Detecting Informed Trading Risk: A Microstructure Approach Using Machine Learning and Signal Processing

Guillaume Bru, Chris Lu, Cathy Shan, Leila Zhou

Abstract

This paper investigates whether high-frequency market microstructure signals can predict short-horizon price impact and facilitate real-time detection of informed trading risk. Using millisecond-level trade and quote data from TAQ for ten U.S. equities between September and November 2011, we construct two core microstructure measures: the Quote Improvement-to-Deterioration ratio (QID), capturing the balance between competitive quoting and trade-driven quote deterioration, and the Book Exhaustion Rate (BER), quantifying the speed at which market orders deplete top-of-book liquidity. We forecast five-minute-ahead absolute price impact using ordinary least squares, autoregressive benchmarks, tree-based ensembles, and recurrent neural networks, augmented with standard liquidity controls and Fourier-based spectral features.

Our analysis yields three main findings. First, among the forecasting approaches we evaluate, tree-based and linear models consistently deliver the strongest and most stable out-of-sample performance, while deep learning methods are more constrained by data limitations and noise in our high-frequency setting. Second, we show that microstructure-based measures such as QID and BER provide incremental information about short-horizon liquidity and informed trading pressure beyond traditional controls. Third, we find that model-agnostic feature enhancements are statistically meaningful and improve predictive accuracy across multiple specifications.

We gratefully acknowledge the guidance of Prof. Leonid Kogan as our project mentor and advisor. We also thank Prof. Hui Chen and Prof. David Thesmar for their instruction in 15.S06, Xin Xiong for exceptional TA support, and Dr. Yashar H. Barardehi for his valuable help regarding the QID and QIDres metrics.

1 Introduction

Informed trading leaves clear footprints in market microstructure, systematically shifting liquidity conditions in real time. When traders possess private information about an asset's fundamental value, their activity generates predictable patterns in order book dynamics, widening spreads, depleting displayed depth, and altering the balance between competitive quote improvements and liquidity-consuming trades. These microstructure signatures manifest most directly in short-horizon price impact, which captures the immediate price response to trading activity and serves as a widely-used proxy for both informed trading intensity and prevailing liquidity conditions.

The ability to detect informed trading risk in real time carries substantial practical importance across multiple market participants. Market makers must continuously assess the toxicity of incoming order flow to manage adverse selection costs and adjust their quoting strategies accordingly. Institutional investors seek to minimize the market impact of their executions by timing trades to avoid periods of elevated information asymmetry. Regulators require tools to identify potentially manipulative or destabilizing trading patterns as they emerge. Despite this broad demand, existing approaches often rely on end-of-day or lower-frequency measures that cannot capture the rapid evolution of information asymmetry within the trading day.

This paper investigates the following research question: To what extent can microstructure-based measures predict short-horizon price impact and thereby help detect informed trading risk in real time? We focus on two recently-proposed metrics that directly capture order book dynamics: the Quote Improvement-to-Deterioration ratio (QID), which measures the relative frequency of competitive quote improvements versus trade-driven quote deteriorations, and the Book Exhaustion Rate (BER), which quantifies how rapidly incoming market orders consume displayed liquidity at the top of the book. Both measures are computable in real time from standard trade and quote data and possess clear economic interpretations grounded in market-making theory.

Our empirical approach combines these microstructure signals with traditional liquidity controls and applies a range of predictive models from ordinary least squares regressions and autoregressive specifications to tree-based ensemble methods (Random Forest and Gradient Boosting) and deep recurrent neural networks (LSTM and GRU). By systematically comparing model performance across this spectrum, we assess not only whether QID and BER contain incremental predictive content for price impact, but also whether complex nonlinear architectures can extract additional signal from the temporal evolution of order flow.

2 Literature Review

2.1 QID and QIDres

Barardehi et al. (2024) use the Quote-Improvement-to-Deterioration ratio (QID) as a quote-based measure linked to informed trading risk. QID is constructed from sequences of NBBO quote updates: quote improvements arise when liquidity providers undercut each other by posting better prices, while deteriorations occur when trades consume the best quote and move the NBBO outward. QID is defined as the ratio of the number of improvement events to the number of trade-driven deteriorations within a short interval. Intuitively, high QID reflects active undercutting and competition for queue priority, when market makers view incoming order flow as relatively uninformed, whereas low QID indicates reluctance to undercut, consistent with elevated toxicity or adverse-selection risk.

Because QID is mechanically related to liquidity conditions (e.g., spreads and quoted depth), the authors also propose a residualized version, QIDres, which removes predictable variation explained by standard liquidity proxies. After regressing QID on relative spreads at the stock-quarter level, the remaining component is scaled and sign-adjusted so that higher QIDres corresponds to higher informed-trading risk. Empirically, QIDres spikes around events and periods where information asymmetry is known to increase, reinforcing its interpretation as an informed-trading indicator.

In our setting, QID provides a high-frequency measure of quoting behavior that directly reflects market makers' willingness to compete for flow, while QIDres offers the economic foundation explaining why quote-improvement dynamics contain information about informed trading. Both metrics therefore connect naturally to our objective of predicting short-horizon price impact as a proxy for informed trading risk.

2.2 Book Exhaustion Rate

We follow Zhao and Linetsky (2021) in using the Book Exhaustion Rate (BER) as another of our high-frequency features. BER measures how fast market order flow is depleting liquidity at the top of the limit order book. Over a look-back window w, the offer-side BER is defined as the average volume of buy market orders executed per second divided by the time-weighted average size of the best ask queue in $[t-w, t)$; the bid-side BER is defined analogously with sell market orders and best bid depth. A two-sided BER uses total market-order volume over total top-of-book depth. Intuitively, BER is the fraction of the queue consumed per second: a value of 0.5 means that, at the current pace, half of the available depth at the best quote would be removed in one second, while 1/BER can be interpreted as the expected lifetime

of the top of the book if the current flow persists and no new limit orders arrive.

The economic interpretation is that BER provides a real-time measure of order-flow toxicity for market makers. In equilibrium, the expected adverse-selection loss per second is approximately one-half of the spread multiplied by the fill probability per second. Since the fill probability is proportional to how quickly the queue ahead of a quote is eaten, BER naturally scales this loss: higher BER implies more frequent adverse fills and therefore a higher expected loss rate for passive liquidity providers. In this sense, BER translates purely mechanical order-book dynamics, how fast trades chew through the best bid or ask, into an interpretable risk metric.

2.3 Amihud Illiquidity Measure

We draw on Amihud (2002) in incorporating the Amihud illiquidity ratio as a measure of the price response to trading volume. Amihud's key insight is that illiquid assets exhibit disproportionately large absolute returns for a given amount of traded volume, reflecting a steep price–liquidity trade-off: when liquidity is thin, even small trades can induce large price movements.

Formally, over a window $[t-w, t)$, the Amihud measure is defined as

$$\mathrm{ILLIQ}(t, w) = \frac{1}{w} \sum_{\tau=t-w}^{t-1} \frac{|r_\tau|}{\mathrm{Volume}_\tau}, \tag{1}$$

where r_τ is the return and Volume_τ is dollar trading volume on interval τ. The statistic captures the average absolute return per unit of volume, providing a reduced-form estimate of the price impact of trades.

Economically, ILLIQ measures the cost of immediacy: in liquid markets, additional trades can be absorbed with minimal price adjustment, leading to low values of ILLIQ; in illiquid markets, comparable trading activity requires larger concessionary price moves. Amihud (2002) documents that this measure is persistent and positively priced in the cross section, indicating that assets with higher illiquidity earn higher expected returns as compensation for bearing liquidity risk.

From a microstructure perspective, ILLIQ aggregates several channels that contribute to illiquidity—limited depth, asymmetric information, and temporary supply–demand imbalances. The Amihud measure provides an interpretable gauge of how aggressively prices respond to contemporaneous trading flow.

3 Data and Preprocessing

3.1 Data Sources

Our empirical analysis draws on millisecond-level trade and quote data from the Wharton Research Data Services (WRDS) Trade and Quote (TAQ) database. We utilize three complementary data sources to construct our microstructure features and prediction targets.

The first source is the TAQ Consolidated Trades dataset, which provides transaction-level records including:

- Transaction prices
- Trade sizes
- Trade correction indicators

These data allow us to identify the timing and magnitude of executed trades, which form the basis for computing price impact and classifying trade direction.

