.3.HFT and latency arbitrage opportunities.

Thus far, our findings indicate that ML-generated HFT measures effectively capture the general dynamics of HFT activity. In this section, we aim to further dissect the nuanced behaviors of liquidity-demanding and liquidity-supplying HFT activities. Our goal is to examine whether the patterns observed in $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ are congruent with theoretical predictions and existing empirical findings in the literature.

In our analysis, the magnitude of the changes in $HFT_{i,t}^{ML,S}$ around informational events and the implementation of the Amex speed bump has thus far consistently exceeded that of $HFT_{i,t}^{ML,D}$. This observation is in line with existing research. First, it indicates that HFTs often act as net liquidity providers during periods of high information flow (e.g., Brogaard et al.

2014). Second, the introduction of a speed bump leads to wider quoted spreads due to a more pronounced effect 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 practices. In particular, in the wake of latency arbitrage opportunities, we expect an increase in liquidity-demanding HFT behavior, 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:

$$HFT_{i,t}^{ML,D} = \alpha_i + \beta_t + \gamma_1 NLAO_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (6)$$

$$HFT_{i,t}^{ML,S} = \alpha_i + \beta_t + \gamma_2 NLAO_{i,t} + \sum_{k=1}^{4} \delta_{i,t}^{k} C_{i,t}^{k} + \varepsilon_{i,t} \quad (7)$$

where $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$ are the ML-generated liquidity-demanding and -supplying HFT activities, $NLAO_{i,t}$ is the number of latency arbitrage opportunities described below. $C_{i,t}^{k}$ 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 ($Volume_{i,t}$). We also include stock (α_i) and day (β_t) fixed effects, with standard errors are double clustered by stock and day.

We identify latency arbitrage opportunities following the approach outlined by Budish et al. (2015). In their approach, Budish et al. (2015) identify quotes as "stale" by examining the magnitude of changes in mid-prices. They define a quote at time $z-1$ as 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.

To identify a latency arbitrage opportunity, we use the first-level quote data obtained from Refinitiv DataScope. The primary challenge in this process is the substantial volume of data required. Therefore, we narrow our focus to a sample of 120 firms used in training our ML model, specifically those firms included in the original NASDAQ HFT data. We calculate $NLAO_{i,t}$ for these 120 firms across our entire sample period, spanning from 2010 to 2023. As reported in Table 2, the average number of latency arbitrage opportunities per stock-day is 67. The standard deviation is high, at 169, and the maximum value reaching 1211, indicating considerable volatility in the occurrence of these opportunities.

INSERT TABLE 6 HERE

The results, as presented in Table 6. Consistent with our expectations, they 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 $NLAO_{i,t}$ also carries economic significance. 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 (6) and (7), as it is not the primary objective, our results indicate that the relationships between latency arbitrage and various HFT strategies are consistent with the existing 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 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, as highlighted by 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).

5. Application: HFT and information acquisition

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

However, there has been limited research specifically addressing the role of HFTs in the acquisition of new information. Theoretical perspectives suggest that HFTs could either enhance or impair the information acquisition process. On the positive side, HFTs can enhance the information acquisition process by increasing market liquidity and diminishing trading costs through their roles as liquidity providers (e.g., Menkveld 2013; Brogaard et al. 2015; Aït-Sahalia and Sağlam 2024). The rationale behind this is straightforward: lower trading costs elevate the profitability of trading, which, in turn, encourages investors to proactively seek out and exploit new information, thereby facilitating the acquisition and dissemination of new information. Conversely, HFTs might engage in aggressive strategies like order anticipation, including front-running, back-running and latency arbitrage, aiming to anticipate 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. This effect could discourage these investors 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 HFTs have a detrimental effect on the information acquisition process.

Weller (2018) significantly advances the understanding of HFTs and their influence on the information acquisition process. Nevertheless, its reliance on trade and order data from the Securities and Exchange Commission's Market Information Data Analytics System (MIDAS) to evaluate HFT activity suggests that it encounters a notable limitation. While the MIDAS dataset provides valuable foundation for investigating HFT behaviors, it only captures the generic aspects of HFT activity and lacks granularity regarding specific HFT strategies. This is a crucial point because, as theoretical frameworks suggest the impact of HFTs on information acquisition is linked to the nature of their trading strategies. Thus, the use of the MIDAS dataset constrains Weller's (2018) ability to investigate the specific mechanisms that might explain the identified adverse relationship between HFT presence and information acquisition. This gap is recognized in the conclusion of the study (p. 2217), where the author underscores the need for further research *"to assess the precise mechanisms by which improved trading technology reduces the information content of prices."*

We respond to this call, by investigating the role of HFTs in the information acquisition process by deploying our ML-generated measures that distinguish between liquidity-demanding and liquidity-supplying HFT activities. The HFT metrics allow us to explore the

specific strategies of HFTs and their impacts on information acquisition. We estimate the following stock-quarter model:

$$JUMP_{i,q} = \alpha_i + \beta_q + \gamma_1 HFT_{i,q}^{ML,D} + \gamma_2 HFT_{i,q}^{ML,S} + \sum_{k=1}^{5} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t} \quad (8)$$

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 the ML-generated liquidity-demanding and liquidity-supplying HFT activities, respectively. These values are calculated as the quarterly averages of the corresponding daily HFT measures. The control variables $C_{i,t}^{k}$ includes a range of control variables, such as volatility ($Volatility_{i,q}$), relative quoted spread ($Spread_{i,q}$), inverse price ($InvPrice_{i,q}$), market value ($MValue_{i,q}$), and the order imbalance for trades over \$20,000 ($OIB20k_{i,q}$) to capture the price impact of institutional traders. The $OIB20k_{i,q}$ data is directly obtained from the TAQ database. $MValue_{i,q}$ is calculated as the quarterly average of the daily market value, where each day's market value is calculated by multiplying that day's closing price with the total number of outstanding shares. The other control variables represent quarterly averages of the previously mentioned daily versions.

Both $HFT_{i,q}^{ML,D}$ and $HFT_{i,q}^{ML,S}$ are incorporated into the same regression model to evaluate their respective impacts on the information acquisition process. The correlation between these two variables is 0.58, indicating that multicollinearity is not a concern. Considering that a higher $JUMP_{i,q}$ value suggests a decrease in information acquisition, $HFT_{i,q}^{ML,D}$ is expected to be positively linked with $JUMP_{i,q}$. As discussed above, this is because aggressive HFTs may elevate trading costs, thereby potentially impeding the information

acquisition process. Conversely, $HFT_{i,q}^{ML,S}$ is expected to have a negative relationship with $JUMP_{i,q}$, as liquidity-providing HFT strategies tend to reduce trading costs in financial markets, thereby facilitating and enhancing the profitability of acquiring new information.

INSERT TABLE 7 HERE

The estimation results of Model (8), reported in Table 7, are consistent with our expectations. First, liquidity-demanding HFT activities ($HFT_{i,q}^{ML,D}$) and $JUMP_{i,q}$ are positively linked, and the relationship is statistically significant. From an economic perspective, a 1% increase in $HFT_{i,q}^{ML,D}$ is associated with a rise of approximately 6% in $JUMP_{i,q}$. Second, the relationship between $JUMP_{i,q}$ and $HFT_{i,q}^{ML,S}$ is negative and statistically significant; a 1% increase in $HFT_{i,q}^{ML,S}$ is associated with a decrease of approximately 0.7% in $JUMP_{i,q}$. Thus, when assessing the relative impacts of HFT strategies on information acquisition by including both $HFT_{i,q}^{ML,D}$ and $HFT_{i,q}^{ML,S}$ in a single model, liquidity-demanding strategies have a more substantial influence (6% vs 0.7%). This finding aligns with Weller (2018), also, however, offers further insight into the observed positive correlation between the generic HFT measures and $JUMP_{i,q}$ in that study. It demonstrates that among the variety of strategies employed by HFTs, those that demand liquidity are likely to have a more significant economic effect on information acquisition when controlling for market quality factors like the bid-ask spread and volatility. The alignment of our results with those of Weller (2018) further strengthens the empirical validity of our HFT measures.

