Quantitative Finance

ISSN: 1469-7688 (Print) 1469-7696 (Online) Journal homepage: www.tandfonline.com/journals/rquf20

The good, the bad, and latency: exploratory trading on Bybit and Binance

Jakob Albers, Mihai Cucuringu, Sam Howison & Alexander Y. Shestopaloff

To cite this article: Jakob Albers, Mihai Cucuringu, Sam Howison & Alexander Y. Shestopaloff (2025) The good, the bad, and latency: exploratory trading on Bybit and Binance, Quantitative Finance, 25:6, 919-947, DOI: 10.1080/14697688.2025.2515933

To link to this article: https://doi.org/10.1080/14697688.2025.2515933

Quantitative Finance, 2025
Vol. 25, No. 6, 919–947, https://doi.org/10.1080/14697688.2025.2515933

The good, the bad, and latency: exploratory trading on Bybit and Binance

JAKOB ALBERS*†, MIHAI CUCURINGU†‡, SAM HOWISON§ and ALEXANDER Y. SHESTOPALOFF¶‖

†Department of Statistics, University of Oxford, Oxford, UK
‡Oxford-Man Institute of Quantitative Finance, University of Oxford, Oxford, UK
§Mathematical Institute, University of Oxford, Oxford, UK
¶School of Mathematical Sciences, Queen Mary University of London, London, UK
‖Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL, Canada

(*Received 1 February 2024; accepted 29 May 2025*)

We present the findings of a large-scale live trading experiment involving the placement of millions of market orders sent at a high frequency on two cryptocurrency exchanges, Bybit and Binance. We analyze the execution outcomes of these orders in comparison to the expected outcome based on the most recent snapshot of the Limit Order Book (LOB) at the time of order submission for two execution modes: one using market orders and the second using marketable limit orders aiming at the best price. Discrepancies between the actual and expected outcomes are due to intermittent LOB updates during a time span resulting from delays on the exchange, delays on the trader's end, or communication delays between the trader and the exchange. We show these discrepancies are strongly correlated with market factors such as volatility, latency, and LOB liquidity. Notably, we find a consistent disadvantage to the trader, pointing to an adverse selection effect for taker orders: profitable orders (as measured by short-term future PnL returns) tend to achieve worse-than-expected outcomes, while unprofitable orders typically achieve their expected (adverse) outcomes. In the case of market orders, this translates to a worsening of fill prices, while marketable limit orders suffer from a substantial probability of failing-to-fill-immediately. Quantitative researchers who fail to take these effects into account face the familiar litany of underperforming in a live trading environment relative to stellar backtests. To address this concern, we propose parsimonious models to estimate an order's probability of failing-to-fill-immediately (in case of a marketable limit order) and the worsening of its fill price (in case of a market order), allowing for greater accuracy when carrying out backtests and minimizing the discrepancy between backtest and realized live PnL.

Keywords: Latency; Limit order book; Order failure; Slippage; Adverse selection; High-frequency trading

JEL Classifications: C55, G14

1. Introduction

Liquidity takers aiming at passive orders in the limit order book (LOB) face a moving target: the LOB in which they observed that liquidity is at least as old as the time it takes to stream information from the exchange, and by the time their orders are processed in the exchange matching engine even more time has passed. During this time gap, the exchange could process instructions by other traders that alter the LOB in such a way that, by the time the taker order (a limit order intended for immediate execution) is processed, it can no longer be matched for immediate execution. This paper sheds light on the circumstances in which such failure-to-fill events occur, their relationship with future returns, and ramifications for the backtesting of high frequency trading strategies. To this end, we conduct a trading experiment which, to the best of our knowledge, is unprecedented in the academic literature. Over a one-week period, we probed the cryptocurrency exchange Bybit with over three million orders of minimal size, and conducted a smaller-scale experiment involving 40 000 market orders on another exchange, Binance. We thus obtained a data set comprising a large quantity of orders and their execution outcomes, which we then analyse, examining the relationship between order outcomes and various microstructural properties.

*Corresponding author. Email: jakob.albers@merton.ox.ac.uk

Market volatility, exchange latency, and LOB liquidity are key microstructural factors associated with the likelihood of failure to execute immediately: as one would expect, we find that increased volatility and latency, and decreased liquidity, all give rise to an elevated probability of such failures. Critically, however, we also find that such failures to fill immediately tend to affect liquidity takers adversely. More specifically, there is an important distinction between orders that are *a-posteriori* 'good' and 'bad', as measured by short-term returns over a variety of holding horizons. Good orders that would achieve positive outcomes for the liquidity taker, have a significant probability of failing to fill as expected. Bad orders, however, which adversely affect the taker's PnL, are almost guaranteed to fill immediately, no matter the volatility regime during which the order happens to be submitted. That is, the LOB changes that occur during the aforementioned latency window are anything but random. Instead, our results suggest that there is a degree of anticipation behind the very short-term processes that happen in the latency window.

Our findings highlight the need for theoretical frameworks to incorporate the adversarial nature of trading and the effects of latency. They are also eminently relevant to practitioners. Backtests of high-frequency trading strategies that fail to account for the execution uncertainty and adverse selection effects highlighted below are bound to inflate results relative to subsequent real-world performance. We propose ways to mitigate this discrepancy, providing a model to conduct more accurate backtests that will better reflect future performance when deployed in a live trading environment.

Summary of Main Contributions

(i) For the first time in the academic literature, we conduct a large-scale trading experiment comprising millions of market orders, and analyze their execution outcomes.

(ii) We present strong empirical evidence of a latency-induced adverse selection effect for taker orders, whereby orders that would have achieved positive outcomes in the absence of latency, experience a worsening of their actual outcome due to intermittent LOB updates in the latency window. For aggressive limit orders, this amounts to the possibility of failing to fill immediately, while for market orders it means achieving a worse-than-expected fill price. Notably, we find that failure to fill as expected is associated with positive future short-term returns.

(iii) We propose parsimonious models to estimate the aforementioned adverse selection effects, with applications to strategy backtesting. For aggressive limit orders, we construct failure probability models, while for market orders we propose a model which estimates the worsening of the fill price.

Paper outline. In section 2 we provide background on the mechanics of trading and different order types, and microstructure-related effects relevant to our study. In section 3 we outline the methodology used in our trading experiment, describe the data we recorded along with some preliminary summary statistics, and introduce central concepts such as order failure and slippage. In section 4 we present our first empirical analysis, investigating volatility as a microstructural factor to explain the circumstances under which limit orders intended for immediate execution and targeting liquidity posted at the touch fail to fill immediately. In section 5, we examine the relationship between an order's failure probability and its short-term returns, introducing a model for failure probability; we also analyze results from a secondary experiment on Binance for comparative insights. In section 6, we extend the scope of our analysis to market orders, examining latency-induced discrepancies between expected and actual fill prices. We then investigate the connection between those price discrepancies and the short-term return of the order. In section 7, we introduce a methodology for more accurate backtesting of taker strategies, accounting for potential deviations from expected execution outcomes, and demonstrate it with two example trading strategies. Section 8 concludes and discusses future avenues of research.

Related Literature

This work was initially motivated by one of the most recalcitrant problems confronting quantitative researchers — the underperformance of their trading strategies in real life compared with stellar backtests. Several lines of work have examined common pitfalls of backtesting: Bailey *et al.* (2014, 2017) warn against the acceptance of strategies based solely on their in-sample performance, and propose a combinatorial approach to quantify and mitigate the risk of spurious in-sample performance leading to false positives. Caccioli *et al.* (2012) introduce an 'impact-adjusted valuation' for investment positions, critiquing traditional mark-to-market accounting for neglecting the effects of market impact in case of liquidation, and hence potentially underestimating systemic risk. A subsequent paper Kolm and Webster (2023) extends this critique to the realm of profit and loss reporting of trading strategies, arguing that PnL reported on the basis of mark-to-market accounting is similarly biased due to potential price impact. Their notion of 'fundamental PnL' adjusts for this impact, and thus provides a more accurate measure of performance. Both papers underscore the necessity of revising financial valuation methods to include the significant, yet often overlooked, price impact of large trades or inventories to liquidate. Harvey and Liu (2015) address the overestimation of investment strategy performance due to the 'multiple testing problem' where the frequent testing of various strategy configurations on the same data set increases the chance of seemingly significant, but erroneous, results. They propose a statistical method to adjust Sharpe ratios, which are commonly used to measure risk-adjusted returns, to more accurately reflect the likelihood of a strategy's true effectiveness, mitigating the distortive effects of this statistical bias.

While much of this literature is framed in relatively general terms, one may also investigate microstructural reasons, at the level of the LOBs, of why live trading tends to yield worse results than backtests. Our paper is, to the best of our knowledge, the first to demonstrate experimentally how an adverse selection effect related to the execution of taker orders contributes to this discrepancy.

General background on LOBs is provided in the foundational survey (Gould *et al.* 2013), while the separate study (Gould *et al.* 2016) notes that an LOB can experience a vast number of updates over a brief period, which has significant implications that play a large role in our paper.

Early investigations of some implications of latency on the performance and execution of marketable limit orders were conducted by Cartea, Sánchez-Betancourt and Jaimungal in a series of works which we briefly summarize. A recurrent theme in their publications is that the LOB can experience updates during delays in the communication between exchange and trader, giving rise to uncertainty about the exact state of the LOB and hence also about order execution. In Cartea *et al.* (2021), the authors present a stochastic control framework in which they solve for a latency-optimal trading strategy, which is designed to improve the execution of marketable limit orders in a dynamic LOB environment subject to random delays. In a subsequent study, Cartea and Sánchez-Betancourt (2021) provide an empirical analysis and quantify the willingness of liquidity takers to pay for reduced latency. We corroborate their empirical findings and add nuance to the picture, highlighting adverse selection effects at play in the execution of marketable limit orders. In a recent paper, Cartea and Sánchez-Betancourt (2023) also tackle the challenge of liquidating large positions in the presence of stochastic latency. They develop and compare optimal strategies for both stochastic and deterministic latencies against traditional benchmarks, demonstrating that latency-optimal strategies can significantly outperform these benchmarks by strategically using speculative marketable limit orders and the price protection they offer. Other papers, such as Maglaras *et al.* (2022) and Arroyo *et al.* (2024), focus on fill probabilities of maker orders, in contrast to our work and the papers mentioned earlier in this paragraph which focus on taker orders.

Several studies have explored how latency differentials significantly impact HFT profitability, with faster traders able to seize lucrative opportunities that slower participants miss. For instance, Baron *et al.* (2014) finds that HFT profitability is largely monopolized by a select group of firms whose speed advantage allows them to consistently outperform through latency-sensitive taker strategies, often at the expense of slower market makers. Similarly, Foucault *et al.* (2012) underscores how quick reaction to news events positively correlates with a trader's expected profitability, as faster traders capitalize on market-moving information before it's fully priced in. Additionally, Budish *et al.* (2013) highlights that the prize pool for arbitrage opportunities in technically correlated instruments like SPY and ES amounts to tens (perhaps hundreds) of millions of dollars annually. This research collectively highlights the competitive imperative of latency in capturing HFT profits. We add to this strand of literature by providing empirical evidence of a latency-induced worsening of taker order execution (hence a reduction of HFT profits) relative to a hypothetical latency-free world; although we do not explicitly discuss latency differentials between traders, the mechanism underpinning our main findings makes it obvious that slower traders suffer from a more substantial profitability reduction than quicker traders.

Our study contributes to the nascent literature on the consequences of latency in the execution of taker orders, but it differs in that we take an observational approach, making no assumptions on the dynamics of latency or the nature of LOB updates occurring during the latency window. Instead, we examine an empirical data set consisting of orders and their actual outcomes, which we assembled by conducting a large scale live trading experiment, an approach as yet unseen in the academic literature, as far as we know. Another difference with the current literature is that we explicitly relate the future return of a limit order and the probability of its failure. This supports the view that failures tend to occur more frequently at times when traders are anticipating positive returns. Although our analysis is framed in a cryptocurrency setting, it is applicable to any LOB-based market with appropriate modifications for factors such as the local fee structure.†

2. Mechanics of trading

2.1. Limit order books

The LOB now serves as the primary trading mechanism across a wide array of financial markets, including equities, spot FX, futures, options, and cryptocurrencies. The LOB is the data structure that, at any time, records all outstanding limit orders (LOs) of a given instrument. It serves as a venue, underpinning an exchange, on which market participants interact continuously to buy and sell the instrument. This continuous interaction leads to the formation of asset prices.

The LOB is displayed as a set of queues (possibly empty) of LOs, waiting to be fulfilled or cancelled, on a regular grid of *price levels* whose increment, typically very small in relation to the asset price, is one *tick*. The displayed LOs fall into two groups (sides):

- *Ask Side*: where sell LOs are placed. The ask side lists the prices (arranged in ascending order), and the total amounts at each price, at which sellers are willing to sell an asset.
- *Bid Side*: where buy LOs are placed. The bid side lists the prices, and the total amounts at each price at which buyers are willing to buy an asset, arranged in descending order.

Market participants have two primary ways to engage with the LOB, which can be broadly categorized as liquidity provision and liquidity taking:

(1) *Liquidity Provision*: Participants specify the price and quantity they wish to buy or sell. A sell order at a

† For more detail on structural characteristics of Bitcoin markets see Albers *et al.* (2021) and Alexander *et al.* (2022).

price higher than the top bid level will be inserted as an outstanding order on the ask side of the LOB. Similarly, a buy order at a price lower than the top ask level becomes an outstanding order on the bid side at the corresponding price level. These outstanding orders are also referred to as 'passive'.

(2) *Liquidity Taking*: Traders can *take liquidity* in two main ways. First, by submitting *market orders*, which require only the specification of quantity. Second, by submitting limit orders, where both quantity and a limit price need to be specified, and where the limit price is chosen to be at least equal to the top ask level for a buy order, or at most equal to the top bid level for a sell order. Both types of taker order are immediately matched with the best available orders on the corresponding side of the book. Market orders start with the liquidity at the best available price, and if that liquidity is insufficient they are matched against the next available level, and so on (known as *walking the book*). For limit orders, matching occurs only against passive orders at price levels that are better than or equal to the specified limit price. If all that liquidity is consumed, any residue of the order is left as a new passive limit order at the limit price. We refer to an LO that executes immediately by taking liquidity as a *marketable* or *aggressive* LO.

It is common in academic circles to use the term 'limit order' to mean only outstanding passive orders in the LOB. This terminology fails to capture an important nuance which is relevant for the present study; this is why we use the more general definition of LOs given above.

In summary, a market order is expected to execute immediately but the price is not guaranteed. A marketable LO may be executed immediately in full as a taker order, or it may be partially filled immediately, leaving a new passive LO. A passive LO, being posted on the opposite side of the book from aggressive LOs or market orders, benefits from a better price but has no guarantee of being executed at all. The different order types reflect the tension between immediacy of trading (but with less price-certainty) and optimizing the price (but with no certainty of execution).

2.2. Exchange fee structures

An important feature of the exchanges we study is that liquidity-taking orders, also referred to as *liquidity-consuming orders*, *taker orders* or *aggressive orders*, incur a taker fee upon execution. Likewise, liquidity-providing orders (*maker orders*) are subject to a maker fee when executed. The taker fee is always larger than the maker fee, which is often set to zero and in some cases is negative: this encourages liquidity provision via passive orders and adds to the price penalty paid for immediate execution. The exchange receives the sum of these fees upon execution of a trade.

Table 1 shows the maker and taker fees on the three largest cryptocurrency exchanges for two types of Bitcoin perpetual futures markets: USDT-margin (USDT-M), where the collateral is in USDT† and the underlying instrument is BTC/USDT; and inverse, where the collateral is in BTC and the underlying instrument is BTC/USD.‡ Most cryptocurrency exchanges employ a tiered fee system, with rates depending on trading volume. The fees in table 1 are the best available public fee rates at the time of writing, accessible only to high-volume traders. Certain traders may, however, have negotiated private agreements with exchanges that provide them with access to more favorable fees, and furthermore, fee structures are subject to frequent changes.

Table 1. Comparison of maker and taker fees (in bps) across exchanges and contract types.

	USDT-M Perpetual		Inverse Perpetual	
Exchange	Taker Fee	Maker Fee	Taker Fee	Maker Fee
Binance	1.7	0.0	2.5	− 1.0
Bybit	3.0	0.0	3.0	0.0
OKX	1.5	− 0.5	2.0	− 1.0

2.3. Order types

We now give more detail of the set of actions available to traders. These comprise an array of order types, each designed to meet specific trading objectives. While some order types, for example iceberg orders or trailing stop orders, are specific to certain markets and may not be universally available, a core set of basic order types is universal to all LOB-based markets, including spot and derivatives markets on cryptocurrencies, as well as equities, futures, and FX markets.

Basic Order Types: An order submitted to an exchange must have a *type*; the main types are

- `LIMIT`: A priced order which is executed when the market reaches the specified price. The order could be executed immediately or it could become an outstanding order in the LOB.
- `MARKET`: An order which executes immediately at the best available market price.
- `STOP_LIMIT`: A Limit Order that remains inactive until a specified stop price is reached, after which it behaves as a regular LO.
- `STOP_MARKET`: A Market Order that remains inactive until a specified stop price is reached, after which it behaves as a regular MO.

Depending on the chosen order type, additional parameters may be required, as seen in table 2.

The *time-in-force* parameter dictates how long an order remains active. The following strategies are supported on all major cryptocurrency exchanges:

† For an explainer on USDT, see for instance (Liao and Caramichael 2022).

‡ Perpetual futures are the most popular and actively traded instruments in cryptocurrency markets, accounting for a significant portion of the overall trading volume; more details on these instruments can be found in Soska *et al.* (2021) for example.

Table 2. Additional mandatory parameters by order type.