The second source is the TAQ Consolidated Quotes dataset, which contains millisecond-stamped updates to the National Best Bid and Offer (NBBO). Each record includes:

- Best bid and ask prices
- Quote sizes
- Quote condition and validity indicators

These data are essential for constructing our spread, depth, and quote-improvement measures.

The third source is the NYSE Official SIP NBBO dataset, which provides the authoritative national best bid and offer as disseminated by the Securities Information Processor. This dataset serves as our primary reference for determining the prevailing quotes against which trades are classified and price impact is measured.

Our sample covers a two-month period from September 12, 2011 to November 18, 2011. We focus on ten companies randomly selected to represent a diverse cross-section of industries:

- Caterpillar (CAT) – Industrials
- Johnson & Johnson (JNJ) – Healthcare
- IBM – Technology

- Microsoft (MSFT) – Technology
- ExxonMobil (XOM) – Energy
- Coca-Cola (KO) – Consumer Staples
- Citigroup (C) – Financials
- Amazon (AMZN) – Consumer Discretionary
- McDonald's (MCD) – Consumer Discretionary
- General Electric (GE) – Industrials

After cleaning and aggregation, our final dataset comprises approximately 35,000 five-minute observations across all stocks.

3.2 Data Cleaning and Filtering

We apply a systematic cleaning procedure to ensure data quality and eliminate observations that could distort our microstructure measures. The cleaning process addresses three categories of potential data issues, as summarized in Table 1.

Table 1: Data Cleaning Procedures

Filter Type	**Procedure**
Time Filter	Keep trades and quotes from 9:45 a.m. – 3:45 p.m. to avoid excessive volatility in opening/closing periods
Quote Filter	• Keep only active bids/asks (with valid condition) • Drop canceled or zero-size quotes • Remove quotes with spread > \$5 or bid/ask deviating > \$2.50 from the previous midpoint
Trade Filter	Keep only original trades (exclude corrections and cancellations)

Temporal Filtering. We restrict our analysis to the core trading hours between 9:45 a.m. and 3:45 p.m. Eastern Time. This window excludes the opening and closing periods of the trading day, which are characterized by elevated volatility, auction mechanisms, and

non-representative price dynamics. The opening period (9:30–9:45 a.m.) often exhibits price discovery effects as overnight information is incorporated, while the closing period (3:45–4:00 p.m.) reflects end-of-day portfolio rebalancing and index-related trading.

Quote Filtering. We apply several filters to the consolidated quote data to retain only valid, economically meaningful observations. First, we keep only quotes with active bid and ask indicators, as indicated by the quote condition field. Second, we remove any quotes that have been subsequently canceled or corrected, as well as quotes with zero size on either side of the book. Third, we eliminate quotes exhibiting anomalous spreads or price levels: specifically, we drop observations where the bid-ask spread exceeds \$5.00 or where either the bid or ask price deviates by more than \$2.50 from the previous midpoint.

Trade Filtering. For the trade data, we retain only original trades and exclude any records flagged as corrections, cancellations, or administrative entries. This ensures that each transaction in our sample represents a genuine executed order rather than a post-hoc adjustment to the tape.

3.3 Standardization

Before estimating any models, we standardize both the predictors and the target to place variables on comparable scales and to avoid inadvertently introducing information from the future.

3.3.1 Feature Standardization (X)

The predictors are standardized using a rolling volatility adjustment applied separately to each stock. All observations are first ordered chronologically, and for each feature we compute a 60-observation rolling standard deviation that uses only past values within the window.

$$X_{i,t}^{\text{scaled}} = \frac{X_{i,t}}{\sigma_{i,t}^{(60)}} \tag{2}$$

Each feature value is then divided by this rolling volatility. This procedure ensures that the standardization respects the time ordering of the data: at any point t, the model has access only to the volatility information that would have been available at time t. As a result, the feature scaling is free of look-ahead bias.

3.3.2 Target Standardization (Y)

The target variable, the next-interval price impact, is scaled using a different logic because it is never used as an input to predict itself. We begin by splitting the dataset into an 80% training sample and a 20% test sample. The standard deviation of the target is computed using only the training data, and both training and test observations are divided by this training-set statistic. The models are estimated in this scaled space, and predictions are converted back to original units by multiplying by the same training-set standard deviation.

$$y_t^{\text{scaled}} = \frac{y_t}{\sigma_{\text{train}}} \tag{3}$$

$$\hat{y}_t = \hat{y}_t^{\text{scaled}} \cdot \sigma_{\text{train}} \tag{4}$$

We use different procedures for X and Y because they serve different roles in the forecasting problem. The predictors must be standardized in a way that preserves the temporal information set; any use of future data in their normalization would immediately contaminate the model through look-ahead bias. By contrast, the target variable only needs to avoid allowing test-period volatility to influence the training process, and scaling by the training-set standard deviation ensures this. Together, these steps produce time-consistent and information-clean inputs for all subsequent estimation.

4 Feature Construction

4.1 Target Variable: Price Impact

Our prediction target is the absolute price impact realized over the subsequent five-minute interval. We define price impact as:

$$y_{i,t} = \text{PriceImpact}_{i,t} = \frac{|P_{i,t} - P_{i,t-1}|}{\text{Volume}_{i,t}} \tag{5}$$

where $P_{i,t}$ denotes the transaction price for stock i at time t, and $\text{Volume}_{i,t}$ is the traded volume in the 5-minute interval.

This definition captures the price movement generated per unit of trading activity, providing a normalized measure that is comparable across stocks with different price levels and trading volumes. The absolute value ensures that we focus on the magnitude of price displacement rather than its direction, consistent with our objective of detecting periods of elevated informed trading regardless of whether the information is positive or negative.

For prediction purposes, we use the contemporaneous features observed at time t to forecast the price impact that will be realized in the interval from t to $t+1$. This forward-looking structure ensures that our models are genuinely predictive rather than explanatory.

4.2 Microstructure Features

We construct two primary microstructure features that capture distinct aspects of order book dynamics: the Quote Improvement-to-Deterioration ratio (QID) and the Book Exhaustion Rate (BER).

4.2.1 Quote Improvement-to-Deterioration Ratio (QID)

The QID measure captures the relative intensity of competitive quoting versus liquidity-consuming trades. We classify each NBBO update into one of two categories based on the nature of the quote change.

A quote improvement occurs when a market maker posts a new quote that improves upon the existing NBBO, either by raising the bid or lowering the ask. Such updates reflect competitive undercutting behavior, where liquidity providers vie for queue priority by offering better prices.

A trade-driven deterioration occurs when an executed trade consumes the quantity available at the best quote, causing the NBBO to widen as the next-best price level becomes the new inside quote.

For each five-minute interval, we compute QID as:

$$\mathrm{QID}_{j,t} = \frac{\#\mathrm{Impr}_{j,t} - \#\mathrm{DeterTrade}_{j,t}}{\#\mathrm{Impr}_{j,t} + \#\mathrm{DeterTrade}_{j,t}} \tag{6}$$

This ratio ranges from -1 to $+1$. High QID values indicate that quote improvements dominate, suggesting active competition among liquidity providers and relatively benign order flow. Low or negative QID values indicate that trade-driven deteriorations dominate, consistent with elevated toxicity or adverse selection risk.

We include QID along with its first two lags (QID_{t-1} and QID_{t-2}) to capture the short-lived persistence in quoting behavior.

We use QID directly rather than QIDres because our specification already includes liquidity controls on the right-hand side. Introducing QIDres would reintroduce the same variables and induce multicollinearity.

4.2.2 Book Exhaustion Rate (BER)

The Book Exhaustion Rate measures how rapidly incoming market orders deplete the liquidity available at the top of the limit order book. We define the two-sided BER as:

$$\mathrm{BER}_{\text{two-sided}}(t,w) = \frac{\begin{array}{c}\text{Average total number of market buy and sell orders}\\ \text{executed per second during } [t-w,t)\end{array}}{\begin{array}{c}\text{Average total top-of-book liquidity}\\ \text{(sum of bids and offers) during } [t-w,t)\end{array}} \tag{7}$$

We compute BER separately for the bid side and offer side as well. Intuitively, BER is the fraction of the queue consumed per second: a value of 0.5 means that, at the current pace, half of the available depth at the best quote would be removed in one second.