The association between HFT and the information acquisition process is jointly endogenous due to potential reverse causality and omitted variables. To address this, Weller (2018) employs the logarithm of the average stock price from 42 to 22 days prior to each earnings announcement as an instrumental variable for HFT activity. The rationale for selecting the pre-announcement stock price as an instrument lies in the assumption that, once factors

such as market capitalization, bid-ask spread, and institutional trading are accounted for, the stock price preceding an announcement should not have a direct influence on the process of information acquisition. Weller (2018) provides a detailed discussion on the empirical validity of using the log of the average stock price as an instrument.

We do not repeat these arguments, because the core focus of our study is not an examination of the link between HFT and information acquisition. Our principal contribution lies in the use of ML techniques to identify and quantify the varied strategies adopted by HFTs from publicly accessible data. Within this context, we consider the information acquisition process as a significant area where our ML-derived HFT metrics can be applied. Moreover, the interplay between HFT and information acquisition remains an underexplored area in the existing literature, presenting a significant economic inquiry. Consequently, we do not argue a definitive causal link between HFT activity and the information acquisition process, acknowledging the complexity and potential endogeneity issues in establishing such causality.

Nevertheless, we see merit in using the same instrumental variable approach as Weller (2018) to determine if the relationship between $HFT_{i,q}^{ML,D}$/$HFT_{i,q}^{ML,S}$ and $JUMP_{i,q}$ holds in the alternative framework. To explore this further, we employ the same instrument to capture exogenous variation in $HFT_{i,q}^{ML,D}$ (Model 9) and $HFT_{i,q}^{ML,S}$ (Model 10) independently:

$$HFT_{i,q}^{ML,D} = \alpha_i + \beta_q + \theta_1 LogPrice_{i,q} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t} \quad (9)$$

$$HFT_{i,q}^{ML,S} = \alpha_i + \beta_q + \theta_1 LogPrice_{i,q} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t} \quad (10)$$

where $LogPrice_{i,q}$ is the logarithm of the average stock price from 42 to 22 days prior to earnings announcements. The fitted values of $HFT_{i,q}^{ML,D}$ and $HFT_{i,q}^{ML,S}$, denoted as $\widehat{HFT_{i,q}^{ML,D}}$ and $\widehat{HFT_{i,q}^{ML,S}}$ respectively, are used in the second stage, as defined in model (11) and (12):

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

$$JUMP_{i,q} = \alpha_i + \beta_q + \gamma_1 \widehat{HFT_{i,q}^{ML,S}} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t} \quad (12)$$

This framework differs from the baseline OLS approach in two key aspects. First, to avoid multicollinearity issues arising from the use of a common instrument, $\widehat{HFT_{i,q}^{ML,D}}$ and $\widehat{HFT_{i,q}^{ML,S}}$ are not included together in the same model. Second, the control variable $InvPrice_{i,q}$ is excluded from this analysis due to its high correlation with the instrument. This approach aligns with Weller (2018), where stock prices are also not controlled for in the instrumental variable framework.

INSERT TABLE 8 HERE

The results for the first and second stages of the instrumental variable approach are presented in Panels A and B of Table 8, respectively. In Panel A, we observe a positive correlation between $LogPrice_{i,q}$ and $HFT_{i,q}^{ML,D}$, and a negative relationship between $LogPrice_{i,q}$ and $HFT_{i,q}^{ML,S}$, with both associations being statistically significant. These findings gain additional interest when compared with the results from Weller (2018), where the same instrument shows a positive correlation with generic HFT measures. This suggests that the generic HFT measures in Weller's (2018) study predominantly reflect liquidity-demanding HFT activities. This insight lends further support to the notion that the positive link between overall HFT activity and $JUMP_{i,q}$ observed in Weller (2018) is predominantly attributed to liquidity-demanding strategies.

The second stage results, detailed in Panel B, are in line with our initial hypothesis. Specifically, we find a positive and statistically significant (at the 0.01 level) relationship between $\widehat{HFT_{i,q}^{ML,D}}$ and $JUMP_{i,q}$, indicating that an increase in aggressive HFT strategies is associated with larger price jumps at the time of information release, suggesting less pre-announcement information acquisition. In contrast, $\widehat{HFT_{i,q}^{ML,S}}$ is negatively linked with $JUMP_{i,q}$ and the relationship is statistically significant at the 0.01 level. This implies that higher liquidity

provision by HFTs correlates with smaller price jumps, potentially reflecting more information acquisition before public disclosures.

In summary, the findings discussed in this section validate the concept that HFT strategies have varied impacts on the information acquisition process, in line with theoretical predictions. Liquidity-demanding HFT strategies appear to impede the information acquisition process, whereas liquidity-supplying strategies facilitate it. Moreover, the OLS results suggesting that the negative effects of liquidity-demanding strategies are more pronounced than the positive contributions of liquidity-supplying strategies are a corroboration of Weller's (2018) findings, which suggest an overall detrimental effect of HFTs on the information acquisition process. Additionally, these results explain Weller (2018) by providing empirical evidence regarding the specific mechanisms through which HFT activity is implicated in the deterioration of information acquisition.

6. Validation using unscaled HFT measures

Thus far, we have demonstrated that our 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, this 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 ($UHFT_{i,t}^{ML,D}$) and liquidity-supplying ($UHFT_{i,t}^{ML,S}$) trading volumes, calculated as the sum of HH and HN (HH and NH) volumes for stock i and day t. We then replicate all our tests using these unscaled values.

The results, reported in the Online Appendix, are consistent with the main findings presented in the paper. Specifically, we document that HFT activity increases before and after earnings announcements, as well as after M&A announcements; it is impacted by speed bumps; distinctly reacts to latency arbitrage opportunities; and while liquidity-demanding HFT strategies are negatively correlated with information acquisition, liquidity-supplying strategies are positively associated with it. Thus, these results confirm that scaling target variables does not distort our findings.

7. Conclusion

Investigating the effects of HFT on various market quality characteristics and other phenomena has been one of the most popular research pursuits within the microstructure literature over the past decade and half. And while most empirical studies find that its effects are largely beneficial, characterizing the mechanisms underpinning the reported effects have been hampered by data limitations. In this paper, we propose a novel ML-based method for generating granular HFT data that accounts for the nature of the HFT strategy(ies) generating it. Specifically, our approach uses trained ensembles to estimate liquidity-demanding and -supplying HFT activity from input variables obtained from the non-proprietary TAQ database. Our method yields a secondary HFT data sample of 9,440,600 stock-day observations for 8,314 US stocks.

We validate the data generation method by conducting a series of validation tests designed to observe whether the ML-generated HFT metrics evolve in line with theoretical and evidence-based understanding of HFT activity. We first exploit the well-documented behavior

of endogenous HFT liquidity suppliers and liquidity-demanding and aggressive HFT snipers around the release of scheduled and unscheduled information events to demonstrate the validity of our ML data generation process. Next, we conduct a natural experiment utilizing the exogenous imposition of a speed bump and document results consistent with theoretical expectations. Consistently, the results indicate, in line with the large number of studies showing that HFT is beneficial for market quality, that the response rates of liquidity-supplying HFTs are higher than those of liquidity-demanding HFTs. Thus, we also test the validation of the metrics in relation to the emergence of latency arbitrage opportunities, a context in which liquidity-demanding HFT activity is expected to increase, while liquidity-supplying HFT activity declines. Our results show, as expected that, a statistically and economically meaningful increase in liquidity-demanding HFT activity is linked with a rise in latency arbitrage opportunities, while a decline in liquidity-supplying HFT activity is observed.

Finally, we provide evidence of how our novel HFT data-generation method shapes the understanding of financial theory by investigating the role of HFT activity in the information acquisition aspect of the price discovery process. We obtain results, which explain existing findings in the literature and, crucially, provide clarity on the mechanism through which HFT activity contributes to information acquisition in financial markets.

References

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

Aliyev, N., Huseynov, F., Rzayev, K., 2021. Algorithmic trading and investment-to-price sensitivity. Available at SSRN 3988764

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

Bhattacharya, N., Chakrabarty, B., Wang, X., 2020. High-frequency traders and price informativeness during earnings announcements. Review of Accounting Studies 25, 1156-1199

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

Biais, B., Glosten, L., Spatt, C., 2005. Market microstructure: A survey of microfoundations, empirical results, and policy implications. Journal of Financial Markets 8, 217-264

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25, 197-227

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

Chakrabarty, B., Moulton, P.C., Wang, X.F., 2022. Attention: How high-frequency trading improves price efficiency following earnings announcements. Journal of Financial Markets 57, 100690

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

DeHaan, E., Shevlin, T., Thornock, J., 2015. Market (in) attention and the strategic scheduling and timing of earnings announcements. Journal of Accounting and Economics 60, 36-55

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

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

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

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

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

Quinlan, J.R., 1986. Induction of decision trees. Machine learning 1, 81-106

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

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., Zheng, M.Y., Zhu, W., 2023. The effect of tick size on managerial learning from stock prices. Journal of Accounting and Economics 75, 101515

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.