Type	Additional Mandatory Parameters
`LIMIT`	quantity, price, time-in-force
`MARKET`	quantity
`STOP_LIMIT`	quantity, price, stop price, time-in-force
`STOP_MARKET`	stop price

- *Good-Til-Cancel (GTC)*: Default setting, where the order remains active until cancelled.
- *Immediate-Or-Cancel (IOC)*: The order is immediately executed at the specified price, and any unfilled quantity is cancelled.
- *Fill-Or-Kill (FOK)*: The order is cancelled if it cannot be fully executed.
- *Post-Only (PO)*: The order is cancelled if it would be immediately filled upon submission.

2.4. Execution uncertainty

The choice of time-in-force strategies serves as a tactical tool for traders, allowing them to align their orders with specific trading objectives and risk profiles. Thus, IOC and FOK orders suit traders who aim to immediately consume specific observed liquidity in the LOB. These orders, by definition, can never become passive orders. This obviates the need for future order management, such as cancellations, thereby reducing operational complexity and potential errors against which they can be thought of as protective mechanisms.

PO orders are tailored for liquidity providers or market makers. These orders are automatically canceled if they would lead to immediate execution, thus ensuring that the trader incurs only the maker fee, which is lower than the taker fee. These orders serve to protect market makers against inadvertently taking liquidity and incurring higher fees.

One may wonder why these protective mechanisms are necessary: cannot a trader simply deduce from the most recent LOB snapshot whether a limit order would lead to immediate execution or become a passive order, and then choose their limit price accordingly? Why do traders often opt for specialized time-in-force strategies like IOC/FOK or PO?

The answer has to do with latency: the state of the LOB at the time where an order is processed internally by the exchange may not coincide with the state of the LOB to which the trader reacted, leading to an uncertain (and potentially adverse) execution, which specialized time-in-force strategies can protect against.

To elucidate further, consider two distinct points in time:

(i) When the LOB snapshot, as observed by the trader, is taken.
(ii) When the trader's order is internally processed in the exchange's matching engine.

The elapsed time between these two points is strictly greater than zero and comprises a sum of internal delays within the exchange and on the trader's end, and communication latencies between the trader and the exchange. We will call this time interval the *latency gap* or *latency window*.

During the latency gap, the LOB may change following actions by other market participants whose orders are processed within the latency gap; in particular, this applies to all orders that arrive at the exchange before the trader's specific order, and are processed after the LOB snapshot time. Such changes may prevent the trader's order from executing as initially intended: an intended passive order may inadvertently execute immediately as a taker order and thus incur the higher taker fee; or, conversely, an order intended for immediate execution may end up as a passive order in the LOB and thus become subject to higher chance of adverse selection.

Our paper is primarily an investigation of latency gap-induced discrepancies between actual and expected executions of taker orders. The mechanical reasons for those discrepancies (the 'how' question) are straightforward. To illustrate, consider the case of a buy taker order targeting liquidity at the top ask price, as seen in the latest LOB snapshot. Under what circumstances does that order fail to fill as targeted? This happens if and only if the liquidity it aims at is removed during the latency gap. Those liquidity removals in turn can occur as a consequence of two types of actions by competing takers and/or market makers: either the target liquidity is consumed first by a quicker taker; or it is cancelled by the maker(s) who posted it; or a combination of the two.

The more interesting questions are the 'when' and 'why' questions. That is, under what circumstances are these types of failures likely to occur and, in those cases, why do makers and takers take actions which make failures likely? We answer those questions in the latter sections of this paper when presenting our main findings.

3. Data and methodology

We now turn to the execution-related outcomes of taker orders, in particular:

- The likelihood that an IOC order targeting liquidity posted at the best price (as seen in the latest LOB snapshot) executes successfully;
- If the order is a market order and does not execute at the target price, how much worse or (less frequently) better is the realized price than the target price?
- How do these outcomes depend on prevailing market conditions?
- Is there any relationship between the execution outcome and the subsequent price changes?

Our approach is experimental. We placed several million minimum-sized market orders on the Bybit cryptocurrency exchange, as well as several tens of thousands on Binance. This approach is necessitated by the absence of alternative ways to determine order execution outcomes accurately: the only way to be sure of an order's execution outcome is to place the order and then review its actual result. Such an experiment can only be done with exploratory trading. Performing it on synthetic data would require a model of market latency conditional on, among other things, future anticipated returns, which is not practically feasible. To the best of our knowledge, the data set we thus assemble is unprecedented

in the realm of academic finance, in that it comprises a massive set of live orders sent to an actual exchange. This was made feasible by the lower barriers to entry in cryptocurrency markets compared with traditional assets. In cryptocurrency markets, trading infrastructure is more accessible - one can simply create an exchange account, rent server space in the same AWS region for co-location, and subscribe to free data feeds like orderbooks and trades.†

Before we discuss the details of our experiment, we summarize our objectives.

Summary of our experiment

(i) We comprehensively probed the Bitcoin market by sending millions of market orders and analyzing the execution outcomes of those orders. To minimize market impact and to control the cost of the experiment, we used the smallest possible order size.
(ii) Our stream of orders was not part of any specific trading strategy; we aimed to be agnostic about the timing of order placements, always sending a buy and a sell order simultaneously. Later in the paper, we consider subsets of these orders that would be triggered when running lead-lag and orderbook imbalance strategies.
(iii) We recorded both the immediate outcome and the return over a variety of holding times, allowing us to examine the relationship between order success and subsequent returns. We also examined circumstances in which different execution outcomes occur.

3.1. Experimental setup & details

3.1.1. Preliminary cost considerations. Our approach was to probe the market with as many market orders as possible, placed independently of any predictive signal, with the goal of retrospectively examining their execution outcomes. As we were direction-agnostic, we always sent pairs of market orders (one buy, one sell); this also avoided inventory control issues. However, we were constrained by the consequent monetary cost. Any market order pair incurs an execution cost equal to the sum of two taker fees and the spread (we buy high and sell low). Expressing this in basis points, the total cost in USD is equal to the total of volume executed across all orders multiplied by the basis point loss per order.

Consider, for example, the Bybit inverse perpetual, where the taker fee is 3 bps, the spread is around 0.16 bps, and the minimum order size is 1 USD. Then for every one million market order pairs (two million individual orders of 1 USD) placed, the cost amounts to

$$1\,000\,000 \times 1\,\text{USD} \times \frac{2 \times 3 + 0.16}{10\,000} = 616\,\text{USD}. \quad (1)$$

In contrast, if one sent these orders on the Binance inverse perpetual, which has a minimum order size of 100 USD, a taker fee of 2.5 bps, and a spread of roughly 0.03 bps, the cost amounts to

$$1\,000\,000 \times 100\,\text{USD} \times \frac{2 \times 2.5 + 0.03}{10\,000} = 50\,300\,\text{USD}. \quad (2)$$

Thus, the cost of assembling a data set of execution outcomes of one million market order pairs differs greatly between exchanges and is determined by minimum order size, taker fee, and spread.

We specifically chose the Bybit inverse perpetual as the venue of our primary large-scale experiment due to its small minimum order size, allowing us to obtain data for a large number of orders. (1 USD is the smallest minimum order size available on any Bitcoin perpetual.) The question of whether our findings might be exchange-specific is addressed in section 5, where we describe conducting a much smaller-scale experiment on Binance, involving around 40 000 orders.

Nonetheless, even when executing market orders of size 1 USD, the total executed volume and therefore the monetary cost of the data set, can spiral out of control. For instance, if we sent one buy and one sell market order every 20 miliseconds (ms) on Bybit over the course of one week, we would execute 30 240 000 such pairs of market orders, resulting in a cost of

$$30\,240\,000 \times 1\,\text{USD} \times \frac{2 \times 3 + 0.16}{10\,000} = 18627.84\,\text{USD}, \quad (3)$$

rendering the experiment infeasible, from a cost perspective, for ill-funded academic settings. We therefore need to further reduce the number of orders sent, as discussed in the next section.

3.1.2. Order placement algorithm. We weighed several options to determine when to send orders, keeping in mind three key considerations: (1) to limit the experimental cost, (2) to gather data for as many orders as possible, particularly during times which are 'interesting' from an HFT perspective (that is, we want our orders to arrive mainly at times when other market participants are also sending orders, rather than during quiet periods); and (3) to avoid exchange limitations such as rate limits and IP bans.

As noted above, sending market order pairs at regular times, close enough to ensure we capture relevant trading activity (say, 20–50 ms apart) is too costly. Moreover, the majority of such orders would arrive in 'quiet' periods and thereby bring little useful information, as market orders submitted at these times virtually always achieve the expected outcome.

We therefore considered either reacting to changes in the LOB or to trade flows. Most passive changes in the LOB (submission or cancellation of passive orders) occur in response

† Note, however, that while setting up basic infrastructure is relatively straightforward, effective trading involves complex optimizations such as handling multiple high-throughput data streams, which requires low-latency and parallelized processing to avoid delays.

to trade flow, a well-established fact (see, for example, van Kervel 2015, Chen *et al.* 2016, Degryse *et al.* 2019). We therefore opted for trade flow, using specific trade volume-based 'trigger conditions', explained below, to determine when we send orders. When a trigger was met, we placed four orders: one buy and one sell market order, and two non-executing IOC orders.†

We sent the non-executing IOC orders merely to confirm that IOC orders experience the same latency as market orders. The sequence of market orders allows us to conduct two experiments at once. First, we can investigate the execution outcome of market orders (discussed in section 6). Second, we can investigate the execution outcomes of IOC orders aiming at liquidity posted at the best price by inferring their outcomes from the observed outcome of the market orders.

To elucidate this further, consider the following pair of orders:

- $\mathrm{MO}_{t,\mathrm{buy}}$ - A min-sized market buy order submitted at time t.
- $\mathrm{LO}_{t,\mathrm{buy},\mathrm{IOC},p}$ - A min-sized IOC buy order with limit price p submitted at the same time t.

If we place only the first order, we can observe the fill price p_{fill} of that order. We note that

$$\mathrm{LO}_{t,\mathrm{buy},\mathrm{IOC},p} \text{ would have filled if it had been sent instead of } \mathrm{MO}_{t,\mathrm{buy}} \iff p \geq p_{\mathrm{fill}} \tag{4}$$

and analogously for sell orders. We can therefore infer all execution related outcomes of all min-sized IOC orders (for any choice of limit price) from the actual outcome of the market order.

Since in Bitcoin markets, trade flow is fragmented (Albers *et al.* 2021), placing orders in reaction to trade flow on our trading market (Bybit inverse perpetual) would be insufficient. We thus decided to react to trade flow cross-sectionally from the most liquid Bitcoin markets. These markets include:

- Linear perpetuals on:
  - Bybit
  - Binance (USDT-margined and BUSD-margined)
  - OKX
  - Huobi
- Inverse perpetuals on:
  - Bybit
  - Binance
- Spot markets:
  - Binance BTC/USDT
  - Binance BTC/BUSD

Altogether, these markets accounted for over 70% of Bitcoin trading volume at the time of the experiment (April 2023)†.

To avoid sending orders in too many 'uninteresting' situations, we avoid reacting to trade flow of negligible size. A new trade with a size of, say, 1 USD is unlikely to have a significant impact on the market and therefore would not activate a trigger. Preliminary experimentation, where we examined the frequency of triggers under different thresholds, suggested 10 000 USD as a reasonable minimum trade size to react to. However, we adjusted this figure slightly to sidestep any potential anomalies or odd effects that might occur at such a round number, settling on a threshold of 9500 USD. In summary, we periodically (once per millisecond) check the above nine markets. Then, if any of them saw more than 9500 USD worth of trade volume executed since the previous time we placed four simultaneous orders, as previously described.

Let us denote the markets we observe by $\mathtt{m}_1, \ldots, \mathtt{m}_9$. The complete order placement algorithm is outlined in Algorithm 1.

3.1.3. Notes on latency and co-location. We make several important remarks regarding latency and co-location in the context of our experiment.

- Most cryptocurrency trading occurs on exchanges hosted in various AWS data centers, including Binance (AWS Tokyo), Bybit (AWS Singapore), and OKX (AWS Hong Kong).‡ These exchanges rely on third-party cloud services, so they do not offer co-location directly; traders can simply (and affordably) co-locate by renting a server in the same data center.
- Since our experiment involved trading on Bybit, we co-located our server to that exchange. This involved renting server space in AWS Singapore, situated close to Bybit's servers, and selecting the availability zone that provides the lowest latency. For a secondary experiment conducted on Binance (described in more detail in subsequent sections), we similarly rented a co-located server in AWS Tokyo.
- To minimize latency, we subscribe to data feeds and place orders using the co-located server.
- Cross-region data transfer delays are unavoidable. For instance, Binance trades data received from our server in Singapore will always be at least as old as the minimum transfer time required between the two regions, though there can be differences in how one chooses to transfer such data. For instance, information between stock exchanges in Frankfurt and London can be transmitted via private microwave tower connections that are much quicker than cable.§ Still, even with undersea cable information transfers, there are subtle differences between approaches one can take, leading to potentially different latencies. Additional details on the networking optimizations we performed are available upon request.
- To avoid internal latency on our end, we distributed processing load across multiple CPUs, ensuring compute load on each CPU remained small. Similarly, we took precautions to ensure sufficient RAM to prevent swapping.

† By 'non-executing', we mean that we set a limit price that ensures the order could not be immediately filled, and thus is cancelled due to its order type.
† See for instance Coingecko.

‡ OKX also has servers in Alicloud HK.
§ Due to the limited range of microwave towers, and the fact that the geodesic path between Singapore and Tokyo largely crosses water, such microwave connections do not yet exist between exchanges located in Singapore and Tokyo. It would require dozens of microwave towers on floating platforms across the ocean.

Algorithm 1 Order Placement Algorithm

```
previous_ts ← None
threshold ← 9500
trigger ← False
while True do
  ts ← currenttimestamp
  if previous_ts ≠ None then
    for i = 1,...,9 do
      volume_diff ← tradevolumeinUSDonm_iintheinterval[previous_ts, ts]
      if volume_diff > threshold then
        trigger ← True
      end if
    end for
  else
    trigger ← True
  end if
  if trigger then
    previous_ts ← ts
    Place two buy and two sell orders (one market and one non-executing IOC in each case)
  end if
end while
```

- More detailed information about the architecture we employed is available upon request (we omit more details here, so as not to detract from the main points of the paper).

3.1.4. Collected data. For each order we sent, we recorded an array of details and information received from the exchange in response to the order submission. The data we gathered includes:

(i) *Order details*:
- `SubmittedTime` - Time of submission
- Limit price
- Amount
- Order type
- Side

(ii) *Exchange response*:
- Order response indicating whether the API request succeeded or failed
- `AcknowledgmentTime` = Time of response receipt

(iii) *Order status update*:
- Updated order status (filled or canceled)
- Fill price
- Two timestamps:
  (a) `CreatedTime` - the time at which the order was created in the matching engine
  (b) `UpdatedTime` - the time at which the order was processed by the matching engine

The exchange response and order status update are received as separate messages, with the status update containing considerably more information than the initial order response. An exploratory analysis of the collected latency data is provided in Supplementary Material B.

It is worth noting that, although the response usually arrives first, the order of arrival between the response and the order status update is not deterministic, meaning it remains unclear which of the two messages will be received first. We hypothesize that the order response is sent to us by the exchange once the order is successfully created within the matching engine, which corresponds to the `CreatedTime`, while the status update is issued after the order has been matched, corresponding to the `UpdatedTime` timestamp. Given that the time interval between these two timestamps is typically quite small, it is possible that the message sent out earlier arrives later than the second message, due to network-related factors.

As well as data from all our submitted orders, we collected LOB data from the Bybit and Binance inverse perpetual markets, comprising Level 2 LOB snapshots which capture the top 50 price levels; Level 3 (market-by-order) data is unavailable on those two exchanges. The interval between LOB snapshots varied from 20 to 200 ms, depending on market activity.

LOB data and private message feeds regarding orders we placed are streamed asynchronously from the exchange to us and it is possible for them to become mistimed relative to one another at times. Such mistimings are likely most common when the exchange is under heavy load (e.g. during periods of high volatility) and one of the feeds becomes delayed. While the core question in our paper pertains to the LOB feed (specifically, whether we can execute at the price seen in the latest available snapshot), one could conceivably use both feeds to construct an 'inferred LOB snapshot' that might be more accurate than the LOB feed alone, particularly when the feeds provide conflicting information due to mistimings. Taker orders aiming at liquidity in such an inferred LOB snapshot are likely to have a higher hit rate than ones reacting to the most recently received actual snapshot. In this paper, we content ourselves with the more elementary analysis involving actual LOB snapshots, reserving its more sophisticated counterpart for future work.

Table 3. Trigger count by market.

Market	Trigger Count (in Thousands)
Binance Linear USDT	797.16
Bybit Linear	200.00
OKX Linear	196.37
Binance Spot USDT	134.78
Binance Linear BUSD	86.28
Binance Inverse	67.69
Bybit Inverse	30.69
Huobi Linear	26.92
Binance Spot BUSD	24.81

3.2. *Scope of our experiment*

We conducted our experiment from from 19:43:04 UTC on April 14, 2023, to 20:10:02 UTC on April 21, 2023. In this period, the price of Bitcoin fluctuated between 27174.5 and 30605.0 USD, with an annualized volatility of 42.49% (computed from hourly log returns); this aligns with the 40%–50% range of implied volatilities for at-the-money Bitcoin options†, suggesting that our experiment spanned a representative week of Bitcoin trading activity. We sent a total of 6 277 628 orders to the exchange. Half of these, 3 138 814, were market orders and therefore were filled immediately. The other half were non-executing taker orders, hence ended up being cancelled. Our trading losses were around 990 USD. For more details, see the Supplementary Material A.

As described earlier, our triggers for placing orders were based on the cross-sectional trading activity across nine major Bitcoin markets. Table 3 provides a breakdown of how often each individual market triggered our order placement. This ranking mirrors the ranking of those nine markets by average daily trading volume, which is unsuprising given the definition of our triggering conditions.

3.3. *Basic definitions*

We begin by introducing two critical definitions pertaining to the outcomes of taker orders. These notions serve as a foundation of our analysis in the remainder of this paper.

