

Fragmentation, Price Formation and Cross-Impact in Bitcoin Markets

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

To cite this article: Jakob Albers, Mihai Cucuringu, Sam Howison & Alexander Y. Shestopaloff (2021) Fragmentation, Price Formation and Cross-Impact in Bitcoin Markets, Applied Mathematical Finance, 28:5, 395-448, DOI: [10.1080/1350486X.2022.2080083](https://doi.org/10.1080/1350486X.2022.2080083)

To link to this article: <https://doi.org/10.1080/1350486X.2022.2080083>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



View supplementary material [↗](#)



Published online: 17 Jun 2022.



Submit your article to this journal [↗](#)



Article views: 585



View related articles [↗](#)



View Crossmark data [↗](#)

Fragmentation, Price Formation and Cross-Impact in Bitcoin Markets

Jakob Albers^a, Mihai Cucuringu^{b,c}, Sam Howison^d and Alexander Y. Shestopaloff^e

^aDepartment of Statistics, University of Oxford, Oxford, UK; ^bDepartment of Statistics and Mathematical Institute, University of Oxford, Oxford, UK; ^cThe Alan Turing Institute, London, UK; ^dMathematical Institute, University of Oxford, Oxford, UK; ^eSchool of Mathematical Sciences, Queen Mary University of London, London, UK

ABSTRACT

In the light of micro-scale inefficiencies due to the highly fragmented bitcoin trading landscape, we use a granular data set comprising orderbook and trades data from the most liquid bitcoin markets, to understand the price formation process at sub-1-second time scales. To this end, we construct a set of features that encapsulate relevant microstructural information over short lookback windows. These features are subsequently leveraged, first to generate a leader–lagger network that quantifies how markets impact one another, and then to train linear models capable of explaining between 10% and 37% of total variation in 500 ms future returns (depending on which market is the prediction target). The results are then compared with those of various PnL calculations that take trading realities, such as transaction costs, into account. The PnL calculations are based on natural taker strategies (meaning they employ market orders) associated with each model. Our findings emphasize the role of a market's fee regime in determining both its propensity to lead or lag, and the profitability of our taker strategy. We further derive a natural *maker* strategy (using only passive limit orders) which, due to the difficulties associated with backtesting maker strategies, we test in a real-world live trading experiment, in which we turned over 1.5 M USD in notional volume. Lending additional confidence to our models, and by extension to the features they are based on, the results indicate a significant improvement over a naive benchmark strategy, which we also deploy in a live trading environment with real capital, for the sake of comparison.

ARTICLE HISTORY

Received 7 January 2022
Accepted 11 May 2022


KEYWORDS

Limit orderbooks; bitcoin; market fragmentation; high frequency data; market microstructure

1. Introduction

Cryptocurrency markets have seen an explosion of trade volumes over the past year, as the price of bitcoin soared to an all-time high of over 60,000 USD in April 2021, sparking interest both from the broad public and academics alike. The trading landscape for bitcoin in particular has matured considerably in recent years, with an ever-greater proportion of

CONTACT Jakob Albers  jakob.albers@merton.ox.ac.uk

 Supplemental data for this article can be accessed here: <https://doi.org/10.1080/1350486X.2022.2080083>

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

trade volume occurring in complex derivatives rather than spot (fiat for bitcoin) markets. With the rise of derivative volumes came a flurry of academic works investigating their role in price formation. For instance, Alexander and Heck (2020), Alexander et al. (2019) examine the price impact of a set of popular unregulated derivative exchanges, Hung, Liu, and Yang (2021) perform a similar analysis, and Aleti and Mizrach (2020) survey market microstructural differences between spot and futures markets. Further work such as Soska et al. (2021) investigates how the proliferation of derivatives has exacerbated price jumps and thus increased volatility.

One of the main distinguishing elements of bitcoin markets compared to traditional stock or bond markets, aside from the much greater volatility of bitcoin, is the high level of *fragmentation* of the bitcoin trading landscape. This fragmentation stems from the existence of 5–10 highly liquid, mutually independent exchanges where the majority of trading takes place. On each of these exchanges, the volume is typically further distributed across a number of different liquid instruments with slightly different specifications and properties. The exchanges are subject to different regulatory environments depending on their base location. They often use different server locations, resulting in cross-sectional arbitrages that can persist for at least as long as the time of information transfer between server locations. There is no cross-collateralization across exchanges, unlike in traditional markets where prime brokers play this role. That is, a trader who enters a long position on one exchange and the equivalent short position on another is market-neutral but must still ensure that both positions are sufficiently collateralized in order to avoid a liquidation event.

Summary of Main Contributions

The main contributions of our work are to identify possible answers to the following questions:

1. Where does new price information tend to arrive first? That is, which venue is most often the originator of price transmission?
2. What are the predominant directions of information flow?
3. In detail, how is a price on one venue affected by the arrival of new information on another?
4. To what extent do time discrepancies in price impact enable price prediction on each venue? That is, can we leverage the cross-section of information from all markets to make reliable price predictions?
5. Can these predictions be leveraged to produce trading strategies in a natural way, which give large PnL values? Do we obtain predictive power sufficient to produce ‘alpha’ in excess of transaction costs?

We add to the rapidly-growing literature on bitcoin price formation in at least the following two ways. First, it is the first of its kind to comprehensively examine the consequences of the aforementioned fragmentation of the bitcoin trading ecosystem. Our study encompasses the largest markets in terms of trade volume while, to the best of our knowledge, all other studies focus on the impact of either a single market or a

limited subset of markets. Often such subsets comprise illiquid spot markets or, for example, the CME bitcoin futures markets, which have little bearing on the price due to their small trade volume. Second, unlike other studies, we use highly granular order-book data which allows us to produce predictive models rather than merely explaining contemporaneous returns.

Paper Outline

In Section 2, we describe the bitcoin trading ecosystem. We also describe the data set used in our study, and the preprocessing needed to render the data amenable to subsequent analysis. In Section 3, we define a set of microstructural features based on orderbook and trade information. In Section 4, we use these features to examine lead-lag effects between the markets considered in this study. Section 5 is devoted to the last two questions from the list above. We compare three different methodologies for developing powerful linear predictive models that make use of all available information. In Section 6, we address a follow-up question arising from the results of the previous section, namely whether the previously-trained linear models are sufficiently powerful to avoid adverse selection for a simple natural maker strategy. In Section 7, we conclude by summarizing our findings and discussing future research directions. Finally, additional numerical experiments and explanations are deferred to the Appendix, which is provided as supplemental data to this article.

2. Background

2.1. Historical Developments and Market Specifications

Crypto markets have matured considerably in recent years. This can be seen by the increasing dominance of derivatives relative to spot volumes. Nowadays, average daily volumes on derivatives markets eclipse those on spot markets by a factor of greater than 5, whereas up until (and including) the year 2018 spot markets (fiat for cryptocurrency) had greater average trade volumes (Tkacik 2021; [Cryptocompare Spot vs Derivatives Volumes](#)). The most traded derivatives are so-called *perpetual swaps* (sometimes simply called a perpetual, and explained below) and futures. Option volumes are still small to negligible, in contrast with traditional markets; this can be interpreted as a sign of lower investor sophistication in bitcoin markets. Derivatives offer two main advantages to traders. First, they greatly expand the ability of traders to express views on future price moves or future volatility. In particular, they make it possible to *short* bitcoin, i.e., to bet on a price decline. Second, they enable the use of leverage, which allows for much more effective use of capital. Due to the absence of prime brokers in crypto markets (as opposed to traditional finance), the precise mechanics of leveraged trading and ensuring trader solvency are different. For more details, see [Spot vs Leveraged Trading](#).

One would expect the aforementioned advantages to give rise to a more efficient marketplace. By and large, this expectation is borne out in reality albeit with notable exceptions. In practice, one often observes the phenomenon of ‘cascading liquidations’ in bitcoin markets (a key puzzle piece in the explanation of the fat tails of their return distributions). That is, large amounts of leverage¹ can exacerbate small price moves and turn them into much larger ones, for example when an initial relatively small price move triggers a set of forced

liquidations, which in turn causes a larger price move and hence might trigger more liquidations, and so on. The price history of bitcoin is littered with interesting case studies of this dynamic playing out in real life. For more details, see Soska et al. (2021).

Terminology. By ‘market’, we mean a pair of *exchange* and *symbol*, where a symbol on a given exchange denotes either a spot market (e.g., BTC/USD) or a derivative contract (e.g., a quarterly futures contract). We define a *limit order* to be a triple consisting of side (buy or sell), limit price, and amount. Such an order can be either entirely executed immediately, partially executed and partially passive, or entirely passive. Further, we define a *market order* to be an ‘aggressive’ order submitted to be executed immediately at the best available price (or, if the volume at that price is insufficient, at whatever set of prices is necessary to complete it: this is often referred to as ‘walking the book’). In contrast to a limit order, no price is specified for a market order, although if it can be executed at the top of the book it is, in effect, the same as a limit order at that price.

Selection of Markets. Motivated by the desire to capture a significant fraction of globally traded volume, we chose to study those bitcoin markets which (to the best of our knowledge) exhibit the highest trading volumes. In Figure 1, we show the ranking of major bitcoin markets by their average daily volumes over the course of February 2021. In the remainder of this paper, we chiefly concern ourselves with the top 14 markets from this bar plot. Next, we provide further context on each of these markets along with their specifications.

Market Specifications. Each of the markets shown in Figure 1 can be classified as either a *spot* market, a *futures* contract market, or a *perpetual contract* market. The latter market type was invented by the crypto exchange Bitmex in 2014, and, to the best of our knowledge, does not exist outside of the crypto trading world at the time of writing (July 2021). Of the three types of market, the perpetual swap is the most popular among crypto traders and it has by far the largest trade volumes. Since its inception, the success and broad popularity of the perpetual contract have compelled many other exchanges to

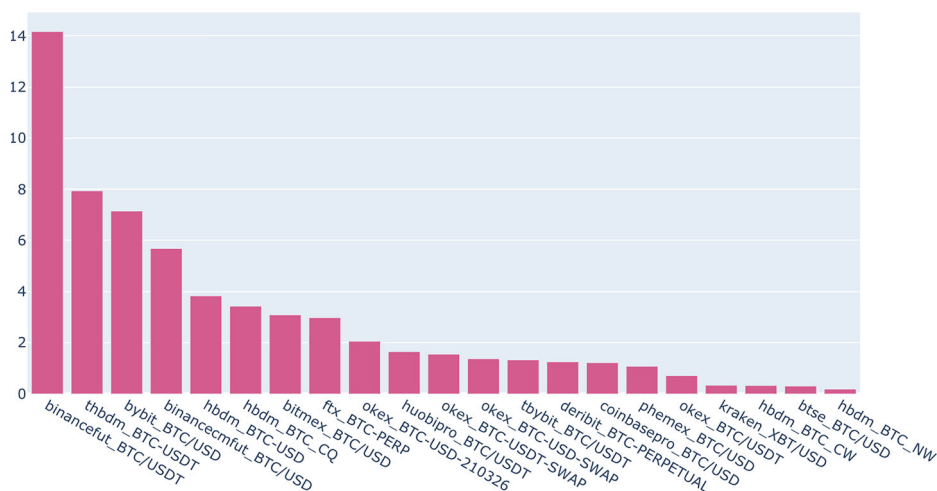


Figure 1. Average daily trading volume in billion USD per market in February 2021.

follow the blueprint originally laid out by Bitmex. Competing exchanges such as Binance, Huobi, Okex, FTX, Deribit, Bybit, and Kraken have either directly copied this derivative contract or launched a nearly identical product with some minor modifications. The combined trade volume across these perpetual contracts routinely exceeds 50 billion USD per day ([Volume Monitor](#)). Given the large role the perpetual swap plays in the crypto markets, we provide some background on its mechanics in Appendix A.1. Further detailed information can also be found in Alexander et al. (2019). Different bitcoin derivatives can use different types of collateral (also known as margin), with the most common ones being bitcoin itself or stablecoins such as Tether. More information on different types of collateralization can be found in Appendix A.2.

The underlying index of a derivative contract is typically a (weighted) average value of a basket of BTC/USD spot markets or BTC/USDT spot markets. The composition of the index can vary from exchange to exchange. The BTC/USD indices typically include at least the Coinbase, Bitstamp and Kraken spot markets, while the BTC/USDT indices typically include spot markets from Binance, Huobi and Okex. The precise mechanism via which the index price is computed from its constituents is generally also different between exchanges. For instance, some exchanges exclude the markets with the minimal and maximal price in the calculation, to achieve both greater robustness of the index price and resistance to price manipulations ([Deribit Index Documentation](#)). Opportunities for the latter can arise because trade volumes on derivative markets far exceed those on the spot markets which make up the derivatives' underlying indices. Colloquially speaking, the much larger volume on derivative markets than spot markets is sometimes referred to as the 'tail wagging the dog'. Index manipulation that seeks to exploit this have been a commonplace occurrence in bitcoin trading history. An interesting case study can be found in [Coin Metrics Market Manipulation Report](#). However, with increasing liquidity, greater market efficiency and more robust index price calculations by derivatives exchanges, such manipulations have become a rarity as of the time of this writing (August 2021).

The 14 most liquid markets from Figure 1 include two futures contracts, 11 perpetuals and only 1 spot market. Table 14 in the Appendix lists some of their properties such as maximum leverage or margin currency. Note that it is not uncommon for exchanges to report fake volumes; this creates challenges in identifying exactly what the most actively traded markets are. For more details, see the discussion in Section A.4 of the Appendix. The most liquid futures markets tend to be the ones whose expiry happens at the end of a quarter. Huobi and Okex, the exchanges whose quarterly futures appear in our set of markets, both list a weekly expiration, a bi-weekly one and a quarterly one. Okex additionally offers a bi-quarterly futures contract.

In the remainder of this paper, we refer to `binancefut_BTC/USDT` as the 'Binance USDT-margined perpetual contract' or simply the 'Binance USDT perpetual'. Similarly, `binancecmfut_BTC/USD` is referred to as the 'Binance BTC-margined perpetual contract' or, for short, the 'Binance BTC perpetual'. Other markets are described in the same fashion using their type, exchange, and when necessary, their margin currency.

2.2. Data and Preprocessing

Our data set comprises orderbook and trades data, mined by subscribing to the exchange's websocket API orderbook and trades endpoints for each of the markets studied; see [Websocket Wikipedia Article](#), though we will give a brief description below.

Websocket API endpoints generally provide the most granular crypto market data one can procure, pushing new data to the subscriber in real time. For trades data, one receives information about all executed trades on the relevant market. For orderbook data, exchanges typically publish orderbook snapshots every 5–200 ms depending on market activity. Note that we therefore do not have order-by-order information (typically referred to as Level 3 data), since exchanges batch changes over a short time window and then publish these updates as part of a new orderbook snapshot. The frequency at which new orderbooks arrive not only depends on market activity but also varies from exchange to exchange. In some cases, there is even variation in the update frequency within an exchange when a market has multiple websocket API endpoints with slightly different properties. For instance, the Bybit BTC-margined perpetual offers a feed that publishes a new snapshot containing 25 orderbook levels every 20 ms, while at the same time offering another slower feed which publishes updates every 100 ms but provides information for the top 200 orderbook levels ([Bybit Websocket API Documentation](#)). A common maximum update frequency used by exchanges (even when they do not document this) is around 20 ms. That is, orderbook updates arrive at a frequency not higher than one per 20 ms, even when market activity is extreme. Exceptions are Deribit and Bitmex where update frequencies are sometimes in the single digit milliseconds when market volatility is high.

There is little uniformity across exchanges in the properties of their data feeds. The number of orderbook levels varies from exchange to exchange (and even from market to market), the update frequencies are different, and the data comes in a different file format for each exchange. For each market studied, we receive between 25 and 75 orderbook levels. In our choice of the websocket API feed, we generally preferred speed over orderbook depth. A single day of orderbook data for all markets combined is over 30 gigabytes in size. Creating uniformity and comparability within our large and disorganized data set posed a major challenge.