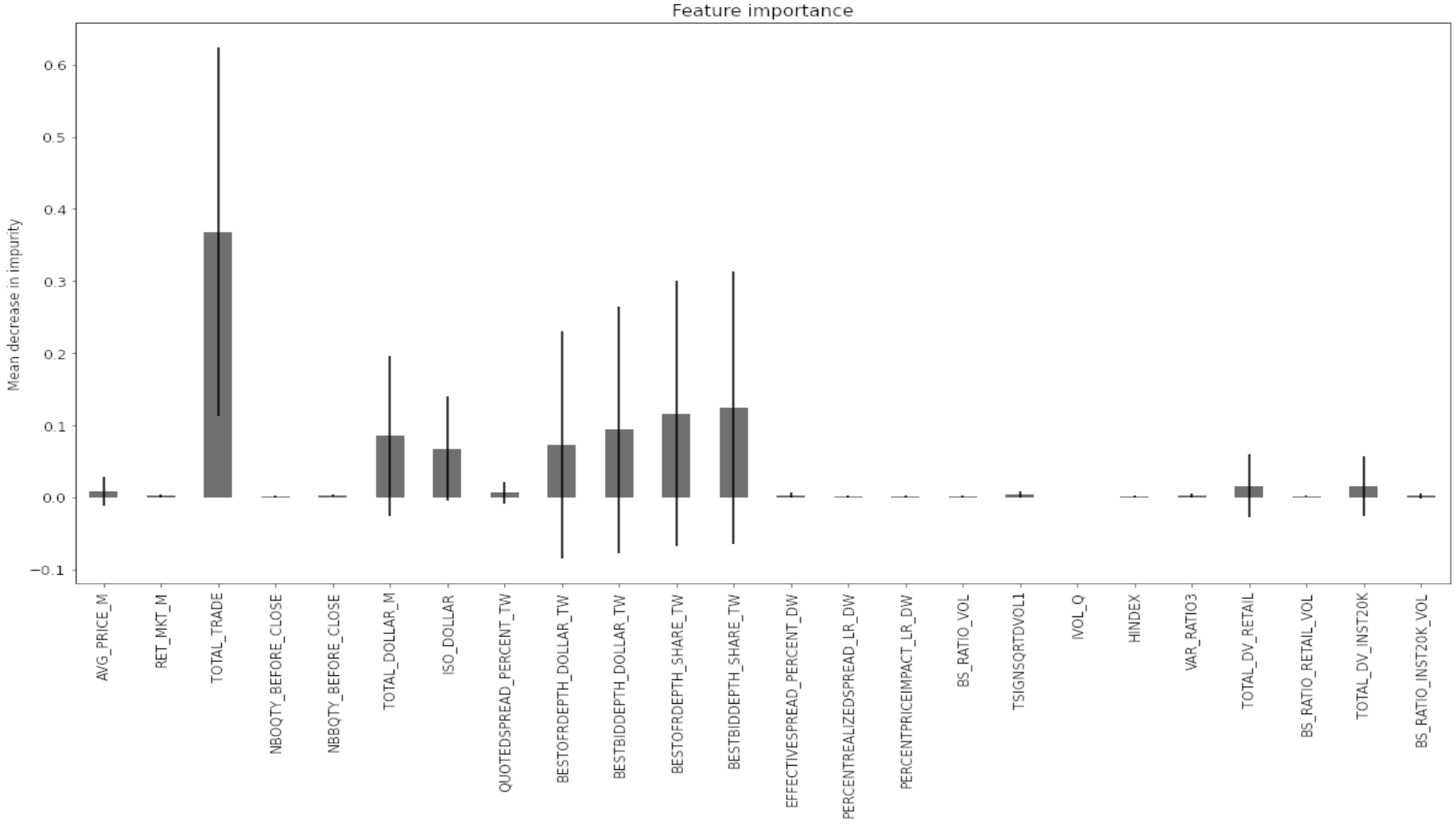


Figure 2

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.

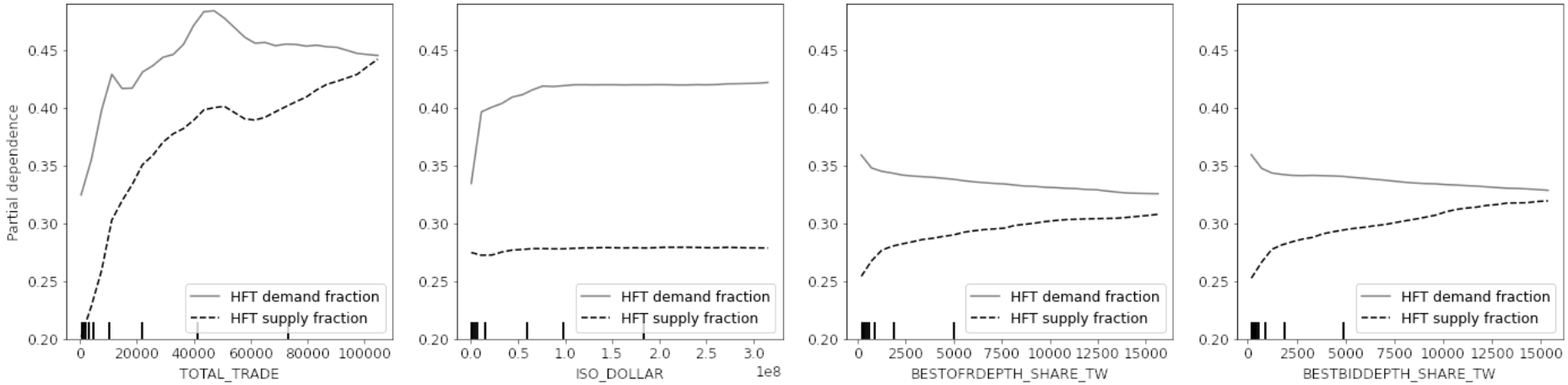
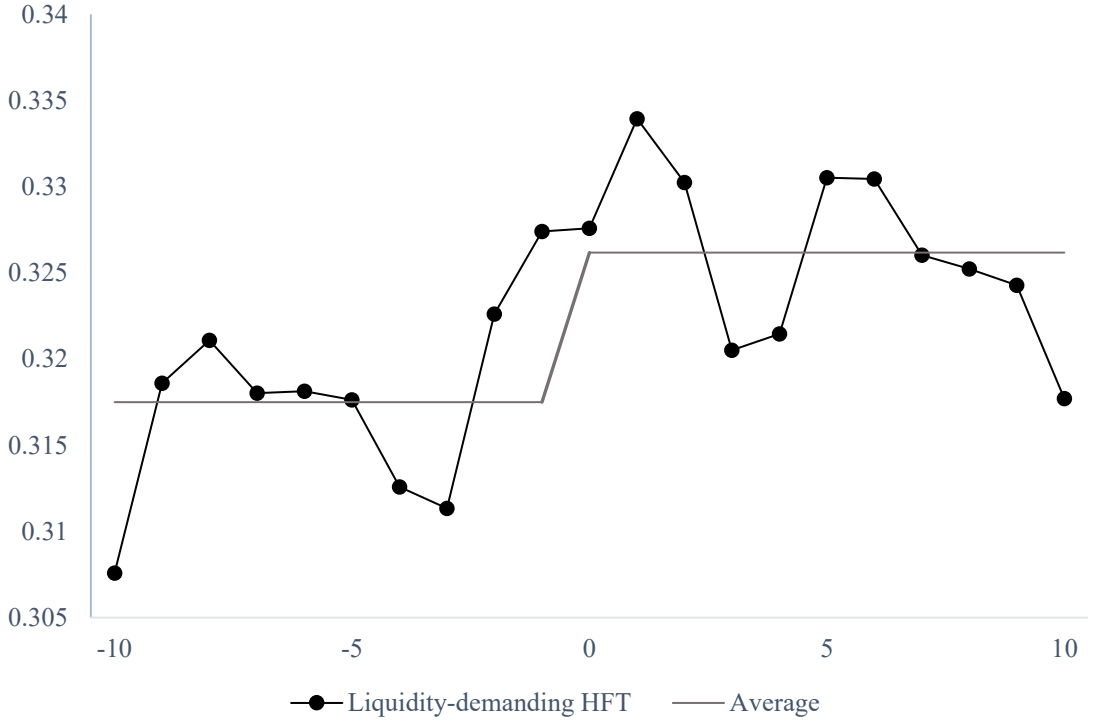