Since our data do not include explicit trade direction flags, we infer buyer- versus seller-initiated trades using the Lee-Ready algorithm by comparing each transaction price to the prevailing NBBO midpoint, and then construct BER on each side of the book using these classified market orders and top-of-book depth.

We include both contemporaneous BER measures and their first lags:

- $\mathrm{BER}^{(o)}_{\text{offer},t}$, $\mathrm{BER}^{(b)}_{\text{bid},t}$ (current period)
- $\mathrm{BER}^{(o)}_{\text{offer},t-1}$, $\mathrm{BER}^{(b)}_{\text{bid},t-1}$ (lagged)

4.3 Spectral Features (FFT)

We incorporate spectral features derived from the Fast Fourier Transform (FFT) to capture periodic patterns in intraday price dynamics. FFT is conceptually similar to the CNN application to price paths by Jiang et al. (2020), which learns frequency components in the price path and extracts patterns. A CNN requires significant training data to effectively extract trends in the price path. To overcome our data shortage, we use FFT, which explicitly computes frequency components.

In a similar set-up to Jiang et al. (2020), for each five-minute interval, we:

1. Build 5-minute, 0.1 second mid-price intervals for each stock and standardize from 0 to 1
2. Compute the discrete FFT to obtain complex coefficients $X[k]$:

$$X[k] = \sum_{t=0}^{N-1} x[t]\, e^{-i2\pi kt/N} \tag{8}$$

Our chosen feature is the low-frequency power ratio:

$$\text{fft\_low\_ratio} = \frac{\sum_{k=1}^{5} |X[k]|^2}{\sum_{k=1}^{N/2} |X[k]|^2} \tag{9}$$

This ratio captures the proportion of spectral power concentrated in the lowest frequency components. Low frequency power (the numerator) captures persistent one sided order pressure - such as from informed traders. This feature complements QID & BER by measuring not just the size of order imbalance, but it's persistence. The economic interpretation is:

- High low_ratio → price path dominated by low-frequency drift → impact is persistent
- Low low_ratio → high-frequency noise dominates → impact is mean-reverting

Overall, this feature defines the current market regime: Is informed trading present? Or are prices noisy and mean-reverting?

4.4 Control Variables

We augment our core microstructure features with a set of standard control variables computed at the five-minute frequency. These controls capture well-documented determinants of price impact and ensure that any predictive power attributed to QID and BER reflects genuine incremental information.

Spread Mean. Average bid-ask spread in the 5-minute bin, proxying for liquidity and transaction costs:

$$\text{spread\_mean\_5m} = \overline{\text{SPREAD}} \tag{10}$$

Depth Imbalance. Imbalance between bid and ask depth, capturing buy versus sell pressure:

$$\text{DEPTH\_IMB} = \frac{\text{bid\_size} - \text{ask\_size}}{\text{bid\_size} + \text{ask\_size}} \tag{11}$$

$$\text{depth\_imb\_mean\_5m} = \overline{\text{DEPTH\_IMB}} \tag{12}$$

Order Flow Imbalance (OFI). Total order-flow imbalance from quote size changes, capturing passive liquidity pressure:

$$\text{OFI} = \Delta(\text{bid\_size}) - \Delta(\text{ask\_size}) \tag{13}$$

$$\text{ofi\_sum\_5m} = \sum \text{OFI} \tag{14}$$

Realized Volatility. Realized volatility of mid-quotes inside the 5-minute bin:

$$\text{LOG\_RET} = \log(M_t) - \log(M_{t-1}) \tag{15}$$

$$\text{rv\_midquote\_5m} = \text{std}(\text{LOG\_RET}) \tag{16}$$

Trade Imbalance. Signed trade volume capturing aggressive order pressure, where sign is based on price movement (↑ implies +volume for buy, ↓ implies −volume for sell):

$$\text{trade\_imbalance\_5m} = \sum \text{signed\_vol} \tag{17}$$

Log Price. We include the log of average transaction price to control for cross-sectional differences in price levels.

5 Methodology

We evaluate a spectrum of models ranging from parametric baselines to deep neural networks.

5.1 OLS Baseline

To establish a benchmark before introducing nonlinear or sequence models, we begin with a set of linear regressions that capture increasingly rich forms of predictability in short-horizon price impact. Across all specifications, the left-hand-side variable is the next-interval price impact for stock i at time t. This ensures that all models are explicitly forecasting future impact rather than partially explaining contemporaneous order-book dynamics.

We build three stages of OLS models.

Sanity-Check Regression. The first is a sanity-check regression that includes the full set of microstructure predictors, which are QID, BER, order-flow variables, and controls, together with stock fixed effects. This provides an initial assessment of whether our two core signals, QID and BER, contain meaningful explanatory power for next-period impact. The specification is:

$$\text{PriceImpact}_{i,t+1} = \alpha + \beta_1 \,\text{QID}_{i,t} + \beta_2 \,\text{BER}_{i,t} + \gamma' X_{i,t} + \delta_i + \varepsilon_{i,t+1}. \tag{18}$$

where $X_{\{i,t\}}$ represents the five-minute control variables (realized volatility, trade imbalance, OFI, depth imbalance, and log-price). This regression serves purely as a diagnostic: we are not yet modeling temporal dependence directly, but rather verifying that the direction and magnitude of the coefficients align with microstructure intuition.

AR(5) Baseline. Next, to isolate the contribution of serial dependence in price impact itself, we estimate a pure autoregressive model. An AR(p) process frames next-period impact as a function of its own past values:

$$\text{PriceImpact}_{i,t+1} = \alpha + \sum_{k=1}^{p} \phi_k \, \text{PriceImpact}_{i,t+1-k} + \delta_i + \varepsilon_{i,t+1}. \tag{19}$$

We determine the optimal lag length p by minimizing AIC and BIC, both of which select p=5. The resulting AR(5) specification captures the well-documented short-horizon persistence in microstructure returns and provides a clean baseline for comparison with our feature-based models (See Table 2).

Table 2: Pure AR(p) Lag Selection Results for Training Sample

k	**nobs**	**AIC**	**BIC**
0	27576	73351.450680	73433.697691
1	27566	69405.719871	69496.187593
2	27556	69365.958640	69464.646347
3	27546	69314.624769	69421.531733
4	27536	69244.533618	69359.659111
5	27526	69120.831717	69244.175011

Augmented Model. Third, we combine the autoregressive structure with microstructure predictors.

$$\text{PriceImpact}_{i,t+1} = \alpha + \sum_{k=1}^{5} \phi_k \, \text{PriceImpact}_{i,t+1-k} + \beta_1 \, \text{QID}_{i,t} + \beta_2 \, \text{BER}_{i,t} + \gamma' X_{i,t} + \delta_i + \varepsilon_{i,t+1}. \tag{20}$$

Here the lag structure captures the mechanical dynamics of price continuation and reversion, while QID, BER, and the five-minute controls capture information about order-book pressure, liquidity consumption, and real-time trading conditions.

Together, these three regressions form the foundation for evaluating whether our microstructure signals provide incremental explanatory power beyond simple autoregressive persistence. They also establish a consistent baseline against which we later compare more flexible nonlinear models.

5.2 Tree-Based Models

We implement Random Forest and Gradient Boosted Trees to capture potential nonlinear relationships between features and price impact. These ensemble methods can identify complex interactions among QID, BER, and control variables that linear models cannot represent.

Random Forests. A Random Forest estimator constructs an ensemble of B regression trees $\{T_b(\cdot)\}_{b=1}^{B}$, each trained on a bootstrap sample of the data with random feature subsampling. For a feature vector x_t, the prediction is the average across trees:

$$\hat{f}_{\mathrm{RF}}(x_t) = \frac{1}{B}\sum_{b=1}^{B} T_b(x_t). \tag{21}$$

Each tree partitions the feature space into regions $\{R_{b,m}\}_{m=1}^{M_b}$ and predicts a constant value within each region:

$$T_b(x) = \sum_{m=1}^{M_b} \gamma_{b,m}\, \mathbf{1}\{x \in R_{b,m}\}. \tag{22}$$

Individual trees have noisy, high-variance predictions, but averaging across the ensemble greatly reduces the variance and preserves the nonlinear structure. This allows the model to detect threshold effects, for example, situations where price impact becomes disproportionately large once QID or OFI exceed certain levels, and to exploit interactions among features that would not appear in an additive linear specification.