Another challenge arises from the fact that different exchanges are generally housed in different cloud data centres across the globe in which they match orders and from which they send out market data. This can give rise to a certain forward-looking bias if one were to conduct an analysis that relies on timestamps provided by the exchange since those timestamps reflect when the event (a trade or orderbook update) occurred on the exchange's server. It would be physically impossible for a trader to observe market events at precisely the timestamp provided with the data by the exchange. This is exacerbated when the trader receives data from multiple exchanges in different locations. We address these concerns by subscribing to all data feeds from a particular AWS server in Singapore, by timestamping the data when we receive it, and then using those timestamps for subsequent analysis. Not only does that alleviate asynchronicity concerns in the empirical portion of our paper, but it also helps us in our subsequent real world trading experiment on the exchange Bybit whose servers are located in the same AWS data center in Singapore from which we gather and timestamp the data used to calibrate our models.

Furthermore, since the data is gathered asynchronously, timestamps of orderbooks generally differ. To more easily compare orderbooks across markets, we resample the data to a 50-ms frequency using the last seen observation ([Pandas Resample Function](#)). Missing values can appear when a market does not publish an orderbook update for, say, 100 ms. In this case, we fill the missing value by propagating the last valid observation forward. For the sake of consistency, we also resample trades data to the same 50 ms frequency. After

this processing step, the buy (or sell) amount in a given 50 ms window is the sum of all buy (sell) trade amounts that occurred in this time window. The trade amounts can have different denominations on different markets. For example, for certain perpetual contracts, the value of one contract is 1 USD on one market while on others it may be 100 USD or 0.001 bitcoin. To obtain like-for-like comparisons between trade volumes, we normalize all trade amounts to a USD amount, using the last-seen bitcoin price to convert BTC to USD.

Since we use a 50-ms resampling frequency, we define the trade price over a 50-ms period as the volume-weighted average price computed over all trades (irrespective of side) which took place during this 50 ms window. If no trade occurs during a time window, we fill the missing price value, as before, with the previous valid price observation.

After preprocessing, our datapoints are equally spaced in time (50 ms apart), making cross-market comparisons easier. For each observation, we have one orderbook per market (possibly with differing numbers of levels), and trade information on total buy amount, sell amount, and average price (aggregated over the previous 50 ms).

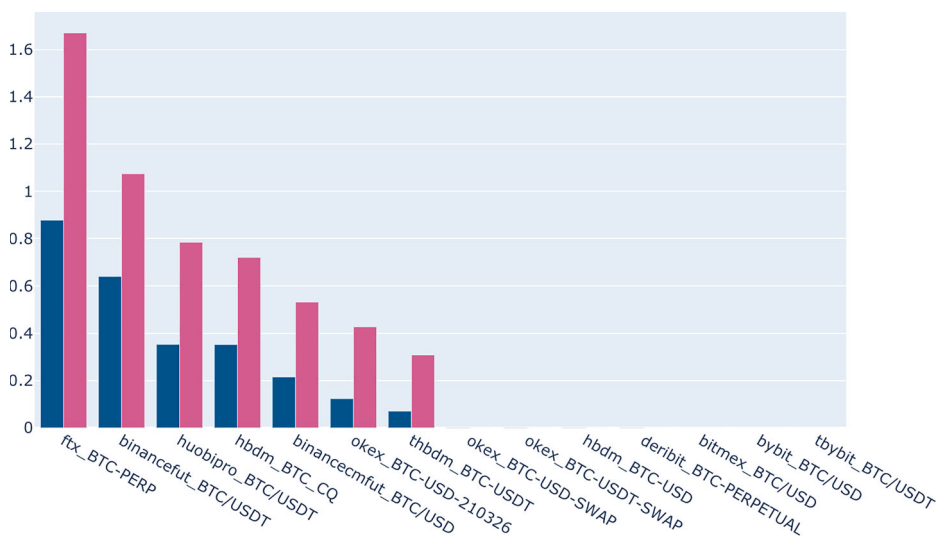
2.3. Remarks on Execution Cost

The cost of execution of a taker order (one which leads to immediate execution) comprises two components. The first is the taker fee charged by the exchange. Fees exhibit great variation across exchanges and even across markets within a single exchange. For instance, the taker fee for a derivatives contract, such as `ftx_BTC-PERP`, is substantially lower than that of a spot market such as Coinbase BTC/USD where the taker fee is up to 50 bpts (this market is not included in our data set because its volume is too small). A number of exchanges additionally offer tiered fee structures, where certain groups of traders receive lower fees than others. These fee rebates can be very substantial. Traders can ascend to a better fee tier by having high transaction volume. The volume cutoffs dictating a trader's fee tier vary between markets. The top tier typically requires 500 million to 1 billion USD of monthly notional volume. Certain exchanges (Binance, FTX and Huobi) have issued their own cryptocurrency and offer additional fee discounts for traders that hold it in their exchange wallet. The magnitude of this additional fee discount is usually proportional to the holdings in the exchange's native cryptocurrency. An alternative (often undocumented and under-the-table) way of obtaining a preferred fee status is through the formation of strategic partnerships with exchanges. Newly launched exchanges, in particular, with little liquidity are typically interested in attracting liquidity, which they can achieve, for example, by partnering with market making firms whom they incentivize to provide the desired liquidity by offering various fee rebates. Table 1 displays the lowest possible (documented) taker fee that a trader in the top fee tier (who also owns a sufficient amount of the exchange's native cryptocurrency, when applicable) receives. Notice that these heavy fee rebates for 'VIP' traders lead to a built-in wealth advantage as VIP traders, who presumably have some measure of wealth by virtue of satisfying the requirements to be a member of the best fee tier, incur far lower trading costs than, say, a novice trader.

The second component of the execution cost can be broadly described by the term 'spread'. We define this cost as the difference between the midprice and the average execution price. An aggressor thus always incurs a spread cost of at least half the bid-ask spread. If the size of their order is greater than the liquidity available at the first orderbook level which is aggressed against, the trader additionally 'walks the book'. That is, part of his

Table 1. Lowest possible and default taker fees per market in basis points.

| | Lowest possible fee | Ordinary fee |
|-----------------------|---------------------|--------------|
| ftx_BTC-PERP | 1.5 | 7 |
| binancefut_BTC/USDT | 1.53 | 4 |
| binancecmfut_BTC/USD | 1.8 | 5 |
| huobipro_BTC/USDT | 1.93 | 4.75 |
| hbdm_BTC_CQ | 2 | 4 |
| okex_BTC-USD-210326 | 2.5 | 5 |
| thbdm_BTC-USDT | 2.7 | 4 |
| okex_BTC-USD-SWAP | 3 | 5 |
| okex_BTC-USDT-SWAP | 3 | 5 |
| hbdm_BTC-USD | 3.7 | 5 |
| deribit_BTC-PERPETUAL | 5 | 5 |
| bitmex_BTC/USD | 7.5 | 7.5 |
| bybit_BTC/USD | 7.5 | 7.5 |
| tbybit_BTC/USDT | 7.5 | 7.5 |

**Figure 2.** Temporary price impact of taker orders of size > 100k USD.

execution occurs at one or multiple prices which are worse than the top quote. In Figure 2, we provide an illustration of the size of the spread cost incurred by an aggressor. In this plot, the set of blue bars represents the median difference in bpts between the minimum and the maximum prices of an order, while the red set of bars represents the median difference in bpts between the best price (minimum for buy orders, maximum for sell orders) and the average price of an order. The sample over which the median was taken consists of all orders of size > 100k USD between February 15–27 (midnight), 2021. If a market has a bar of height zero in Figure 2, this means that the median temporary price impact (as defined above) of orders > 100k USD is zero. This points to a large amount of top of the book liquidity, capable of absorbing most orders of size larger than 100k USD.

It is interesting to note that there seems to be an inverse relationship between a market's VIP taker fee and the price impact (in other words, spread cost) of moderately large to large

orders (greater than 100k USD). Writing

$$\text{execution cost} = \text{taker fee} + \text{spread}, \quad (1)$$

in practice, the terms on the right-hand side roughly compensate each other across exchanges. That is, if a market has a high taker fee, the orderbook is usually more liquid and hence the spread cost is small. Conversely, a market with a low taker fee has little liquidity at or near the top of the book, and hence the spread cost is large. One explanation for this phenomenon is that a large taker fee typically also implies a large maker rebate, which in turn incentivizes market makers to quote a tighter spread since part of the adverse selection experienced by the market maker is recompensed by the rebate. We will later see what ramifications the fee structure has on the lead-lag behaviour of a market. An alternative and equivalent perspective is that of the taker: a larger taker fee means that high frequency takers require a stronger signal to overcome the fee, and thus they have a higher threshold for sending market orders. Thus the twin effects of incentivizing market makers to keep their quotes for longer on the one hand and disincentivizing takers from sending taker orders on the other hand both contribute to the same effect: spreads remain tight longer.

3. Features and Preprocessing

The fundamental microstructural determinant of price movements in a limit orderbook market is *order flow*, which comes in two forms. The first one is *aggressive*, comprising market orders or limit orders that lead to immediate execution.² The second one is *passive*, comprising limit orders that do not lead to immediate execution or cancellation of limit orders. Our objective in this section is to develop a set of base features that build on order flow, and which achieve high efficacy for certain tasks tackled later in this paper, such as predicting future returns or characterizing cancellation behaviour of HFT market makers. We discuss our base features one by one and provide a motivation and justification for their inclusion. In later sections, we discuss ways of transforming and optimizing the base features for the tasks that lay ahead.

3.1. Orderbook Imbalances

It is well established in the literature that the shape of the orderbook has significant impact on the distribution of future returns (Bouchaud, Mézard, and Potters 2002; Alfonsi, Fruth, and Schied 2009; Cont, Kukanov, and Stoikov 2014; Bouchaud, Farmer, and Lillo 2008). For example, if the liquidity on the bid side far outweighs that of the ask side, the volume required for a price move up is much smaller than the volume required for a price move down. Thus if we assume that the arrival of a sell order is not much more likely than the arrival of a buy order, the probability of a future price move up is larger than that of a price move down. Other liquidity profiles can similarly be interpreted and increasing the likelihood of an appropriate future price evolution. The empirical evidence cited above strongly suggests the inclusion of an orderbook imbalance-based feature in our set of features. A commonly-used measure of orderbook imbalance is given by

$$\frac{v_b - v_a}{v_b + v_a} \in [-1, 1], \quad (2)$$

where v_a and v_b represent the top of the ask and bid liquidity, respectively Cartea, Jaimungal, and Penalva (2015). However, this simple ('classical') quantity only takes into account the liquidity at the two tops of the book: information from deeper orderbook levels is discarded. This is particularly problematic for markets where the liquidity at the top of the book tends to be very small, with liquidity instead concentrated more strongly around deeper orderbook levels.

Extensions of this simple orderbook imbalance typically account for liquidity on deeper levels via a weighting scheme, replaces the quantity a in Equation (2) by $\sum_j w_j v_{a,j}$, where $v_{a,j}$ denotes the volume on orderbook level j and $w_j \in \mathbb{R}_{>0}$ is a weight, and similarly for v_b ; see, e.g. Xu, Gould, and Howison (2018). The weight w_j is usually constructed in a way which reflects the probability of execution of an order on level j . It can, however, be difficult to estimate the execution probability at each orderbook level. Conventional models of orderbook dynamics imply that the probability of execution decays exponentially with distance to the midprice. In practice, however, it is not only the distance to midprice that matters, but also, for example, the distribution of liquidity in the orderbook or the fee structure. We saw evidence of this in Figure (2), where we noted that a low taker fee seems to imply more liberal 'walking of the book' (i.e., greater trade impact and hence a larger execution probability on deeper levels compared to markets where the taker fee is large). This makes it difficult in practice to choose the weights, since they must be chosen individually for each market. To further complicate matters, different markets often have different tick sizes; for instance, the tick size on FTX is 1 USD, while Binance uses a tick size of 0.1 USD.³

Another concern with the basic orderbook imbalance (2) (and most of its extensions) is that it does not adequately reflect uncertainty of future returns in times of high volatility when (effective) spreads are large. To illustrate this point, consider a sample orderbook during non-volatile times (Figure 3a) and one shortly after a volatility spike (Figure 3). The classical orderbook imbalance for the first orderbook has the value -0.98 , reflecting well the intuitive observation, based on inspection, that the ask liquidity is overwhelmingly large compared with the bid liquidity. We would expect, all else being equal, a price move down to be much more likely than a price move up. The second orderbook gives a classical imbalance of 0.69 which one would interpret as a relatively strong predictive price signal of a price move up. However, a glance at deeper levels suggests that, in fact, a price move down should be more likely than a price move up. The discrepancy between the interpretation of the classical imbalance value and our intuition occurs because the liquidity at the best bid and ask prices is much smaller than that deeper in the orderbook, and so has little bearing on the distribution of future returns. A submission of a mere 5000 USD to the top ask would result in a vastly different orderbook imbalance of -0.42 . Such small top-of-the-book amounts are a common occurrence during volatility breakouts and they often persist for several hundred milliseconds; in effect, they represent a widening of the spread.

This phenomenon makes the classical orderbook imbalance values extremely noisy, rendering it unsuitable for practical purposes. We therefore propose an alternative pair of orderbook-based features (one for each of the ask and bid sides) which are more robust during volatility breakouts, and which do not depend on a complicated choice of weights. Let us define the following quantities for each market $i = 1, \dots, 14$ and for fixed quantities

| Orderbook (XBTUSD) | | | ⚙️ ↗️ ✕ |
|-------------------------------|-----------|-----------|---------|
| Price | Size | Total | |
| 32829.0 | 83,000 | 3,817,200 | |
| 32828.5 | 67,500 | 3,734,200 | |
| 32828.0 | 41,800 | 3,666,700 | |
| 32827.5 | 201,900 | 3,624,900 | |
| 32827.0 | 63,700 | 3,423,000 | |
| 32826.5 | 61,000 | 3,359,300 | |
| 32826.0 | 253,000 | 3,298,300 | |
| 32825.5 | 113,900 | 3,045,300 | |
| 32825.0 | 74,000 | 2,931,400 | |
| 32824.5 | 64,400 | 2,857,400 | |
| 32823.0 | 8,600 | 2,793,000 | |
| 32822.5 | 300,000 | 2,784,400 | |
| 32822.0 | 226,200 | 2,484,400 | |
| 32821.5 | 131,800 | 2,258,200 | |
| 32821.0 | 155,300 | 2,126,400 | |
| 32820.5 | 1,971,100 | 1,971,100 | |
| 32817.5 ↓ | | | |
| 🔍 32823.37 / 32819.88 □□□□□ | | | |
| 32820.0 | 20,900 | 20,900 | |
| 32818.5 | 1,300 | 22,200 | |
| 32817.5 | 500 | 22,700 | |
| 32815.5 | 100 | 22,800 | |
| 32814.5 | 1,000 | 23,800 | |
| 32811.5 | 25,000 | 48,800 | |
| 32810.0 | 10,100 | 58,900 | |
| 32806.0 | 200 | 59,100 | |
| 32803.0 | 5,800 | 64,900 | |
| 32802.5 | 100 | 65,000 | |
| 32802.0 | 400 | 65,400 | |
| 32800.5 | 100 | 65,500 | |
| 32800.0 | 15,700 | 81,200 | |
| 32799.0 | 21,600 | 102,800 | |
| 32798.5 | 9,300 | 112,100 | |
| 32797.5 | 100 | 112,200 | |

(a) Low volatility regime.

| Orderbook (XBTUSD) | | | ⚙️ ↗️ ✕ |
|-------------------------------|-----------|-----------|---------|
| Price | Size | Total | |
| 32731.5 | 25,000 | 3,387,600 | |
| 32731.0 | 66,100 | 3,362,600 | |
| 32730.5 | 94,500 | 3,296,500 | |
| 32730.0 | 118,400 | 3,202,000 | |
| 32729.5 | 110,000 | 3,083,600 | |
| 32727.5 | 102,500 | 2,973,600 | |
| 32727.0 | 72,300 | 2,871,100 | |
| 32725.5 | 133,900 | 2,798,800 | |
| 32725.0 | 331,300 | 2,664,900 | |
| 32724.5 | 50,400 | 2,333,600 | |
| 32722.0 | 1,250,000 | 2,283,200 | |
| 32721.5 | 39,900 | 1,033,200 | |
| 32720.5 | 304,300 | 993,300 | |
| 32720.0 | 593,800 | 689,000 | |
| 32719.5 | 94,800 | 95,200 | |
| 32719.0 | 400 | 400 | |
| 32719.5 ▼ | | | |
| 🔍 32743.85 / 32740.38 □□□□□ | | | |
| 32718.5 | 2,200 | 2,200 | |
| 32717.5 | 1,000 | 3,200 | |
| 32717.0 | 100 | 3,300 | |
| 32716.0 | 100 | 3,400 | |
| 32713.0 | 100 | 3,500 | |
| 32712.5 | 100 | 3,600 | |
| 32712.0 | 20,600 | 24,200 | |
| 32711.5 | 100 | 24,300 | |
| 32711.0 | 190,300 | 214,600 | |
| 32710.5 | 5,100 | 219,700 | |
| 32710.0 | 7,100 | 226,800 | |
| 32709.5 | 500 | 227,300 | |
| 32709.0 | 200 | 227,500 | |
| 32708.5 | 100 | 227,600 | |
| 32708.0 | 6,300 | 233,900 | |
| 32707.5 | 19,600 | 253,500 | |

(b) High volatility regime.