Figure 3
HFT around earnings announcements
This figure illustrates the changes in ML-generated HFT measures surrounding scheduled events, specifically earnings announcements. The event window spans 10 days before and after the announcement dates, which are sourced from the I/B/E/S database. The analysis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

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



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

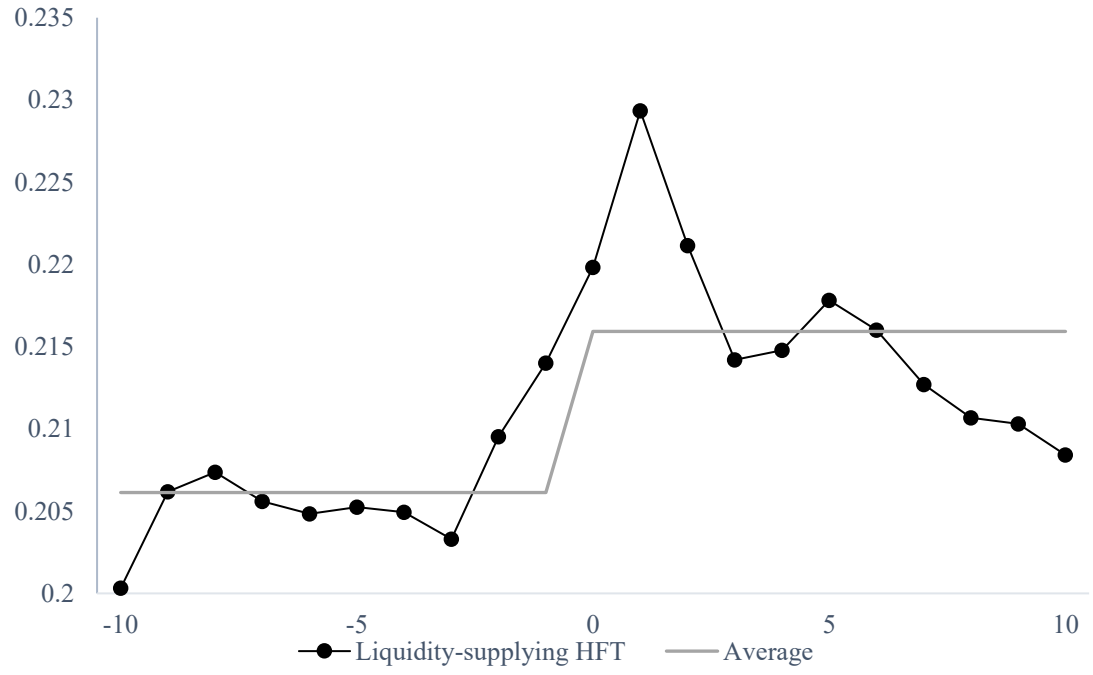
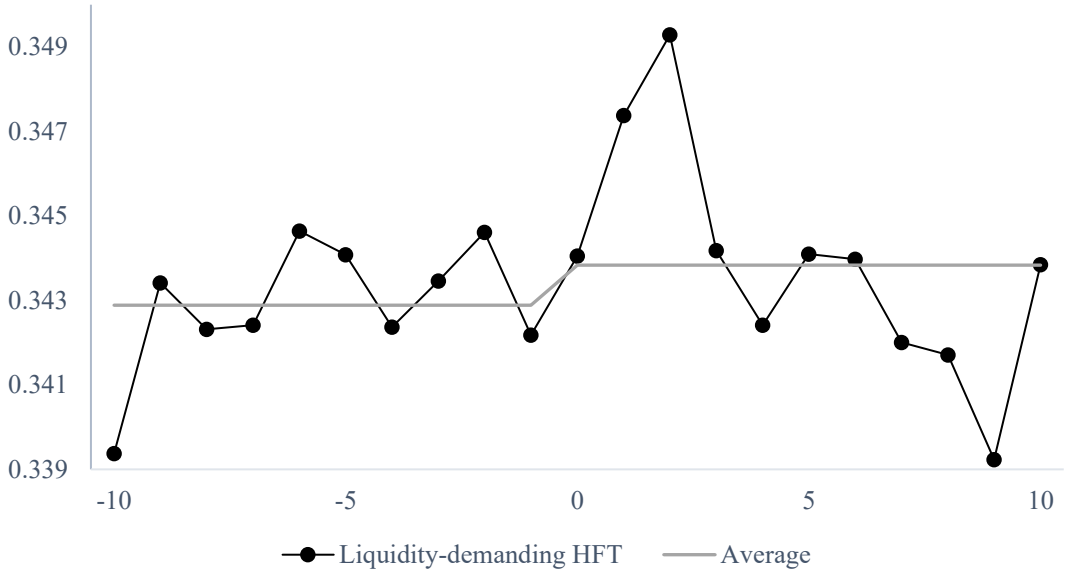


Figure 4
HFT around M&A announcements
This figure illustrates the changes in ML-generated HFT measures surrounding unscheduled events, specifically mergers and acquisitions (M&A) announcements. The event window spans 10 days before and after the announcement dates, which are sourced from the Thomson 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: $HFT_{i,t}^{ML,D}$ around M&A announcements.



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

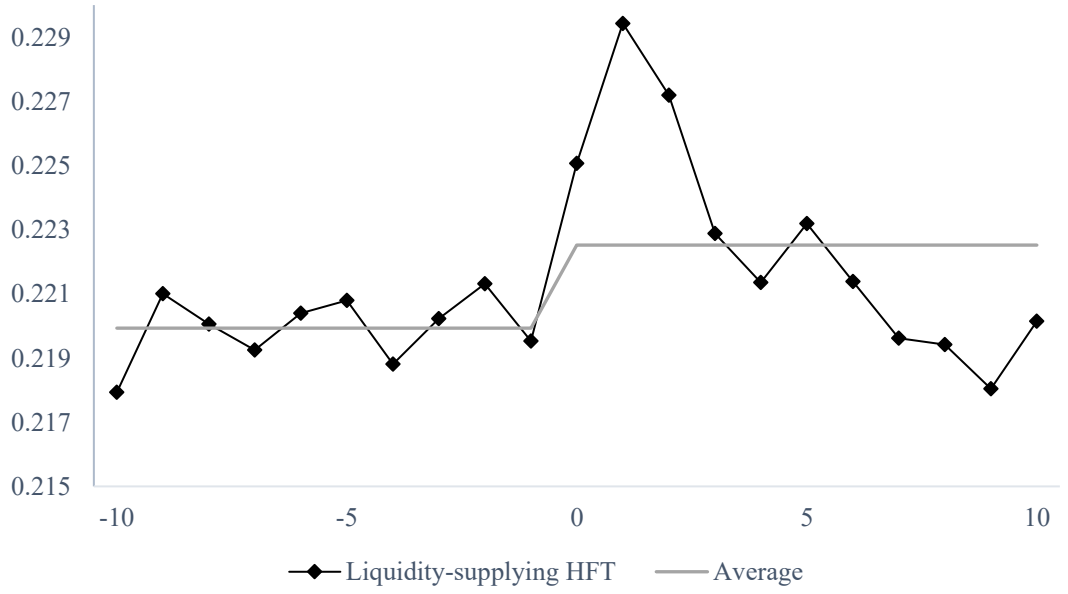


Table 1
Definitions of variables
This table outlines the notation, descriptions, and data sources of the variables used in our study. The input variables for training the machine learning model, derived from the TAQ database, are defined according to the WRDS TAQ Data Manual. To facilitate ease of reference for readers wishing to locate the specific variables within the TAQ database, we use the same variable labels as those specified in the WRDS TAQ Data Manual.