Gradient Boosted Trees. Gradient Boosted Trees approximate the unknown price-impact function $f(x)$ by iteratively adding shallow regression trees that correct the errors of the current model. Beginning with an initial estimate $f_0(x)$, the model updates according to

$$f_m(x) = f_{m-1}(x) + \nu\, T_m(x), \tag{23}$$

where $T_m(\cdot)$ is the tree fitted to the pseudo-residuals

$$r_t^{(m)} = -\left.\frac{\partial \ell(y_t, f(x_t))}{\partial f}\right|_{f=f_{m-1}}, \tag{24}$$

with $\ell(y,f) = \frac{1}{2}(y-f)^2$ for squared-error loss and $\nu \in (0,1)$ a learning rate. Thus each tree solves

$$T_m = \arg\min_{T} \sum_{t} \left(r_t^{(m)} - T(x_t)\right)^2, \tag{25}$$

and the final model is

$$\hat{f}_{\mathrm{GBT}}(x) = f_0(x) + \nu \sum_{m=1}^{M} T_m(x). \tag{26}$$

GBTs can therefore capture subtle microstructure dynamics such as asymmetric responses and nonlinear liquidity depletion effects.

Compared with OLS, both tree-based approaches naturally handle interactions, nonlinearities, and heteroskedastic behavior without requiring feature engineering. However, they differ in bias–variance trade-offs: Random Forests typically yield smoother, more stable predictions, while Gradient Boosted Trees may better capture fine-grained structure at the cost of higher sensitivity to noise and hyperparameter settings. Together, these models provide a flexible benchmark for assessing whether nonlinear microstructure patterns materially improve predictability of short-horizon price impact.

5.3 Deep Sequence Models

Traditional regression and tree-based models treat each observation as independent, ignoring the sequential structure inherent in market microstructure data. However, features such as QID, BER, depth imbalance, and order-flow imbalance evolve over time, and their temporal trajectories often contain predictive information that cannot be captured by static models. To incorporate this path dependency, we implement Recurrent Neural Networks (RNNs), which are specifically designed for sequence modeling.

5.3.1 Recurrent Neural Network

An RNN processes a sequence of inputs $(x_1, \ldots, x_T)$ one step at a time. Its core mechanism is the hidden state, h_t, which acts as an internal memory summarizing past information. At each time step, the model updates this memory through a nonlinear transformation:

$$h_t = \tanh(W_{\mathrm{input}} x_t + W_{\mathrm{memory}} h_{t-1}), \tag{27}$$

allowing the network to encode the evolution of order-flow features over the lookback window rather than relying solely on the most recent snapshot.

However, the vanilla RNN architecture suffers from the vanishing gradient problem: as gradients are propagated backward through many time steps, they decay exponentially, inhibiting the model's ability to learn long-range dependencies. In practice, this causes the network to overweight the most recent observations while effectively "forgetting" earlier portions of the sequence. This limitation motivates the use of gated architectures, which are

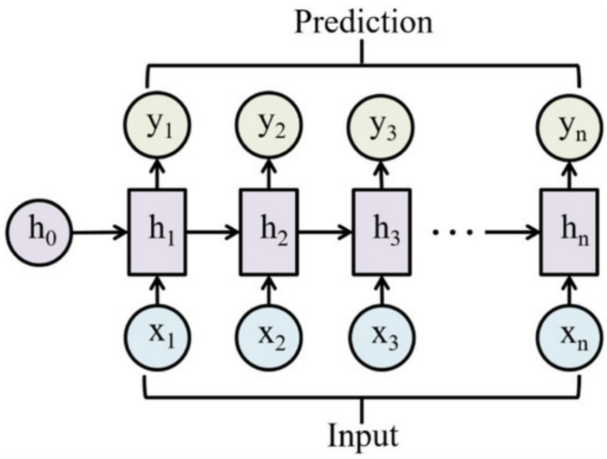

Figure 1: RNN Architecture

designed to stabilize gradient flow and retain relevant information over longer horizons.

5.3.2 Long Short-Term Memory (LSTM)

To address the limitations of vanilla RNNs, we employ Long Short-Term Memory (LSTM) networks, which introduce gating mechanisms to maintain stable memory over extended sequences. Unlike RNNs, which maintain a single hidden state h_t, LSTMs maintain both a cell state c_t and a hidden state h_t, enabling more controlled information flow across time.

Each LSTM unit contains three gates:

- **Input gate**: determines how much new information enters the cell.
- **Forget gate**: regulates how much past information is retained.
- **Output gate**: controls how much of the internal state is exposed as output.

These gates collectively mitigate the vanishing gradient issue by preserving gradients through the cell state, allowing the network to capture longer historical patterns in order-flow evolution. This makes LSTMs particularly expressive for modeling microstructure signals, where persistent liquidity shifts or accumulation of informed trading pressure may unfold gradually. However, this expressiveness comes at the cost of computational complexity. LSTMs have more parameters than RNNs or GRUs and therefore require more training data and computational resources. In data-constrained settings such as ours, this additional flexibility may not always translate into improved predictive performance, but LSTMs nevertheless provide a strong benchmark for sequence modeling.

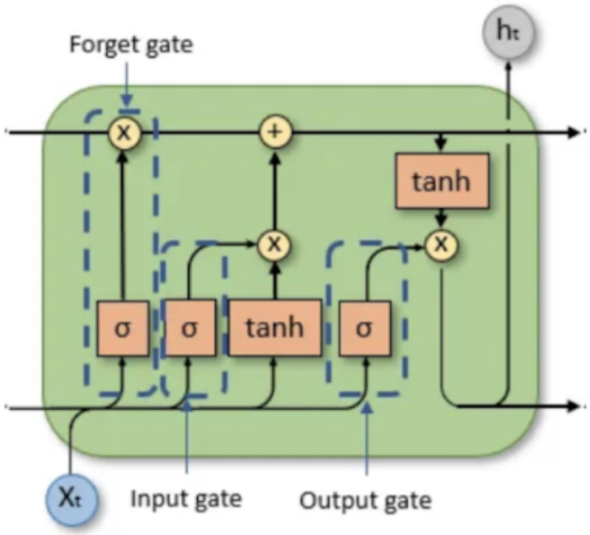

Figure 2: LSTM Architecture

5.3.3 Gated Recurrent Unit (GRU)

We also evaluate the Gated Recurrent Unit (GRU) architecture, which offers a streamlined alternative to LSTMs while retaining the benefits of gated memory. GRUs merge the cell state and hidden state into a single state vector and simplify the gating structure to two gates:

- **Update gate**: determines how much of the previous state should be carried forward.
- **Reset gate**: controls how new input is combined with past information.

This reduction in complexity leads to fewer parameters and lower training costs, making GRUs computationally efficient and often easier to tune—an important advantage when the sample size is limited. Despite their simpler structure, GRUs typically achieve performance comparable to LSTMs and frequently generalize better when datasets are not sufficiently large.

Conceptually, GRUs strike a balance between expressiveness and parsimony: they mitigate vanishing gradients with a lighter memory architecture while avoiding the overhead associated with LSTM's three-gate mechanism. For our application, GRUs offer an effective way to capture medium-range dependencies in order-flow dynamics without incurring the risk of overfitting.

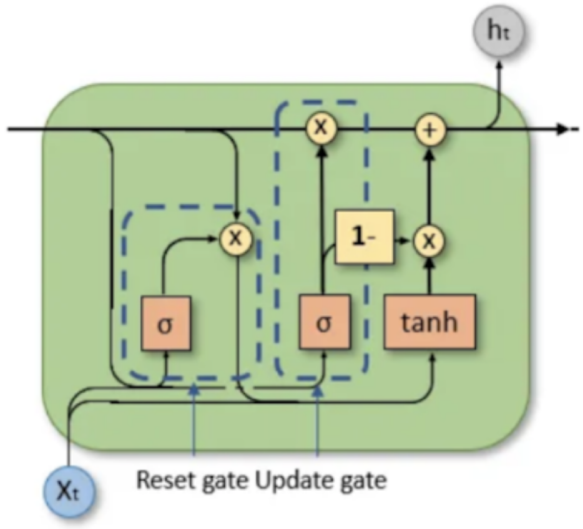

Figure 3: GRU Architecture

5.3.4 Hyperparameter Tuning

To ensure that our recurrent architectures (RNN, LSTM, and GRU) were appropriately calibrated for the microstructure forecasting task, we conducted a systematic grid search over three key hyperparameters controlling the temporal scope, representational capacity, and architectural depth of the models. All configurations were evaluated using out-of-sample performance to mitigate overfitting and to select models that generalize reliably to unseen market conditions.