Figure 3. Bitmex orderbooks for both volatile and non-volatile times. Sizes are in USD: (a) low volatility regime and (b) high volatility regime.

$$N_1, \dots, N_{14} \in \mathbb{R}_{\geq 1}$$

$$\text{IMB}_t^{a,i} := \left(\frac{p_{a,t}^i(N_i)}{p_{a,t}^i(1)} - 1 \right) \cdot 10,000, \quad \text{IMB}_t^{b,i} := \left(\frac{p_{b,t}^i(1)}{p_{b,t}^i(N_i)} - 1 \right) \cdot 10,000, \quad (3)$$

where $p_{a,t}^i(x)$ denotes the average price one would pay at time t for a market *buy* order of size x USD on market i (recall that in our preprocessing we normalized all amounts to USD). In particular the quantity $p_{a,t}^i(1)$ denotes the price of a market buy order of size 1USD, which is simply the top ask price since 1USD is the smallest permitted size. Similarly $p_{b,t}^i(x)$ is the average price for a market *sell* order of size $x \in \mathbb{R}_{\geq 1}$, so that $p_{b,t}^i(1)$ is just the top bid price. In order for this definition to cover the case where the total volume on one side (bid or ask) of the orderbook is insufficient (i.e., less than x), we augment the orderbook by prices $+\infty$ and $-\infty$, where the volume is infinite.

The quantity $\text{IMB}_t^{a,i}$ can be described as the difference in basis points between the top ask price on market i and the average price of a market order of size N_i , and analogously for

the bid version $\text{IMB}_t^{b,i}$. We now need to consider the choice of N_i , noting that $N_i = 1$ always yields $\text{IMB}_t^{a,i} = 0$, while letting $N_i \rightarrow \infty$ we have $\text{IMB}_t^{a,i} \rightarrow \infty$; thus the two extreme ends of the spectrum offer no signal. Based on preliminary experiments, we set N_i to be the median liquidity within the top five basis points of the top of the book on market i , calculated as follows: we compute this median value for the ask side and the bid side separately, and then take the average of the two resulting values. Our analysis indicate this to be a sensible choice, in the sense that the resulting features (3) appear to contain much signal and little noise. We leave it to future work to calibrate the choice of N_i in a more rigorous fashion.

Let us revisit the two orderbooks considered above. The orderbook from Figure 3(a) has $(\text{IMB}_t^{a,i}, \text{IMB}_t^{b,i}) \approx (0, 7)$. The small $\text{IMB}_t^{a,i}$ value means that the ask side is heavily populated, while the large $\text{IMB}_t^{b,i}$ points at the sparsity of the bid side. Comparing these two values, we would unequivocally conclude that a price drop is far more likely than an increase. For the orderbook from Figure 3(b) we have $(\text{IMB}_t^{a,i}, \text{IMB}_t^{b,i}) \approx (1, 7.4)$. As in the previous case, the far larger $\text{IMB}_t^{b,i}$ value suggests a much greater probability of a price decrease, which agrees with our intuition from observing the liquidity distribution in the orderbook, and which is in contrast with the greater likelihood of a price increase suggested by the classical orderbook imbalance.

Moreover, our pair of features retains its high fidelity during highly volatile times when liquidities on both sides of the book are sparse. In such cases, the greater uncertainty in the distribution of future returns is represented by values $\text{IMB}_t^{a,i}$ and $\text{IMB}_t^{b,i}$ which are both large.

As further empirical evidence of the suitability of the features defined in (3), we compare their explanatory power with that of the classical orderbook imbalance. Specifically, for each market $i = 1, \dots, 14$, we fit the following pair of univariate linear models:

$$\text{fret}_t = \alpha + \beta(\text{IMB}_t^{a,i} - \text{IMB}_t^{b,i}) + \epsilon_t, \quad (4)$$

$$\text{fret}_t = \alpha' + \beta' \left(\frac{v_{b,t}^i - v_{a,t}^i}{v_{b,t}^i + v_{a,t}^i} \right) + \epsilon'_t, \quad (5)$$

where $v_{a,t}^i$ and $v_{b,t}^i$ are the top ask and bid on market i , respectively, at time t . The prediction target fret_t used here consists of 500 ms future returns on the Bybit BTC perpetual. We fixed this target market for the sake of concreteness; other markets give similar results. The above linear models are fitted using OLS on training data spanning February 22nd until February 27th (midnight UTC), 2021. Over this time period, the price of bitcoin moved from approximately 57,500 USD to 46,200 USD. See Table 2 for a comparison of the corresponding R^2 values. With the new imbalance measure, we find an average increase in R^2 of 0.9557% compared to the classical one. The new imbalance measure outperforms the classical one in 11 out of 14 cases.

3.2. Trade Imbalances

Let us define the *trade flow imbalance* indicators, for each market $i = 1, \dots, 14$, over time horizons $\delta \in \{100 \text{ ms}, 250 \text{ ms}, 500 \text{ ms}, 1000 \text{ ms}, 2000 \text{ ms}\}$, by

$$\text{TFI}_t^{i,\delta} := B_{[t-\delta,t]}^i - S_{[t-\delta,t]}^i, \quad (6)$$

Table 2. Comparison of R^2 values of classical order-book imbalance and our proposed imbalance.

| | $R^2_{\text{classical}}$ | R^2_{new} |
|-----------------------|--------------------------|--------------------|
| bybit_BTC/USD | 0.1242 | 0.1628 |
| hbdm_BTC-USD | 0.0898 | 0.1201 |
| thbdm_BTC-USDT | 0.0784 | 0.0925 |
| tbybit_BTC/USDT | 0.0673 | 0.0679 |
| hbdm_BTC_CQ | 0.0162 | 0.0583 |
| ftx_BTC-PERP | 0.0195 | 0.0392 |
| okex_BTC-USD-SWAP | 0.0359 | 0.0541 |
| bitmex_BTC/USD | 0.067 | 0.0421 |
| deribit_BTC-PERPETUAL | 0.0106 | 0.0159 |
| binancecmfut_BTC/USD | 0.059 | 0.0607 |
| huobipro_BTC/USDT | 0.0167 | 0.0205 |
| okex_BTC-USDT-SWAP | 0.022 | 0.0103 |
| okex_BTC-USD-210326 | 0.0138 | 0.0083 |
| binancefut_BTC/USDT | 0.0037 | 0.0052 |

where $B^i_{[t-\delta, t]}$ represents the total volume of buy trades⁴ on market i in the time interval $[t - \delta, t]$. Similarly, $S^i_{[t-\delta, t]}$ represents the total volume of sell trades on market i over the same time interval.

This feature can be construed as the ‘aggressive’ component (relating to taker flow rather than submissions or cancellations of passive limit orders) of the order flow imbalance indicator defined by Cont, Kukanov, and Stoikov (2014)

$$\begin{aligned}
 \text{ofi}_t &= [\text{buy flow} - \text{sell flow}]_{[t-\delta, t]} \\
 &= \underbrace{[(\text{buy trades} - \text{sell trades})]}_{\text{‘aggressive’ component}} + (\text{bid submissions} - \text{ask submissions}) \\
 &\quad + (\text{ask cancels} - \text{bid cancels})_{[t-\delta, t]}.
 \end{aligned} \tag{7}$$

The order flow imbalance indicator was demonstrated in Cont, Kukanov, and Stoikov (2014) to have significant explanatory power over *contemporaneous* returns. In Cont, Cucuringu, and Zhang (2022), similar results are demonstrated for future returns, when employing a variation of the order flow imbalance from Cont, Kukanov, and Stoikov (2014), showcasing the usefulness of this family of indicators.

The fragmented nature of the bitcoin trading universe means that volumes are scattered across several venues. This implies an asynchronous information arrival: trade flow typically arrives first on one venue and then ‘trickles across’ to other venues in the form of cancellations or taker orders. We therefore expect the cross-section of trade flows also have some explanatory power for future returns, so we include the trade flow imbalance indicators (6) in our set of base features. We also make the following remarks.

- We believe that the aggressive component in (7) has greater importance (more signal) than the remaining two ‘passive’ terms, because passive flow (submissions or cancellations) are almost exclusively the work of HFT market makers who are largely only reacting to the cross-section of aggressive flow and orderbook imbalances, rather than contain genuine new information. A rigorous investigation of this topic is left for future work.

- One might consider a relative measure of the form $(B_{[t-\delta,t]}^i - S_{[t-\delta,t]}^i)/(B_{[t-\delta,t]}^i + S_{[t-\delta,t]}^i)$, instead of the trade flow imbalance $B_{[t-\delta,t]}^i - S_{[t-\delta,t]}^i$. However, this relative indicator suffers from some of the same drawbacks as the classical orderbook imbalance, namely a far noisier predictive signal than its non-relative USD denominated analogue. For instance, if the sell volume is zero, a trivial buy volume of size 1 USD gives rise to the same imbalance value as a significant buy volume of size, say, 10 million USD.
- The nature of our orderbook data (snapshots data) makes it impossible to precisely compute the remaining two ‘passive’ terms that make up the order flow quantity (7). A precise computation requires order-by-order feeds, and is left as future work, for the case of equity data, for which such feeds are available.⁵

3.3. Past Returns

Arguing as we did for the trade flow imbalance, we expect a drastic price change on one market to cross-impact short-term future returns on other markets. It is, therefore, natural to include past returns on each market in our feature set. We use the same time horizons here as for the trade flow imbalances. For each market $i = 1, \dots, 14$ and time horizon $\delta \in \{100 \text{ ms}, 250 \text{ ms}, 500 \text{ ms}, 1000 \text{ ms}, 2000 \text{ ms}\}$, we define the following indicator at each time t :

$$\text{PRET}_t^{i,\delta} := \left(\frac{p_t^i}{p_{t-\delta}^i} - 1 \right) \cdot 10,000, \quad (8)$$

where p_t^i denotes the price on market i at time t . This quantity is the price change in basis points on market i over the time interval $[t - \delta, t]$. We should clarify what notion of price p_t^i represents, since there are many candidates (mid price, last trade price, and so on). We use the last trade price, modified to deal with the ambiguity arising when a number of trades are executed simultaneously at different price levels (e.g. for ‘sweeping’ orders which consume multiple orderbook levels). To circumvent this concern, we compute the average price over a short lookback window of 50 ms (which is the minimum time step in the preprocessing of our data). That is, we define p_t^i as

$$p_t^i := \frac{1}{V} \sum_{(p,v) \in T_t^i} p \cdot v, \quad (9)$$

where T_t^i is the set of all trades, i.e., pairs (p, v) of price and amount, executed on market i in the time period $[t - 50 \text{ ms}, t]$ and $V = \sum_{(p,v) \in T_t^i} v$. Note that we include both buy and sell trades in T_t^i .

3.4. Mean Divergence

The instruments traded on the markets that we consider are closely related but not identical. The similarity stems from the fact that they are all either perpetuals or futures contracts on a bitcoin underlying (a BTC/USDT or BTC/USDT index), or a BTC/USDT spot market. We naturally expect price differences between markets to be mean-reverting processes since arbitrage opportunities would otherwise apparently arise. This motivates

the inclusion of the price differences between pairs of markets in our set of base features, as indicators for future returns. One would expect, for instance, that if a particular market is cheap relative to some (or all) other markets (e.g., due to a price move up on the other markets) there should be a short-term price increase on the cheap market, as arbitrageurs and market makers rush to profit from the discrepancy. However, several assumptions underlying this hypothesized market behaviour do not hold in practice, making the actual dynamics more complicated. These confounding factors include the following.

- (1) The presence of trading fees restricts the set of arbitrages. That is, price differences that would represent arbitrages in a world without trading fees are not in fact arbitrage opportunities. It follows that the price difference between a pair of markets can fluctuate within a ‘no arbitrage band’ whose size is given by the sum of the taker fees of the two markets. Only when the price difference exceeds this sum is there an arbitrage that can be immediately profited from with a pair of taker orders.
- (2) The futures premium of, say, quarterly futures contracts is (mostly) nonzero, so that the price difference between a futures contract and another market does not *a priori* indicate where the price will move in the near future. Rather, what matters is how far the price difference between the futures contract and the other market strays from its ‘equilibrium’ level. This equilibrium is difficult to assess in practice, since it depends on future expected funding rate payments whose magnitude is uncertain. To see why this is, consider the standard arbitrage argument underpinning the spot-futures relations via a long spot position, a short futures position, and a bond. In our case, the role of the bond is played by a perpetual; however, unlike for a bond, the cash flows on the perpetual are not predictable farther into the future. Hence, the futures premium, which is determined by the expected cash flows of the perpetual, generally fluctuates with changing expectations of future funding rate payments until the expiration of the futures contract. In particular, since the number of funding payments converges to zero as once approaches maturity, one ultimately obtains a convergence of the futures price to that of the underlying.
- (3) The price difference between a BTC/USDT based market (e.g., a perpetual whose underlying is a basket of BTC/USDT spot markets) and a BTC/USD based market alone cannot be expected to be indicative of future returns since it depends on the USD/USDT rate, which is generally only *approximately* 1, but can fluctuate by several basis points, and in rare cases even percentage points. As we discussed before, such fluctuations are usually related to varying investor confidence in the peg between USD and USDT which can decrease the value of USDT (see Griffin and Shams 2020). The value of one USDT can also temporarily exceed 1 USD during periods where investors are willing to pay a premium for the convenience of transacting in USDT rather than USD (usually when sentiment is extremely bullish). Deviations of the USD/USDT rate from 1 cannot usually be arbitrated immediately since the conversion typically takes several business days and involves a bank transfer; however, the credit risk that led to the original discrepancy may remain over this period, ruling out an arbitrage.
- (4) The dominant trader demographics differ from exchange to exchange. For instance, US citizens are barred from trading on Bitmex and must therefore resort to other exchanges. These types of restrictions can result in persistent price differences between, say, Bitmex and another exchange which is popular with US traders, perhaps

arising if prevailing investor sentiment and expectations of future returns differ between distinct demographics, for example due to divergent macroeconomic backdrops. Other reasons for persistently large price divergences include differences in capital controls, financial regulation, ‘know-your-customer’ (KYC) requirements, etc. across different countries and exchanges. A prominent example is the premium of the bitcoin price on Korean exchanges relative to US exchanges, dubbed as the *Kimchi premium*, which at various points in the past was notoriously large for extended periods. More details can be found in [Kimchi Premium Investopedia](#).

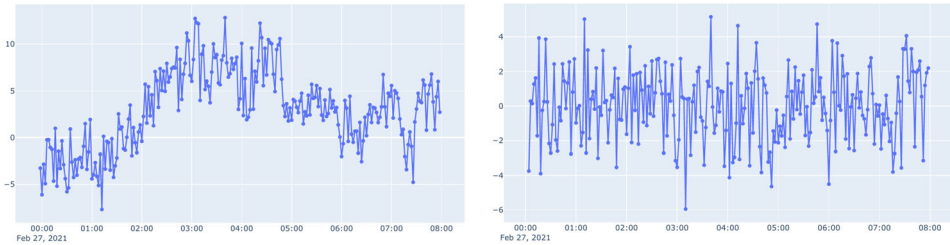
Taken together, these factors diminish the predictive signal of price differences. For visual evidence of this, consider Figure 4(a), in which we show the price difference between a pair of perpetuals over an eight-hour period; here one uses BTC/USD as its underlying, while the other uses a BTC/USDT index.