3.3.1. Slippage. Consider a market order $\mathrm{MO}_{t,\mathrm{side}}$ with side $\in \{\mathrm{buy}, \mathrm{sell}\}$, submission time t, and of the minimum order size. Orders of these types comprised part of our trading experiment. We write $p^{\mathrm{fill}}(\mathrm{MO}_{t,\mathrm{side}})$ for the fill price of the market order. We further denote by $p_{t,a}$ and $p_{t,b}$ the most recently observed top ask and top bid prices at time t, respectively.

Naively, one would expect the market order $\mathrm{MO}_{t,\mathrm{side}}$ to fill at price $p_{t,a}$ or price $p_{t,b}$ for a buy or sell order, respectively.‡ However, as noted earlier, by the time the order is processed in the matching engine, the top of the book may have moved up or down from its value observed as at submission time t. In such cases, the actual fill price will have worsened or improved. We quantify this price difference, dubbed *slippage*, as follows

$$\mathrm{slip}\left(\mathrm{MO}_{t,\mathrm{side}}\right) := \begin{cases} p_{t,a} - p^{\mathrm{fill}}\left(\mathrm{MO}_{t,\mathrm{side}}\right) & \text{if side} = \text{buy} \\ p^{\mathrm{fill}}\left(\mathrm{MO}_{t,\mathrm{side}}\right) - p_{t,b} & \text{if side} = \text{sell} \end{cases} \tag{5}$$

Slippage measures the deviation between an order's actual and anticipated fill prices. Irrespective of the order's side, positive values represent a price improvement, while negative values signify a worsening. Slippage can be seen as an extra cost incurred by liquidity takers due to latency.

3.3.2. Order failure. Closely related to slippage is the notion of order failure. We say that a limit order submitted with time-in-force parameter FOK or IOC has failed if its filled quantity is zero. Typically such orders aim at outstanding orders in the LOB. Failure means the order does not execute against the targeted order (due to intervening updates of the LOB), and is therefore immediately cancelled by the exchange without having filled any amount.

In case of our trading experiment, if instead of a min-sized market order $\mathrm{MO}_{t,\mathrm{side}}$ we had sent an IOC or FOK order§ with limit price

$$p := \begin{cases} p_{t,a} & \text{if side} = \text{buy} \\ p_{t,b} & \text{if side} = \text{sell}, \end{cases} \tag{6}$$

targeting liquidity at the best price for immediate execution, then this order would have failed if and only if $\mathrm{slip}(\mathrm{MO}_{t,\mathrm{side}}) < 0$.

This motivates us to say that a market order $\mathrm{MO}_{t,\mathrm{side}}$ has *failed* whenever it satisfies $\mathrm{slip}(\mathrm{MO}_{t,\mathrm{side}}) < 0$. In our empirical data set from our experiment, the global failure rate across all orders was 1.354%.

3.3.3. Markouts. To analyze the economic performance of our orders, we first need to decide how to measure future returns. While conventional methods may involve setting a fixed time horizon, we argue that examining returns in *event time* is more robust. Event time here refers to a (variable in calendar time) window spanning some fixed number N of future trading events, for example number of trades, traded volume over a threshold, or orderbook updates. An event-time-based unit of measurement for future returns inherently incorporates volatility and therefore provides a built-in adjustment for varying market dynamics.

We therefore define the *future returns* (interchangably referred to as *markout returns*) of a market order $\mathrm{MO}_{t,\mathrm{side}}$ as follows:

$$\mathrm{fret}\left(\mathrm{MO}_{t,\mathrm{side}}\right) := \begin{cases} \left(\dfrac{p_{a,t+\tau}}{p_{a,t}} - 1\right) \cdot 10\,000 & \text{if side} = \text{buy} \\ \left(\dfrac{p_{b,t}}{p_{b,t+\tau}} - 1\right) \cdot 10\,000 & \text{if side} = \text{sell} \end{cases} \tag{7}$$

† Implied volatilities of Bitcoin options prices can be seen on Deribit.

‡ This is under the assumption that the liquidity at the best price is at least as large as the quantity of the market order. If this does not hold, the market order would be expected to walk the book.

§ Note that IOC and FOK orders are equivalent when using the minimum order size.

Table 4. Summary statistics for the return horizon τ based on a fixed number of ten trigger events.

Statistic	Value (in seconds)
Average	1.48
Std Deviation	1.08
Minimum	0.10
25th Percentile	0.65
Median	1.14
75th Percentile	2.01
Maximum	5.00

giving a positive return from a favorable price movement and a negative return from the converse, irrespective of the side (buy or sell).† Note that this definition of future return does not account for the trading fees. The term $t + \tau$ represents the earliest point after t at which ten new triggers (as per Algorithm 1) have occurred. The time gap τ varies depending on cross-sectional Bitcoin trading activity, typically between 650 milliseconds (the 25th percentile) to 2.01 seconds (the 75th percentile). Summary statistics can be seen in table 4. An added benefit of using an event-based future return horizon is that at each such event we sent both a buy and a sell market order as part of our experiment. This allows us to compare an order's outcome with the actual outcome of a subsequent opposite-sided order. We use the term *markout* to refer to the future return over the particular horizon specified above.

4. Correlation between order failure and volatility

We begin our analysis by investigating the relationship between an order's probability of failure and volatility. Our findings corroborate those of Cartea and Sánchez-Betancourt (2021) by empirically supporting their assertion that market volatility is positively correlated with MLO failures. This may seem intuitive because volatility is generated by top-of-book changes and order failures are a consequence of price-changing LOB actions (taker orders and/or cancellations) during an order's latency gap. In subsequent sections we offer evidence for a less intuitive modified version of the hypothesis: volatility and order failures are essentially uncorrelated if those orders are ex-post loss-making in an HFT sense (the orders always fill, to the detriment of the trader).

Figure 1 presents a price time series alongside failure rates, aggregated into 30-minute bins. A visual inspection of figure 1 supports the hypothesis that periods of higher volatility are associated with elevated failure rates.

The notion of volatility we adopt mirrors that of Cartea and Sánchez-Betancourt (2021), facilitating comparison of our results with their FX market analysis. We define the *microprice* at time t as $m_t := (p_{a,t}q_{a,t} + p_{b,t}q_{b,t})/(q_{a,t} + q_{b,t})$, where $p_{a,t}$ and $q_{a,t}$ represent the top ask price and liquidity at time t, respectively, with analogous terms for the bid side.

† Some refer to this as the 'PnL future returns'; however, for simplicity and consistency, we use the term 'future returns' or or 'markout returns' throughout the paper.

The microprice is sampled every 500 ms based on the latest LOB snapshot. Log returns are computed as $\text{ret}_t := \log m_t/m_{t-500\,\text{ms}}$. For a time interval $[t, T]$ where $T = t + k \cdot 500\,\text{ms}$ for some $k \in \mathbb{N}$, price volatility is defined as the standard deviation of log returns

$$\text{vol}_{[t,T]} := \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(\text{ret}_{t+i\cdot 500\,\text{ms}} - \overline{\text{ret}})^2},$$

where $\overline{\text{ret}}$ represents the average return over the time interval. Note that we use calendar time rather than event time in this definition.

4.1. Contemporaneous and lagged correlations

We first explore the contemporaneous and lagged correlations between volatility and the likelihood of failure across various time intervals.

For a given (calendar time) interval duration $\Delta > 0$, we segment the sample period into non-overlapping intervals $t_0 < t_1 = t_0 + \Delta < \cdots < t_k$ of equal length Δ. Within each interval $[t_{i-1}, t_i]$, we consider all orders sent to the exchange, denoted as $\text{order}_1, \ldots, \text{order}_{n_i}$. The failure rate of orders executed in this interval is denoted as ϕ_i, and is defined by

$$\phi_i := \frac{\sum_{j=1}^{n_i} \mathbf{1}_{\{\text{order}_j \text{ fails}\}}}{n_i}. \tag{8}$$

The volatility in each interval $[t_{i-1}, t_i]$ is represented by $\text{vol}_i := \text{vol}_{[t_{i-1},t_i]}$.

We compute two types of correlations: contemporaneous and lagged. The contemporaneous correlation, denoted as $\text{Corr}_{\text{cont}}$, is the Pearson correlation coefficient between the sequences of volatility (vol_i) and failure rates (ϕ_i) for each interval i. This is expressed as

$$\text{Corr}_{\text{cont}} = \text{Corr}\,(\text{vol}_i, \phi_i) \quad \text{for } i = 0, \ldots, k \tag{9}$$

The lagged correlation, denoted as Corr_{lag}, follows from the relationship between the volatility of the preceding interval and the failure rates of the current interval. It is given by

$$\text{Corr}_{\text{lag}} = \text{Corr}\,(\text{vol}_{i-1}, \phi_i) \quad \text{for } i = 1, \ldots, k \tag{10}$$

The contemporaneous correlation provides insights into the role volatility plays in driving failure rates, while the lagged correlation shows how well past volatility (over the previous non-overlapping time interval) predicts failure rates in the subsequent interval. Table 5 presents the results.

Contemporaneous correlations are high across all time intervals, with its maximum value achieved for the one hour window. The high correlations are consistent with the economic intuition that during volatile periods LOB updates are more frequent than at quiet times, thus increasing the chance of such an update occurring during an order's latency gap and resulting in its failure. Lagged correlations are also large at time scales of several minutes to hours, with maximum achieved at the 10 minute lookback window, indicating that past volatility can be used as a predictor of future likelihood of

Figure 1. Bitcoin price and order failure rates: This figure shows the Bitcoin price (left y-axis) and 30-minute aggregated order failure rates during our experiment (right y-axis).

Table 5. Contemporaneous and lagged correlation over time intervals.

Window	$Corr_{\text{cont}}$	$Corr_{\text{lag}}$	Pts
1s	0.4481	0.1312	413 050
10s	0.5843	0.2550	55 714
1min	0.6757	0.3899	10 037
10min	0.7913	0.5458	1009
1h	0.8009	0.4012	168
2h	0.7634	0.4049	84
4h	0.7509	0.4085	42

failure. For small lookback windows lagged correlations are substantially lower, likely stemming from the fact that volatility estimates based on high-frequency data are noisy due to the Epps effect and microstructure noise; see for instance Hansen and Lunde (2012) and Zhang (2011).

4.2. *Failure rates conditioned on volatility*

Consider the probability that an order submitted at time t fails, conditional on the rolling 10-minute microprice volatility taking the value x: $\mathrm{P}(\text{order}_t \text{ fails} \mid \text{vol}_{[t-10\,\text{min},t]} = x)$. We estimate this conditional probability from our data set.

Let $\mathcal{O}$ denote the set of market orders executed during our experiment. We define the function $\text{vol} : \text{order}_t \mapsto \text{vol}_{[t-10\,\text{min},t]}$, which maps $\text{order}_t \in \mathcal{O}$, an order with submission time t, to the microprice volatility over the preceding 10 minutes.

For any volatility level $x > 0$ and bin size $\epsilon > 0$, we define the following subset of orders:

$$\mathcal{O}_{x,\epsilon} := \{o \in \mathcal{O} : \text{vol}(o) \in [x, x+\epsilon)\}, \tag{11}$$

and its associated (empirical) failure rate:

$$\phi_{x,\epsilon} := \frac{\sum_{o \in \mathcal{O}_{x,\epsilon}} \mathbf{1}_{\{o\text{fails}\}}}{|\mathcal{O}_{x,\epsilon}|}. \tag{12}$$

The subset $\mathcal{O}_{x,\epsilon}$ comprises orders executed when volatility was approximately x, while its associated failure rate approximates the conditional probability:

$$\phi_{x,\epsilon} \approx \mathrm{P}\left(\text{order}_t \text{ fails} \mid \text{vol}_{[t-10\,\text{min},t]} = x\right) \tag{13}$$

We implement these definitions as follows. First we compute the minimum and 99th percentile† of volatility values seen in the sample period of our experiment. Next, we partition this range of volatility values into 100 equal-length subintervals with grid points $x_1, \ldots, x_{100}$, and we set $\epsilon := x_2 - x_1$ in order to align with the scale of our observations and ensure that the sets $O_{x_i,\epsilon}$ and $O_{x_j,\epsilon}$ are disjoint whenever $i \neq j$. One needs to be careful to avoid aliasing effects in the choice of ϵ, as an incautious choice can lead to erroneous results or misleading plots - in our analysis, we were aware of this issue and carefully chose ϵ to mitigate such risks.

Figure 2 presents a plot of pairs $(x_i, \phi_{x_i,\epsilon})_{i=1,\ldots,100}$ along with an overlaid OLS regression line. The regression fit yields an R^2 value of 72.2%, suggesting a strong linear relationship between volatility and failure rate. However, the plot becomes noisier at large levels of volatility, which can be explained by the smaller number of orders in the sets $\mathcal{O}_{x,\epsilon}$ for large x, leading to noisier estimates of $\phi_{x,\epsilon}$, as is evident in the plot.

In Supplementary Material C we look at the autocorrelation properties of order failures of successive trade attempts, finding that order failures exhibit significant autocorrelation. We also examine the connection between order failures and two further factors, LOB liquidity and latency, showing a significant degree of dependence.

† We do not use maximum here to avoid skewed results due to rare outlier values.

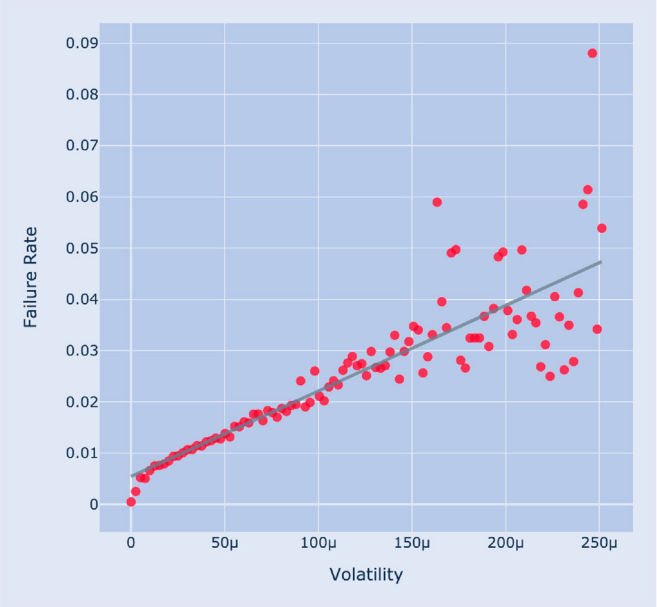

Figure 2. Failure rates conditioned on 10-minute microprice volatility.

5. Top-of-book execution

5.1. *Good orders can fail*

In section 3.3, we made precise the notions of order failure and future returns. Equipped with these foundational definitions, we now explore our empirical order data, in particular shedding light on the relationship between a taker order's probability of execution and its future returns. To this end, we bucket orders in our data set $\mathcal{O}$, which consists of over 3 million market orders, based on their future returns in basis points (bps); we then determine failure rates for each bucket.

To make this precise, let us introduce some notation. We define the variable ξ_{bb} as the tick size (0.5 USD on Bybit) converted to bps at the maximum observed price of 30605.0 USD during our experiment, resulting in $\xi_{bb} = 0.184$ bps.†

Next, we introduce a function φ that maps the future return $\mathrm{fret}(\mathrm{O})$ of an order O to its nearest multiple of ξ_{bb}, denoted as $\varphi(\mathrm{fret}(\mathrm{O}); \xi_{bb})$. We then partition $\mathcal{O}$ as $\mathcal{O} = \bigsqcup_{i\in\mathbb{Z}} \mathcal{O}_i$, where

$$\mathcal{O}_i = \{\mathrm{O} \in \mathcal{O} \mid \varphi(\mathrm{fret}\,(\mathrm{O})\,; \xi_{bb}) = i \cdot \xi_{bb}\}. \tag{14}$$

In other words, $\mathcal{O}_i$ is the set of orders whose rounded future return in bps is equal to $i \cdot \xi_{bb}$.

For each bucket $\mathcal{O}_i$ we then compute the empirical failure probability, defined as

$$\gamma_i := \frac{|\mathcal{O}_{i,fail}|}{|\mathcal{O}_i|}, \tag{15}$$

where $\mathcal{O}_{i,fail} := \{\mathrm{O} \in \mathcal{O}_i \mid \mathrm{O} \text{ failed}\}$ represents the subset of taker orders that failed (according to our definition from section 3.3).

† Over the span of our Bybit experiment, the price ranged between 27174.5 and 30605.0 USD, implying that the tick size expressed in bps exhibited only small variation from 0.164 to 0.184.

The vast majority of orders in our data set are concentrated at small future returns (in absolute terms), while orders achieving extreme return values are rare. The most negative future return value we observed was −72.78 bps, while the most positive was 51.14 bps. However, if we restrict ourselves to only buckets with at least 10 orders, then the return values range from −14.54 to 14.54 bps (this symmetry appears to be coincidental). As failure rates associated with buckets containing only very few orders are imbued with a higher degree of uncertainty, we ignore buckets $\mathcal{O}_i$ with $|\mathcal{O}_i| < 10$ in the subsequent analysis. This reduces the noise in our data at the extreme values.

We were careful to avoid aliasing effects with our bucketing choice above. When these plots are compared with their analogues based on the number of tick returns, which offer maximum granularity, they appear largely identical to the bucketed basis point return plots considered here.

A scatter plot exhibiting the relationship between future returns and failure rate can be seen in figure 3. To visually represent the different number of orders in each bucket, we colored each data point based on the (base 10) logarithm of the order count associated to it.

Several patterns emerge upon inspection of the scatter plot:

(i) *Consistent 'Success' at Negative Returns*: Taker orders achieving negative future returns (i.e. 'bad bets') almost always succeed in their execution.
(ii) *Jump at the Origin*: There is a noticeable jump in failure rate as returns cross zero. Orders with subsequent positive short-term returns have a non-negligible risk of failure; those with negative future returns do not.
(iii) *Correlation of Success with Future Returns*: As the future returns increase, so does the failure rate. This growth appears to be monotonic.
(iv) *Noise in Extreme Values*: Empirical estimates linked with larger future returns (in absolute value) tend exhibit a larger degree of variance, indicating that our sample size may not be robust enough for extreme return values.