Lookback Window The lookback window determines how many past observations of QID, BER, and other microstructure features are fed into the sequence model at each training step. We tested windows of 10, 20, and 40 time steps, corresponding to progressively longer histories of order-flow evolution. Shorter windows restrict the model to very recent liquidity conditions, emphasizing immediate shocks, while longer windows allow the network to learn more persistent temporal patterns such as sustained liquidity depletion or multi-interval buildup of imbalance. Longer windows, however, increase model complexity and heighten the risk of overfitting in a noisy, limited dataset.

Hidden State Dimension The hidden dimension controls the number of latent features the network uses to summarize the historical sequence. We evaluated dimensions of 32, 64, and 128. Larger hidden states increase expressive power, enabling the model to capture

more complex interactions between QID, BER, and other covariates across time. At the same time, excessive dimensionality can lead to instability or overfitting.

Number of Layers We considered architectures with 1, 2, and 3 stacked recurrent layers. Additional layers allow the model to learn hierarchical temporal representations, for example, separating short-horizon shocks from slower structural shifts in order flow. However, deeper networks require more data to train effectively and can exacerbate vanishing-gradient or optimization challenges.

$$\text{Lookback} \in \{10, 20, 40\}, \quad \text{Hidden Size} \in \{32, 64, 128\}, \quad \text{Layers} \in \{1, 2, 3\}.$$

Overall, the grid search across the configuration space allowed us to identify the architecture that minimized out-of-sample error for each model family. This tuning process ensured that the chosen sequence models achieved an appropriate balance between temporal coverage, model capacity, and generalization performance.

5.4 Evaluation Metrics

To assess the predictive performance of our models, we focus on two complementary out-of-sample metrics: the coefficient of determination (R^2) and the mean squared error (MSE). Evaluating models on held-out data is critical in a high-frequency microstructure setting, where signals are noisy and overfitting can occur easily. Out-of-sample performance therefore provides a more reliable gauge of whether a model captures genuine structure in the data rather than spurious patterns specific to the training sample.

The out-of-sample R^2 measures the fraction of variation in next-interval price impact that the model is able to explain relative to a naïve benchmark. Formally, it compares the model's prediction error to the variance of the target in the test set. Because price impact series are volatile and strongly mean-reverting, the level of attainable R^2 is typically low; even small improvements can be economically meaningful. In our context, out-of-sample R^2 provides a natural way to evaluate whether features such as QID, BER, and lagged impact contain incremental predictive content beyond simple autoregressive dynamics.

The MSE complements the R^2 by measuring the average squared magnitude of forecasting errors in the original scale of the dependent variable. Since our target variable (absolute next-interval price impact) is very small in magnitude and highly skewed, MSE allows us to quantify how closely predicted values align with realized price movements. A lower MSE indicates not only better predictive accuracy but also fewer large misspecifications, which is particularly important when modeling microstructure quantities that can exhibit occasional

bursts of extreme activity.

6 Results

6.1 Sanity-check Regression

As a first step after constructing and standardizing the panel, we estimate a "sanity-check" regression to assess whether our two core microstructure signals (QID and BER) exhibit explanatory power for short-horizon price impact. The specification includes three groups of predictors. First, we incorporate the contemporaneous QID measure together with its first two lags. These terms capture how relative quote improvements (versus trade-driven quote deteriorations) contribute to subsequent price moves. Second, we include both bid-side and offer-side BER variables, each of which measures the rate at which incoming market orders consume top-of-book liquidity. Finally, we add a standard set of microstructure controls (spread, depth imbalance, order-flow imbalance, realized mid-quote volatility, and trade imbalance) along with stock fixed effects.

The regression results, shown in Appendix A, indicate that both QID and BER exhibit statistically significant relationships with next-period price impact. The contemporaneous QID term is positive and significant at conventional levels, implying that intervals with a higher proportion of quote improvements tend to be followed by larger absolute price moves. This aligns with the interpretation in Barardehi et al. (2024): when liquidity providers revise quotes more aggressively on the improvement side, it often reflects informed pressure around an anticipated price revaluation. Among the lagged QID terms, the first and second lags are also significant, though the magnitudes are smaller. This pattern reflects the short-lived nature of informed quoting: predictive power persists briefly but decays quickly across bins.

The BER variables also behave consistently with microstructure theory. Both bid-side and offer-side BER enter with negative and significant coefficients. Because BER captures the speed at which market orders consume displayed liquidity, high BER reflects localized bursts of toxicity and rapid inventory shocks for market makers. The negative sign is consistent with these events materializing within the current bin: once liquidity is consumed aggressively, the immediate price impact is absorbed contemporaneously rather than spilling into the next interval. This interpretation aligns with the Zhao and Linetsky (2021) view that toxicity-driven jumps tend to occur at the moment liquidity is taken, leaving less residual movement for the following period.

The remaining controls behave as expected. Spread and volatility enter with positive coefficients, consistent with wider spreads and higher uncertainty being associated with greater

subsequent price impact. Trade imbalance also shows a strong positive and highly significant coefficient, reflecting the mechanical relationship between directional order flow and short-run returns.

These results confirm that both QID and BER contain meaningful information about short-horizon price formation, even after controlling for standard liquidity and order-flow variables. The presence of consistent, interpretable, and statistically significant coefficients provides a strong justification for including these variables in the more sophisticated autoregressive and feature-augmented models presented in the next sections.

6.2 Baseline Autoregressive Model: AR(5)

Estimation of the AR(5) model reveals a clear structure in the dynamics of five-minute price impact, as shown in Appendix B. The first lag, ϕ_1, is by far the dominant term: it is large in magnitude and highly statistically significant, indicating strong short-lived persistence. When a bin experiences a sizable price impact, the subsequent bin tends to retain part of that movement. The remaining lags, ϕ_2, remain statistically significant but are much smaller in magnitude, suggesting that persistence decays quickly after the first interval. This pattern aligns with the well-documented behavior of high-frequency returns and impact measures: most serial dependence is concentrated at very short horizons, while longer-lag predictability dissipates rapidly.

In terms of fit, the model performs substantially better than a naïve no-memory benchmark. The AR(5) model explains approximately 28% of the in-sample variation in next-period price impact. More importantly, the model generalizes well: the out-of-sample R^2 remains around 26%, and the associated out-of-sample mean squared error is extremely small due to the scale of the standardized target. These results show that past price impact alone captures a meaningful portion of short-horizon dynamics.

This autoregressive benchmark provides two essential insights for the subsequent analysis. First, price impact exhibits strong but rapidly decaying persistence, supporting the use of lagged impact terms in richer models. Second, the AR(5) model establishes a strong baseline level of predictability: any microstructure-based or machine-learning model must outperform this standard to demonstrate incremental explanatory power. Consequently, the AR(5) specification serves as both a conceptual and empirical foundation for the feature-augmented regressions and nonlinear sequence models examined in later sections.

6.3 Feature-Augmented Autoregressive Model

After establishing the benchmark AR(5) specification, we next examine whether microstructure-based features add incremental explanatory power when combined with the same autoregressive structure. To do this, we estimate a feature-augmented AR model where the dependent variable is the next-interval price impact, and the regressors include five lags of past price impact, the QID and BER measures and their lags, as well as the five-minute control variables.

According to results in Appendix C, the addition of these features produces a modest but systematic improvement in model performance. The in-sample R2 rises from 0.281 in the pure AR(5) model to 0.285 in the augmented regression, and the out-of-sample R^2 increases from roughly 0.263 to approximately 0.268. The mean squared error also declines. While the gains are small, they are meaningful given the difficulty of predicting high-frequency price movements and confirm that the microstructure variables carry nontrivial incremental signals.