We address these concerns by correcting for average price differences by using their deviation from the mean price difference over a moving time window, of length ranging from several seconds to several minutes, ending at the current time. We anticipate that persistent large price differences as mentioned above (e.g., nonzero futures premium, USD/USDT rate different from 1, or differing investor outlook between distinct groups of traders) are usually approximately constant over such short time horizons. For instance, barring an extreme event, the repricing of a futures premium most often happens over several hours or days as investors gradually adjust expectations of future interest rates. Our proposed indicator therefore serves to center the price difference of such confounding longer time scale effects relative to their slowly-varying (instantaneous) mean. In Figure 4(b), we show the same price difference depicted in Figure 4(a) with its five minute rolling mean subtracted. Note the much clearer mean-reversion.

In sum, this motivates the definition of the following features for pairs of markets $(i, j) \in \{1, \dots, 14\}^2$ and time horizons $\Delta \in \{5 \text{ s}, 9 \text{ s}, 19 \text{ s}, 38 \text{ s}, 75 \text{ s}, 150 \text{ s}, 300 \text{ s}, 600 \text{ s}\}$

$$\text{DIV}_t^{i,j,\Delta} = d(p_t^i, p_t^j) - \text{rolling}^\Delta(d(p_t^i, p_t^j)), \quad (10)$$

where $d(p, q) = (p - q)/q \cdot 10,000$ is the relative difference in basis points between $p, q \in (0, \infty)$, and the $\text{rolling}^\Delta(\cdot)$ function returns the rolling mean of samples from its input over



(a) Price difference between the Binance USDT perpetual and the Huobi BTC perpetual (b) Mean divergence feature for the Binance USDT perpetual and the Huobi BTC perpetual

Figure 4. To avoid an excessively large number of samples (and hence to facilitate legibility), the data was resampled to a 2 minute frequency: (a) price difference between the Binance USDT perpetual and the Huobi BTC perpetual and (b) mean divergence feature for the Binance USDT perpetual and the Huobi BTC perpetual.

Table 3. Comparison of R^2 values of vanilla price difference (left column) with divergence from mean (right column).

| | R^2_{diff} | R^2_{div} |
|-----------------------|---------------------|--------------------|
| binancecmfut_BTC/USD | 0.057 | 0.1426 |
| binancefut_BTC/USDT | 0.0252 | 0.1331 |
| hbdm_BTC_CQ | 0.0017 | 0.1224 |
| okex_BTC-USD-210326 | 0.0015 | 0.119 |
| okex_BTC-USD-SWAP | 0.0337 | 0.1169 |
| okex_BTC-USDT-SWAP | 0.0175 | 0.1113 |
| thbdm_BTC-USDT | 0.0099 | 0.0961 |
| hbdm_BTC-USD | 0.0175 | 0.0778 |
| ftx_BTC-PERP | 0.031 | 0.0754 |
| deribit_BTC-PERPETUAL | 0.0149 | 0.0744 |
| huobipro_BTC/USDT | 0.0118 | 0.0661 |
| tbybit_BTC/USDT | 0.0032 | 0.0393 |
| bitmex_BTC/USD | 0.0038 | 0.016 |

the past Δ seconds. We shall call $\text{DIV}_t^{i,j,\Delta}$ the *mean divergence* feature between markets i and j .

As further evidence of the superiority of this indicator compared to the vanilla price difference, we display in Table 3 a comparison of R^2 values corresponding to the univariate regression models

$$\text{fret}_t = \alpha + \beta(d(p_t^{i_0}, p_t^j)) + \epsilon_t, \quad \text{and} \quad (11)$$

$$\text{fret}_t = \alpha' + \beta'(\text{DIV}_t^{i_0,j,150\text{ ms}}) + \epsilon'_t, \quad (12)$$

for each $j = 1, \dots, 14$. For sake of concreteness, we fix a reference market i_0 , namely the Bybit BTC perpetual, whose future returns are used in the above models. That is, we are comparing the explanatory power over Bybit's future returns of, one the one hand, the 'plain' price difference between Bybit and other markets and, on the other hand, the mean divergence feature of Bybit with other markets. On average, the explanatory power of our indicator exceeds that of its raw version by 0.074, representing an average 15-fold increase. The improvement is particularly drastic for the quarterly futures contracts hbdm_BTC_CQ and okex_BTC-USD-210326 where we see a more than 70-fold increase.

3.5. Eliminating Nonlinearities

Our objective is to train linear price prediction models using the features defined above. The performance of a linear model is, however, limited by the extent to which the future returns vary linearly with the feature variables. If the relationship between variates and covariates is nonlinear, we cannot, in general, expect to achieve good performance of our model. In this section, we demonstrate the nonlinear nature of the dependence between future returns and some of our indicators. We address these concerns by constructing bespoke nonlinear transformations of the features, which we then use in linear models instead of the raw features.

To illustrate the presence of some of the aforementioned nonlinearities consider Figure 5(a). This scatter plot shows the 500 ms trade flow imbalance of the Binance USDT

perpetual plotted against the 500 ms future returns of the Huobi BTC perpetual. The horizontal axis corresponds to trade flow imbalance, split into intervals

$$B_n := [n - 1, n) \cdot 10,000\text{USD},$$

for integers $n \in \{-100, -99, -98, \dots, 99, 100\}$. The plotted value (vertical axis) associated with the bucket B_n was obtained by computing the mean 500 ms future returns on the Huobi perpetual over all samples where the Binance trade flow lies in B_n . Inspection of Figure 5(a) suggests fitting to a sigmoid function rather than a linear one (although samples for intervals B_n with large $|n|$ are sparse, hence the data is relatively noisy). The sigmoid-like relationship between trade flow imbalance and future returns agrees well with prior work, which noted the same empirical observation (Potters and Bouchaud 2003; Plerou et al. 2002).

Why does the price impact decrease as trade flow imbalance grows? One possible economic interpretation is that extreme trade flow values (large in absolute value) may be correlated negatively with trader sophistication and positively with the ‘sloppiness’ of execution. That is, an informed trader wishing to trade a large position will rarely do so by using, for example, one large market order or a large ‘meta order’ executed over a time window as small as 500 ms. It is perhaps more likely that extreme trade flow values originate from uninformed traders, or, in the extreme case, are the result of forced liquidations where traders involuntarily liquidate their entire positions as a consequence of a margin call. If this hypothesis is correct, the relative lack of sophistication reflected by extreme trade flow imbalances may be recognized by other (more astute) market participants who would then only minimally adjust their estimation of the ‘fair price’, and hence contribute towards a more limited price impact. Another possible explanation for the relatively low impact of large market orders is that traders sending those orders only do so at points in time where the liquidity on the opposing side of the orderbook is large, thereby constraining their market impact.

The details of our feature transformation, not directly relevant to the remainder of this paper, are given in Section B in the Appendix. In Figure 5(b), we plot the transformed Binance trade flow imbalance feature against Huobi returns. Note the stronger degree of linearity in the dependence of future returns on the transformed feature compared with the raw feature, plotted in Figure 5(a).

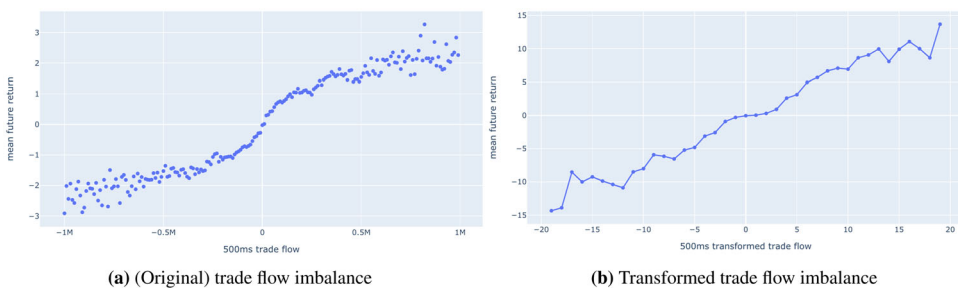


Figure 5. 500 ms future returns of the Huobi BTC perpetual as a function of Binance USDT perpetual 500 ms original and transformed trade flow imbalance: (a) (Original) trade flow imbalance and (b) transformed trade flow imbalance.

3.6. Time Horizon Selection

With the exception of the orderbook imbalance features, all of our base features are defined over multiple time horizons. However, to reduce complexity it is desirable to select only one time horizon per base feature. This is because base features of a single feature category at different time horizons have a high degree of collinearity. For instance, 500 ms past returns are highly correlated with 1s past returns (the values often coincide). We now set out how we select a single time horizon for any feature defined over multiple time horizons. With an eye to eventually training large composite models that employ the full feature set, this additional feature selection step helps in ameliorating concerns of overfitting (it is common knowledge that OLS regression is prone to overfitting when the covariates are large in number and exhibit high cross-correlations). Additionally, it is an interesting research question *per se* to investigate which lookback windows for which features are most predictive of, say, 500 ms future returns.

Our methodology for selecting the ‘optimal’ time horizon is a simple process of computing an average predictability score based on R^2 values for each time horizon, and then choosing the one with the highest score. Let us denote the time horizons corresponding to a (generic) feature f by $\delta_1, \dots, \delta_{K_f}$. Suppose we wish to determine the optimal time horizon δ_{k^*} corresponding to the set of features $f^{\delta_1}, \dots, f^{\delta_{K_f}}$. We proceed in two steps. The first step consists of using OLS regression to fit the following model for every time horizon $k \in \{1, \dots, K_f\}$ and target market $j = 1, \dots, 14$

$$\text{fret}_t^{500 \text{ ms}, j} = \alpha + \beta f_t^{\delta_k} + \epsilon_t, \quad (13)$$

whose coefficient of determination we denote by $R_{k,j}^2$. For the second step, we compute the following averages for each time horizon $k \in \{1, \dots, K_f\}$

$$\bar{R}_k^2 := \frac{1}{14} \sum_{j=1}^{14} R_{k,j}^2, \quad (14)$$

and then set $k^* := \operatorname{argmax}_k \bar{R}_k^2$. We call δ_{k^*} the *optimal time horizon* for the feature f ; it is the one whose average explanatory power over 500 ms future returns is largest.

Empirical results. We now examine the empirical results of the procedure described above.

(1) *Trade flow imbalance (TFI) indicators.* We use the transformed TFI, since as noted above it yields a considerable improvement over the raw one (for brevity, we generally omit the qualifier ‘transformed’). In Figure 6, we illustrate for each market’s (transformed) TFI the quantities $\bar{R}_1^2, \dots, \bar{R}_4^2$ corresponding to the five time horizons.

Neither the smallest nor the largest time horizons is selected for any market, indicating that the optimal value falls somewhere in between. The 500 ms horizon is preferred in eight cases, compared with two cases for the 250 ms horizon and four for the 1000 ms horizon. Finally, it is noteworthy that the five markets whose trade flows have the most predictive power are all on either Binance or Huobi; this set of five markets comprises the two perpetual contracts (one of them margined with USDT, the other one with bitcoin) on Binance and Huobi, as well as the Huobi quarterly futures contract. This furnishes our

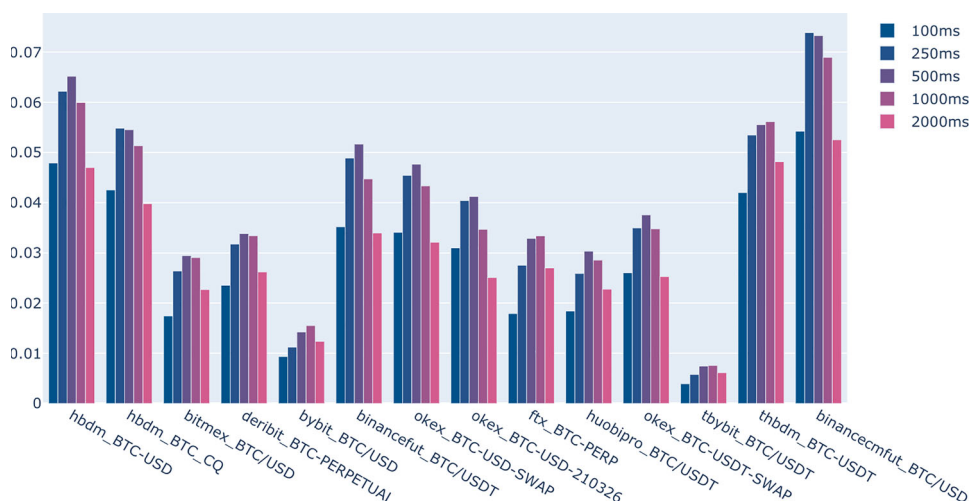


Figure 6. Average explanatory power of TFI features for each time horizon.

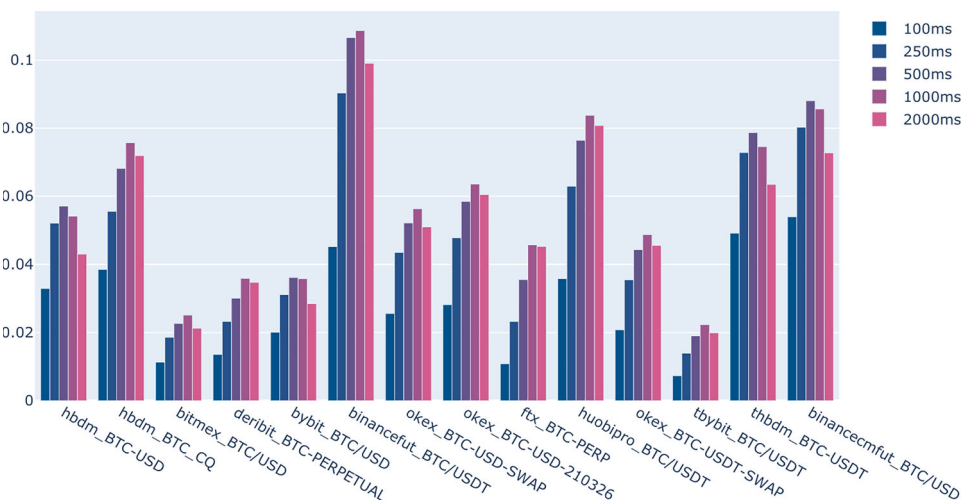


Figure 7. Average explanatory power of past returns features for each time horizon.

first indication of the leading behaviour of Binance and Huobi, as we shall see repeatedly below.

(2) *Past returns indicators.* For these features, we previously also found an improvement when using the transformed feature instead of the original one, so we again use the transformed version in the following analysis. In Figure 7, we show for each market the average explanatory powers $\overline{R}_1^2, \dots, \overline{R}_5^2$ corresponding to the five time horizons.

As for the TFI indicators, all the optimal time horizons lie strictly between the minimum and maximum value. The 500 ms horizon is selected in six cases, compared with four cases for 1000 ms and two for 250 ms. The most commonly selected optimal horizon is 500 ms for six markets. It is noteworthy that the two markets where the smaller time horizon (250 ms) is preferred are also among the markets where the smaller time horizon is preferred for the

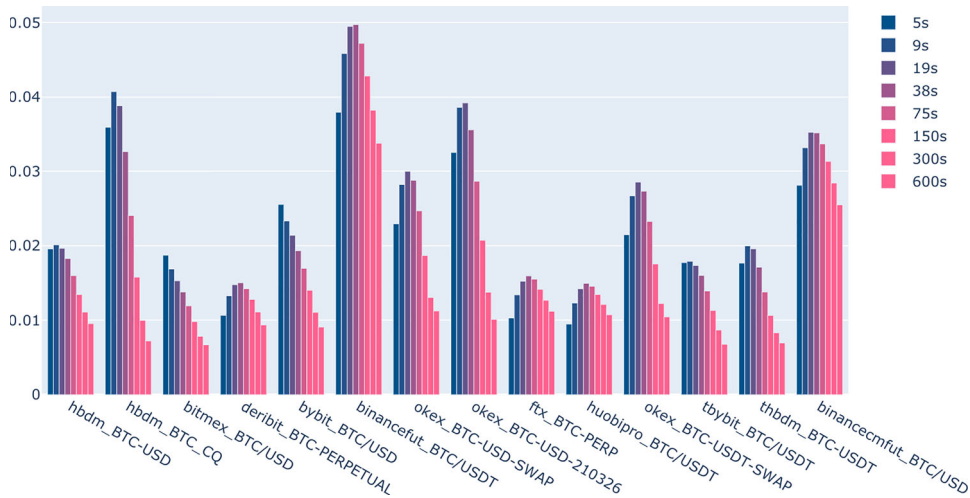


Figure 8. Average explanatory power of mean divergence features for each time horizon.