Variable	Description	Data source
Panel B: Output variables used in the ML model.		
$HFT^{D}_{i,t}$	Liquidity-demanding HFT activities for stock i in day t is computed as the daily number of shares traded by liquidity - demanding HFTs (HH and HN) divided by the total number of shares (HH, HN, NH, and NN) trading in day *t*.	NASDAQ HFT
$HFT^{S}_{i,t}$	Liquidity-supplying HFT activities for stock i in day t is computed as the daily number of shares traded by liquidity - supplying HFTs (HH and HN) divided by the total number of shares (HH, HN, NH, and NN) trading in day *t*.	NASDAQ HFT
Panel B: Input variables (features) used in the ML model.		
$AVG_PRICE_M_{i,t}$	Average trade price during market hours (Open to Close) for stock i in day t.	TAQ
$RET_MKT_M_{i,t}$	Open to close return for stock i in day t is computed as the log return of the official opening price over the official closing price.	TAQ
$TOTAL_TRADE_{i,t}$	The total number of trades for stock i in day t.	TAQ
$NBOQTY_BEFORE_CLOSE_{i,t}$	The best offer size of the last quote before market close for stock i in day t.	TAQ
$NBBQTY_BEFORE_CLOSE_{i,t}$	The best bid size of the last quote before market close for stock i in day t.	TAQ
$TOTAL_DOLLAR_M_{i,t}$	The total trade value in dollars during market hours for stock i in day t.	TAQ
$ISO_DOLLAR_{i,t}$	The sum of intermarket sweep order trade dollar value (during market hours) for stock i in day t.	TAQ
$QUOTEDSPREAD_PERCENT_TW_{i,t}$	The time-weighted percentage quoted spread (during market hours) for stock i in day t. The quoted spread is calculated as the difference between ask and bid prices for each transaction divided by the mid-price (the average of ask and bid prices).	TAQ
$BESTOFRDEPTH_DOLLAR_TW_{i,t}$	The time-weighted best offer dollar depth (during market hours) for stock i in day t is determined based on the size of the best ask price.	TAQ

(continued)

$BESTBIDDEPTH_DOLLAR_TW_{i,t}$	The time-weighted best bid dollar depth (during market hours) for stock i in day t is determined as the size of the best bid price.	TAQ
$BESTOFRDEPTH_SHARE_TW_{i,t}$	The time-weighted best offer share depth (during market hours) for stock i in day t is determined based on the size of the best ask price.	TAQ
$BESTBIDDEPTH_SHARE_TW_{i,t}$	The time-weighted best bid share depth (during market hours) for stock i in day t is determined based on the size of the best bid price.	TAQ
$EFFECTIVESPREAD_PERCENT_DW_{i,t}$	The dollar value-weighted percentage effective spread for stock i in day t. The effective spread is calculated using the following equation: $Effective\ Spread = 2D_k(P_k - M_k)/M_k$, where k denotes transaction, D_k denotes the sign of transaction (-1 for sale and +1 for buy), P_k is the transaction price, and M_k is the prevailing mid-price for each transaction. Lee and Ready (1991) algorithm is used for trade classification.	TAQ
$PERCENTREALIZEDSPREAD_LR_DW_{i,t}$	The dollar value-weighted percentage realized spread for stock i in day t. The realized spread is calculated using the following equation: $Realized\ Spread = 2D_k(P_k - M_{k+5})/M_k$, where M_{k+5} is the bid-ask mid-point five minutes after the kth trade, and all other variables are as previously defined. Lee and Ready (1991) algorithm is used for trade classification.	TAQ
$PERCENTPRICEIMPACT_LR_DW_{i,t}$	The dollar value-weighted percentage price impact for stock i in day t. The price impact is calculated using the following equation: $Percent\ Price\ Impact = 2D_k(M_{k+5} - M_k)/M_k$, where all variables are as previously defined. Lee and Ready (1991) algorithm is used for trade classification.	TAQ
$BS_RATIO_VOL_{i,t}$	The absolute percentage order imbalance for stock i in day t is calculated as the absolute value of buy volume minus sell volume divided by the total trade volume. Lee and Ready (1991) algorithm is used for trade classification.	TAQ
$TSIGNSQRTDVOL1_{i,t}$	The lambda (price impact coefficient) with intercept for stock i in day t is calculated using the following equation: $Ln\frac{M_{i,s}}{M_{i,s-300}} = \alpha + \lambda * \text{SSqrtDvol} + \epsilon$, where $\text{SSqrtDvol} = Sgn(\sum_{s-300}^{s} BuyDollar - \sum_{s-300}^{s} SellDollar) \times \sqrt{\lvert\sum_{s-300}^{s} BuyDollar - \sum_{s-300}^{s} SellDollar\rvert}$, 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
Summary statistics
This table reports the summary statistics for the variables incorporated in the main regression models. The units for each variable are indicated in parentheses next to the variable names in the first column. All variables have been winsorized at the 1% level on both tails.

Panel A: Variables list and definitions

Variable	*Variable definition*
$HFT_{i,t}^{ML,D}$	ML – generated liquidity – demanding HFT activities for stock i in day t is estimated by using the ML model described in Section 3.
$HFT_{i,t}^{ML,S}$	ML – generated liquidity – supplying HFT activities for stock i in day t is estimated by using the ML model described in Section 3.
$Volatility_{i,t}$ (1/00,000)	The proxy of volatility is calculated as the daily (t) standard deviation of the transactional-level returns for stock i.
$Spread_{i,t}$ (%)	The proxy for liquidity for stock i in day t calculated as 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.
$InvPrice_{i,t}$	The inverse of the stock price for stock i in day t.
$Volume_{i,t}$ ($'000,000,00)	Trading volume in dollars for stock i in day t.
$NLAO_{i,t}$ (000)	$NLAO_{i,t}$ is the number of latency arbitrage opportunities for stock i in day t. A latency arbitrage opportunity is identified by using the approach described in Section 4.3.
$JUMP_{i,q}$	Proxy for information acquisition for stock i in quarter q calculated as 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
$MValue_{i,q}$ ($'000,000)	The market value for stock i in a quarter q calculated as the quarterly average of the daily market value, where each day's market value is calculated by multiplying that day's closing price with the total number of outstanding shares.
$OIB20k_{i,q}$	The order imbalance for trades over $20,000 for stock i in a quarter q directly obtained from the TAQ database. This is a proxy for institutional traders' order imbalances.

Panel B: Summary statistics

Variable	*Mean*	*Std.*	*Min*	*p.25*	*p.50*	*p.75*	*Max*
$HFT_{i,t}^{ML,D}$	0.316	0.112	0.025	0.222	0.335	0.414	0.602
$HFT_{i,t}^{ML,S}$	0.208	0.101	0.036	0.131	0.174	0.259	0.626
$Volatility_{i,t}$ (1/00,000)	0.008	0.018	0.000	0.001	0.002	0.007	0.123
$Spread_{i,t}$ (%)	0.142	0.154	0.012	0.037	0.090	0.189	0.886
$InvPrice_{i,t}$	0.039	0.050	0.001	0.013	0.024	0.047	0.344
$Volume_{i,t}$ ($'000,000,00)	2.614	6.305	0.007	0.070	0.330	2.556	47.391
$NLAO_{i,t}$ (000)	0.067	0.169	0.001	0.006	0.017	0.046	1.211
$JUMP_{i,q}$	-1.950	5.774	-13.798	-0.943	0.174	0.673	13.991
$MValue_{i,q}$ ($'000,000)	10.304	48.106	0.000	0.583	1.775	5.468	2743.790
$OIB20k_{i,q}$	0.348	0.259	0.015	0.145	0.276	0.495	1.00

Table 3
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 4
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 5
HFT and speed bumps.
This table reports the results for the estimation of the impact of the NYSE Amex speed bump on the HFT activities using the following differences-in-difference (DiD) model:

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

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

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. $Post_{i,t}$ is an indicator variable, taking the value of 1 after the implementation of the speed bump on July 24, 2017, and 0 before and $Amex_{i,t}$ is an indicator variable that equals 1 for NYSE American-listed stocks and 0 for NYSE- and NASDAQ-listed firms. $C_{i,t}^{k}$ 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 determined as 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. Panel A (B) presents the results for $HFT_{i,t}^{ML,D}$ ($HFT_{i,t}^{ML,S}$). Columns (i) and (iii) present the results for the windows of the 30 days before and after the implementation of the speed bump, and Columns (ii) and (iv) present the results for the windows of the 60 days before and after the implementation of the speed bump. The sample includes all the U.S.-listed common stocks. The standard errors used to compute the t-statistics (in brackets) are double clustered by stock and day. *, **, and *** denote the significance at 10%, 5%, and 1%, respectively.