In section 2.4, we noted *how* taker order failures arise: the mechanism immediately responsible is liquidity removal, within the latency gap, by competing makers canceling their orders and takers filling theirs. The results presented above provide answers to the question of *when* order failures tend to happen: orders with a positive future return fail with a probability significantly greater than zero, while those with a negative return are almost guaranteed to fill. But *why* do failures occur almost only for ex-post profit-generating orders (at markout time) and only very rarely for loss-making ones? Are liquidity takers cursed? And why does the data exhibit the peculiar jump at the origin?

General Discussion. An MLO fails if, and only if, the price moves in the order's direction during the latency gap, i.e. the order has an instantaneous positive return.‡

For this positive return to turn into a non-positive markout return, the price would need to fully retrace by at least

‡ This refers to the return with respect to the (possibly unobserved) LOB state immediately after the fill.

Figure 3. Failure rate by future returns with area representation. The colour scale on the right of the plot is the (base 10) log of the order count.

the magnitude of the initial move. This would be rare if price moves were independent, and is made even less likely by the strong autocorrelation of returns at the small timescales considered here. Thus, the great majority of failures occur at positive markout returns (the positive segment in the plot) and here are very few at non-positive markout returns.

Conversely, orders that succeed (i.e. they fill) experience either a zero instantaneous return (no price-changing action in the latency gap) or a negative instantaneous return (where the fill occurs at a more favorable price than targeted), with the former being far more common. Among filled orders with zero instantaneous returns, most also exhibit zero markout returns (no price change by the markout time), while some show positive markout returns (corresponding to the right-hand side of the plot) or negative markout returns (corresponding to the left-hand side). Filled orders with negative instantaneous returns almost always have negative markout returns, as a retrace of the immediate negative return during the markout window is highly unlikely for the same reason described above. These orders thus populate almost exclusively the left-hand side of the plot, contributing to the suppression of failure rates for non-positive markout returns.

These arguments provide a mechanical explanation for the low (close to zero) failure rates observed at non-positive markout return† and the significantly greater-than-zero failure rates at positive returns.

The vast majority of orders have a zero instantaneous and markout return, i.e. there is no price change (see the large yellow circle in the above figure, or the colour scale indicating order count), and there are so many of those cases that this is why we have an especially small (close to zero) failure rate at zero.

We note that a positive instantaneous return leading to an MLO's failure can only result from competing taker orders, maker cancellations, or a combination of both. Analyzing the respective contributions of each is an interesting avenue for future work that could provide insights into the adverse selection costs faced by makers and the HFT profits realized by takers.

To deepen our discussion and build further intuition, let us now consider an important class of examples: trading with signals.

A simple toy model. Suppose there is a public signal of the asset's value, observed by most or all high frequency takers and market makers. This could for instance be something as simple as the asset's price on a much more liquid venue. Consider what happens when there is a jump in that public signal resulting in a substantial (temporary) discrepancy between the asset's exchange-published price and its signal-implied price: at this moment, takers send their MLOs in the direction of the signal and makers cancel their stale opposite-direction orders (which collectively comprise the same liquidity targeted by the takers).

Two things (or a combination thereof) then happen: (1) some taker order(s) may fill, perhaps also (by removing the top queue) creating a positive instantaneous return which, all else being equal, is positively correlated with a positive short-term markout return. Any maker(s) who failed to cancel in time have a corresponding negative instantaneous return (or opportunity cost); (2) some market makers canceled their stale quotes in time, some takers failed to fill and capture the positive short-term return.

To make things more concrete, consider a scenario where a jump in the public signal prompts simultaneous submissions

† Note that since the vast majority of all orders in our data set exhibit zero instantaneous and markout returns (evident from the large yellow circle in figure 3 and the color scale indicating order count), we obtain an especially small failure rate at a zero bps return.

of the following messages, processed in a random order by the exchange:

- We submit a minimum-sized MLO targeting the top-of-book liquidity.
- $N_T \geq 0$ additional takers submit MLOs, each exceeding the available top-of-book liquidity.
- $N_M \geq 1$ makers, collectively responsible for the liquidity at that level, each send a cancellation request.
- The arrival sequence of these arders is random.

What is the the probability that our MLO fills? Our order must be processed before any other taker order and at least one cancellation request must arrive after it. The probability that no taker order arrives before the queue clears is $1/\binom{N_T+1+N_M}{N_M}$. Thus, the probability that at least one taker order is processed before the queue clears is $1 - 1/\binom{N_T+1+N_M}{N_M}$. Since all taker orders are equally likely to arrive first, the probability that our MLO is the first to fill is:

$$P(N_T, N_M) := \frac{1}{N_T + 1} \cdot \left(1 - \frac{1}{\binom{N_T+N_M+1}{N_M}}\right) \tag{16}$$

As a sanity check:

- When $N_T = 0$ and $N_M = 1$, we are the only taker with one maker, yielding $P(0, 1) = \frac{1}{2}$.
- For $N_T = 1$ and $N_M = 1$, the exchange randomly orders the two taker orders (ours and one competing one) and the cancellation in six possible ways, of which two have our order first and thereby filling.
- With $N_M = 1$ and $N_T \to \infty$, the fill probability approaches zero.
- With $N_T = 1$ and $N_M \to \infty$, the fill probability approaches $\frac{1}{2}$.

A larger jump in the signal (corresponding to a larger future return) is likely to prompt more takers to react, increasing N_T. With N_M fixed, this causes $P(N_T, N_M)$ to decrease, resulting in a higher failure rate. as we observe empirically.

We note that no rational HFT taker would trade against a strong signal, for example submitting a buy order in response to a sell signal, nor would a market maker have any reason to cancel their sell order. The taker would be virtually certain to fill, highly likely to achieve an adverse markout return, and would pay the taker fee, while the maker order would only be filled by such irrational takers so it becomes irrelevant and can be canceled later if it becomes exposed.

Finally, in the absence of a strong signal (zero or near-zero future return predicted), we expect an MLO to fill with probability close to 1 since we have $N_T = N_M = 0$ most of the time (with occasional exceptions, such as a long-term holder submitting a taker order at that moment), as takers would avoid submitting orders incurring taker fee without any compensating return, while makers would see little incentive to cancel without the risk of adverse price moves.

Assumptions in our toy model. Our computations rely on several simplifying assumptions—why, despite these, is our result realistic?

In essence, it comes down to transmission time between traders and the exchange; its average and standard deviation dominate differences in traders' reaction times when responding to an "obvious" signal (one shared among traders in highly correlated though not identical forms).

Specifically, even from a co-located server (see section 3.1.3), the time to send a message (e.g. an MLO or cancellation request) to the exchange is random, with an average around 2 ms and a standard deviation around 1 ms. By contrast, the slight differences in traders' reaction times—due to minor differences in signals or processing time (including compute time)—are likely only a few microseconds. As a result, it effectively becomes a lottery, much like in our toy model, as to which trader's message reaches the exchange first when multiple traders are sending their message at nearly the same time (within a few microseconds of one another).

Predictions. Our model yields a set of testable predictions that researchers with access to the right type of data could investigate:

- *Takers*: Holding the number of makers and their cancellation attempts constant, an increase in the number of takers leads to a higher failure rate.
- *Makers*: With the number of takers fixed, an increase in cancellation attempts per maker order also results in a higher failure rate. Fewer makers at the top-of-book, while holding their cancellation attempts and the number of takers constant, produces higher failure rates.

Extending the mechanism behind the toy model, we further predict:

- *Signal type*: More obvious signals attract more takers and maker cancellations, leading to higher failure rates compared to subtler signals. Would-be HFTs looking to maximize execution probabilities should therefore focus on devising signals based on more concealed patterns that are not widely acted upon.
- *Single maker*: Failures associated with negative markout returns from the taker's perspective represent missed fills with positive returns for the maker if the failure occurred as the result of a cancellation. Such failures become more likely when all the liquidity at the top-of-book is provided by exactly one maker because a single maker's cancellation signal is far more likely to misfire compared to multiple makers with different cancellation signals (even if they are highly correlated). Simultaneous misfires across many makers are rare, reducing the likelihood of such failures in a more diversified top-of-book.
- *Signal-agnostic takers*: Taker orders submitted agnostic to short-term returns (e.g. by a long-term holder trading without a short-term signal) have a substantial probability of yielding a negative markout return. When this occurs, any other MLO submitted at roughly the same time and in the same direction will have a significantly positive failure rate, despite its negative markout return. Thus, a stronger presence of signal-agnostic takers increases failure rates at negative returns.
- *Signal strength*: Stronger signals not only attract more takers at the touch, resulting in higher failure rates, but

also prompt takers to target deeper levels in the book (e.g. using MLOs with deeper limit prices). This leads to greater instantaneous returns (or slippage for market orders, as detailed in section 6 and supported by table 7).

5.2. *Secondary experiments: Binance and Ethereum*

To evaluate the robustness of our findings across different exchanges, coins, and their persistence over time, we conducted two additional experiments:

- *Binance Bitcoin Inverse Perpetual*: This experiment aimed to test whether our findings would be replicated on a different exchange with differing fee structure, tick size, and trading volume. It involved 40,588 market orders executed over approximately one hour, from 14:02:39 to 15:00:00 UTC on July 20th, 2023.† Despite the substantially smaller number of orders, the notional trading volume was larger and more costly than in the Bybit experiment because the minimum order size on Binance's inverse perpetual is 100 USD compared twith Bybit's 1 USD. The experimental methodology was identical to that of the Bybit experiment. We co-located with Binance in AWS Tokyo, similar to our setup with Bybit in AWS Singapore, by renting a server in the same data center.‡
- *Bybit Ethereum Inverse Perpetual*: Conducted from 10:12:01 UTC on October 9th, 2024, to 08:49:57 UTC on October 11th, 2024, this experiment involved 1 056 203 market orders, all of which were min-sized with an order size of 1 USD, equal to that of Bybit's Bitcoin inverse perpetual. This experiment, conducted over a year after our initial BTC experiments, followed the same methodology with minor adjustments for the different coin. The resulting data allows us to ascertain (1) whether our findings exist across different coins; and (2) their persistence over time, given the large time gap between this and the Bitcoin experiments.

Figure 4 provides an overview of the microstructural properties of the three markets we studied. Figures 5 and 6 illustrate the relationship between failure rate and future returns for Binance and Ethereum, respectively, using the same bining procedure as previously described, with bin sizes $\xi_{bn} = 0.034$ and $\xi_{eth} = 0.215$.§

The plots reveal a general pattern consistent with the original Bybit Bitcoin experiment: failure rates near zero for negative returns, followed by a discontinuity and a monotonically increasing trend. However, there are notable differences in the details: the magnitude of the jump (5% for Bybit Bitcoin, 6% for Bybit Ethereum, and 12% on Binance), the steepness of growth in the positive segment, and significantly larger failure rates on the negative segment in the Ethereum data compared to the two Bitcoin experiments.

Explaining the differences. What accounts for the notable discrepancies in failure rates across exchanges and coins? Our hypothesis draws on the toy model from the previous subsection: we attribute these differences to variations in the trader pools.

- *Binance is more competitive than Bybit:* One possibility is that the number of taker orders on Binance sent in response to a signal of a given magnitude (denoted N_T^{bn}) is greater than that on Bybit (N_T^{bb}), while maker behavior is the same on both exchanges (N_M cancellation requests on each). Our toy model (see equation (16)) then implies a higher failure rate on Binance, which is consistent with our empirical data. Beyond a larger number of takers, other variations in trader pools could also explain the observed results. For instance, even if $N_T^{bb} = N_T^{bn}$, Binance makers might, in the presence of a strong signal, submit a larger number of cancellation requests per order than their counterparts on Bybit, effectively reducing the fill probability of our MLO. While we lack the data to pinpoint the exact underlying differences in trader pools, all possibilities suggest fiercer competition on Binance: this is also consistent with the much larger trading volumes, see figure 4.
- *Bybit's Ethereum market has fewer makers*: Bybit's Ethereum LOB is relatively illiquid compared to those of the two Bitcoin markets, particularly at the touch (see figure 4), suggesting fewer maker orders (observational data suggests there may often be just a single maker order at the touch). This can explain the differences we observe between the two coins:
  (i) As noted in our predictions in section 5.1, failures with negative returns—representing missed opportunities for makers and favorable misses for takers—are much more likely with fewer market makers responsible for the liquidity at the touch and especially likely if there is just one.
  (ii) The higher failure rates on Bybit's Ethereum market compared to its Bitcoin market can again be attributed to fewer makers at the touch, as discussed in the second prediction at the bottom of section 5.1.

In summary, the observed differences between exchanges and coins are plausibly explained as a consequence of differences in trader pool composition. Validating this hypothesis would require a dataset with full LOB actions and trader identification, including failed attempts to take or cancel liquidity—information not even visible in L3 feeds. Unfortunately, such data is unavailable to us and would need to be

† Given that our Binance sample is considerably smaller than the Bybit sample, questions may arise regarding its representativeness; in Supplementary Material D we conducted a detailed analysis, which indicates that a sample size of 40 000 orders on Bybit sufficiently captures the true relationship between future returns and failure probability. Moreover, our results suggest that even a small-scale probing experiment involving tens of thousands of orders can effectively infer the failure rate curve — which we consider in the next section — in various markets, offering a cost-effective approach for meaningful analysis.

‡ Latency statistics available upon request.

§ During the Binance experiment the price ranged from 29603.1 to 30,230.1 USD and the tick size of 0.1 USD took values between 0.033 and 0.034 bps, while for the Ethereum experiment the price ranged from 2325.75 to 2,469.65 USD so the tick size ranged between 0.2025 and 0.215 bps.

	Bybit ETH	Bybit BTC	Binance BTC
Volatility	Very High (65% IV)	High (55% IV)	High (55% IV)
Trading Volume	Low (100m/day)	Medium (1bn/day)	High (2bn/day)
LOB Liquidity	Very Thin	Medium	Medium
Tick Size	0.05 USD (0.2 bp)	0.5 USD (0.17 bp)	0.5 USD (0.03 bp)
Taker/Maker Fee	1.8 / -0.5 bp	3.0 / 0.0 bp	2.5 / -1.0 bp

Figure 4. Comparison of key attributes across the three markets. Daily trading volumes, tick size in basis points (bp), and fees are from the time of the experiment (they change over time). Implied volatility (IV) is taken from at-the-money options.

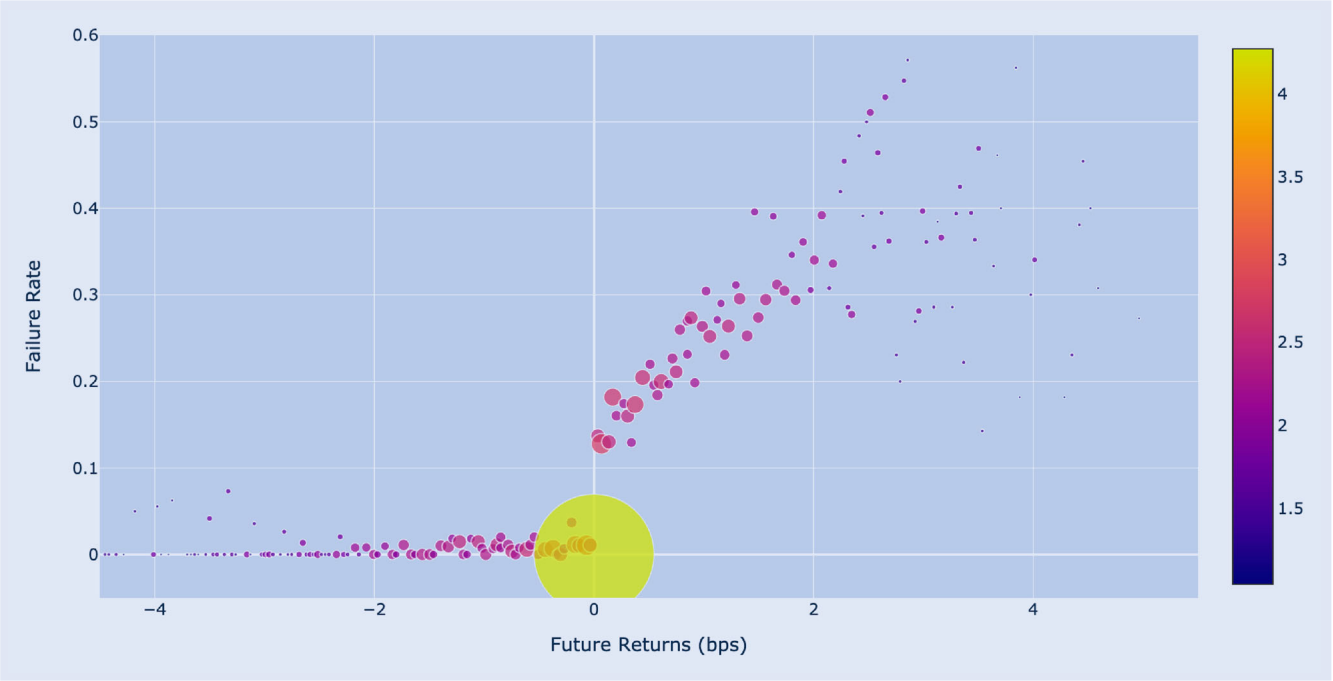

Figure 5. Failure rate by future returns with area representation, for the Binance experiment. The colour scale on the right of the plot is the (base 10) log of the order count.

sourced directly from the exchange or, in traditional finance, from regulators who hold this type of data.