TFI feature. Furthermore, the markets which appeared to be leaders based on their comparatively large R^2 values for the TFI features also have among the highest explanatory powers in terms of their past returns features. This lends additional confidence to the previously conjectured leading behavior of the two Binance and Huobi perpetual contracts, as well as the Huobi quarterly futures market.

(3) *Mean divergence feature.* Here, the feature transformation do not yield any improvement, so we use the original feature. A plot of each market's average R^2 values is shown in Figure 8. The smallest time horizon (5 s) appears twice as the optimal choice; while 9 s, 19 s and 38 s each appear for four markets; longer horizons are not selected at all. The previously noted pattern of leading markets persists in this feature class, as does the Binance USDT-margined perpetual achieving the largest average R^2 , followed by the Huobi futures contract, and the Binance BTC perpetual.

For each feature, the time horizon we shall henceforth use in this study is the one which was determined as the optimal one. For instance, the TFI indicator of the Huobi BTC perpetual will be computed over a 500 ms horizon while that of the Huobi quarterly futures contract will employ the 250 ms lookback window.

4. Network Effects in Bitcoin Markets

Having finalized our feature set, we now examine network effects between the markets. We investigate the presence and strength of *lead-lag relationships* between pairs of markets, and we examine how they relate to the realities of trading.

4.1. Leader-Lagging Network

What does it mean for a market j to lead another market i ? Intuitively, one would say that a market j *leads* a (lagging) market i if future returns on the lagging market i can reliably be anticipated based on information from market j . In our case, this information consists of microstructural data encapsulated in the features we previously calculated.

Let us make this more precise. A market i is said lag another market j if a large portion of the total variation in future returns of market i can be explained by features of the leading market j . It is, therefore, natural to fit the following linear models for each pair of markets $(i, j) \in \{1, \dots, 14\}$ for the future returns $\text{fret}_t^{\delta,i}$ over a lookahead window of length δ :

$$\text{fret}_t^{\delta,i} = \mu_{ij} + \beta_{ij,1}\text{IMB}_j^{a,j} + \beta_{ij,2}\text{IMB}_t^{b,j} + \beta_{ij,3}\text{TFI}_t^j + \beta_{ij,4}\text{PRET}_t^j + \beta_{ij,5}\text{DIV}_t^{ij} + \epsilon_{ij,t}; \quad (15)$$

in these models, the features are taken to be in their transformed form and using the optimal lookback window (and we suppress these facts in our notation).

The models are fitted for both future returns lookahead windows $\delta \in \{500 \text{ ms}, 1000 \text{ ms}\}$, although as the results turn out to be similar, we report results for the 500 ms horizon unless stated otherwise. We use OLS regression with training data spanning the period February 22–27 (midnight UTC), 2021. The feature data was normalized by subtracting the mean and dividing by the standard deviation. The out-of-sample data was normalized using the in-sample parameters.

Let us denote the coefficient of determination of the above fitted linear model by R_{ij}^2 . This is the total variation in future returns on market i explained with features from market j , and it is our first indicator of the lead-lag relationship between the two markets. We depict the matrix $R := (R_{ij}^2)$ in Figure 9 as a pair of heatmaps, where the first one uses an ordering by column sum, and the second uses an ordering by row sum.

The largest R^2 value, of about 25.9%, is achieved when predicting FTX using features from the Binance USDT perpetual. The smallest R^2 value is about 1% for predicting the Binance USDT perpetual using features of the Bybit USDT perpetual. Overall, we achieve an average R^2 of 9.6% and a median of 9%. The results are similar when we use $\delta = 1000 \text{ ms}$ as our future returns time horizon, where we find average and median values of 10.4% and 9.2%, respectively. This is quite remarkable considering the fact that only information of a single market was used in these regression models. In other words, even information from just a single market can be quite effective in explaining future returns. It demonstrates that it is possible and indeed quite feasible to anticipate future price moves in bitcoin markets on sub-1s time scales, and it is a reassuring demonstration of the utility of our features.

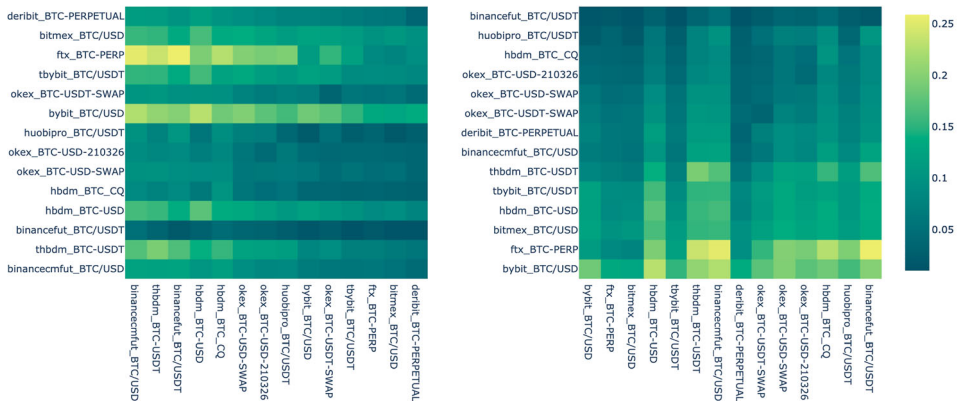


Figure 9. Entries in the matrix R illustrating the lead-lag relationships between pairs of markets.

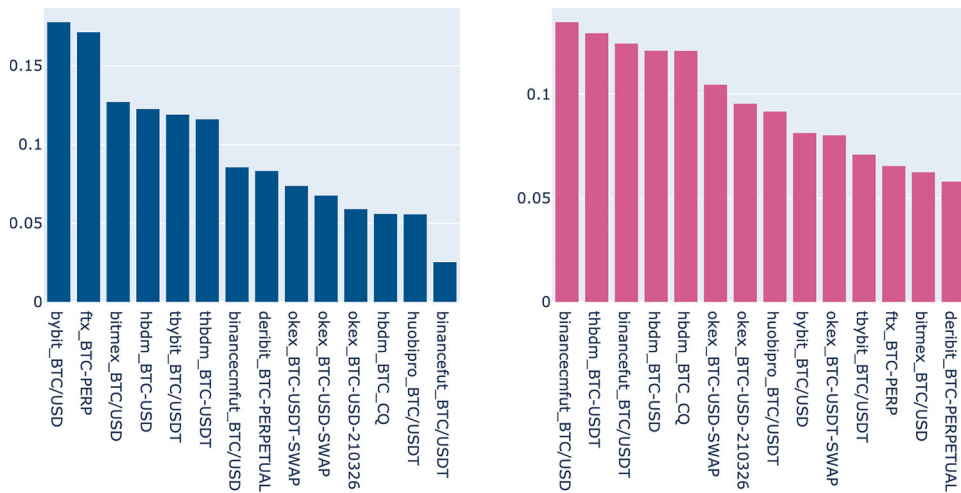


Figure 10. The left bar plot shows row averages of the matrix R , while the right bar plot shows column averages.

One can compare the R^2 values we obtain here with those reported in Cont, Kukanov, and Stoikov (2014) where the authors achieve values of around 60% using order flow imbalance to explain *contemporaneous* returns (i.e., computed over the same time bucket as the features).

Our next set of observations pertains to the hierarchy of markets implied by our results. Figure 10 shows the column and row averages. A large column average shows that the market is easily predicted by others (hence a lagger), while a large row average shows that a market is particularly useful in predicting price action elsewhere. Notably high are the column averages of the Bybit, FTX and Bitmex perptuals. Many of the markets that are especially useful in predicting Bybit and Bitmex are from the exchanges Binance and Huobi, as can be seen in the above heatmap by, for instance, the cells in the Bybit row corresponding to the Binance USDt perpetual or the Huobi perpetual. Indeed, the top five markets by row average are all from Binance and Huobi, providing strong evidence for the central role occupied by these exchanges in the price formation process. Note also that this agrees well with the ranking of markets by volume: the Binance USDt perpetual has the largest average daily volume in February by a sizable margin, as seen in Figure 1.

4.2. Accounting for Trading Realities

So far we have assessed the goodness-of-fit of our models only by an in-sample measure, namely their R^2 values. This measure suffers from two complications. First, the R^2 of a model is an inherently dimensionless quantity, and second, it is not clear how an R^2 value translates to the realities of trading. Out-of-sample measures like RMSE or misclassification error can give a more interpretable quantity having a dimension, but they too suffer from the latter complication mentioned above. It is unclear how, for example, an accuracy score maps to PnL. One can imagine cases where a model has an extremely high accuracy score which maps to a large number of slightly profitable trades, but where a single

misclassification results in a disastrous loss that wipes out any prior profits and more. To address these concerns, we propose an alternative out-of-sample measure of goodness-of-fit which is more in line with trading realities. Our measure is defined as the PnL of a natural trading strategy associated with the model from Equation (15). This PnL value is computed from a synthetic walk-forward on an out-of-sample data set, consisting of the two days following our training period (February 22–27 midnight UTC, 2021); the test period includes 28th February and 1st March 2021.

Mapping a linear model to a trading strategy. We now describe how we map each model from Equation (15) to a trading strategy. The basic idea is very simple: we buy when the prediction is large and positive, and sell when it is large and negative. More precisely, the starting point is the strategy, which generates predictions on out-of-sample observations and places a hypothetical buy order at the top ask price when the model predicts a value greater than a threshold T (> 0). When the prediction is less than $-T$, the strategy places a hypothetical sell order. Clearly, this strategy involves the choice of T . However, there is a relatively canonical choice: we set T as a quantile value of the set of predictions that the model produced on the training period. Let us make this more precise. We denote by M_{ij} the model from Equation (15), and write $\text{preds}_{\text{in-sample}}(M_{ij})$ for the set of in-sample predictions produced by M_{ij} . These are predictions of future returns. Then we define

$$T_{ij} := 95\text{th percentile of the set } \text{preds}_{\text{in-sample}}(M_{ij}). \quad (16)$$

The strategy in its current state still has (at least) two drawbacks which we address with a simple additional constraint. The first problem is that when, for example, we place a hypothetical buy order any time the prediction exceeds T_{ij} , for the time period in which the prediction remains large, we may see many buy orders of which, because of price impact, only the first order is actionable at the best price. The second issue is that the PnL calculation becomes noisier when we have frequent bursts of buy or sell orders. For these reasons, we impose the additional constraint on the strategy that the hypothetical trader's position never exceeds 'one unit', by which we mean that, in our sequence of hypothetical trades, a buy order may only be followed by a sell order (and vice versa).

We summarize the strategy associated with M_{ij} as follows:

- (1) Initialize the trader's position Π to $\Pi = 0$.
- (2) If $\Pi \leq 0$ and the prediction of M_{ij} on an unseen observation exceeds T_{ij} , execute a hypothetical buy order at the top ask price and set $\Pi = 1$.
- (3) If $\Pi \geq 0$ and the prediction of M_{ij} on an unseen observation is less than $-T_{ij}$, execute a hypothetical sell order at the top bid price and set $\Pi = -1$.

Applying this to the out-of-sample period (28th February and 1st March), we obtain an ordered sequence of hypothetical trade prices, from which we subsequently compute various measures of PnL, of the following form:

$$\mathcal{S}_{ij} = (\dots, p_k^{ij, \text{buy}}, p_{k+1}^{ij, \text{sell}}, p_{k+2}^{ij, \text{buy}}, \dots). \quad (17)$$

Here $p_k^{ij, \text{buy}}$ denotes the price of the k th buy trade of the strategy associated with M_{ij} (and similarly for sell trades). All trades take place at the top bid or ask price (for sell and buy orders respectively) but it is important to observe that the top quote often contains only a

fleetingly small volume, quite commonly less than 1000 USD and occasionally just pennies. This implies that the strategy in its current form is often not scalable.

Results without execution fees. The first PnL we calculate from the sequence (17) is the simplest, namely the total PnL over the test period without accounting for the exchange's execution fees. Note, however, that we already implicitly incorporated the spread component of transaction cost since buys occur at the top ask price and sells at the top bid price.

With this caveat, we define our first PnL measure by

$$\text{PnL}_{1,ij} := \sum_{k=1}^{|\mathcal{S}_{ij}|} \left(\frac{p_{k+1}^{ij, \text{sell}}}{p_k^{ij, \text{buy}}} - 1 \right) \cdot 10,000, \quad (18)$$

which can be regarded as the difference in basis points between the average buy price and the average sell price over the sequence \mathcal{S}_{ij} , multiplied by the total number of trades.

The resulting matrix $\text{PnL}_1 = (\text{PnL}_{1,ij})_{i,j=1,\dots,14}$ is visualized as a pair of heatmaps in Figure 11, where the left heatmap is ordered by row sum and the right heatmap by column sum.

Row and column averages are provided in Figure 12. We find strong agreement of the results here with those based on R^2 in the previous subsection. This, reassuringly, provides evidence of the consistency of our approaches. Markets that yield among the largest PnL values are the Bybit and Bitmex perptuals, which were previously identified as some of the most predictable ones. The most useful markets in producing high PnL values on other markets are the two Binance perptuals and the Huobi quarterly futures contract, as well as the Huobi USDt perptual.

The mean and median values of the matrix PnL_1 are both approximately 11,000. Note that this number is the total number of basis points ‘accumulated’ by the strategy. This basically says that, on average, we accumulate 11,000 basis points over the course of the two day test period, or, in other words, we roughly double the initial capital. Of course, this is not realistic at any scale, nor does it account for the impact of execution fees, to which we turn next.

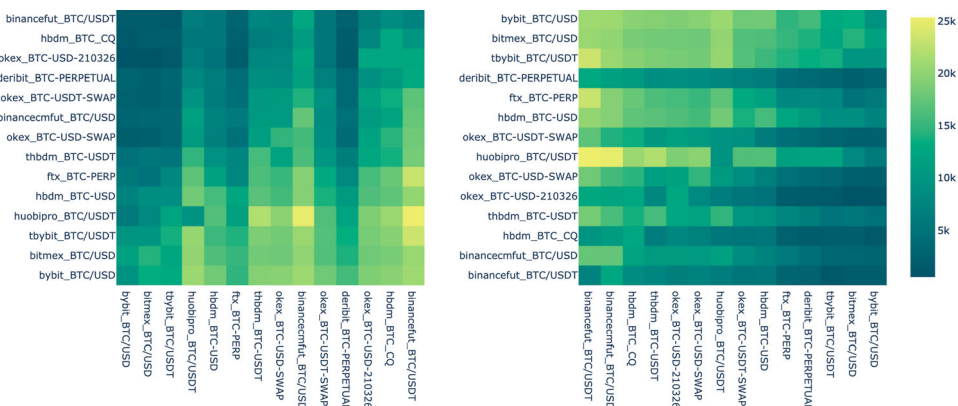


Figure 11. Visualization of the matrix PnL_1 with two orderings, one by row sum and the other by column sum.

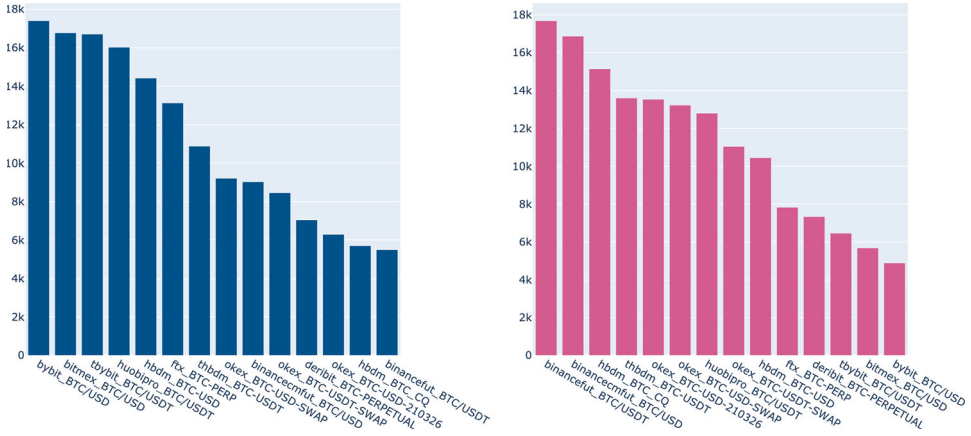


Figure 12. Row and column averages of the matrix PnL_1 .