Turning to the coefficient estimates, QID continues to load with the expected sign, and its second lag remains statistically significant even after controlling for the autoregressive structure. This indicates that quote-improvement pressure still contains short-lived predictive information once we account for past price impact. In contrast, several BER terms lose significance compared with the earlier sanity-check regression. This attenuation is consistent with the fact that many of the short-horizon toxicity effects captured by BER are now absorbed by the lagged dependent variables and the richer set of contemporaneous liquidity controls. The microstructure controls themselves behave as expected: log price and trade imbalance remain strongly significant, and realized volatility continues to play an important role in shaping next-interval price impact.

These results show that the AR(5) structure continues to explain the majority of the predictability, but augmenting it with QID and liquidity measures produces clear, measurable improvement. The findings also highlight that QID retains informational content even in the presence of autoregressive dynamics, while BER's predictive power is more sensitive to the conditioning set. This motivates the move toward nonlinear and sequence-based models, which may better capture interactions and temporal structure beyond what a linear regression can represent.

6.4 FFT Feature Addition

Appendix D demonstrates the result after adding our Fast Fourier Transform features. Running our feature-augmented autoregressive model, our FFT ratio feature is significant with

a positive coefficient ($p = 0.049$), confirming that spectral characteristics of the price path contain predictive information for future impact.

6.5 Tree-Based Models

Table 3: Out-of-Sample Tree Comparison

Model	**OOS R^2**	**OOS MSE**
Random Forest	0.2907	1.010×10^{-11}
Gradient Boosting	0.1901	1.154×10^{-11}

From the results in Table 3, Random Forests are able to capture non-linearity and interactions that OLS does not, yielding a greater out-of-sample R^2. This model is also very robust to noise which is highly prevalent in intraday price data, as it average through bootstrapping. Its' ability to randomly select features correct any multi-collinearity issues present in our selected features.

Gradient Boosting performs less well than OLS. The primary cause for this is too little data, which results in the model over-fitting. In-sample R^2 is almost 0.6, which illustrates this clearly. The model's inability to handle correlated features effectively further erodes it's predictive power.

6.6 Sequential Models

We estimate GRU and LSTM architectures across the structured hyperparameter grid. Model selection is based on validation loss (MSE) using the scaled target variable. The complete hyperparameter tables are reported in Appendix E , where validation losses should be interpreted relative to the scaling procedure described there.

6.6.1 Long Short-Term Memory

The LSTM architecture achieves its strongest performance with the longest lookback window (40 steps), indicating that extended histories of QID, BER, and other microstructure features provide valuable information for predicting short-horizon price impact. The optimal configuration further uses a 128-dimensional hidden state and two stacked recurrent layers.

LSTM Optimal Hyperparameters

- Lookback window: 40

- Hidden dimension: 128
- Layers: 2

Out-of-sample test performance

- $R^2 = 0.1854$
- MSE $= 1.157 \times 10^{-11}$

This design balances representational capacity with stability: the deeper memory enabled by the LSTM's cell state benefits from a larger hidden dimension, while a two-layer design provides sufficient depth without introducing the optimization and overfitting issues common in deeper LSTM architectures.

6.6.2 Gated Recurrent Unit

The GRU architecture also converges on the 40-step lookback window, and performs best with a 32-dimensional hidden state and three recurrent layers.

GRU Optimal Hyperparameters

- Lookback window: 40
- Hidden dimension: 32
- Layers: 3

Out-of-sample test performance

- $R^2 = 0.1800$
- MSE $= 1.165 \times 10^{-11}$

This configuration indicates that GRUs benefit from additional depth to extract multi-interval temporal patterns, while maintaining a compact hidden representation to avoid instability.

A common observation across both architectures is the selection of the maximum lookback window tested (40 steps). This pattern underscores that short-horizon price impact is not driven solely by recent shocks, but by longer-term evolution in liquidity and order flow, such as sustained queue depletion, prolonged shifts in depth imbalance, or persistent informed trading pressure. Longer windows allow the models to internalize these broader dynamics.

The two architectures differ in how they allocate model capacity. The LSTM performs best with a larger hidden dimension and fewer layers, which is consistent with its richer internal gating (input, forget, output gates and a cell state): each layer can store and process a substantial amount of temporal information, so increasing the hidden size is more effective than stacking many layers. By contrast, the GRU performs best with more layers and a smaller hidden state. With a simpler two-gate design, it extracts temporal patterns by stacking layers rather than by enlarging each one, effectively spreading capacity across depth instead of within a single layer. These choices reflect how each architecture organizes temporal information while still generalizing in a data-limited setting.

6.7 Model Performance Comparison

Table 4: Out-of-Sample Model Performance Comparison

Model	**Out-of-Sample R^2**	**Out-of-Sample MSE ($\times 10^{-11}$)**
Random Forest	0.2907	1.001
OLS	0.2683	1.018
Gradient Boosting	0.1901	1.154
LSTM	0.1854	1.157
GRU	0.1800	1.165

Table 4 summarizes the out-of-sample performance across all empirical models in our study. Two findings stand out. First, the simplest models, OLS and Random Forest, deliver the strongest predictive accuracy. Random Forest achieves the highest out-of-sample R^2 (0.2907) and lowest MSE (1.001×10^{-11}), outperforming both linear and deep learning approaches. The OLS model also performs competitively.

The fact that Random Forest surpasses OLS suggests that the relationship between order book conditions (QID, BER, depth imbalance, OFI, etc.) and short-horizon price impact is not purely linear. The ensemble's ability to capture nonlinear feature interactions, such as conditional liquidity depletion or asymmetric order-flow pressure, appears to add incremental explanatory power. At the same time, Random Forest's robustness stems from its bagging framework: averaging many decorrelated trees stabilizes predictions in environments where the dependent variable is extremely noisy, as is typical in high-frequency microstructure data.

In contrast, Gradient Boosting and the sequential deep learning models (GRU, LSTM) underperform relative to both OLS and Random Forest, with out-of-sample R^2 values closer to 0.18–0.19. This reflects two structural challenges. Gradient Boosting is more sensitive

to noise and hyperparameter misspecification because it fits trees sequentially, which can lead to overfitting when labels are noisy. In contrast, Random Forest's averaging makes it inherently noise-resistant.

For the deep models, performance is limited by sample size considerations. Although recurrent architectures are designed to learn temporal dependencies in order-flow evolution, they require large datasets to avoid overfitting when modeling complex nonlinear patterns. Our dataset contains roughly 30–40,000 five-minute intervals, substantial by econometric standards but small for high-capacity RNNs. In this regime, simpler models that impose fewer parameters and converge more quickly, such as OLS and Random Forest, generalize more effectively.

Within the sequential models, GRU and LSTM perform similarly, with the LSTM showing a slight advantage. This difference is modest and likely attributable to architectural nuances rather than fundamentally stronger temporal signal extraction. Overall, the sequential models capture some structure in the data but do not outperform simpler alternatives in this sample.

7 Conclusions and Contributions

This study investigates whether microstructure-based features, particularly QID and BER, can help forecast short-horizon price impact and detect informed trading risk in real time. Using millisecond-level trade and quote data aggregated to five-minute intervals, we evaluate a broad set of models ranging from linear regressions to nonlinear tree ensembles and deep recurrent architectures.

From a forecasting perspective, our main conclusion is that well-regularized linear and simple nonlinear models perform best in this setting. Across all specifications, OLS and Random Forest deliver the strongest out-of-sample performance. Given the sample size and the noise inherent in short-horizon price impact, these models strike the most effective balance between flexibility and stability. Random Forest's improvement over OLS indicates that the relationship between order-book conditions and price impact contains meaningful nonlinearities. More complex architectures do not yet deliver incremental predictive gains in this dataset, largely due to the sparsity of data.

Our contributions to the literature are twofold. First, we provide evidence that QID and BER are informative, microstructure-based liquidity measures. Although the autoregressive component accounts for the largest share of explained variation, both QID and BER contribute statistically significant predictive power for short-horizon price impact. Built from quote improvements, deteriorations, and execution flow, these measures reflect a dif-

ferent theoretical motivation than traditional impact models, yet they consistently encode meaningful information about near-term liquidity conditions. Our results demonstrate that microstructure-driven indicators, even when noisy at high frequency, carry complementary information about liquidity stress and order-flow toxicity.