5.3. *Failure probability models*

We now develop a parsimonious model to characterize the relationship between probability of order failure and future returns, with a view towards applying it to the problem of improving backtest accuracy. Traditional backtesting often fails to account for the potential failure of marketable limit orders, or assumes a random failure rate, neglecting the adverse selection effects we observe in our empirical data, whereby 'good' orders are more likely to fail than 'bad' ones. By incorporating these insights into backtesting procedures, one can enhance their accuracy, leading to 'paper simulation results' that more accurately reflect future potential live trading outcomes.

Recognizing that failure probabilities exhibit differences across exchanges and coins, we constructed and fitted separate models for each of our three markets—Bybit Bitcoin, Binance Bitcoin, and Bybit Ethereum—and then compared them.

As we aim for a simple, yet reasonably accurate model of probability of failure, and motivated by the discontinuity discussed above, we adopt a simple piecewise approach. For negative future returns, motivated by the observation that the failure probability within this range varies little, we fit a constant. For positive returns in $(0, \infty)$, we have chosen to fit a simple linear function. Given that the probabilities are generally small, a linear approximation is sufficient to capture the trend without introducing unnecessary complexity. To enhance the robustness of our model and mitigate the potential noise and bias introduced by extreme future return values, which are associated with a smaller number of orders, we employ a weighted variant of ordinary least squares (OLS) that accounts for the varying levels of certainty across

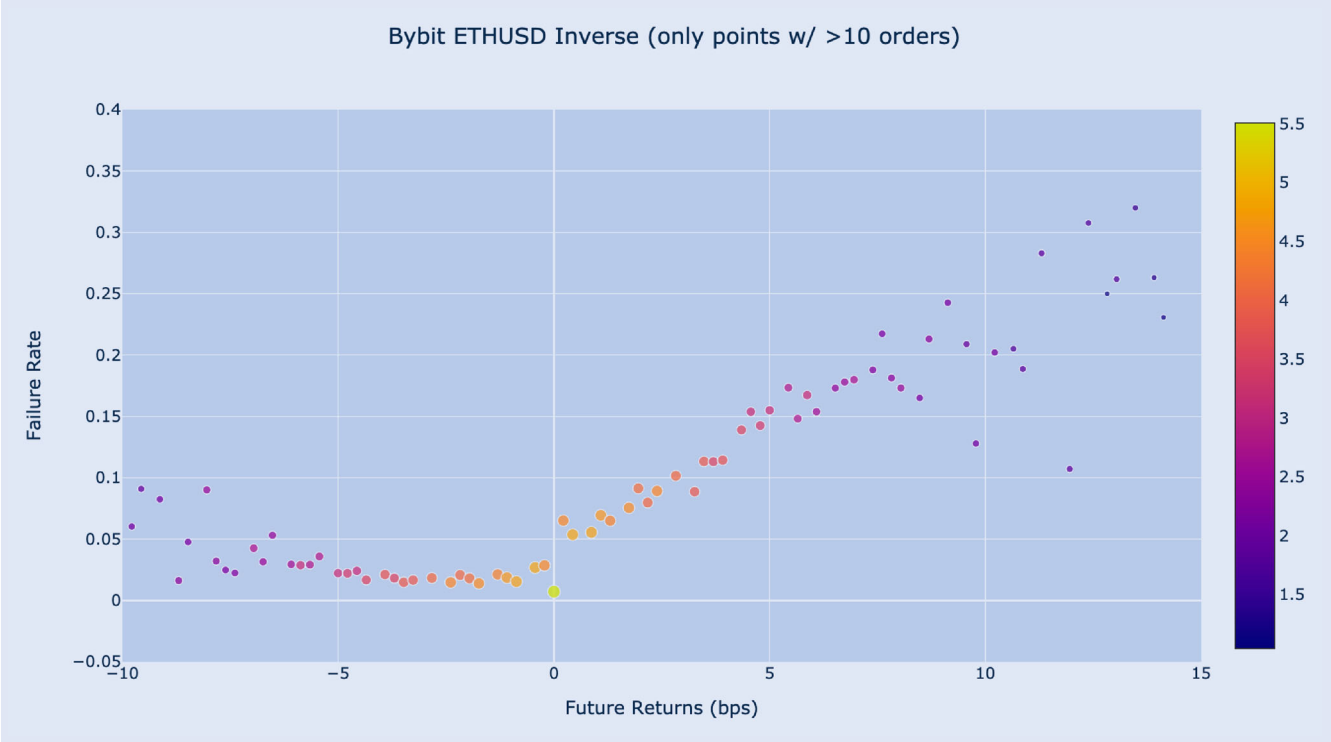

Figure 6. Failure rate by future returns on the Bybit ETH Inverse Perpetual, with the colour scale on the right indicating the (base 10) log of the order count.

different observations. Our modeling approach is consistent with our objective of developing a parsimonious yet effective model of failure probability that aligns closely with the stylized facts of the observed data.

We now turn to our model for estimating order failure probability given a basis point return, initially described for Bybit. The process for the Binance and Ethereum models is nearly identical, fitted on their respective datasets with constants replaced as appropriate, e.g. ξ_{bb} with ξ_{bn} for Binance and ξ_{bb} with ξ_{eth} for Ethereum.

Keeping with the notation introduced earlier in equation (15), let γ_i denote the empirical failure rate for all orders with a rounded future return of $i \cdot \xi_{bb}$ bps, where ξ_{bb} is the basis point increment. On the non-positive segment $(\infty, 0]$, we fit the optimal constant $c^* \in \mathbb{R}$ which minimizes the weighted sum of squared residuals, $\min_c \sum_{i\leq 0}(c-\gamma_i)^2 w_i$. On the positive segment $(0, \infty)$, we search for intercept $\alpha^* \in \mathbb{R}$ and slope $\beta^* \in \mathbb{R}$ parameters such that the corresponding linear function $x \mapsto \alpha + \beta x$ minimizes the quantity $\min_{\alpha,\beta} \sum_{i>0}(\alpha+\beta x_i - \gamma_i)^2 w_i$.

The failure probability of an order O conditioned on its future return value is then modeled as

$$\mathrm{P}\left(\mathrm{O} \text{ fails} \mid \mathrm{fret}(\mathrm{O}) = x\right) = \begin{cases} c^* & \text{if } x \leq 0 \\ \alpha^* + \beta^* x & \text{else} \end{cases}. \quad (17)$$

This modeling strategy offers a straightforward yet effective means of capturing the essential properties observed in our empirical data. The slope parameter β can be interpreted as the increase in failure probability per basis point of (conditional) return.

The resulting parameter values, after applying the above modeling strategy on both exchanges, are presented in table 6,

Table 6. Comparison of failure rate model parameters and R^2.

Market	c	α	β	R^2
Bybit BTC	0.00088	0.0424	0.02187	0.928
Binance BTC	0.00037	0.1658	0.07432	0.637
Bybit ETH	0.01407	0.04691	0.01790	0.895

with a bootstrap sampling-based standard error estimation for each parameter conducted in Supplementary Material E. In assessing the fit quality of our models, we utilized the coefficient of determination, R^2, as a metric. Given that our fitting process was based on weighted least squares, the weights, corresponding to the number of orders for each future return value, were incorporated into the R^2 calculation. The formula for the weighted R^2 is given by

$$R^2 = 1 - \frac{\sum_{i=1}^n w_i(\gamma_i - \hat{\gamma}_i)^2}{\sum_{i=1}^n w_i(\gamma_i - \bar{\gamma})^2},$$

where w_i denotes the (number of orders) for the i^{th} data point, γ_i is the observed value, $\hat{\gamma}_i$ is the predicted value, and $\bar{\gamma}$ is the weighted mean of the observed values. We report R^2 on the positive part of the fit, as by definition the R^2 of the constant fit on the negative part is 0, with values presented in table 6.

The fitted models are further illustrated in figures 7–9, allowing a visual assessment of the fit quality on Bybit (Bitcoin and Ethereum) and Binance.

All three models exhibit high R^2 values, indicating a strong fit. The lower R^2 on Binance compared to the two Bybit experiments is likely due to the significantly smaller sample size. In Supplementary Material E, we conducted a detailed

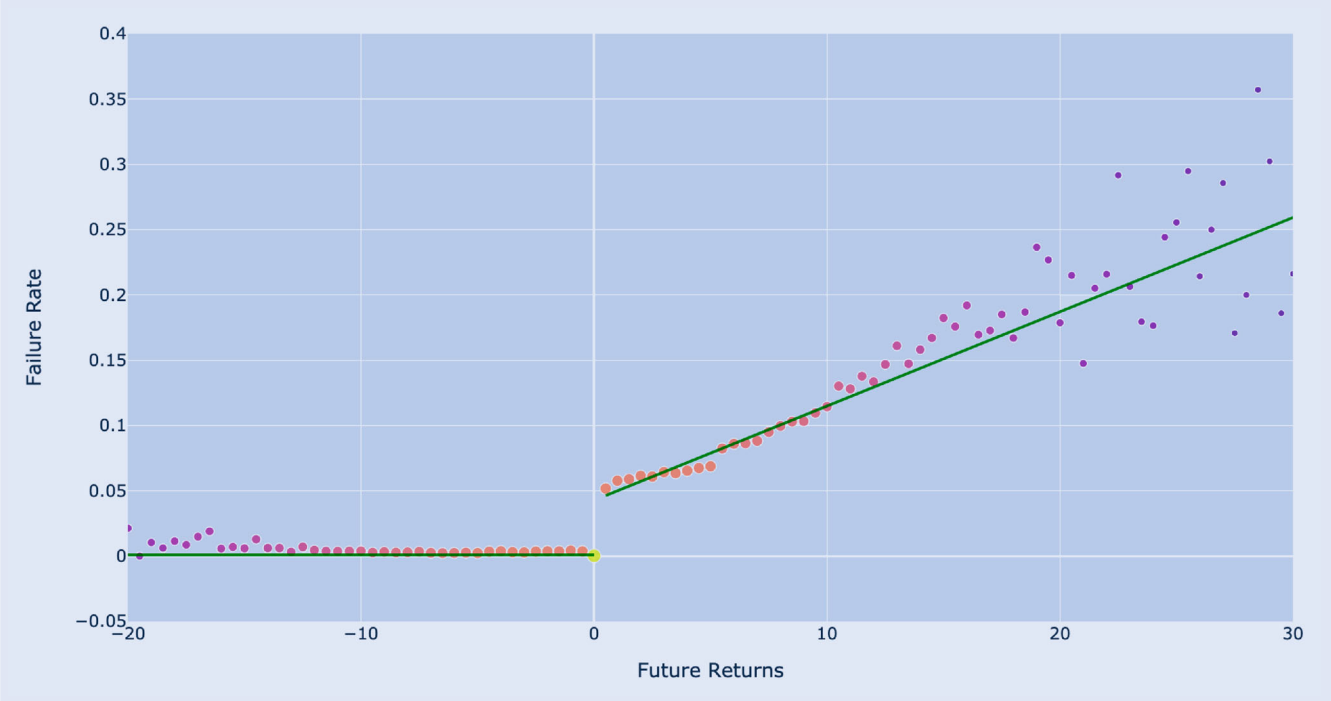

Figure 7. Failure rate by future returns on Bybit's Bitcoin perpetual with fitted model (green curve).

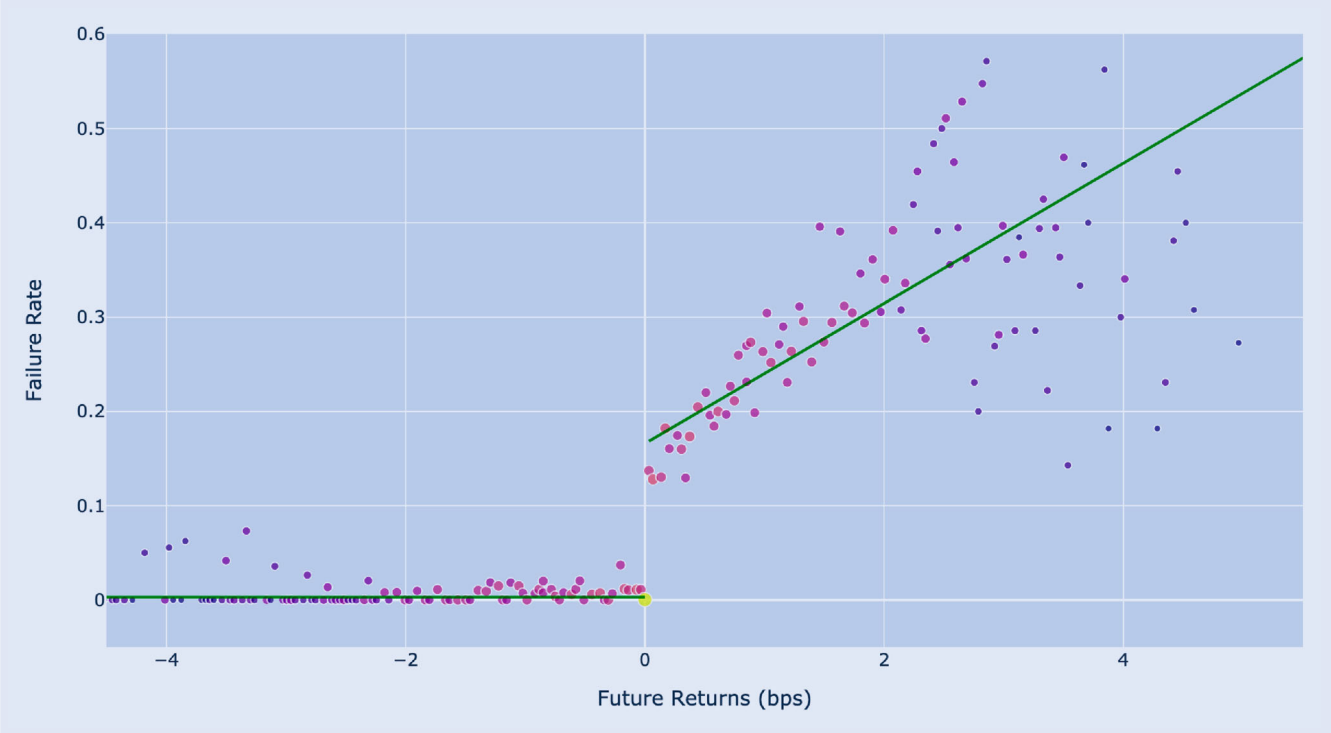

Figure 8. Failure rate by future returns on Binance with fitted model (green curve).

analysis of sample sizes; in particular, given the relatively high volatility during the Binance experiment, we tested whether volatility within small sample periods materially affects the resulting parameters. Our findings show that the model parameters remain stable across smaller subsamples (comparable to the size of our Binance data), regardless of sample period volatility.

The differences in parameter values across our three models align with the observations made in the previous section: Ethereum shows the highest c parameter, reflecting higher failure rates at negative returns; the intercept values in the two Bybit experiments are similar, consistent with the comparable jump sizes, while the Binance intercept is significantly higher, corresponding to its larger jump. Growth on the positive segment is steeper on Binance than on the other two markets, likely reflecting the more intense competition observed there.

If our hypothesis is correct that cross-market differences in failure probabilities are a result of differences in traders pools

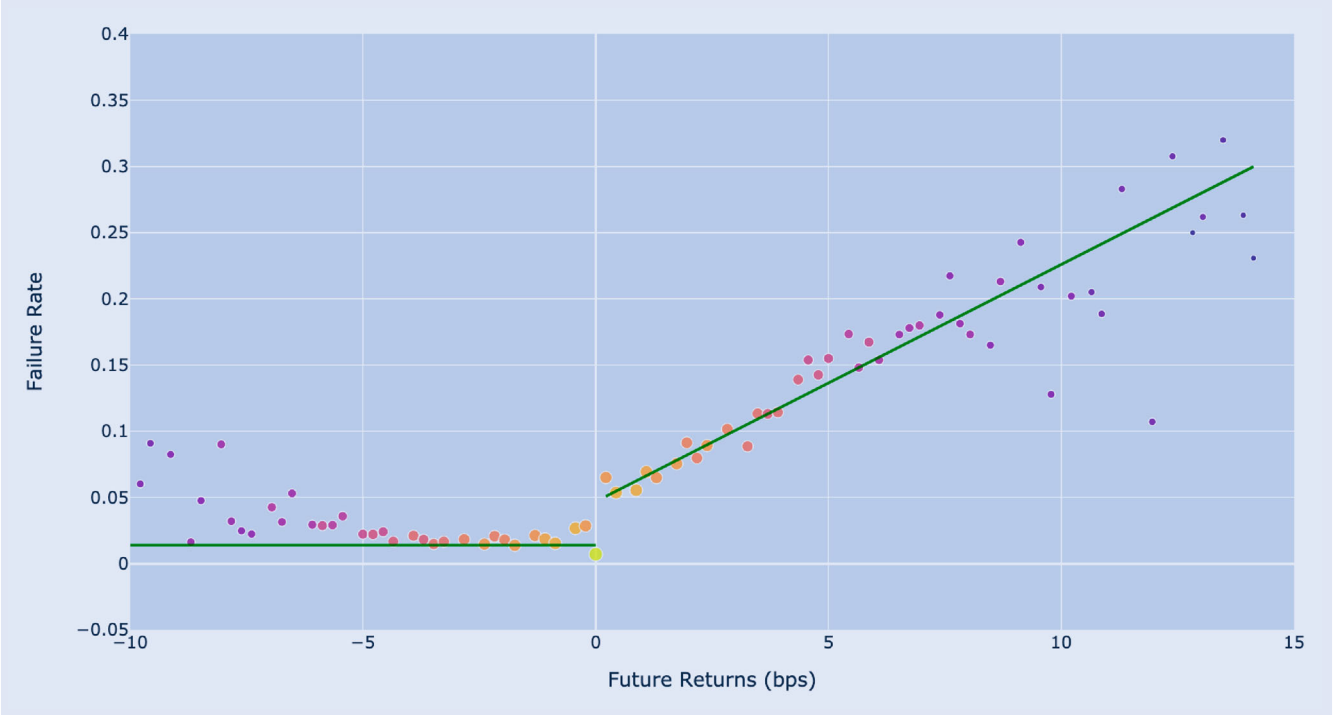

Figure 9. Failure rate by future returns on Bybit's Ethereum perpetual with fitted model (green lines).