Results with VIP execution fees. We examine a matrix $\text{PnL}_{3,ij}$ where we correct the PnL value from the preceding passage by the lowest available execution fee on each exchange. For the sake of brevity, we deferred a discussion of the matrix $\text{PnL}_{2,ij}$ where we account for the ordinary fee to Appendix C.

This VIP fee typically requires on the scale of 500 m USD monthly notional volume, and in some cases significant holdings of the exchange's native cryptocurrency in the trader's exchange wallet. Only a small number of wealthy individual traders or significant trading firms qualify for this fee. We define

$$\text{PnL}_{3,ij} := \text{PnL}_{1,ij} - |\mathcal{S}_{ij}| f_i^{\text{VIP}}, \quad (19)$$

where f_i^{VIP} is the lowest possible execution fee on market i . The corresponding matrix $\text{PnL}_3 = (\text{PnL}_{3,ij})_{i,j=1,\dots,14}$ is visualized in Figure 13, along with its row and column averages in Figure 14.

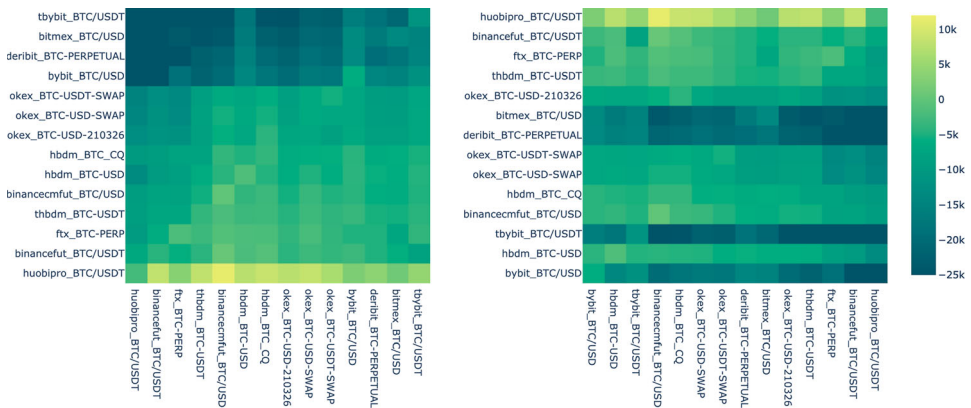


Figure 13. Visualization of the matrix PnL_3 with two orderings, one by row sum and the other by column sum.

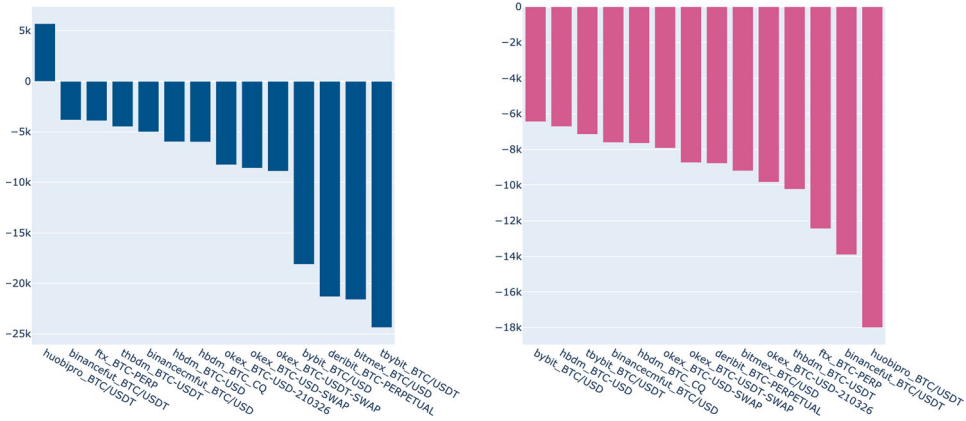


Figure 14. Row and column averages of the matrix PnL_3 .

Surprisingly, the most leading markets are also those where the largest PnL can be achieved. For instance, the Binance USDt perpetual has the second largest row average. Conversely, many of the most lagging markets are those where the lowest PnL is achieved, for example the Bitmex perpetual which has the second lowest row average. The Huobi spot market seems to be an outlier where, even when averaged across markets, a positive PnL can be achieved. It should be noted that amounts at the top of the book amounts are often small on this market, so the capital that might be deployed in our top-of-the-book strategy is likewise small (perhaps a single digit USD amount).

Overall, in the PnL_3 matrix, we see an average loss of 9621 bpts. The highest value is observed for trading on Huobi spot using information from Binance BTC perpetual, where the PnL is 11,426 bpts. The largest loss (of size $-42,435$ bpts) is incurred by the model trading on the Bybit USDt perpetual using features of the Huobi spot market.

The Role of Exchange Fees. We have identified a set of leading markets and a set of lagging markets using an approach based on R^2 values of the models M_{ij} . In our first comparison, of PnL results ignoring execution fees, we found strong correspondence with the R^2 based lead-lag analysis. In line with intuition, we found that the most lagging markets yielded the largest PnL values. Accounting for execution fees, however, we obtained a very different picture. Some of the most leading markets attained the highest PnL values, while some of the most lagging markets attained among the lowest PnL values. How may we explain this phenomenon? We argue below that a market's fee regime naturally predisposes it to being either a leader or a lagger (or something in between). For another study which examines the role of exchange fees in price formation see Malinova and Park (2015). This natural tendency does not, *per se*, imply trading 'alpha' (positive PnL). Only when a market exhibits 'leadingness' in excess of its natural disposition to being a leader, can we extract alpha to produce positive PnLs on other markets. By the same token, a market's 'laggardness' only implies alpha (positive PnL) if the lag effect is greater than what is implied by the market's fee regime.

So how does the fee structure on a market correspond to a natural disposition to being a leader of laggard? Take a pair of markets (1, 2) with differing fee structures and consider how these affect their lead-lag behaviour. Recall that the taker fee is the execution cost in bpts for a market order or a marketable limit order, while the maker rebate is the rebate,

also in bpts, paid to the market maker following a fill for a passive limit order. The maker rebate can also be negative, in which case the market maker incurs a cost for getting a fill. Let us assume that the difference between taker fee and maker rebate is the same on both exchanges. That is, writing fee_i and rebate_i for the taker fee and maker rebate respectively on market $i \in \{1, 2\}$, there is a scalar c such that

$$\text{fee}_i - \text{rebate}_i = c. \quad (20)$$

In practice, one finds a nearly uniform difference

$$\text{fee} - \text{rebate} \approx 5, \quad (21)$$

across exchanges (VIP traders often enjoy a smaller difference, hence paying less in net fees). Note that Equation (20) imposes no restrictions on the values of fee_i or rebate_i themselves. Suppose that $c = 5$ and that the first market has fee structure $(\text{fee}_1, \text{rebate}_1) = (10, 5)$ while the second one uses $(\text{fee}_1, \text{rebate}_1) = (5, 0)$.

An informed market taker (a trader contemplating placing a taker order) should rationally place a buy order on market i if and only if they expect an upward price move of at least fee_i basis points over their anticipated holding period, denoted h . In other words, the trader sends a taker buy order if and only if

$$\mathbb{E}_t \left[\left(\frac{p_{t+h,i}}{p_{t,i}} - 1 \right) \cdot 10,000 \right] > \text{fee}_i, \quad (22)$$

where $p_{t,i}$ is the price on market i at time t and $\mathbb{E}_t[\cdot]$ denotes the expectation conditional on all information up to time t . Since $\text{fee}_1 > \text{fee}_2$ in our example, the trader requires a stronger signal to send a taker order to market 1 and is therefore disincentivized from sending taker orders there. For the market maker, the larger rebate rebate_1 on market 1 implies a disincentive to cancel passive orders, since a larger amount of adverse selection can be compensated for by the maker rebate on market 1 relative to market 2. Taken together, these effects imply a natural tendency of market 1 to be a laggard. In the limiting case $\text{fee}_1, \text{rebate}_1 \rightarrow \infty$, the orderbook would be completely static with no transactions ever occurring, as market makers could tolerate an arbitrary amount of adverse selection, and takers would require an infinitely strong price signal to overcome the taker fee. In conclusion, market 1 is naturally disposed towards lagging behaviour when compared to market 2 (which is consequently a leader in this comparison).

We conclude from this argument that the extent to which a market is a leader or a lagger does not, *per se*, have any implications regarding trading ‘alpha’ that can be achieved for trading on it, nor for using its information to trade on other markets. The reality is more subtle due to potentially different fee regimes (even when the difference between taker fee and maker rebate is uniform across exchanges). We therefore believe that the inherent degree of leadingness or laggardness induced by the fee structure carries no information. Only when a market exhibits lagging behaviour *in excess* of its inherent fee-driven laggardness do we obtain trading ‘alpha’ that can be exploited to generate positive PnL values on the market in question. For instance, counter-intuitively, the Binance USDT perpetual, while generally a strong leading market, is nevertheless sufficiently predictable, owing to its fee structure, to give rise to positive PnL values.

5. Combining Information Cross-Sectionally

Given that the results above indicate a single market does not offer a sufficiently strong signal to produce ‘alpha’ in excess of execution cost (with rare exceptions for traders with large fee discounts), a natural next step is to consider a larger subset of relevant markets to leverage information from other venues for subsequent prediction models. Market makers that consistently remain profitable must necessarily take into account all or most relevant information (at least enough to avoid getting adversely selected regularly enough to become unprofitable), implying that the bulk of liquidity in the orderbook is dictated by actors whose signal is based on a sufficiently large set of significant markets. Therefore, if our goal is to reliably predict order flow and price change on sub-second time scales, our model should span a similar range of markets as the ones considered by liquidity providers. A natural idea is to fit linear regression models (as before with OLS) using the full set of features from all markets considered in this study. This is the starting point and a baseline for comparison of our analysis in this section. One concern with this approach is that the model might be overfitted due to a significant number of irrelevant features and/or a strong degree of collinearity between features. We therefore compare the baseline approach with the following two alternative approaches.

The first alternative approach is to deal with the large number of features by using L^1 -regularization in the regression model. That is, instead of fitting a simple OLS model on the full feature set, we will train a LASSO regression model which ends up selecting a subset of the features.

The second alternative approach involves linearly combining features across the markets to form a set of ‘meta features’ which are subsequently leveraged to generate parsimonious models for price prediction on each market. The linear combinations are formed using coefficients that are proportional to predictive power, as measured by R^2 .

5.1. Baseline Approach

Model specification. Our baseline model consists of a linear regression trained on the full feature set. Each market has 5 features and there are 14 markets, resulting in a total of 70 covariates. We use OLS regression to fit the following linear models for all $i \in \{1, \dots, 14\}$ and $\delta \in \{500 \text{ ms}, 1000 \text{ ms}\}$

$$\text{fret}_t^{\delta,i} = \mu_i + \sum_{j=1}^{14} (\beta_{ij,1} \text{IMB}_t^{a,j} + \beta_{ij,2} \text{IMB}_t^{b,j} + \beta_{ij,3} \text{TFL}_t^j + \beta_{ij,4} \text{PRET}_t^j + \beta_{ij,5} \text{DIV}_t^{ij}) + \epsilon_{i,t}. \quad (23)$$

As usual, to keep notation simple, it is implicit that we use the previously found optimal time horizon for each feature, and that we use the transformed features rather than the raw versions, whenever an improvement was found for the feature. We denote the model corresponding to Equation (23) by $M_{i,\delta}^{\text{baseline}}$.

Results. We now examine the coefficients of determination of the models $M_{i,\delta}^{\text{baseline}}$ for $i = 1, \dots, 14$ and $\delta \in \{500 \text{ ms}, 1000 \text{ ms}\}$. We compute this number both in-sample and out-of-sample. In other words, we report the portion of total variation in future returns explained by each model on the training period, as well as on the two-day test period

Table 4. R^2 values for baseline models.

| | R^2 | | R^2_{OOS} | |
|-----------------------|--------|---------|--------------------|---------|
| | 500 ms | 1000 ms | 500 ms | 1000 ms |
| ftx_BTC-PERP | 0.364 | 0.342 | 0.273 | 0.264 |
| bybit_BTC/USD | 0.294 | 0.331 | 0.289 | 0.32 |
| thbdtm_BTC-USDT | 0.273 | 0.231 | 0.25 | 0.215 |
| hbdm_BTC-USD | 0.229 | 0.262 | 0.214 | 0.243 |
| huobipro_BTC/USDT | 0.219 | 0.136 | 0.191 | 0.126 |
| bitmex_BTC/USD | 0.217 | 0.292 | 0.193 | 0.27 |
| tbybit_BTC/USDT | 0.214 | 0.279 | 0.219 | 0.28 |
| binancecmfut_BTC/USD | 0.209 | 0.193 | 0.187 | 0.177 |
| deribit_BTC-PERPETUAL | 0.187 | 0.212 | 0.178 | 0.214 |
| hbdm_BTC_CQ | 0.186 | 0.152 | 0.171 | 0.144 |
| okex_BTC-USDT-SWAP | 0.173 | 0.196 | 0.156 | 0.184 |
| okex_BTC-USD-210326 | 0.17 | 0.164 | 0.158 | 0.155 |
| okex_BTC-USD-SWAP | 0.164 | 0.18 | 0.151 | 0.167 |
| binancefut_BTC/USDT | 0.106 | 0.071 | 0.094 | 0.069 |

following the training period. The results can be found in Table 4. Significance values and p -values for all models are reported in Tables 16 and 17 in the Appendix.

We observe that the difference between the two time horizons 500 ms and 1000 ms is generally rather small. As one would expect, the 500 ms time horizon is somewhat more easily predictable. On average, the in-sample R^2 values for this horizon are 4.6% larger than their 1000 ms counterparts. When we perform the same computation on the out-of-sample values, we find that the R^2 on the 500 ms horizon is, on average, greater than that the 1000 ms R^2 by 1.4%. There are, however, some noteworthy outliers. Notably larger values on the 500 ms horizon are found on the Binance USDT perpetual and the Huobi spot market, while the Bitmex and Bybit USDT perpetuals exhibit considerably larger R^2 values for the 1000 ms horizons. What are some possible explanations of these phenomena?

When a market's 1000 ms R^2 is much greater than the 500 ms R^2 it means that the features at time t contain information that is not yet reflected in the price at time $t + 500$ ms but is priced in at $t + 1000$ ms. This suggests the possibility that cross-sectional arbitrages can survive for more than 500 ms between a strong leading market and a strong lagging market. Note that this agrees well with our previous identification of Bitmex and Bybit as lagging exchanges, and Binance and Huobi as leading exchanges.

Conversely, when the 500 ms R^2 is much larger than its 1000 ms counterpart on a given market, this indicates that this market is prone to the arrival of new information in the time window $t + 500$ ms to $t + 1000$ ms which is not yet encapsulated in the feature values at time t . We would predict this to be the case for strongly leading markets, as indeed appears to be the case when we compare the R^2 values with our previous findings on the leader-lagger network.

Our next observation pertains to the comparison between in-sample and out-of-sample values. On average, we find that on the 500 ms time horizon, the out-of-sample R^2 is around 91% of the in-sample one. The largest discrepancy, both in relative and absolute terms, can be seen for the FTX perpetual where the out-of-sample R^2 is 25% lower than the in-sample R^2 . Overall, the small to moderate loss in accuracy as we pass from the in-sample to the out-of-sample values ameliorates concerns of overfitting to an extent.

Next, we note that by using a market's actual 'leadingness' or 'laggarness', as measured by R^2 , in conjunction with its fee structure, one can form a set of expectations about what

PnL values could be obtained. See Table 1 for an overview of each market's lowest possible taker fee (usually accessible only to highly active traders). As we reasoned in the previous section, we would a-priori expect a market with a low taker fee to be a leading market, which should express itself in a low R^2 value. If we find that a market which 'should' be a leader based on its fee structure is in fact a lagger, we would predict a high PnL value for this market. When we compare the VIP fee from Table 1 with Table 4 we observe that, remarkably, the FTX perpetual has both the highest R^2 (indicating that its returns can be anticipated well) and the lowest taker fee. On this basis, one would predict a large PnL value for those VIP traders with access to this fee.

The Bybit BTC perpetual is among the most lagging markets, although this is in agreement with the fact that it has the highest taker fee, so we would not necessarily foresee a high PnL value for this market, unless its lagging nature is so significant that it outweighs the large fee.