Second, we introduce Fourier-based features as a model-agnostic way to capture intraday structure in liquidity. Incorporating Fourier-based features yields statistically significant coefficients, suggesting that liquidity conditions exhibit intraday temporal structure not captured by conventional microstructure variables. This result highlights that augmenting standard feature sets with structural representations of intraday price paths can enrich predictive modeling, even in relatively simple frameworks.

8 Future Directions

Several directions offer promising opportunities to extend this research and deepen the understanding of short-horizon liquidity and informed trading risk.

Non-Parametric Feature Extraction (CNNs) Our current approach introduces structure through Fourier-based features, which encode assumed frequency components of the price impact series. A natural extension is to treat the price impact series and the limit order book as two-dimensional objects and apply Convolutional Neural Networks (CNNs) for non-parametric feature extraction. By viewing these data as "images," CNNs could automatically learn spatial–temporal patterns that are difficult to engineer through economics. Because CNNs are high-capacity models, this direction would require a substantially larger dataset to avoid overfitting.

Advanced Sequential Architectures With a larger and richer dataset, we could revisit sequential modeling using more expressive architectures. This includes deeper and better-regularized RNNs, GRUs, and LSTMs, as well as hybrid models that combine recurrent layers with convolutional or attention-based components. Transformer models with attention mechanisms could be explored to handle very long sequences and assign dynamic weights to different parts of the order-flow history, potentially uncovering structural breaks or regime shifts that standard RNNs miss.

Market-Making Strategy Simulation with L2 Data Access to full-depth limit order book (L2) data would enable a more direct link between forecasting and trading outcomes.

With detailed queue and depth information, one could build a realistic market-making simulator and evaluate PnL under different quoting and inventory policies. This would allow us to test whether the model's price-impact forecasts help avoid adverse selection, manage inventory risk, and respond to liquidity shocks. Such a simulation-based study would move beyond statistical fit and assess the economic value of microstructure-based forecasts in a setting that incorporates execution costs and market frictions.

References

Barardehi, Y. H., Dixon, B., & Liu, W. (2024). Quote improvement and trade deterioration. *Working Paper*.

Zhao, S., & Linetsky, V. (2021). Book exhaustion rate. *Working Paper*.

Jiang, J., Kelly, B. T., & Xiu, D. (2020). (Re-)Imag(in)ing Price Trends. *Chicago Booth Research Paper No. 21-01*.

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, 5(1), 31–56.

APPENDIX

A Sanity-Check Regression Results

OLS Regression Results

Dep. Variable:	y_scaled	R-squared:	0.177
Model:	OLS	Adj. R-squared:	0.177
Method:	Least Squares	F-statistic:	504.5
Date:	Tue, 09 Dec 2025	Prob (F-statistic):	0.00
Time:	09:00:32	Log-Likelihood:	-36435.
No. Observations:	27576	AIC:	7.291e+04
Df Residuals:	27554	BIC:	7.309e+04
Df Model:	21		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.7251	0.067	10.762	0.000	0.593	0.857
C(STOCK)[T.C]	-0.4033	0.079	-5.086	0.000	-0.559	-0.248
C(STOCK)[T.CAT]	-0.7945	0.038	-20.845	0.000	-0.869	-0.720
C(STOCK)[T.GE]	-0.6924	0.060	-11.497	0.000	-0.810	-0.574
C(STOCK)[T.IBM]	-0.3130	0.043	-7.285	0.000	-0.397	-0.229
C(STOCK)[T.JNJ]	-0.7969	0.051	-15.595	0.000	-0.897	-0.697
C(STOCK)[T.KO]	-0.7555	0.053	-14.361	0.000	-0.859	-0.652
C(STOCK)[T.MCD]	-0.7287	0.045	-16.363	0.000	-0.816	-0.641
C(STOCK)[T.MSFT]	-0.8399	0.056	-14.892	0.000	-0.950	-0.729
C(STOCK)[T.XOM]	-0.9036	0.047	-19.316	0.000	-0.995	-0.812
QID_z	0.0268	0.007	4.079	0.000	0.014	0.040
QID_lag1_z	0.0146	0.007	2.006	0.045	0.000	0.029
QID_lag2_z	0.0266	0.007	3.765	0.000	0.013	0.040
BER_offer_5min_z	-0.0266	0.006	-4.530	0.000	-0.038	-0.015
BER_bid_5min_z	-0.0196	0.007	-2.833	0.005	-0.033	-0.006
BER_offer_5min_lag1_z	-0.0084	0.006	-1.371	0.170	-0.020	0.004
BER_bid_5min_lag1_z	-0.0081	0.006	-1.330	0.183	-0.020	0.004
spread_mean_5m_z	-0.0008	0.001	-0.994	0.320	-0.003	0.001
depth_imb_mean_5m_z	0.0039	0.005	0.774	0.439	-0.006	0.014
ofi_sum_5m_z	-0.0007	0.009	-0.076	0.939	-0.018	0.017
rv_midquote_5m_z	0.0357	0.007	4.807	0.000	0.021	0.050
trade_imbalance_5m_z	-0.0526	0.004	-11.839	0.000	-0.061	-0.044

B AR(5) Regression Results

OLS Regression Results

Dep. Variable:	y_scaled	R-squared:	0.281
Model:	OLS	Adj. R-squared:	0.280
Method:	Least Squares	F-statistic:	766.8
Date:	Tue, 09 Dec 2025	Prob (F-statistic):	0.00
Time:	09:07:16	Log-Likelihood:	-34545.
No. Observations:	27526	AIC:	6.912e+04
Df Residuals:	27511	BIC:	6.924e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.7368	0.021	35.827	0.000	0.697	0.777
C(STOCK)[T.C]	-0.4637	0.024	-19.135	0.000	-0.511	-0.416
C(STOCK)[T.CAT]	-0.5711	0.025	-22.932	0.000	-0.620	-0.522
C(STOCK)[T.GE]	-0.7258	0.026	-27.797	0.000	-0.777	-0.675
C(STOCK)[T.IBM]	-0.3621	0.024	-15.269	0.000	-0.409	-0.316
C(STOCK)[T.JNJ]	-0.6784	0.026	-26.403	0.000	-0.729	-0.628
C(STOCK)[T.KO]	-0.6676	0.026	-26.070	0.000	-0.718	-0.617
C(STOCK)[T.MCD]	-0.5971	0.025	-23.803	0.000	-0.646	-0.548
C(STOCK)[T.MSFT]	-0.7303	0.026	-27.964	0.000	-0.781	-0.679
C(STOCK)[T.XOM]	-0.6950	0.026	-26.914	0.000	-0.746	-0.644
y_scaled_lag1	0.3478	0.006	57.806	0.000	0.336	0.360
y_scaled_lag2	0.0158	0.006	2.483	0.013	0.003	0.028
y_scaled_lag3	0.0200	0.006	3.139	0.002	0.008	0.032
y_scaled_lag4	0.0218	0.006	3.420	0.001	0.009	0.034
y_scaled_lag5	0.0631	0.006	10.487	0.000	0.051	0.075