	Panel A: $HFT_{i,t}^{ML,D}$		Panel B: $HFT_{i,t}^{ML,S}$	
	(i) $[-30;+30]$	(ii) $[-60;+60]$	(iii) $[-30;+30]$	(iv) $[-60;+60]$
$Post_{i,t} * Amex_{i,t}$	-0.005***	-0.005**	-0.008***	-0.009***
	(-2.61)	(-2.48)	(-3.80)	(-3.02)
$Volatility_{i,t}$	-0.006***	-0.006***	-0.002	-0.004***
	(-5.08)	(-5.30)	(-1.64)	(-2.76)
$Spread_{i,t}$	-0.010***	-0.016***	-0.009***	-0.010***
	(-3.08)	(-4.22)	(-5.66)	(-5.43)
$InvPrice_{i,t}$	-0.108***	-0.116***	0.025	0.018
	(-4.86)	(-6.75)	(0.79)	(0.55)
$Volume_{i,t}$	0.012***	0.011***	0.024***	0.022***
	(11.35)	(13.48)	(22.68)	(23.52)
Stock and Day FE	Yes	Yes	Yes	Yes
N obs.	101,579	203,241	101,579	203,241
R^2	4%	5%	10%	9%

Table 6
HFT and latency arbitrage opportunities.
This table reports the results for the estimation of the impact of latency arbitrage opportunities on the HFT activities using the following ordinary least squares (OLS) model:

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

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

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. $C_{i,t}^{k}$ 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 determined as 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. Columns (i) and (ii) present the results for $HFT_{i,t}^{ML,D}$ and $HFT_{i,t}^{ML,S}$, respectively. The sample includes 120 randomly selected NASDAQ- and NYSE-listed firms. The standard errors used to compute the t-statistics (in brackets) are double clustered by stock and day. *, **, and *** denote the significance at 10%, 5%, and 1%, respectively.

	(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)	(7.87)
Stock and Day FE	Yes	Yes
N obs.	246,139	246,139
R^2	17%	12%

Table 7
HFT and information acquisition – OLS
This table reports the results for the estimation of the impact of HFT activities on information acquisition using the following ordinary least squares (OLS) model:

$$JUMP_{i,q} = \alpha_i + \beta_q + \gamma_1 HFT_{i,q}^{ML,D} + \gamma_2 HFT_{i,q}^{ML,S} + \sum_{k=1}^{5} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

where $JUMP_{i,q}$ is the ratio of cumulative abnormal returns during trading days [-1, 1] surrounding (quarterly) earnings announcements divided by the cumulative abnormal returns during trading days [-21, 1] surrounding earnings announcements. $HFT_{i,q}^{ML,D}$ and $HFT_{i,q}^{ML,S}$ represent the quarterly averages of daily measures of ML – generated liquidity – demanding ($HFT_{i,t}^{ML,D}$) and – supplying ($HFT_{i,t}^{ML,S}$) HFT activities for stock i. α_i and β_q capture stock and quarter fixed effects, respectively. $C_{i,q}^{k}$ includes a range of control variables, such as volatility ($Volatility_{i,q}$), relative quoted spread ($Spread_{i,q}$), inverse price ($InvPrice_{i,q}$), market value ($MValue_{i,q}$), and the order imbalance for trades over $20,000 ($OIB20k_{i,q}$) to capture the price impact of institutional traders. The $OIB20k_{i,q}$ data is directly obtained from the TAQ database. The $MValue_{i,q}$ is computed as the product of the price and the number of outstanding shares. The other control variables represent quarterly averages of the previously mentioned daily versions. The sample includes all the U.S.-listed common stocks. The standard errors used to compute the t-statistics (in brackets) are double clustered by stock and quarter. *, **, and *** denote the significance at 10%, 5%, and 1%, respectively.

	$JUMP_{i,q}$
$HFT_{i,q}^{ML,D}$	11.709***
	(14.62)
$HFT_{i,q}^{ML,S}$	-1.409***
	(-3.15)
$Volatility_{i,q}$	0.017
	(0.21)
$Spread_{i,q}$	-2.592***
	(-14.30)
$InvPrice_{i,q}$	0.321
	(0.81)
$Volume_{i,t}$	-0.250***
	(-7.19)
$MValue_{i,q}$	0.002***
	(3.35)
$OIB20k_{i,q}$	-0.593***
	(-5.39)
Stock and Quarter FE	Yes
N obs.	110,555
R^2	7%

Table 8
HFT and information acquisition – 2 SLS
This table reports the results for the estimation of the impact of HFT activities on information acquisition using the following two-stage least square (2 SLS) model:

$$HFT_{i,q}^{ML,D} = \alpha_i + \beta_q + \theta_1 LogPrice_{i,q} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

$$HFT_{i,q}^{ML,S} = \alpha_i + \beta_q + \theta_1 LogPrice_{i,q} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

$$JUMP_{i,q} = \alpha_i + \beta_q + \gamma_1 \widehat{HFT_{\iota,q}^{ML,D}} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

$$JUMP_{i,q} = \alpha_i + \beta_q + \gamma_1 \widehat{HFT_{\iota,q}^{ML,S}} + \sum_{k=1}^{4} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

where $LogPrice_{i,q}$ is the logarithm of the average stock price from 42 to 22 days prior to earnings announcements, $JUMP_{i,q}$ is the ratio of cumulative abnormal returns during trading days [-1, 1] surrounding (quarterly) earnings announcements divided by the cumulative abnormal returns during trading days [-21, 1] surrounding earnings announcements. $\widehat{HFT_{\iota,q}^{ML,D}}$ and $\widehat{HFT_{\iota,q}^{ML,S}}$ are the fitted values of $HFT_{i,q}^{ML,D}$ and $HFT_{i,q}^{ML,S}$, where $HFT_{i,q}^{ML,D}$ and $HFT_{i,q}^{ML,S}$ represent the quarterly averages of daily measures of ML – generated liquidity – demanding ($HFT_{i,t}^{ML,D}$) and – supplying ($HFT_{i,t}^{ML,S}$) HFT activities for stock i. α_i and β_q capture stock and quarter fixed effects, respectively. $C_{i,q}^{k}$ includes a range of control variables, such as volatility ($Volatility_{i,q}$), relative quoted spread ($Spread_{i,q}$), market value ($MValue_{i,q}$), and the order imbalance for trades over $20,000 ($OIB20k_{i,q}$) to capture the price impact of institutional traders. The $OIB20k_{i,q}$ data is directly obtained from the TAQ database. The $MValue_{i,q}$ is computed as the product of the price and the number of outstanding shares. The other control variables represent quarterly averages of the previously mentioned daily versions. Panels A and B present the results for the first and second stages, respectively. The sample includes all the U.S.-listed common stocks. The standard errors used to compute the t-statistics (in brackets) are double clustered by stock and quarter. *, **, and *** denote the significance at 10%, 5%, and 1%, respectively.