(as outlined in the preceding section), it might be possible, with the right type of data, to obtain a 'universal' model of order failures by scaling model parameters with appropriately defined metrics quantifying relevant attributes of the trader pool. Examples of what those attributes might include are: the number of active co-located high-frequency takers, the number of market makers posting orders at the touch, the number of cancellation attempts of single orders, the proportion of taker orders agnostic to short-term returns. We leave it to others with access to the requisite data to further investigate those questions.

6. Market order execution

Thus far, we have focused on the failure rates for taker orders aiming at the best price. This is one end of a spectrum of execution strategies: the emphasis is on achieving the best possible price, albeit with a risk of not always getting filled. However, traders may choose to send their taker orders at a worse limit price if they wish to improve their probability of receiving a fill. The spectrum of trading strategies can be understood as follows:

- At one extreme end, traders send taker orders at the best price. This strategy prioritizes achieving the best price but may result in a lower fill ratio.
- At the other extreme end, traders can send a market order. This approach guarantees a fill but runs the risk of receiving a bad fill price.

Having examined the first extreme, we turn to the other end of the spectrum by considering market orders (recall that the data we collected for our experiment enables us to conduct analysis on both types of order placement strategies). By exploring both ends of this spectrum, we shed light on the range of execution strategies available to traders and the underlying trade-off between execution price and fill probability. Market orders, while guaranteed to execute immediately, may fill at a price that is worse (or, more rarely, better) than the target price observed in the latest LOB snapshot. This discrepancy is quantified as *slippage*, as defined earlier in equation (5).

6.1. Good market orders incur slippage

We begin our analysis by exploring the relationship between the expected slippage of an order and its future returns. This approach parallels our previous examination in section 5, where we investigated the relationship of failure probability on future returns.

Recall that we defined $\mathcal{O}_i$ as the set of all orders with a (rounded) future return of $x_i := \xi_{bb} \cdot i$ bps (see equation (14)). For each distinct future return value x_i, we construct the corresponding set of slippage values $S_{x_i} := \{\text{slip}(o) \mid o \in \mathcal{O}_i\}$, representing the slippage values for orders with a future return of x_i bps. We proceed to compute the average slippage, denoted by $\overline{S_{x_i}}$, for each set S_{x_i}. This average serves as our empirical estimate of the expected slippage conditional on an order realizing a future return of x_i bps. The methodology applied to the Bybit Bitcoin experiment is replicated identically for the Ethereum and Binance experiments, with adjustments made to relevant constants.

Recall that our analysis is based on minimum-sized orders. Traders utilizing larger orders should expect to encounter slippage that is worse than the values reported in our study. The slippage figures presented herein should therefore be considered a conservative estimate, serving as a lower bound for

the actual slippage experienced for any order exceeding the minimum size.

Figure 10 displays a scatter plot of the pairs $(x_i, \overline{S_{x_i}})$ derived from the Bybit Bitcoin data, over all i for which $\mathcal{O}_i$ contains at least 10 orders. Similarly, we apply this analysis to the Binance and Ethereum datasets, yielding comparable scatter plots shown in figures 11 and 12.

Our empirical findings indicate a monotonic relationship between an order's future returns and its expected slippage: orders with more favorable (positive) future returns tend to experience greater slippage on average—a trend observed consistently across Bybit (Bitcoin and Ethereum) and Binance. This relationship is remarkably well captured by a linear function (solid green lines), fitted using weighted least squares, with weights proportional to the number of orders at each future return value. The parameter values of these linear functions, along with their R^2 values for all three markets, are presented in table 7. The high R^2 values—0.659 for Bybit Bitcoin, 0.830 for Binance Bitcoin, and 0.942 for Bybit Ethereum—indicate that a linear function effectively explains the relationship between an order's future returns and expected slippage across markets.

Expected slippage per basis point return is noticeably worse on Binance than on Bybit (Bitcoin and Ethereum). For instance, an order that would have achieved a 2 bps return if executed at the targeted price experiences slippage due to the target moving during the latency gap of: around 0.1 bps on Bybit Bitcoin, 0.07 bps on Bybit Ethereum, and 0.3 bps on Binance.

These differences are consistent with our previous observations on MLO failures: more frequent MLO failures imply more frequent negative slippage for an MO. While this does not, by itself, guarantee the observed slippage differences—since slippage on Bybit could, hypothetically, be infrequent but severe, or on Binance frequent but small—our data confirms that Binance slippage is indeed worse than the two Bybit markets, which are roughly comparable.

An extended version of our explanation for differing MLO failure rates across markets can also explain differences in average slippage. While the earlier arguments focused on the trader pools active at the touch (specifically, HFT takers targeting liquidity and market makers attempting to cancel it), we can extend this inquiry to consider trader pools active within a certain range of basis points from the touch, raising key questions that can shed light on the observed differences in slippage: In response to a signal, how many HFT takers attempt to take liquidity within $\leq x$ basis points of the touch, for some scalar $x > 0$? What are their order sizes, how many makers are responsible for that liquidity, and how many cancellation requests do they for each order?

A market's attributes can influence the answers to those questions. For example, low overall trading volume may reduce the number of market makers providing liquidity, while high asset volatility may lead makers to post fewer or smaller orders at the touch, instead posting ones deeper in the book. Both of these things apply to our Ethereum experiment (see figure 4), explaining the higher failure rates yet less severe slippage on Bybit's Ethereum market compared to its Bitcoin market.

Across all markets, latency-induced slippage is a significant trading cost for MOs aiming for positive short-term returns and cannot be ignored in real-world trading, where latency is unavoidable—especially for HFT traders focused on short-term gains.

6.2. *Slippage distributions*

Our analysis thus far only provides insights into *average* slippage values conditioned on a given future return level. But are there any further insights that can be gleaned from a closer look at the full conditional distribution, rather than just computing its average? In this subsection, we tackle that question by examining those distributions across a range of different future return levels.

To begin with, we provide visualizations that help us build some intuition for the patterns our data exhibits. We utilize violin plots to visualize the conditional distributions of slippage values across a range of future return levels. These violin plots employ Kernel Density Estimation (KDE) to provide a smooth representation of the distribution of slippage values across different future return levels, offering a more comprehensive view than average slippage alone. Figures 13(a) and 13(b) display violin plots for Bybit Bitcoin and Binance data, respectively. Bybit Ethereum results are omitted, as they are similar to Bybit Bitcoin and do not offer materially different insights.

The plots reveal several key insights, summarized as follows.

(i) *Zero-Centric Density*: The density peaks at zero across all conditional distributions, indicating that most orders, even those with large future return values, typically experience zero slippage, executing at the expected price.
(ii) *Left-Skewness*: As the future returns increase (i.e. moving to higher distributions on the y-axis), the left-skewness of the distributions becomes more pronounced. This suggests that a growing proportion of orders execute with negative slippage, meaning they fill at a worse price compared to the best price from the most recent LOB snapshot at submission time.
(iii) *Secondary Density Peaks*: For the distributions conditioned on strictly positive future return values, secondary density peaks emerge left of the main peak at zero. These secondary peaks shift further left as future returns increase, indicating an increased magnitude of negative slippage. This is particularly visible in the Bybit plot.
(iv) *Exchange Variability*: Both exchanges show similar trends, but Binance generally exhibits worse slippage. Additionally, the data for Binance appears noisier, likely due to a smaller sample size of orders.

The zero-centricity and left-skewness observed in the violin plots align with our earlier finding that while most market orders fill at the expected price, a fraction of orders, increasing with future returns, fail to do so. This failure results in negative slippage, where the magnitude of deviation from the target price increases with higher future returns, illustrating

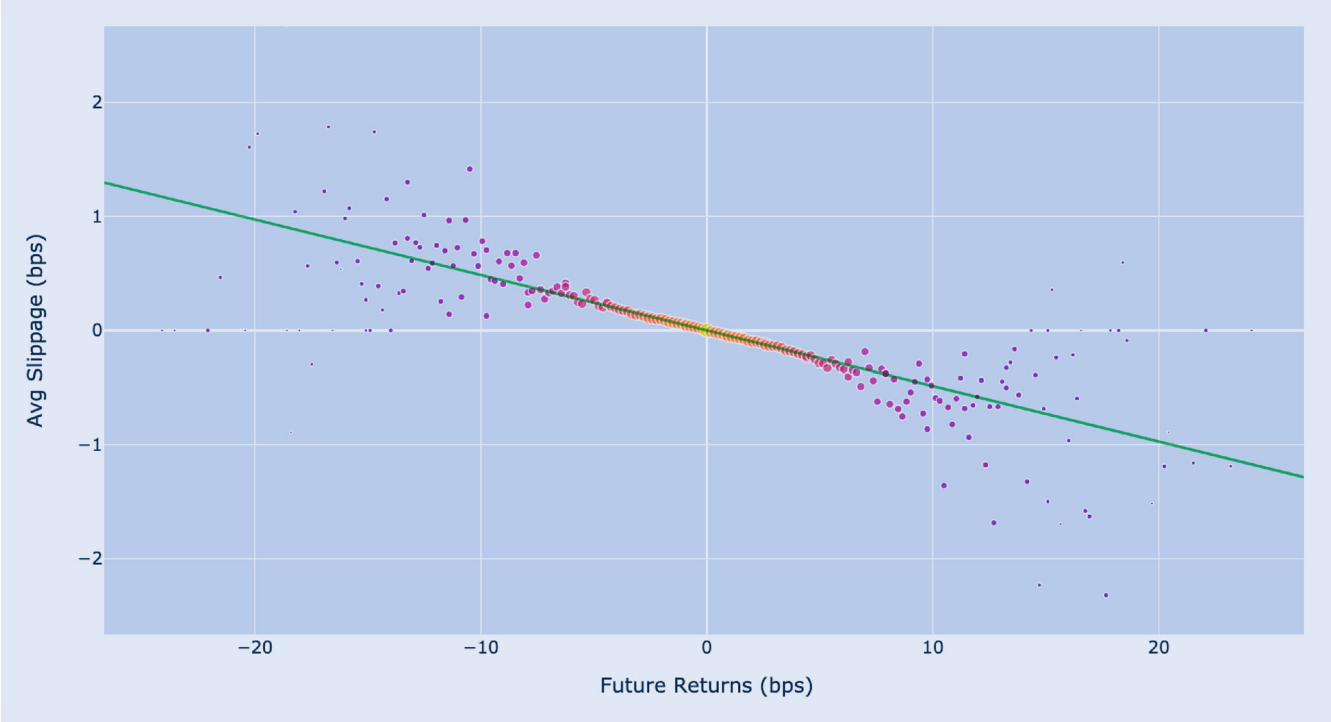

Figure 10. Average slippage vs future returns on the Bybit Bitcoin perpetual.

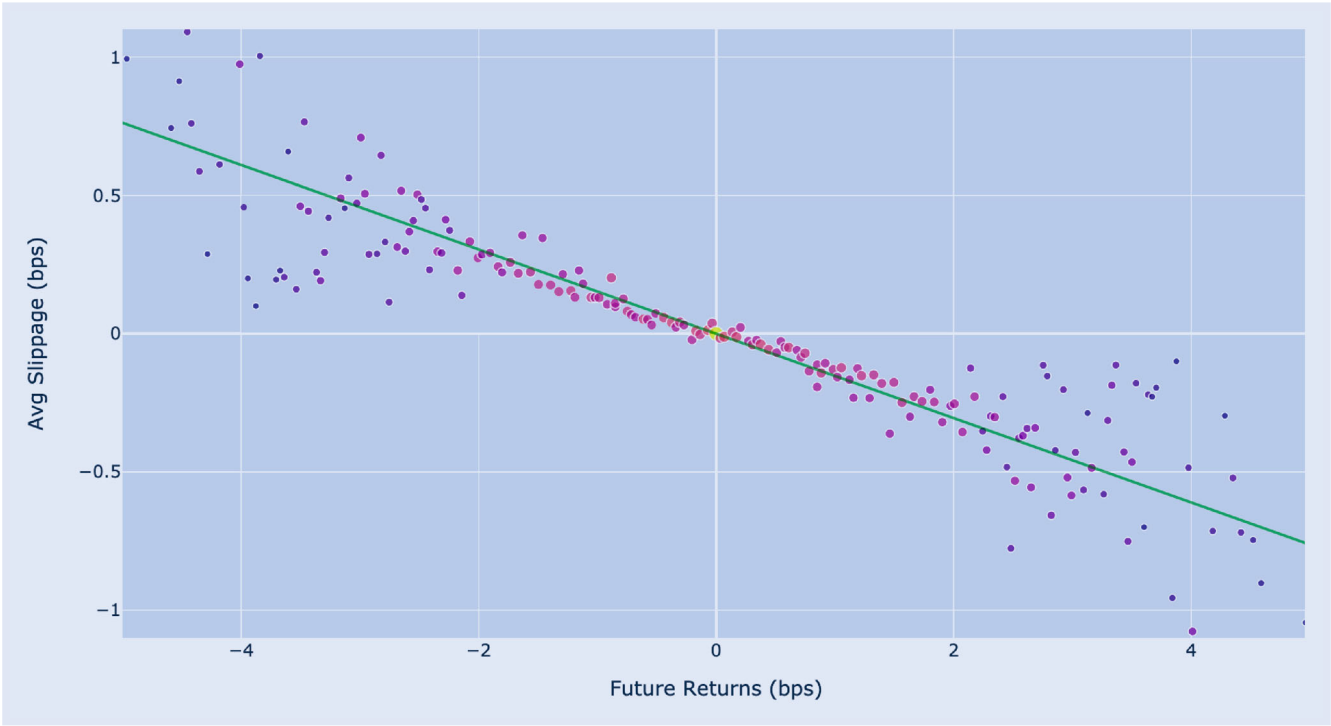

Figure 11. Average slippage vs future returns on Binance.

the analogue of the adverse selection effects for taker orders which we observed earlier in the paper (for IOC and FOK orders).

The secondary density peaks in the distributions offer new insights. Consider the Bybit plot at the 2 bps future return level, where a secondary peak at around -2.2 bps slippage suggests that a significant fraction of orders orders o with $\text{fret}(o) = 2$ not only fail to fully realize the anticipated return, but actually incur slippage greater than the anticipated return.

For a more granular and detailed view, figure 14 presents histograms at tick size granularity for the 2 bps future return level, and for comparison, also at the zero bps future return

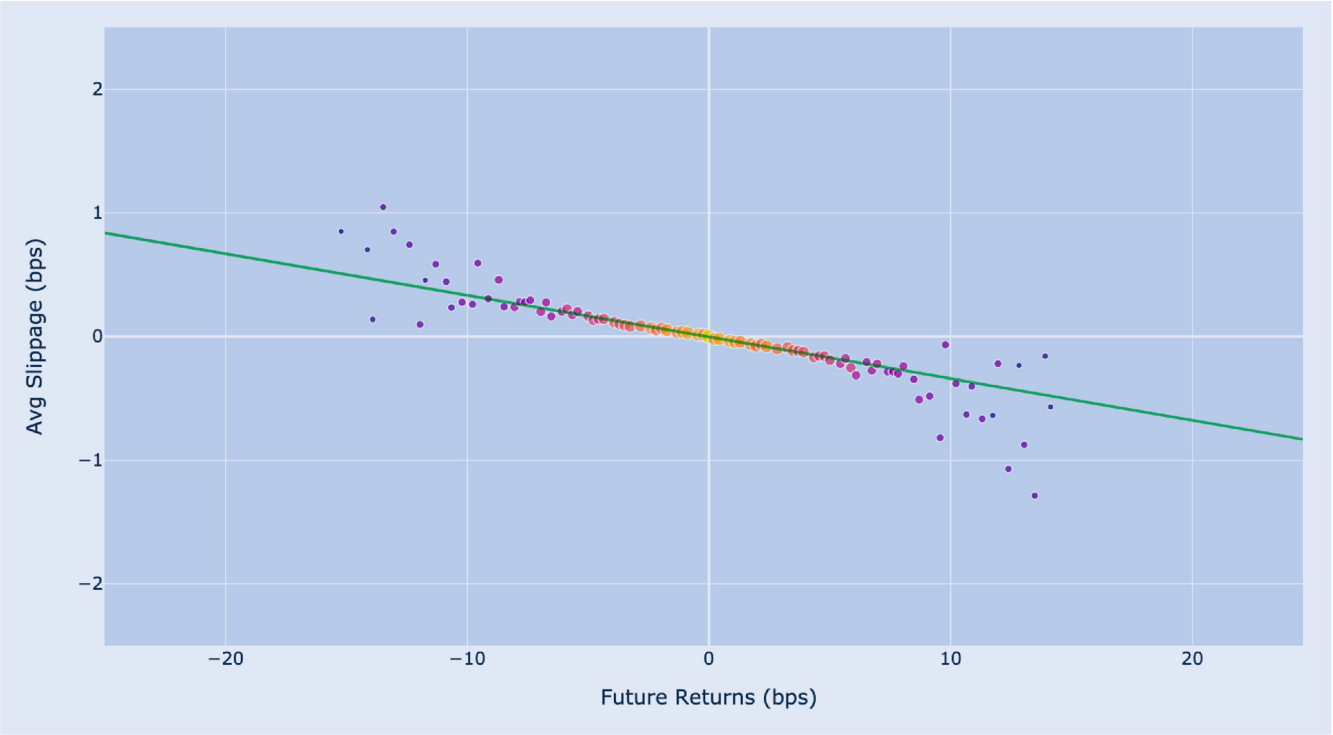

Figure 12. Average slippage vs future returns on the Bybit Ethereum perpetual.

level, offering a more detailed perspective of the slippage distributions than the KDE-based smooth approximations provided by the violin plots.

The observation of a secondary peak at -2.2 bps slippage in the Bybit plot, at the 2 bps future return level, indicates a substantial liquidity presence up to 2.2 bps deep in the book, with a notable concentration around this level.