On the other end of the spectrum, the most leading market, the Binance USDT perpetual, also has one of the lowest taker fees, so it is not straightforward to predict what the PnL might be on this market. Whether a trader subject to the lowest fee is capable of producing a positive PnL boils down to the question of whether the general leading nature of this market can be overcome by its low taker fee.

In Table 5, we display the PnL values produced by the trading strategies corresponding to the models $M_{i,500\text{ ms}}^{\text{baseline}}$ for $i = 1, \dots, 14$. The values are computed by the same procedure we introduced in Section 4.2. That is, we map a model to a natural trading strategy and calculate the PnL of this strategy in a synthetic walk-forward over the out-of-sample period. As before, we compute the PnL in three different ways: (1) ignoring execution fees, (2) adjusting for the default execution fee, (3) adjusting for the lowest possible ('VIP') execution fee. The PnL formulae are precise analogues of the ones from Equations (18), (36), and (19). Note that, as before, the PnL therefore tells us the total number of accumulated basis points over the 2-day test period.

In line with the expectations, we formed on the basis of predictability and fee regime, we find that FTX gives the largest PnL₃ value. Somewhat surprisingly, even the most leading market, the Binance USDT perpetual, exhibits a positive PnL₃ value. This can be interpreted as a suggestion that the extremely low taker fee is sufficient to 'overpower' the

Table 5. PnLs of baseline models.

| | PnL ₁ | PnL ₂ | PnL ₃ |
|-----------------------|------------------|------------------|------------------|
| ftx_BTC-PERP | 19792.4 | -27527.6 | 9652.4 |
| huobipro_BTC/USDT | 28844.0 | -28251.0 | 5645.4 |
| binancecmfut_BTC/USD | 16519.1 | -22010.9 | 2648.3 |
| binancefut_BTC/USDT | 12107.6 | -17724.4 | 696.8 |
| thbdrm_BTC-USDT | 20784.9 | -10703.1 | -469.5 |
| hbdm_BTC-USD | 17703.1 | -8186.9 | -1455.5 |
| hbdm_BTC_CQ | 16432.0 | -23400.0 | -3484.0 |
| okex_BTC-USD-SWAP | 18041.0 | -19319.0 | -4375.0 |
| okex_BTC-USD-210326 | 17436.4 | -27273.6 | -4918.6 |
| okex_BTC-USDT-SWAP | 15012.4 | -20097.6 | -6053.6 |
| bybit_BTC/USD | 19811.2 | -11508.8 | -11508.8 |
| tbybit_BTC/USDT | 17036.4 | -15828.6 | -15828.6 |
| bitmex_BTC/USD | 17804.5 | -16140.5 | -16140.5 |
| deribit_BTC-PERPETUAL | 10401.1 | -19738.9 | -19738.9 |

market's relative lack in predictability. The lowest PnL is found on Deribit which has one of the highest taker fee and middling R^2 values. Similarly, Bitmex has slightly larger R^2 values but also a larger taker fee, so it is not surprising to see it having the second smallest PnL₃ value.

Note that all PnL₂ numbers are negative. This hints at 'the rich get richer' phenomenon in the crypto markets, whereby traders with access to the lowest taker fee have a much greater ability to turn profits than novice traders. Since a trader's fee regime is determined by trade volume, and in some cases, holdings in the exchange's native cryptocurrency, the privilege of being in the best fee tier clearly correlates highly with the trader's net worth. On the other hand, a more novice trader wishing to ascend to the best fee tier to boost profits might have to 'burn through' some losses in lower fee tiers until they have accumulated enough volume to be granted a better fee.

5.2. LASSO Regression Approach

The models we trained in the preceding subsection used a total of 70 features from 14 different markets. Many of these features exhibit large cross-correlations. One would certainly expect, for instance, past returns between different markets to be highly correlated for simple no-arbitrage reasons. In Table 18 in the Appendix, we display the cross-correlations amongst the 500 ms past returns features across markets.

In this section, we pursue an alternative methodology for fitting the parameters of Equation (23) based on *LASSO regression*. The difference between an OLS fit and a LASSO fit resides in an additional regularization term introduced in the objective function of the latter. Specifically, OLS regression minimizes the standard least-squares objective function, while LASSO regression leads to the problem

$$\min_{w \in \mathbb{R}^p} (\|y - \mathbf{X}w\|_2^2 + \lambda \|w\|_1), \quad (24)$$

for a regularized objective function, where $\mathbf{X} \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}_{>0}$ is the regularization parameter. The additional ℓ_1 regularization term has the effect of zeroing a subset of the parameters. This subset grows in size as the regularization parameter λ is increased. For a full discussion on LASSO regression, the reader is referred to Hastie, Tibshirani, and Friedman (2001).

We fit the coefficients in Equation (23) using a LASSO regression with regularization parameters from the set $\{0.001 \cdot 2^k \mid k = 0, \dots, 8\}$. The lower bound $\lambda = 0.01$ of this set was chosen since it was found to be the smallest (round) number which, when used as the regularization parameter in the LASSO regression, results in a positive number of zero coefficients for each market's model. Similarly, the upper bound $\lambda = 0.256 = 0.01 \cdot 2^8$ was selected since it was the largest doubling of the lower bound $\lambda = 0.01$ for which a positive number of nonzero coefficients persist.

For each market and each of the nine regularization parameters, we then examine the number of nonzero coefficients, their in-sample and out-of-sample explanatory power over future returns, and the PnL values of their naturally associated trading strategies. Furthermore, we inspect exactly which coefficients tend to survive as λ is increased. Comparisons of the results in this subsection with the results of baseline approach are deferred until we

Table 6. Number of surviving coefficients, as we vary the λ regularization coefficient.

| | 0.001 | 0.002 | 0.004 | 0.008 | 0.016 | 0.032 | 0.064 | 0.128 | 0.256 |
|-----------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| deribit_BTC-PERPETUAL | 67 | 64 | 61 | 53 | 44 | 36 | 25 | 9 | 4 |
| binancefut_BTC/USDT | 67 | 65 | 62 | 54 | 43 | 26 | 16 | 6 | 1 |
| ftx_BTC-PERP | 67 | 64 | 57 | 50 | 45 | 40 | 26 | 14 | 6 |
| huobipro_BTC/USDT | 67 | 67 | 61 | 51 | 41 | 28 | 21 | 9 | 2 |
| binancecmfut_BTC/USD | 67 | 62 | 59 | 49 | 35 | 27 | 17 | 9 | 2 |
| hbdm_BTC-USD | 66 | 64 | 56 | 43 | 34 | 27 | 15 | 6 | 6 |
| okex_BTC-USD-SWAP | 66 | 62 | 53 | 43 | 38 | 33 | 20 | 10 | 5 |
| okex_BTC-USDT-SWAP | 66 | 64 | 58 | 47 | 41 | 35 | 21 | 9 | 4 |
| thbdm_BTC-USDT | 66 | 61 | 56 | 51 | 33 | 20 | 14 | 10 | 3 |
| hbdm_BTC_CQ | 65 | 63 | 58 | 52 | 32 | 26 | 22 | 6 | 3 |
| okex_BTC-USD-210326 | 65 | 61 | 58 | 48 | 37 | 30 | 24 | 9 | 2 |
| bitmex_BTC/USD | 62 | 60 | 54 | 48 | 38 | 28 | 22 | 15 | 4 |
| bybit_BTC/USD | 62 | 60 | 55 | 33 | 28 | 23 | 17 | 9 | 5 |
| tbybit_BTC/USDT | 62 | 61 | 51 | 43 | 39 | 26 | 19 | 10 | 5 |

introduce our third alternative methodology for training powerful linear models to explain short-term future returns.

Model Selection Results. First, we inspect the number of nonzero coefficients of each of the LASSO fits, as the regularization parameter λ is increased. The results can be seen in Table 6. With $\lambda = 0.001$, the majority of the 70 coefficients are found to be nonzero. In line with expectation, as λ is increased the number of surviving coefficients decreases. With our largest choice of $\lambda = 0.256$, we typically do not retain more than a handful of features in the model.

Our next set of observations pertains to the explanatory power of each of the LASSO models. It would be reasonable to hypothesize that explanatory power (particularly out-of-sample) might increase as we discard superfluous features in a LASSO model. However, this turns out to be the case only very rarely, as evidenced in Table 7 where we show out-of-sample R^2 values for each model. In fact, R^2 is generally decreasing in the regularization parameter λ . That said, the decrease is quite slow, especially for $\lambda \leq 0.032$. For example, the LASSO models with $\lambda = 0.016$ retain on average 97.2% of the out-of-sample R^2 compared with the LASSO models with $\lambda = 0.001$. When we compare the $\lambda = 0.001$ models with the

Table 7. Out-of-sample R^2 values for LASSO models using 500 ms future return horizon, as we vary the λ regularization coefficient.

| | 0.001 | 0.002 | 0.004 | 0.008 | 0.016 | 0.032 | 0.064 | 0.128 | 0.256 |
|-----------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| bybit_BTC/USD | 0.289 | 0.289 | 0.287 | 0.283 | 0.281 | 0.279 | 0.272 | 0.263 | 0.236 |
| ftx_BTC-PERP | 0.273 | 0.272 | 0.272 | 0.271 | 0.269 | 0.265 | 0.25 | 0.211 | 0.193 |
| thbdm_BTC-USDT | 0.25 | 0.25 | 0.249 | 0.247 | 0.242 | 0.234 | 0.223 | 0.211 | 0.149 |
| tbybit_BTC/USDT | 0.219 | 0.219 | 0.218 | 0.217 | 0.216 | 0.212 | 0.202 | 0.194 | 0.167 |
| hbdm_BTC-USD | 0.213 | 0.213 | 0.212 | 0.21 | 0.206 | 0.198 | 0.193 | 0.188 | 0.176 |
| bitmex_BTC/USD | 0.193 | 0.193 | 0.192 | 0.192 | 0.191 | 0.188 | 0.181 | 0.171 | 0.141 |
| huobipro_BTC/USDT | 0.191 | 0.19 | 0.19 | 0.188 | 0.183 | 0.176 | 0.159 | 0.095 | 0.073 |
| binancecmfut_BTC/USD | 0.187 | 0.187 | 0.186 | 0.184 | 0.181 | 0.173 | 0.153 | 0.136 | 0.101 |
| deribit_BTC-PERPETUAL | 0.178 | 0.178 | 0.178 | 0.177 | 0.176 | 0.173 | 0.167 | 0.143 | 0.122 |
| hbdm_BTC_CQ | 0.171 | 0.171 | 0.17 | 0.168 | 0.164 | 0.16 | 0.146 | 0.107 | 0.082 |
| okex_BTC-USD-210326 | 0.158 | 0.158 | 0.158 | 0.156 | 0.154 | 0.151 | 0.14 | 0.103 | 0.083 |
| okex_BTC-USDT-SWAP | 0.156 | 0.156 | 0.156 | 0.155 | 0.153 | 0.15 | 0.138 | 0.115 | 0.096 |
| okex_BTC-USD-SWAP | 0.151 | 0.151 | 0.15 | 0.149 | 0.148 | 0.144 | 0.131 | 0.111 | 0.09 |
| binancefut_BTC/USDT | 0.094 | 0.094 | 0.093 | 0.092 | 0.087 | 0.076 | 0.051 | 0.038 | 0.026 |

Table 8. PnL₃ values of LASSO models as regularization parameter varies.

| | 0.001 | 0.002 | 0.004 | 0.008 | 0.016 | 0.032 | 0.064 | 0.128 | 0.256 |
|-----------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| ftx_BTC-PERP | 9576 | 9439 | 9371 | 9253 | 9444 | 9407 | 9370 | 8847 | 6987 |
| huobipro_BTC/USDT | 5738 | 5679 | 5654 | 5637 | 5610 | 6196 | 7264 | 8908 | 3598 |
| binancecmfut_BTC/USD | 2576 | 2497 | 2229 | 1802 | 1991 | 1855 | 1511 | -179 | -2188 |
| binancefut_BTC/USDT | 766 | 824 | 844 | 1274 | 1627 | 2335 | 2095 | 1340 | 1297 |
| thbdm_BTC-USDT | -502 | -516 | -732 | -1013 | -1246 | -1756 | -2527 | -3722 | -8110 |
| hbdm_BTC-USDT | -1517 | -1572 | -1648 | -1551 | -1816 | -2104 | -2483 | -3197 | -4461 |
| hbdm_BTC_CQ | -3579 | -3669 | -3711 | -4104 | -4232 | -4061 | -4129 | -5479 | -9427 |
| okex_BTC-USDT-SWAP | -4380 | -4376 | -4603 | -4979 | -5314 | -5929 | -6544 | -7829 | -12021 |
| okex_BTC-USDT-210326 | -4951 | -5006 | -5327 | -5784 | -5966 | -6776 | -7439 | -8122 | -11823 |
| okex_BTC-USDT-SWAP | -6050 | -6105 | -6418 | -6770 | -6951 | -7605 | -7752 | -8334 | -11518 |
| bybit_BTC/USD | -11631 | -11440 | -11504 | -11808 | -11895 | -12510 | -13771 | -15709 | -19746 |
| tbybit_BTC/USDT | -15855 | -15917 | -16264 | -16521 | -16601 | -17025 | -17592 | -19935 | -25232 |
| bitmex_BTC/USD | -16209 | -16091 | -16341 | -16386 | -16311 | -15813 | -16328 | -18577 | -23969 |
| deribit_BTC-PERPETUAL | -19666 | -19621 | -19784 | -20189 | -20368 | -20438 | -20441 | -20387 | -25174 |

$\lambda = 0.032$ ones, we still find an average decrease in out-of-sample R^2 of only 5.9%. Since the number of features retained in the latter models ranges between 20 and 40, around half of the original 70 features suffice to retain the vast majority of explanatory power. When we increase the regularization parameter to $\lambda = 0.064$, 0.128 and 0.256, we lose on average 13.3%, 25.97% and 39% of explanatory power relative to the $\lambda = 0.001$ models, respectively. The in-sample analogue of Table 7 can be found in the Appendix in Table 19. The conclusions are identical in both cases. Both of the aforementioned tables were created using the 500 ms future returns horizon. For the 1000 ms versions, we refer to Tables 20 and 21 in the Appendix. Here again, the results are very similar.

Next, we examine the PnL attained by the LASSO models. In Tables 22 and 23 of the Appendix, we show the PnL₁ and PnL₂ values (respectively) for each model. Most interesting are the PnL₃ values which we display in Table 8. We note that in most cases the maximal PnL is achieved with either $\lambda = 0.001$ or $\lambda = 0.002$, although the difference between the maximum PnL and any other PnL with $\lambda \leq 0.032$ is marginal. This is in agreement with our prior observation that explanatory power is almost identical across all models with $\lambda \leq 0.032$. The only markets where the maximal PnL is not attained with $\lambda \in \{0.001, 0.002\}$ are the Huobi spot market and the Binance USDT perpetual. The former achieves its maximum with $\lambda = 0.128$ and nine features while the latter does so with $\lambda = 0.032$ and 26 features.

We have seen that the PnL and explanatory power of a market's LASSO models are almost unchanged for any regularization parameters $\lambda \leq 0.032$. It is interesting to examine which of the 70 original features are retained in these models. In Figure 15(a) we depict the number of markets whose $\lambda = 0.032$ model uses the features listed on the x -axis. For ease of legibility, we include in this plot only those features which appear in the models of at least six markets.

It is striking in Figure 15(b) that the features used by any of the $\lambda = 0.256$ models are *exclusively* from a market on Binance or Huobi. This underlines the outsized role that these two exchanges play in price discovery. Let us next inspect Figure 15(a). Here we observe the same dominance of Binance and Huobi with notably large counts also occurring from Okex markets. We note that the most common type of feature is a (transformed) trade flow imbalance feature. This suggests signed volume is the most powerful of our indicators. Of the orderbook imbalance features, the one corresponding to the Huobi BTC perpetual is the

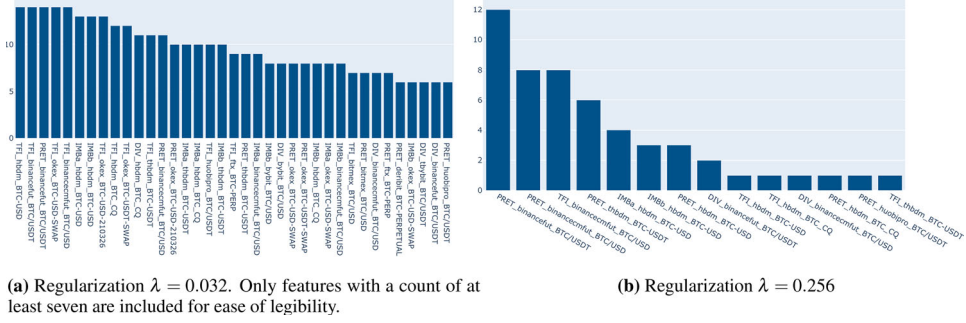


Figure 15. Count of nonzero features. (a) Regularization $\lambda = 0.032$. Only features with a count of at least seven are included for ease of legibility. (b) Regularization $\lambda = 0.256$.

most commonly selected. The most frequent DIV feature comes from the Huobi quarterly futures contract, which indicates that the price difference with this market is particularly useful to predict price action.