Panel A: First stage results		
	(i) $HFT_{i,t}^{ML,D}$	(ii) $HFT_{i,t}^{ML,S}$
$LogPrice_{i,q}$	0.035***	-0.021***
	(26.58)	(-12.07)
$Controls$	Yes	Yes
Stock and Quarter FE	Yes	Yes
N obs.	82,041	82,041
R^2	33%	12%
Panel B: Second stage results		
	(i) $JUMP_{i,q}$	(ii) $JUMP_{i,q}$
$\widehat{HFT_{\iota,q}^{ML,D}}$	3.466***	
	(3.28)	
$\widehat{HFT_{\iota,q}^{ML,S}}$		-10.185***
		(-6.07)
$Controls$	Yes	Yes
Stock and Quarter FE	Yes	Yes
N obs.	82,041	82,041
R^2	7%	7%

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

September 2024

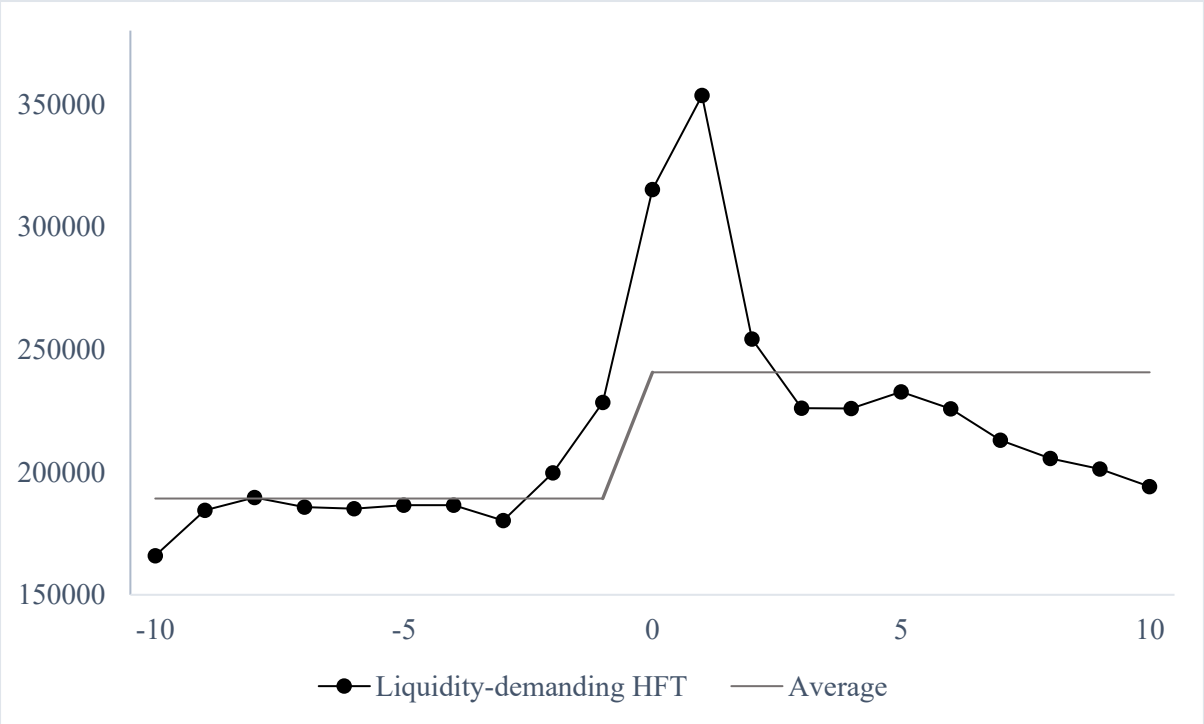
Introduction

This online appendix provides supplementary results to the findings presented in Ibikunle et al. (2024). 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

Figure OA.1
HFT around earnings announcements
This figure illustrates the changes in ML-generated unscaled HFT measures 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: $UHFT_{i,t}^{ML,D}$ around earning announcements.



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

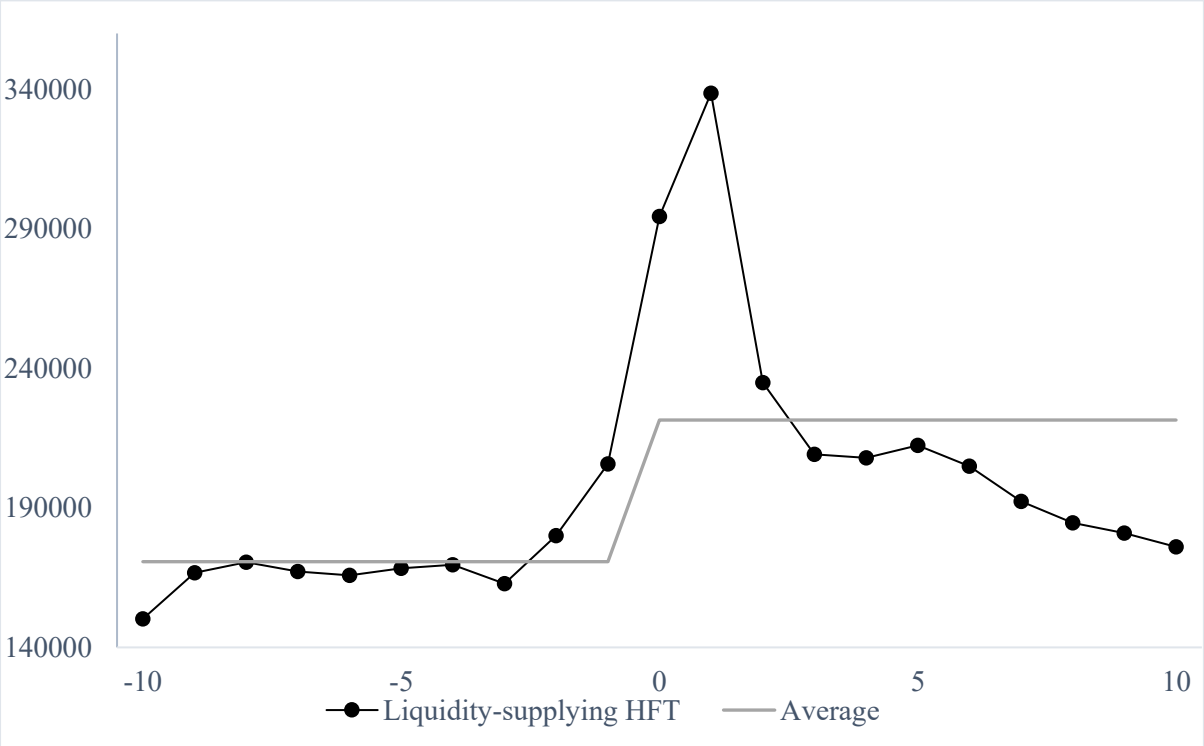
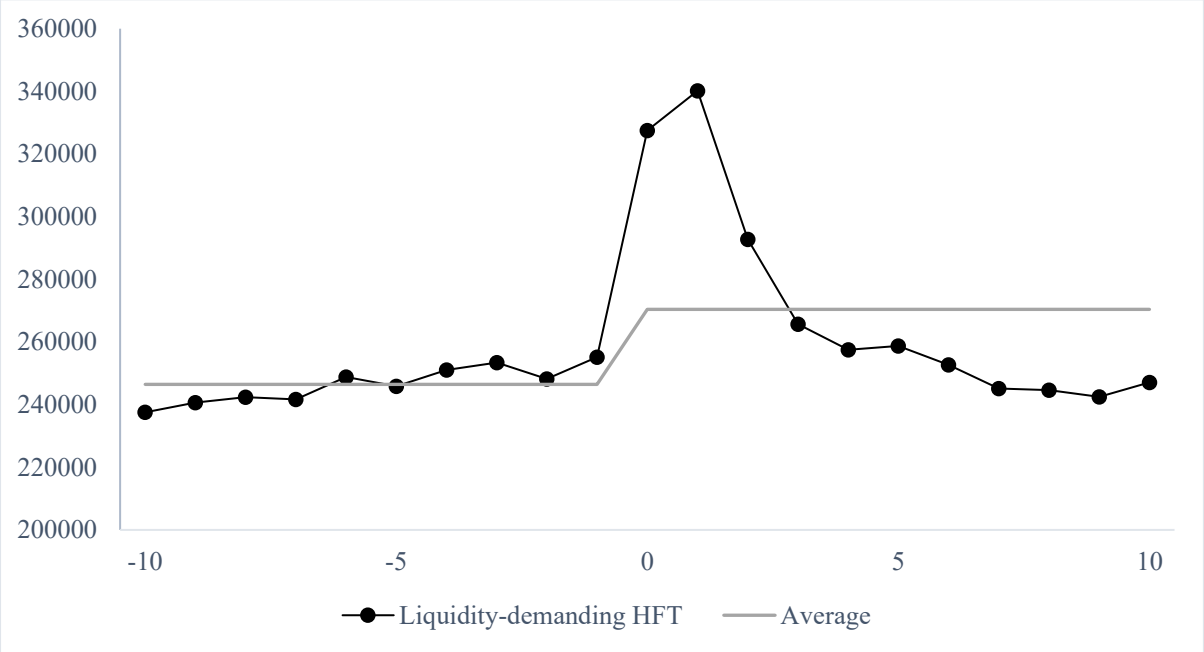


Figure OA.2
HFT around M&A announcements
This figure illustrates the changes in ML-generated unscaled HFT measures surrounding unscheduled events, specifically mergers and acquisitions (M&A) announcements. The event window spans 10 days before and after the announcement dates, which are sourced from 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: $UHFT_{i,t}^{ML,D}$ around M&A announcements.