The secondary peak thus sheds light on market makers' quoting behavior, revealing their liquidity positioning in anticipation of potential adverse movements. As future returns increase, these secondary peaks shift leftward, indicating deeper liquidity placement by market makers in response to anticipated positive returns for takers.

7. Implications for backtesting

This section aims to equip quantitative researchers with practical guidelines for conducting backtests that incorporate the toxicity effects previously discussed, employing our models of order failure and insights into slippage in order to dispel illusions about the profitability of certain strategies. Altogether, this allows quantitative researchers to focus on researching better alpha instead of iterating through endless backtests and execution scenarios.

Our approach is illustrated with two distinct trading strategies: one based on orderbook imbalance measures and a lead-lag strategy with latency advantage. Each strategy is examined under two different modes of trade execution, in order to elucidate their impact on performance metrics.

The first execution mode targets liquidity at the top of the book, and thus employs marketable limit orders priced at the top ask for buy orders and the top bid for sell orders. Orders under this mode either fill at the most favorable price or fail, with the likelihood of failure being modeled as per the methodologies developed in prior sections of this paper.

The second execution mode employs market orders, thereby eliminating the possibility of order failure but introducing variability in the fill price. Hence, the focus here shifts to quantifying the deterioration in execution price relative to the best available price at the time of order submission.

In the remainder of this section, we outline the two trading strategies, detail our backtest PnL adjustment procedures in each execution mode, and empirically analyze the performance of the strategies, drawing on insights from Sections 5 and 6.

7.1. Trading strategy specifications

Before delving into the specifics of each trading strategy, it is important to note some overarching aspects that apply to both. We focus on trading strategies on the Bybit Bitcoin inverse perpetual market, where our data is most extensive. Firstly, order execution is contingent on the chosen mode: in the first execution mode, orders are sent at the best price, namely the top ask (bid) for buy (sell) orders respectively, while under the second execution mode, orders are submitted as market orders. Secondly, it is assumed that all orders are minimum-size: this gives a conservative estimate of strategy performance, establishing a lower bound on underperformance. In practice, larger order sizes will typically result in greater underperformance than reported here. Lastly, to streamline post-trade analysis and avoid double-reacting to the same market signal, the strategies are constrained to preclude the submission of multiple consecutive orders on the same side. This restriction not only

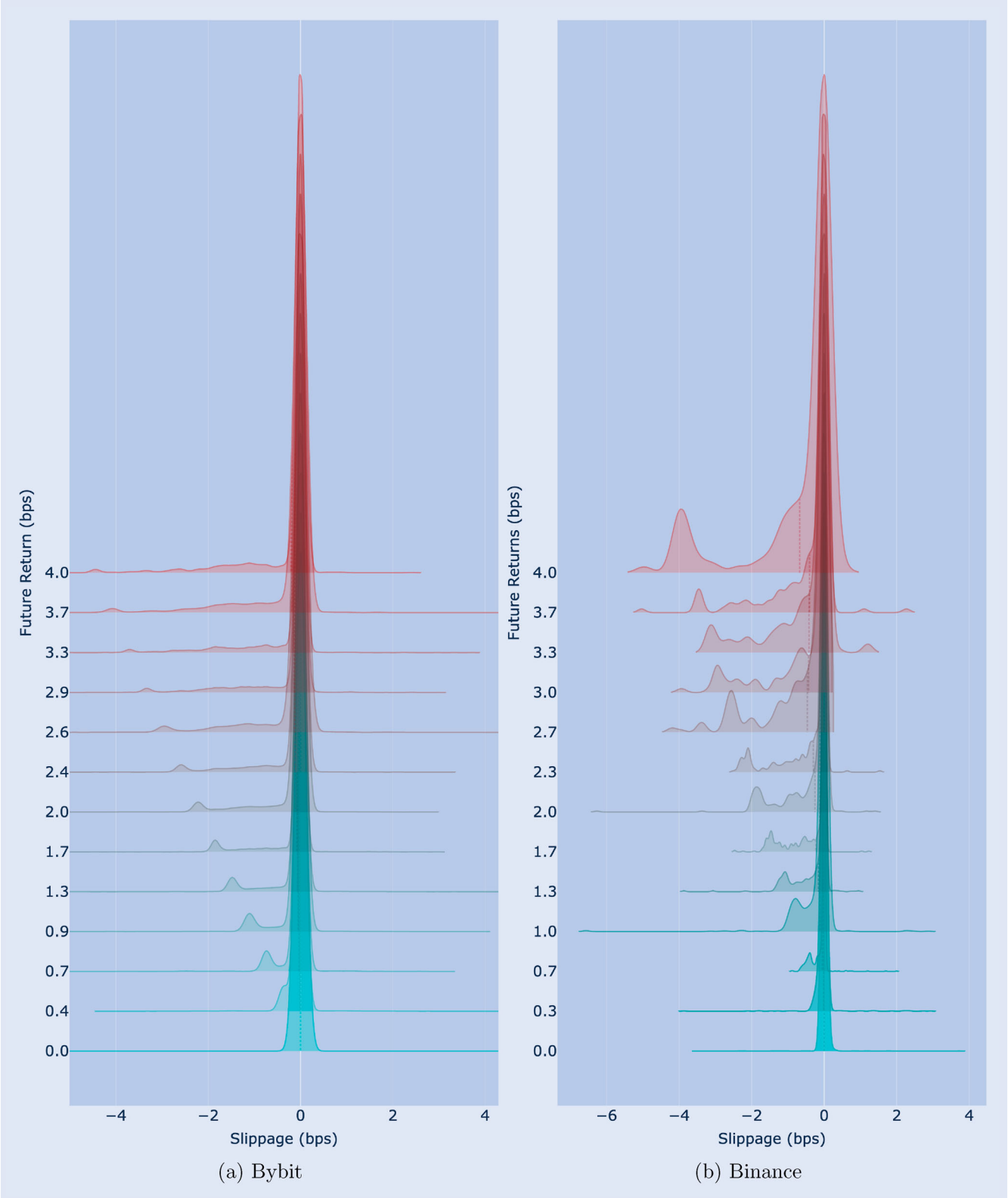

Figure 13. Distributions of slippage, conditioned on various future return levels for Bybit and Binance.

simplifies the post-analysis by reducing the variability in the PnL metric (introduced below) that computes price differences between consecutive trades, but it also avoids the complication of managing inventory risk and accounting for unrealized PnL due to unequal buy and sell quantities. This ensures a clear measure of the strategy's realized financial performance.

Strategy 1: Orderbook Imbalance. This strategy sends a buy (resp., sell) order when the top bid (ask) has high liquidity and the ask (bid) side is illiquid. To make this precise, we calculate the volume-weighted average price (VWAP) for both the ask and bid sides, denoted as $\text{VWAP}_{a,t}(q)$ and $\text{VWAP}_{b,t}(q)$, respectively. These metrics compute the average price at which an order of quantity q would be executed given

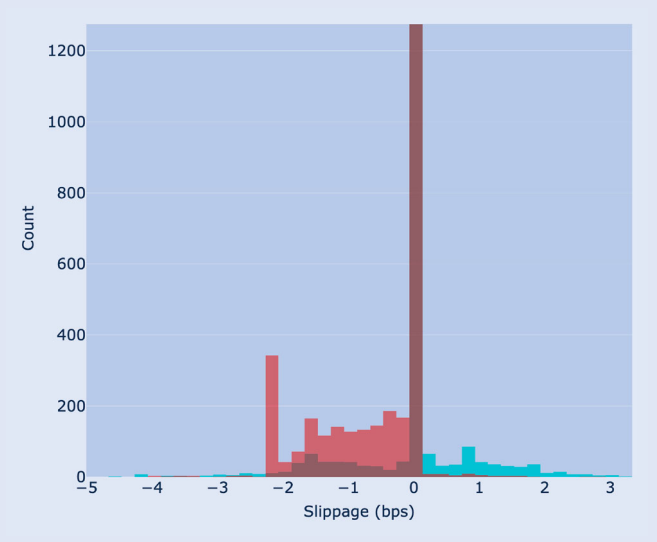

Figure 14. Granular histograms of slippage for 0 bps (teal) and 2 bps (red) future return values.

the current state of the LOB (using the most recent snapshot, at time t). Next, we define liquidity measures in basis points for the ask side $\mathrm{LIQ}_{a,t}(q) := (\mathrm{VWAP}_{a,t}(q)/p_{a,t} - 1) \cdot 10\,000$, and $\mathrm{LIQ}_{b,t}(q) := (p_{b,t}/\,\mathrm{VWAP}_{b,t}(q) - 1) \cdot 10\,000$, for the bid side. These measures capture the price difference between the top-level quotes and the VWAP prices. Larger values indicate a less liquid book, as one would need to walk deeper into the orderbook to fill an order of size q.† Preliminary analysis of typical LOB liquidity revealed $q = 400\,000$ USD to be a reasonable choice for this threshold - we henceforth fix this choice for the remainder of the section. Our findings are robust across different choices of q, ranging from 100 000 to 1 000 000 USD.

The strategy specification is as follows:

(i) At each trigger event (as defined in Algorithm 1), compute the orderbook metrics $\mathrm{LIQ}_{a,t}(q)$ and $\mathrm{LIQ}_{b,t}(q)$ based on the most recently observed Bybit LOB snapshot.

(ii) Submit a buy order if the previous order was a sell order or if no order has been submitted yet, and $\mathrm{LIQ}_{b,t}(q) = 0$ and $\mathrm{LIQ}_{a,t}(q) > 3.5$. Conversely, submit a sell order if the previous order was a buy order, and $\mathrm{LIQ}_{a,t}(q) = 0$ and $\mathrm{LIQ}_{b,t}(q) > 3.5$. Depending on the execution mode, orders are sent at the last-seen top price or as market orders.

The liquidity conditions $\mathrm{LIQ}_{b,t}(q) = 0$ and $\mathrm{LIQ}_{a,t}(q) > 3.5$ are signals for a buy order, indicating high liquidity with at least 400k USD at the top bid and significant illiquidity on the ask side, hinting at a potential price increase. For a sell order, $\mathrm{LIQ}_{a,t}(q) = 0$ and $\mathrm{LIQ}_{b,t}(q) > 3.5$ suggest high liquidity at the top ask and illiquidity at the bid, hinting at a likely price drop.† We chose the 3.5 bps threshold based on the 0.99 quantile of the historical distribution of $\mathrm{LIQ}_{-,t}(q)$ indicators.

† A more detailed discussion of these features and their relationship with order failures can be found in **Supplementary Material C**.

† The decision to use our liquidity measure over the classical orderbook imbalance is based on its robustness, as it accounts for the total liquidity available, rather than providing a relative measure that does not indicate the actual volume of liquidity at or near the best bid and ask prices.

Strategy 2: Lead-Lag. Our second strategy is a lead-lag strategy that aims to exploit the fact that price moves on Bybit are often led by those on Binance, a relationship that is well-established, even within the academic literature (Albers *et al.* 2021, Alexander *et al.* 2022). The strategy sends a buy (resp., sell) order when the market on which we are trading (Bybit inverse perpetual) is cheap (expensive) with respect to a leading market (Binance linear perpetual), and when the price on Binance just moved up (down), while the one on Bybit has not yet. More precisely:

(i) At each trigger event (as defined in Algorithm 1), we update the following quantities
- On every Bybit or Binance orderbook update, we compute the latest price difference δ between the two exchanges, denominated in bps. This calculation uses microprices.
- We also keep track of the baseline microprice difference between Bybit and Binance, denoted by $\bar{\delta}$. This is the 10-minute rolling-window median microprice difference between Bybit and Binance, updated every 500 ms, and denominated in bps.
- On every incoming trade from Binance, received via the trades feed, we obtain the last trade price on that exchange and compute the 500 ms return $\mathrm{fret}_{500\,\mathrm{ms,binance}}$ in bps.

(ii) Submit a buy order if the previous order was a sell order or if no order has been submitted yet, and the current price difference δ indicates that Bybit is cheaper than Binance by at least 2.5 bps relative to the baseline $\bar{\delta}$, and the 500 ms return $\mathrm{fret}_{500\,\mathrm{ms,binance}}$ shows a price increase on Binance of at least 1.5 bps. Submit a sell order if the previous order was a buy order, and δ indicates that Bybit is more expensive than Binance by at least 2.5 bps relative to $\bar{\delta}$, and $\mathrm{fret}_{500\,\mathrm{ms,binance}}$ shows a price drop on Binance of at least 1.5 bps. Depending on the execution mode, orders are sent at the last-seen top price or as market orders.

7.2. *PnL metrics*

Consider a taker strategy using marketable limit orders or market orders. Over a historical backtest sample ranging from t to T, it produces a number N of orders

$$\mathrm{order}_i = \left(t_i, p_i, q_i, \mathrm{side}_i, p_{i,f}\right) \quad \text{for } i = 1, \ldots, N, \tag{18}$$

submitted at time $t_i \in [t, T]$, limit price p_i (equal to $\pm\infty$ in case of a market order), quantity q_i, trade direction $\mathrm{side}_i \in \{\mathrm{buy}, \mathrm{sell}\}$, and achieving fill price $p_{i,f}$.

We consider two canonical approaches for mapping the set of orders to a PnL value.

(i) *Short-term returns*: We compute short-term markouts for each order_i relative to its fill price over return

horizon τ_i (defined as in equation (7)) as follows:

$$\text{fret}(\text{order}_i) := \begin{cases} ((p_{a,t_i+\tau_i}/p_{i,f}) - 1) \cdot 10\,000 & \text{if side}_i = \text{buy} \\ ((p_{i,f}/p_{b,t_i+\tau_i}) - 1) \cdot 10\,000 & \text{if side}_i = \text{sell} \end{cases} \tag{19}$$

We then calculate the average and standard deviation of these values across orders $i = 1, \ldots, N$, denoted as PnL_1 and StdDev_1, respectively.

(ii) *Realized PnL*: The second approach considers the price difference (in bps) between consecutive fill prices $(p_{i+1,f}/p_{i,f} - 1) \cdot 10\,000$ for $i = 1, \ldots, N-1$. For this to yield meaningful results, we need to restrict ourselves to sequences of orders that are alternating in side and for which the first fill is a buy order. Given such a sequence, we calculate the average and standard deviation of the price differences, denoted as PnL_2 and StdDev_2, respectively.

The variance of realized PnL tends to be larger than that of the average short-term return, especially for strategies that trade infrequently. This is because the first metric compares an order's price with a markout price that is very close to it in time, while the second metric compares prices of consecutive orders, which can be minutes or hours apart from each other, depending on the strategy. Therefore, quant trading firms typically focus on optimizing the average short-term basis point edge, as it provides a more robust metric.

7.3. First execution mode: top of the book

Backtest PnL Adjustment: Building on our previous work, we introduce a methodology for conducting more accurate backtests that account for the potential failure of orders. The procedure is summarized as follows.

(i) For each order order_i where $i = 1, \ldots, N$, compute the short-term future return $\text{fret}(\text{order}_i)$ and estimate its fill probability $\text{P}(\text{order}_i \text{ fails})$ using the fill probability model as described in equation (17), with parameters tailored to the specific exchange.

(ii) Conduct multiple simulated historical runs, each time utilizing our estimate of $\text{P}(\text{order}_i \text{ fails})$ to simulate whether order_i fails, for every $i = 1, \ldots, N$. This simulation yields a subset $I \subset \{1, \ldots, N\}$, corresponding to the orders that succeed in each run. Then, compute the above PnL metrics over the set $\{\text{order}_i\}_{i \in I}$ and aggregate these metrics across all runs to obtain statistical measures such as average, median, and standard deviation.

To assist in understanding this methodology, we illustrate its application via the two trading strategies defined earlier, when the first execution mode is employed, targeting liquidity at the best price.

We proceed to analyze the strategy PnL performance using both metrics PnL_1 and PnL_2 previously defined. Specifically, for each of these metrics, we consider three quantities of interest:

Table 7. Comparison of slippage model parameters and R^2.

Market	Slope	Intercept	R^2
Bybit BTC	− 0.0486	0.0000	0.659
Binance BTC	− 0.1526	0.0000	0.830
Bybit ETH	− 0.0337	0.0000	0.942

(i) *Naive PnL*, where the possibility of order failure is not taken into account.

(ii) *Actual PnL* achieved over the sample period of our trading experiment, based on known order outcomes.

(iii) *Simulated PnL*, where we apply our backtest PnL adjustment procedure, simulating order outcomes according to our fill probability model. We then compute statistical measures such as the average, median, and standard deviation of the PnL metric.

To avoid data overlap between the model training and backtest periods, we fit the failure probability model using only the first day of experimental data. The subsequent six days are reserved for backtesting, ensuring an unbiased evaluation. In Supplementary Material D, we show that even data sets of even a few hours can yield comparable parameter values, confirming the robustness of our approach.

It is important to note that we are able to compute the 'Actual PnL' quantity due to our unique data set where we sent market orders to the exchange at a high frequency, such that the orders the strategy would have sent over the course of the week-long sample period of our trading experiment *are actually a subset* of the orders that we did in fact send. That is, we use the subset of the orders that would have been triggered by the strategies that we study. This means we have full insight into order outcomes over the sample period, and hence know exactly how the strategy would have performed in practice during this time.

We evaluate the performance of the Orderbook Imbalance and Lead-Lag strategies using the short-term returns and realized PnL metrics defined above. The results are summarized in table 8, where teal-colored values are associated with the Orderbook Imbalance strategy, and olive-colored values with the Lead-Lag strategy.

The 3 bps taker fee on Bybit determines the profitability threshold for both metrics. For PnL_1, a strategy can be considered profitable if it achieves an average short-term markout greater than this taker fee. Profitability with the PnL_2 metric means achieving a roundtrip bps price difference greater than *twice* the taker fee, since a roundtrip involves the execution of two trades. Neither of our proposed strategies is profitable when accounting for these transaction costs.