C Feature-Augmented AR Model Results

OLS Regression Results

Dep. Variable:	y_scaled	R-squared:	0.285
Model:	OLS	Adj. R-squared:	0.284
Method:	Least Squares	F-statistic:	471.9
Date:	Wed, 10 Dec 2025	Prob (F-statistic):	0.00
Time:	20:08:03	Log-Likelihood:	-34467.
No. Observations:	27526	AIC:	6.899e+04
Df Residuals:	27498	BIC:	6.922e+04
Df Model:	27		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.5899	0.102	5.798	0.000	0.390	0.789
C(STOCK)[T.C]	-0.3289	0.096	-3.440	0.001	-0.516	-0.141
C(STOCK)[T.CAT]	-0.4850	0.090	-5.414	0.000	-0.661	-0.309
C(STOCK)[T.GE]	-0.5187	0.101	-5.129	0.000	-0.717	-0.320
C(STOCK)[T.IBM]	-0.2844	0.055	-5.129	0.000	-0.393	-0.176
C(STOCK)[T.JNJ]	-0.5264	0.098	-5.365	0.000	-0.719	-0.334
C(STOCK)[T.KO]	-0.4957	0.094	-5.250	0.000	-0.681	-0.311
C(STOCK)[T.MCD]	-0.4735	0.088	-5.351	0.000	-0.647	-0.300
C(STOCK)[T.MSFT]	-0.5537	0.105	-5.285	0.000	-0.759	-0.348
C(STOCK)[T.XOM]	-0.5496	0.105	-5.230	0.000	-0.756	-0.344
QID_z	0.0067	0.005	1.394	0.163	-0.003	0.016
QID_lag1_z	0.0032	0.006	0.502	0.616	-0.009	0.016
QID_lag2_z	0.0096	0.005	2.048	0.041	0.000	0.019
BER_offer_5min_z	-0.0040	0.004	-0.909	0.363	-0.013	0.005
BER_bid_5min_z	-0.0072	0.005	-1.376	0.169	-0.017	0.003
BER_offer_5min_lag1_z	0.0023	0.004	0.535	0.592	-0.006	0.011
BER_bid_5min_lag1_z	0.0005	0.005	0.098	0.922	-0.010	0.011
spread_mean_5m_z	0.0003	0.000	0.894	0.371	-0.000	0.001
depth_imb_mean_5m_z	0.0021	0.005	0.459	0.646	-0.007	0.011
ofi_sum_5m_z	-0.0063	0.005	-1.217	0.224	-0.017	0.004
rv_midquote_5m_z	0.0030	0.002	1.326	0.185	-0.001	0.007
trade_imbalance_5m_z	-0.0466	0.003	-13.405	0.000	-0.053	-0.040
LOG_PRICE_z	1.358e-05	5.12e-06	2.654	0.008	3.55e-06	2.36e-05
y_scaled_lag1	0.3441	0.083	4.162	0.000	0.182	0.506
y_scaled_lag2	0.0141	0.065	0.216	0.829	-0.114	0.142
y_scaled_lag3	0.0173	0.056	0.308	0.758	-0.092	0.127
y_scaled_lag4	0.0194	0.041	0.467	0.640	-0.062	0.101
y_scaled_lag5	0.0606	0.044	1.379	0.168	-0.025	0.147

D Fourier Features Model Results

OLS Regression Results

Dep. Variable:	y_scaled	R-squared:	0.284
Model:	OLS	Adj. R-squared:	0.283
Method:	Least Squares	F-statistic:	433.1
Date:	Wed, 10 Dec 2025	Prob (F-statistic):	0.00
Time:	19:17:39	Log-Likelihood:	-33838.
No. Observations:	26809	AIC:	6.774e+04
Df Residuals:	26779	BIC:	6.798e+04
Df Model:	29		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.5784	0.102	5.644	0.000	0.378	0.779
C(STOCK)[T.C]	−0.3157	0.097	−3.265	0.001	−0.505	−0.126
C(STOCK)[T.CAT]	−0.4908	0.090	−5.440	0.000	−0.668	−0.314
C(STOCK)[T.GE]	−0.5172	0.102	−5.091	0.000	−0.716	−0.318
C(STOCK)[T.IBM]	−0.2886	0.057	−5.107	0.000	−0.399	−0.178
C(STOCK)[T.JNJ]	−0.5286	0.098	−5.377	0.000	−0.721	−0.336
C(STOCK)[T.KO]	−0.4989	0.095	−5.267	0.000	−0.685	−0.313
C(STOCK)[T.MCD]	−0.4763	0.089	−5.359	0.000	−0.651	−0.302
C(STOCK)[T.MSFT]	−0.5523	0.105	−5.285	0.000	−0.757	−0.348
C(STOCK)[T.XOM]	−0.5462	0.105	−5.225	0.000	−0.751	−0.341
QID_z	0.0066	0.005	1.360	0.174	−0.003	0.016
QID_lag1_z	0.0029	0.006	0.453	0.651	−0.010	0.016
QID_lag2_z	0.0102	0.005	2.143	0.032	0.001	0.020
BER_offer_5min_z	−0.0033	0.005	−0.718	0.473	−0.012	0.006
BER_bid_5min_z	−0.0080	0.005	−1.490	0.136	−0.019	0.003
BER_offer_5min_lag1_z	0.0021	0.005	0.451	0.652	−0.007	0.011
BER_bid_5min_lag1_z	0.0014	0.005	0.258	0.796	−0.009	0.012
spread_mean_5m_z	3.278e−05	0.000	0.094	0.925	−0.001	0.001
depth_imb_mean_5m_z	0.0017	0.005	0.362	0.717	−0.007	0.011
ofi_sum_5m_z	−0.0060	0.005	−1.119	0.263	−0.016	0.005
rv_midquote_5m_z	0.0029	0.002	1.242	0.214	−0.002	0.008
trade_imbalance_5m_z	−0.0487	0.004	−13.470	0.000	−0.056	−0.042
fft_low_ratio_z	0.0051	0.003	1.965	0.049	1.34e−05	0.010
fft_low_power_z	−0.0012	0.001	−1.473	0.141	−0.003	0.000
LOG_PRICE_z	1.321e−05	5.22e−06	2.529	0.011	2.97e−06	2.35e−05
y_scaled_lag1	0.3440	0.083	4.141	0.000	0.181	0.507
y_scaled_lag2	0.0138	0.066	0.210	0.834	−0.115	0.143
y_scaled_lag3	0.0169	0.056	0.298	0.765	−0.094	0.128
y_scaled_lag4	0.0196	0.042	0.469	0.639	−0.062	0.102
y_scaled_lag5	0.0596	0.044	1.353	0.176	−0.027	0.146

E Hyperparameter Search Results

Note: This appendix reports the full hyperparameter search for LSTM and GRU. Validation MSE values are computed on the scaled dependent variable and should be interpreted only for relative comparison across configurations, not as direct measures of prediction error.

Table E.1: GRU Hyperparameter Search Results (Validation MSE)

Layers	Hidden Size	Lookback	Validation MSE
1	32	10	1.385×10^{-2}
1	32	20	1.862×10^{-2}
1	32	40	1.900×10^{-2}
1	64	10	1.148×10^{-2}
1	64	20	9.588×10^{-3}
1	64	40	1.918×10^{-2}
1	128	10	5.182×10^{-2}
1	128	20	3.037×10^{-2}
1	128	40	6.185×10^{-2}
2	32	10	7.084×10^{-3}
2	32	20	7.533×10^{-3}
2	32	40	9.200×10^{-3}
2	64	10	7.804×10^{-3}
2	64	20	8.667×10^{-3}
2	64	40	1.490×10^{-2}
2	128	10	1.162×10^{-2}
2	128	20	1.143×10^{-2}
2	128	40	6.821×10^{-3}
3	32	10	6.988×10^{-3}
3	32	20	6.457×10^{-3}
3	32	40	6.019×10^{-3}
3	64	10	8.233×10^{-2}
3	64	20	1.021×10^{-1}
3	64	40	1.055×10^{-1}
3	128	10	3.257×10^{-2}
3	128	20	2.806×10^{-2}
3	128	40	1.812×10^{-2}

Optimal configuration: Lookback = 40, Hidden = 32, Layers = 3.

Table E.2: LSTM Hyperparameter Search Results (Validation MSE)

Layers	Hidden Size	Lookback	Validation MSE
1	32	10	1.596×10^{-2}
1	32	20	8.304×10^{-3}
1	32	40	1.015×10^{-2}
1	64	10	1.521×10^{-2}
1	64	20	7.979×10^{-3}
1	64	40	8.348×10^{-3}
1	128	10	7.375×10^{-3}
1	128	20	7.610×10^{-3}
1	128	40	9.693×10^{-3}
2	32	10	6.917×10^{-3}
2	32	20	5.785×10^{-3}
2	32	40	6.265×10^{-3}
2	64	10	6.471×10^{-3}
2	64	20	5.281×10^{-3}
2	64	40	5.369×10^{-3}
2	128	10	5.420×10^{-3}
2	128	20	4.648×10^{-3}
2	128	40	3.584×10^{-3}
3	32	10	6.281×10^{-3}
3	32	20	4.973×10^{-3}
3	32	40	4.988×10^{-3}
3	64	10	4.713×10^{-3}
3	64	20	5.118×10^{-3}
3	64	40	5.393×10^{-3}
3	128	10	7.249×10^{-3}
3	128	20	7.200×10^{-3}
3	128	40	5.561×10^{-3}

Optimal configuration: Lookback = 40, Hidden = 128, Layers = 2.