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

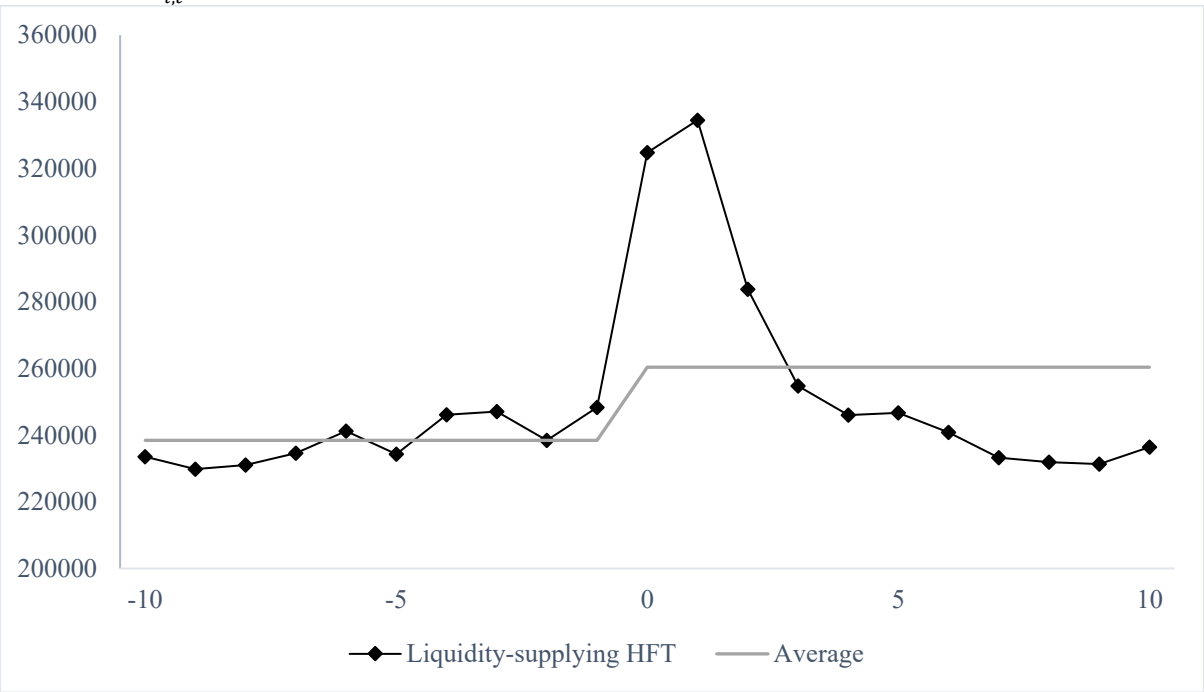


Table OA.1
HFT and speed bumps.
This table reports the results for the estimation of the impact of the NYSE Amex speed bump on the HFT activities using the following differences-in-difference (DiD) model:

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

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

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 and day t. α_i and β_t capture stock and day fixed effects, respectively. $Post_{i,t}$ is an indicator variable, taking the value of 1 after the implementation of the speed bump on July 24, 2017, and 0 before and $Amex_{i,t}$ is an indicator variable that equals 1 for NYSE American-listed stocks and 0 for NYSE- and NASDAQ-listed firms. $C_{i,t}^{k}$ 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 determined as 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. Panel A (B) presents the results for $UHFT_{i,t}^{ML,D}$ ($UHFT_{i,t}^{ML,S}$). Columns (i) and (iii) present the results for the windows of the 30 days before and after the implementation of the speed bump, and Columns (ii) and (iv) present the results for the windows of the 60 days before and after the implementation of the speed bump. The sample includes all the U.S.-listed common stocks. The standard errors used to compute the t-statistics (in brackets) are double clustered by stock and day. *, **, and *** denote the significance at 10%, 5%, and 1%, respectively.

	Panel A: $UHFT_{i,t}^{ML,D}$		Panel B: $UHFT_{i,t}^{ML,S}$	
	(i) $[-30;+30]$	(ii) $[-60;+60]$	(iii) $[-30;+30]$	(iv) $[-60;+60]$
$Post_{i,t} * Amex_{i,t}$	-0.684***	-0.434***	-0.729***	-0.531***
	(-3.02)	(-3.02)	(-2.91)	(-2.91)
Controls	Yes	Yes	Yes	Yes
Stock and Day FE	Yes	Yes	Yes	Yes
N obs.	101,579	203,241	101,579	203,241
R^2	35%	34%	20%	21%

Table OA.2
HFT and latency arbitrage opportunities.
This table reports the results for the estimation of the impact of latency arbitrage opportunities on the HFT activities using the following ordinary least squares (OLS) model:

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

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

where $UHFT_{i,t}^{ML,D}$ and $UHFT_{i,t}^{ML,S}$ represent the ML – generated unscaled liquidity – demanding and – supplying HFT activities for stock i and day t. α_i and β_t capture stock and day fixed effects, respectively. $NLAO_{i,t}$ is the number of latency arbitrage opportunities. $C_{i,t}^{k}$ 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 determined as 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. Columns (i) and (ii) present the results for $UHFT_{i,t}^{ML,D}$ and $UHFT_{i,t}^{ML,S}$, respectively. The sample includes 120 randomly selected NASDAQ- and NYSE-listed firms. The standard errors used to compute the t-statistics (in brackets) are double clustered by stock and day. *, **, and *** denote the significance at 10%, 5%, and 1%, respectively.

	(i) $UHFT_{i,t}^{ML,D}$	(ii) $UHFT_{i,t}^{ML,S}$
$NLAO_{i,t}$	66.266***	-150.517**
	(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
HFT and information acquisition – OLS
This table reports the results for the estimation of the impact of HFT activities on information acquisition using the following ordinary least squares (OLS) model:

$$JUMP_{i,q} = \alpha_i + \beta_q + \gamma_1 UHFT_{i,q}^{ML,D} + \gamma_2 UHFT_{i,q}^{ML,S} + \sum_{k=1}^{5} \delta_{i,q}^{k} C_{i,q}^{k} + \varepsilon_{i,t}$$

where $JUMP_{i,q}$ is the ratio of cumulative abnormal returns during trading days [-1, 1] surrounding (quarterly) earnings announcements divided by the cumulative abnormal returns during trading days [-21, 1] surrounding earnings announcements. $UHFT_{i,q}^{ML,D}$ and $UHFT_{i,q}^{ML,S}$ represent the quarterly averages of daily measures of ML – generated unscaled liquidity – demanding ($UHFT_{i,t}^{ML,D}$) and – supplying ($UHFT_{i,t}^{ML,S}$) HFT activities for stock i. α_i and β_q capture stock and quarter fixed effects, respectively. $C_{i,q}^{k}$ includes a range of control variables, such as volatility ($Volatility_{i,q}$), relative quoted spread ($Spread_{i,q}$), inverse price ($InvPrice_{i,q}$), market value ($MValue_{i,q}$), and the order imbalance for trades over \$20,000 ($OIB20k_{i,q}$) to capture the price impact of institutional traders. The $OIB20k_{i,q}$ data is directly obtained from the TAQ database. The $MValue_{i,q}$ is computed as the product of the price and the number of outstanding shares. The other control variables represent quarterly averages of the previously mentioned daily versions. The sample includes all the U.S.-listed common stocks. The standard errors used to compute the t-statistics (in brackets) are double clustered by stock and quarter. *, **, and *** denote the significance at 10%, 5%, and 1%, respectively.

	$JUMP_{i,q}$
$UHFT_{i,q}^{ML,D}$	0.008***
	(2.88)
$UHFT_{i,q}^{ML,S}$	-0.006***
	(-3.12)
Controls	Yes
Stock and Quarter FE	Yes
N obs.	110,555
R^2	6%