It is also worth noting that the variability in the PnL_2 metric (see StdDev_2 values in table 9) is substantially larger due to the nature of the metric, which compares prices of consecutive opposite-side trades. For example, in the context of the Orderbook Imbalance strategy, the average time between consecutive trades is approximately 23 minutes, with a maximum time gap of more than 8 hours. This wide temporal spacing means that this metric inevitably includes variance of price drift, making it less suitable if one wants to isolate the short-term effects of order flow that lead to order failures, which are

Table 8. Metrics of PnL_1, $StdDev_1$, PnL_2, and $StdDev_2$ for the Orderbook Imbalance strategy and the Lead-Lag strategy. Values in Simulated PnL column are averaged across 100 simulated runs, with standard deviation provided in brackets.

Metrics	Naive		Actual		Simulated	
PnL_1	1.834	2.086	1.523	1.455	1.612 (0.043)	1.773 (0.058)
$StdDev_1$	1.941	2.330	1.759	1.701	1.934 (0.087)	2.168 (0.148)
PnL_2	3.873	3.689	1.882	0.003	2.251 (0.658)	2.185 (0.495)
$StdDev_2$	21.869	24.598	22.836	28.612	22.201 (1.223)	26.223 (0.713)
Number of Trades	466	624	352	381	388 (7)	531 (11)

better captured by our first metric. Despite its limitations, the realized PnL metric remains relevant as it directly quantifies the financial outcome over the trading period.

Both strategies exhibit a significant erosion in PnL when transitioning from naive to actual outcomes. For the Orderbook Imbalance strategy, the PnL erosion is around 17% for PnL_1 and 51% for PnL_2. For the Lead-Lag strategy, the decrease is even more pronounced, with a PnL_1 decrease of approximately 30% and nearly 100% for PnL_2, where the actual PnL is negligible.

The simulated PnL values are substantially lower than the naive PnL for both strategies and across both PnL_1 and PnL_2 metrics, thus better aligning with the actual realized PnL. For PnL_1, the simulation predicts an expected erosion of approximately 12% for the Orderbook Imbalance strategy and 15% for the Lead-Lag strategy. In the case of PnL_2, the simulated PnL suggests a decrease of about 42% for the Orderbook Imbalance strategy and 41% for the Lead-Lag strategy.

Remarks and Takeaways. Based on the analysis of both PnL metrics across the two strategies, we make a few important remarks and draw some conclusions from the data.

- The large discrepancy between naive and actual PnL highlights the critical importance of accounting for execution-related toxicity effects. Relying on backtesting under the naive assumption of taker order execution at prices from the last seen orderbook update proves to be a deeply flawed approach that can lead to inflated profitability expectations.
- Only ex-post profit-generating orders face the risk of failure (or slippage, discussed in the next section). Thus, money-losing strategies (e.g. ones that trade only during quiet times, agnostic to short-term returns) are unlikely to experience a further reduction of their already negative PnL; their orders will have a failure rate or approximately zero. In fact, strategies that are a priori destined to lose money fast (for instance, ones that counter-trade a profitable HFT signal) may actually fill at better-than-expected price, and thus experience an improvement of their PnL: they lose money slightly less fast.
- The fact that the simulated PnL, as per our methodology, constitutes a significant discount on the naive PnL (and therefore more closely resembles the actual PnL) validates the utility of our proposed procedure for enhancing backtest accuracy.
- While the simulated PnL serves as a useful practical guideline for quantitative researchers, it does not capture all factors that influence failure rates. For instance, it is to be expected that "obvious" trading alphas are more crowded than more nuanced or idiosyncratic alphas, and therefore experience higher failure rates. This is likely the underlying reason why the simulated PnL of our proposed trading strategies, which are based on highly obvious and widely employed signals, underestimates the erosive impact of order failures. This is particularly true for the lead-lag strategy where the actual PnL represents almost a 100% discount compared with the naive PnL, while our simulated PnL only discounts it by around 42%. We advise researchers to keep these caveats in mind when applying our methodology.
- The performance deterioration due to the toxicity effects discussed in this paper can likely be mitigated to an extent by devising more idiosyncratic trading alphas that fewer other traders perceive or react to. Strategies based on less crowded and more subtle alphas are likely to experience lower failure rates. However, quantifying the idiosyncrasy of a trading alpha a-priori is a challenging (perhaps intractable) task.
- Another factor pertaining to the lead-lag strategy is that it relies on cross-region data transfer (from Binance in Tokyo to Bybit in Singapore) and it is very difficult to assess whether we have made all the optimizations necessary to minimize the latency of this data transfer. If a quant firm built microwave towers connecting the two locations, they would likely always react quicker and experience substantially lower failure rates than we do, but we might never find out that they did so.†

7.4. *Second execution mode: market orders*

We now turn our focus to the second execution mode involving the placement of market orders. As with the previous execution mode, we consider three variants for each PnL metric:

(i) *Naive PnL*, assuming execution at the last observed best price, providing a baseline without slippage or order failure.
(ii) *Actual PnL*, using known fill prices, reflecting empirical performance including slippage.

† When placed on the geodesic path between the two locations, this would entail placing tens of microwave towers on open ocean, given the reach limitations of microwave towers.

Table 9. Metrics for Execution Mode 2 with the Orderbook Imbalance strategy and the Lead-Lag strategy.

Metrics	Naive		Actual		Simulated	
$\text{PnL}_{1,\text{market}}$	1.834	2.086	1.439	1.374	1.577	1.794
$\text{StdDev}_{1,\text{market}}$	1.941	2.330	1.876	1.985	1.670	2.004
$\text{PnL}_{2,\text{market}}$	3.873	3.689	2.317	1.782	2.576	2.604
$\text{StdDev}_{2,\text{market}}$	21.869	24.598	20.208	24.138	20.267	23.950
Number of Trades	466	624	466	624	466	624

(iii) *Simulated PnL*, using a model to estimate expected slippage, diverging from the previous mode by focusing on slippage instead of order failure probability.

Backtest PnL Adjustment: To compute the simulated PnL for the second execution mode, we employ our previously established findings on slippage (see equation (5)). We use our result that a linear function gives an accurate approximation of expected slippage as a function of an order's future return, as evidenced by figure 10 and table 7. We proceed with the following steps:

(i) For each market order order_i, $i = 1, \ldots, N$, we compute its short-term future return $\text{fret}(\text{order}_i)$ as per equation (7) (with respect to the best price at the time of submission).
(ii) We then estimate the expected slippage using a linear model of the form $\text{slip}(\text{order}_i) \approx \beta_0 + \beta_1 \cdot \text{fret}(\text{order}_i)$.
(iii) The slippage estimate is then applied to the best price at the time of order submission to obtain a simulated fill price. Using this simulated fill price, we compute the metrics $\text{PnL}_{1,\text{market}}$ and $\text{PnL}_{2,\text{market}}$.

To maintain the integrity of our backtesting process, we continue to use a time-separated approach for parameter estimation and backtesting, similar to the previous execution mode. We calibrate the linear model's parameters using data from the first day and reserve the following six days for backtesting. The parameters derived from this limited data set align closely with those in table 7, corroborating our observation (see Supplementary Material D) that even smaller sample sizes than ours are adequate for robust analysis.

The consolidated results for $\text{PnL}_{1,\text{market}}$ and $\text{PnL}_{2,\text{market}}$ in the second execution mode, presented in table 9, are discussed alongside the first execution mode to provide insights into the relative performance of trading at the top of the book versus using market orders.

While the naive PnL metrics are (by definition) identical in both execution modes, the actual PnL values diverge significantly. We begin by examining the $\text{PnL}_{1,\text{market}}$ row. For the Orderbook Imbalance strategy, the actual performance deteriorates by approximately 22% relative to the naive backtest PnL, a marginally larger decline than seen in the first execution mode. In the case of the Lead-Lag strategy, this decline is more pronounced at around 34%, again slightly exceeding the decrease seen in the first execution mode. This suggests that, perhaps unsurprisingly, short-term markouts end up being slightly worse when market orders are employed as opposed to IOC orders targeting the last seen best price. The simulated $\text{PnL}_{1,\text{market}}$ metric represents a 14% decrease relative to the naive $\text{PnL}_{1,\text{market}}$ for both the Orderbook Imbalance and the Lead-Lag strategies.

Let us now focus on the $\text{PnL}_{2,\text{market}}$ row. Here we find that actual performance represents a 40% and 52% decrease over the naive one for the Orderbook Imbalance and Lead-Lag strategies, respectively. This is notably much better than with the first execution mode, implying that the actual realized PnL in case of our strategies would have been much better with market orders than with IOC orders. Our simulated $\text{PnL}_{2,\text{market}}$ values constitute a 33% and 29% discount compared with the naive versions, respectively for the Orderbook Imbalance and the Lead-Lag strategies.

In our analysis of both execution modes, the actual PnL constitutes a substantial discount on the naive PnL, underscoring the necessity for traders to account for our finding when performing backtests. The simulated PnL also represents a significant reduction compared to the naive PnL, which illustrates that our methodology can help improve the accuracy of backtest results. However, as observed also in the first execution mode, the simulated PnL slightly overestimates actual performance. This discrepancy may arise from the fact that the trading strategies under consideration are very 'obvious' with the underlying signal likely acted upon by many other market participants, hence contributing to larger failure rates and worse slippage. Other factors, such as latency in cross-regional arbitrage strategies, should also be kept in mind when applicable. However, while these factors are perhaps impossible to estimate a-priori with any degree of certainty, our simulated PnL serves as a solid foundational baseline that traders can incorporate in their backtesting procedure. Put differently, our adjustment represents a PnL haircut over a broad universe of trading strategies. Depending on the sub-universe in which one attempts to devise a trading strategy, our adjustment may be more or less accurate, largely depending on the 'crowdedness' of that sub-universe.

8. Conclusion

We probed the Bybit Bitcoin market with several million market orders placed at high frequency over a one-week period, the Bybit Ethereum market with just over one million market orders, and the Binance market with several tens of thousands of orders over a few hours. We subsequently analyzed the execution-related outcomes of those orders. Our data set allowed us to analyze the empirical outcomes of two modes of execution: firstly, marketable limit orders targeting specific price levels, in which case the order can fail to fill; secondly, market orders, where the fill price is uncertain.

We showed that the likelihood of failure of a marketable limit order targeting the best price is highly correlated with volatility, latency, and LOB liquidity. We also showed that the outcomes of successive orders in the same direction are serially correlated.

Emerging from our analysis of the resulting empirical data is the unequivocal conclusion that taker orders in practice suffer from an adverse selection effect due to delays in the communication between exchange and trader. This adverse selection expresses itself in different ways for marketable limit orders (IOC or FOK orders with a limit price) versus market orders.

In the former case, profitable taker orders (as measured by short-term markouts) have notable probability of failing to execute due to intervening LOB updates; unprofitable orders almost always fill. We proposed a parsimonious three-parameter model that accurately captures the correspondence between an order's profitability (quantified by its future return) and its failure probability.

In the latter case, market orders that would have been profitable had they executed at the expected fill price (based on the most recent LOB snapshot), in practice tend to achieve a worse fill price, resulting in an erosion of profitability. Unprofitable market orders, on the other hand, fill at the expected price in almost all cases, again to the detriment of the traders who submitted them. The relationship between expected slippage and the profitability of an order (quantified by its short term future PnL return) is remarkably well described by a linear function.

Those adverse selection effects have wide-ranging consequences, including ones that apply to the backtesting of trading strategies involving taker orders. When these effects are unaccounted for in such backtests (particularly of HFT strategies optimizing for short-term markouts), the backtest results will overestimate the strategy profitability in the real world. We proposed a simple methodology of accounting for this execution uncertainty in order to conduct more accurate backtests.

8.1. Future work

The incipient field of cryptocurrencies offers a unique opportunity for empirical research, thanks to the widespread availability of market data and the low barriers to entry into trading. We hope that this environment contributes to a paradigm shift in academic finance and encourages finance scholars to take an experimental approach or to leverage the wealth of data available for insightful empirical studies.

We conclude with describing future avenues of studies, more directly related to failures of marketable limit orders aiming at liquidity at the touch. In this work, we showed that probability of failure for marketable limit orders is highly correlated with volatility, latency, and LOB liquidity. However, we expect that the probability of failure is also correlated with other quantities, such as past price action and volume patterns. For instance, at the beginning of a price trend upwards, failures of buy trade attempts are likely, whereas at a later point of that trend, when the trend begins to show signs of exhaustion, buy attempts are hypothesized to be substantially more likely to fill. Our intuition awaits testing.

Another promising avenue is the development of predictive models for execution outcomes. While our research identifies the key drivers of execution outcomes, constructing a comprehensive model that accurately predicts these outcomes remains an open problem. Such a model would need to account for the distinct probabilities of failure for buy and sell orders. The development of a robust predictive model for taker orders, encompassing both failure probabilities for marketable limit orders and slippage for market orders, would represent a significant step forward in the field.

Finally, while broad patterns in failure rates and slippage are consistent across different markets, we observed some differences in specifics. We proposed that these variations stem from differences in trader pools, with factors such as the number of active HFT takers and market makers, the sizes of targeted orders (to take or cancel), and the number of take or cancel attempts playing a role. Testing this hypothesis would require access to data on unsuccessful attempts to take or cancel liquidity—information not included even in L3 feeds. Researchers with access to such data, perhaps from an exchange, could further investigate and test our hypothesis.

Acknowledgments

We are grateful to the anonymous reviewers for their insightful comments and constructive suggestions, which have helped us improve the clarity and quality of this paper. For the purpose of open access, the author has applied a CC BY public copyright licence to any author accepted manuscript arising from this submission.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

No funding was received for conducting this study.

Supplemental data

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/14697688.2025.2515933.

References

Albers, J., Cucuringu, M., Howison, S. and Shestopaloff, A.Y., Fragmentation, price formation and cross-impact in Bitcoin markets. *Appl. Math. Finance*, 2021, **28**, 395–448.

Alexander, C., Heck, D.F. and Kaeck, A., The role of Binance in Bitcoin volatility transmission. *Appl. Math. Finance*, 2022, **29**, 1–32.

Arroyo, Á., Cartea, Á., Moreno-Pino, F. and Zohren, S., Deep attentive survival analysis in limit order books: Estimating fill probabilities with convolutional-transformers. *Quant. Finance*, 2024, **24**, 35–57.

Bailey, D.H., Borwein, J.M., de Prado, M.L. and Zhu, Q.J., Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the AMS*, 2014, **61**, 458–471.

Bailey, D.H., Borwein, J.M., de Prado, M.L. and Zhu, Q.J., The probability of backtest overfitting. *J. Compu. Finan.*, 2017, **20**, 36–39.

Baron, M., Brogaard, J. and Kirilenko, A.A., Risk and return in high frequency trading. *SSRN Electron. J.*, 2014, **54**, 993–1024.

Budish, E.B., Cramton, P. and Shim, J.J., The high-frequency trading arms race: Frequent batch auctions as a market design response. *SSRN Electron. J.*, 2013, **130**, 1547–1621.

Caccioli, F., Bouchaud, J.P. and Farmer, J.D., A proposal for impact-adjusted valuation: Critical leverage and execution risk. arXiv: General Finance, 2012. https://doi.org/10.48550/arXiv.1204.0922

Cartea, Á., Jaimungal, S. and Sánchez-Betancourt, L., Latency and liquidity risk. *Int. J. Theor. Appl. Finance*, 2021, **24**, 2150035.

Cartea, A. and Sánchez-Betancourt, L., The shadow price of latency: Improving intraday fill ratios in foreign exchange markets. *SIAM J. Financ. Math.*, 2021, **12**, 254–294.

Cartea, Á. and Sánchez-Betancourt, L., Optimal execution with stochastic delay. *Finance Stochastics*, 2023, **27**, 1–47.

Chen, H., Foley, S. and Ruf, T., The value of a millisecond: Harnessing information in fast, fragmented markets. *SSRN Electron. J.*, 2016.

Degryse, H., Winne, R.D., Gresse, C. and Payne, R.G., Cross-venue liquidity provision: High frequency trading and ghost liquidity. *SSRN Electron. J.*, 2019.

Foucault, T., Hombert, J. and Rosu, I., News trading and speed. *SSRN Electron. J.*, 2012, **71**, 335–382.

Gould, M.D., Porter, M.A., Williams, S., McDonald, M., Fenn, D.J. and Howison, S.D., Limit order books. *Quant. Finance*, 2013, **13**, 1709–1742.

Gould, M.D., Porter, M.A. and Howison, S.D., The long memory of order flow in the foreign exchange spot market. *Mark. Microstruct. Liquidity*, 2016, **2**, 1650001.

Hansen, P.R. and Lunde, A., Forecasting volatility using high-frequency data. In *The Oxford Handbook of Economic Forecasting*, pp. 525–556, 2011 (Oxford University Press: Oxford).

Harvey, C.R. and Liu, Y., Backtesting. *J. Portfolio Manage.*, 2015, **42**, 13–28.

Kolm, P.N. and Webster, K., Do you really know your PnL? The importance of impact-adjusting the PnL. *SSRN Electron. J.*, 2023.

Liao, G.Y. and Caramichael, J., Stablecoins: Growth Potential and Impact on Banking. *International Finance Discussion Paper*, 2022, **2022**, 1–26.

Maglaras, C., Moallemi, C.C. and Wang, M., A deep learning approach to estimating fill probabilities in a limit order book. *Quant. Finance*, 2022, **22**, 1989–2003.

Soska, K., Dong, J.D., Khodaverdian, A., Zetlin-Jones, A., Routledge, B. and Christin, N., Towards understanding cryptocurrency derivatives: A case study of BitMEX. In *Proceedings of the Proceedings of the Web Conference 2021, WWW '21*, April, 2021 (ACM).

van Kervel, V., Competition for order flow with fast and slow traders. *Rev. Financ. Stud.*, 2015, **28**, 2094–2127.

Zhang, L., Estimating covariation: Epps effect, microstructure noise. *J. Econom.*, 2011, **160**, 33–47.