5.3. Meta-Features Approach

In the previous section, we found that a much smaller subset, retaining approximately 30 of the original 70 features used in the baseline model, is sufficient to produce models whose explanatory power and PnL values are only marginally lower than those of the baseline models. In this section, we pursue an alternative methodology for drastically reducing dimensionality of our linear models. The procedure involves an additional feature processing step, which cuts down the number of features from 70 to five. The resulting models turn out to achieve a greater average PnL than the baseline models, even though average explanatory power is diminished.

We observed in Table 18 that past returns features exhibit large cross-correlations across exchanges. This is of course not surprising, since the markets we are considering are either bitcoin spot markets themselves or derivative contracts whose underlying is an index composed of bitcoin spot markets. Motivated by this consideration, we devise a methodology by which we form a single ‘meta’ past returns feature from the set of all individual past returns features. That is, we will define for each market $i \in \{1, \dots, 14\}$ a meta-feature mPRET_i using

$$\text{mPRET}_t^i := \sum_{j=1}^{14} \alpha_{ij}^{\text{PRET}} \text{PRET}_t^j, \quad (25)$$

where the coefficient $\alpha_{ij}^{\text{PRET}}$ is chosen in proportion to the predictive power that the past returns feature of market j has over the future returns of market i . Other meta-features (orderbook imbalances, trade flow imbalance, mean divergence) are computed analogously. More specifically, we define a set of so-called *meta-features* for each market $i \in \{1, \dots, 14\}$

$$\begin{aligned} \text{mIMB}_t^{a,i} &:= \sum_{j=1}^{14} \alpha_{ij}^{\text{IMB}^a} \text{IMB}_t^{a,j}, & \text{mIMB}_t^{b,i} &:= \sum_{j=1}^{14} \alpha_{ij}^{\text{IMB}^b} \text{IMB}_t^{b,j}, \\ \text{mTFI}_t^i &:= \sum_{j=1}^{14} \alpha_{ij}^{\text{TFI}} \text{TFI}_t^j, & \text{mDIV}_t^i &:= \sum_{j=1}^{14} \alpha_{ij}^{\text{DIV}} \text{DIV}_t^{ij}, \end{aligned}$$

where the coefficients are determined by the procedure described below. We will demonstrate this process on the example of the past returns feature. Let us fix $i \in \{1, \dots, 14\}$ and $\delta \in \{500 \text{ ms}, 1000 \text{ ms}\}$. The steps are then as follows:

- (1) For every $j \in \{1, \dots, 14\}$, we train the univariate models $\text{fret}_t^{\delta,i} = \mu_{ij} + \beta_{ij}\text{PRET}_t^j + \epsilon_{ij,t}$ using an OLS regression. We denote its coefficient of determination by R_{ij}^2 .
- (2) We set $\alpha_{ij}^{\text{mPRET}} := \frac{R_{ij}^2}{\max_j R_{ij}^2}$.

The procedure is completely analogous for other features. Our normalization procedure involves division by a maximum value. Note that there are other possible weight normalizations, such as division by $\sum_j R_{ij}^2$; we leave it for future work to compare these approaches.

Now that we have defined the set of meta features, we can define the following linear models for every market $i \in \{1, \dots, 14\}$ and the two future returns time horizons $\delta \in \{500 \text{ ms}, 1000 \text{ ms}\}$:

$$\text{fret}_t^{\delta,i} = \mu_i + \beta_{i,1}\text{mIMB}_t^{a,j} + \beta_{i,2}\text{mIMB}_t^{b,j} + \beta_{i,3}\text{mTFI}_t^i + \beta_{i,4}\text{mPRET}_t^i + \beta_{i,5}\text{mDIV}_t^i + \epsilon_{i,t} \quad (26)$$

We fit these models using an OLS regression and analyse their performance along the same axes we have previously considered (explanatory power and PnL).

Results. One main benefit of the meta feature models is their simplicity. Compared to the baseline models, we have cut down the number of covariates from 70 to five. In the appendix in Table 24, we report t -statistics and p -values for each of the coefficients in all of the 14 models we trained. We find strong statistical significance for all meta features in each of the models. All p -values are close to zero. The smallest absolute t -statistic corresponds to the meta past returns feature for the model predicting future returns on the Okex quarterly futures contract. It is perhaps not too surprising that we would see a comparably small t -statistic in this case, since the futures premium can fluctuate somewhat freely in the sense that it is not linked by any strong arbitrage bounds to the price of its underlying. This argument applies more strongly the more distant the expiration of the futures contract is (note that in our case the expiration date lays approximately one month in the future).

The coefficients of determination for the meta models defined by Equation (26) are given in Table 9, where we show the in-sample and out-of-sample R^2 values for the two time horizons under consideration. Here we find that a similar set of observations holds true as in the baseline and LASSO models. First, the average difference between the 500 ms and 1000 ms R^2 values tends to be quite small. For example, in case of the in-sample R^2 values we observe that the 500 ms version is on average 7% larger than the 1000 ms one, in line with the expectation that shorter time horizons would be more easily predicted. Some outliers are the lagging markets Bybit USDT perpetual and Bitmex perpetual, where passing from the 500 ms to the 1000 ms future returns horizon yields an increased R^2 of 32.6% and 37%, respectively. On the other hand, markets that see a notably high decrease in accuracy in passing from the 500 ms to the 1000 ms horizon are leading markets such as the Binance USDT perpetual or the Huobi spot market, where we see decreases of 39.7% and 40.5%, respectively. A similar phenomenon was noted for the baseline and LASSO models.

Table 9. R^2 values for the meta models.

| | R^2 | | R^2_{oos} | |
|-----------------------|--------|---------|-------------|---------|
| | 500 ms | 1000 ms | 500 ms | 1000 ms |
| ftx_BTC-PERP | 0.306 | 0.288 | 0.213 | 0.212 |
| bybit_BTC/USD | 0.262 | 0.3 | 0.261 | 0.294 |
| thbdtm_BTC-USDT | 0.215 | 0.18 | 0.202 | 0.173 |
| hbdm_BTC-USD | 0.201 | 0.229 | 0.19 | 0.217 |
| bitmex_BTC/USD | 0.192 | 0.263 | 0.174 | 0.249 |
| tbybit_BTC/USDT | 0.19 | 0.252 | 0.198 | 0.26 |
| binancecmfut_BTC/USD | 0.156 | 0.139 | 0.137 | 0.13 |
| deribit_BTC-PERPETUAL | 0.141 | 0.166 | 0.152 | 0.186 |
| okex_BTC-USDT-SWAP | 0.13 | 0.144 | 0.117 | 0.139 |
| huobipro_BTC/USDT | 0.126 | 0.075 | 0.106 | 0.069 |
| okex_BTC-USD-SWAP | 0.122 | 0.131 | 0.112 | 0.124 |
| hbdm_BTC_CQ | 0.116 | 0.091 | 0.104 | 0.09 |
| okex_BTC-USD-210326 | 0.112 | 0.105 | 0.105 | 0.102 |
| binancefut_BTC/USDT | 0.068 | 0.041 | 0.055 | 0.041 |

Table 10. PnLs of meta models.

| | PnL ₁ | PnL ₂ | PnL ₃ |
|-----------------------|------------------|------------------|------------------|
| huobipro_BTC/USDT | 21220.4 | −6377.1 | 10007.1 |
| ftx_BTC-PERP | 17140.6 | −19357.4 | 9319.6 |
| binancefut_BTC/USDT | 10790.5 | −12097.5 | 2035.8 |
| binancecmfut_BTC/USD | 10085.4 | −13284.6 | 1672.2 |
| thbdtm_BTC-USDT | 12393.6 | −6910.4 | −636.6 |
| hbdm_BTC_CQ | 8051.1 | −10916.9 | −1432.9 |
| hbdm_BTC-USD | 15057.7 | −8112.3 | −2088.1 |
| okex_BTC-USD-210326 | 8739.4 | −15220.6 | −3240.6 |
| okex_BTC-USD-SWAP | 10398.4 | −13091.6 | −3695.6 |
| okex_BTC-USDT-SWAP | 9803.9 | −13946.1 | −4446.1 |
| bybit_BTC/USD | 19613.6 | −14451.4 | −14451.4 |
| deribit_BTC-PERPETUAL | 9116.7 | −14623.3 | −14623.3 |
| bitmex_BTC/USD | 17561.6 | −17433.4 | −17433.4 |
| tbybit_BTC/USDT | 17659.5 | −18580.5 | −18580.5 |

Second, the in-sample R^2 is not significantly larger than its out-of-sample counterpart, suggesting the absence of significant overfitting. We find average decreases of 10.7% for the 500 ms horizon and 4.9% for the 1000 ms horizon.

In Table 10, we report PnL values for the meta models (26). As before, we compute three separate values: PnL₁ which is the total number of basis points accumulated over the test period by the strategy associated with the respective meta model, PnL₂ where we adjust PnL₁ by the default fee, and PnL₃ where we adjust PnL₁ by the VIP fee.

As before with the other models, we find particularly large PnL₃ values for FTX where the fee structure implies that it ‘should be’ a leader, but its large R^2 value points at lagging behaviour. We also note that the Binance USDT perpetual yields a positive PnL₃ value despite the fact that it has (by quite a margin) the lowest predictability score (as measured by R^2). An interpretation of this result could be that the extremely low VIP fee on this market is sufficiently small to overcome its general leadingness behaviour.

The Huobi spot market has one of the lowest VIP taker fees and its corresponding meta model has a middling R^2 value. Based on these observations, we would predict a relatively large PnL₃ value although we would not anticipate it having the highest PnL₃, as it turns out to be the case. As mentioned before, this might be partially explained by the fact that

this market tends to have extremely low top of the book liquidity, implying that the strategy whose PnL we are computing is not immediately scalable to the deployment of substantial amounts of trading capital. The additional cost of slippage could become an important consideration at that point.

Surprisingly, the Binance BTC-margined perpetual, which has a similarly low taker fee as the USDT-margined one (see Table 1), exhibits a lower PnL_3 when we compare the two (a relative difference of 18%), despite its considerably higher R^2 values. At the other end of the spectrum, the four markets whose PnL_3 is lowest are the precisely ones with the largest taker fees. This is a similar finding to what we noted for other models: the greater predictability is not sufficient to overcome the larger fee.

5.4. Comparisons of Models

We have described three ways to fit linear models of anticipated price changes on each market: baseline models using all 70 features; ℓ_1 regularizations of the baseline model with nine different regularization parameters to obtain more parsimonious models with fewer non-zero coefficients; and a methodology to reduce the number of features down to only five using an additional feature-processing step. We now compare the respective explanatory powers and PnLs of each of the trained models.

R^2 values of our models compare with one another. Figure 16 shows the out-of-sample R^2 values of each model. The in-sample and 1000 ms versions of the same quantities are deferred to Figure 30, Figures 31 and 32 of the Appendix since all of the R^2 -based plots strongly resemble one another. In examining Figure 16, we first note that the maximal R^2 is attained by the baseline model in each case. LASSO models with small regularization parameter λ perform similarly. For example, the LASSO models with $\lambda = 0.016$ show an average drop of only 2.8% relative to the baseline model and the $\lambda = 0.032$ models show an average decrease of 5.9%. The average number of nonzero coefficients retained is 38 for $\lambda = 0.016$ and 29 for $\lambda = 0.032$. As noted earlier, this suggests that of the 70 original

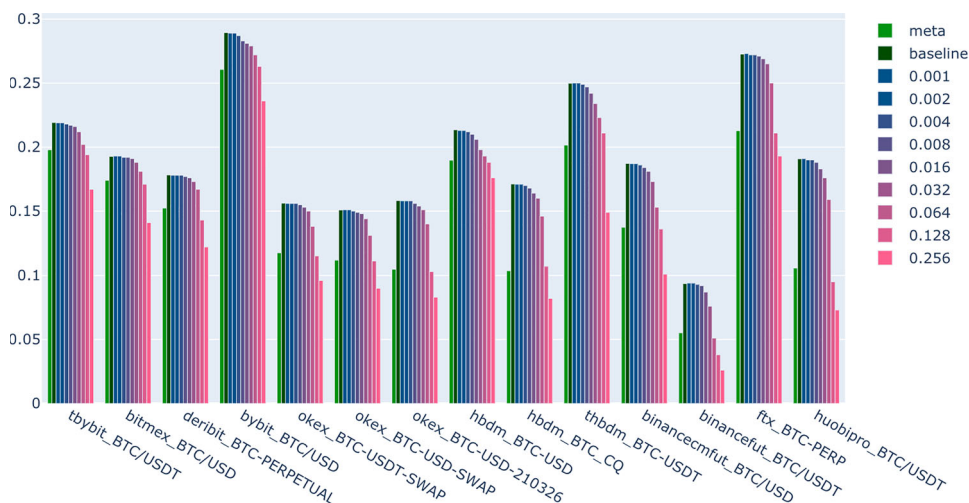


Figure 16. 500 ms out-of-sample R^2 values.

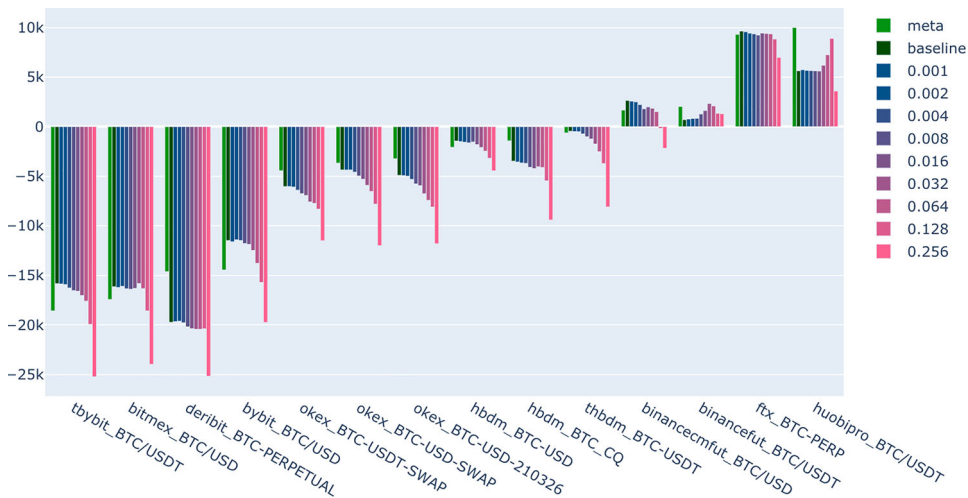


Figure 17. PnL_3 values for each model.

features used by the baseline model, around 30–40 contribute very little to the model's performance. When λ is increased beyond 0.032, and as more and more coefficients are thus set to zero, we find progressively larger drops in explanatory power.

The meta-model, using only five features, is by far the most parsimonious in terms of explanatory power per feature. To illustrate this point, we can compare the results of the meta-model with those of the $\lambda = 0.256$ LASSO models which have a similar number of features (namely an average of four). We find a lower average R^2 of 21.2% for the LASSO models. The R^2 values of the meta-models are actually quite similar to those of the $\lambda = 0.128$ models which retain an average of nine features. The average difference in R^2 values between the two is 3.5%. Comparing the meta-models with the baseline models, we note a substantial reduction in explanatory power, that averages to 23.7%.

Let us now compare the PnL_3 values of each of the models, shown in Figure 17. Remarkably, the meta-models exhibit the largest average PnL_3 of -4114 bpts compared with -4666 bpts for the baseline models, which have the second largest average. These findings are unexpected in the sense that they do not agree with the earlier findings on explanatory power, which may hint at the power of the meta-models and their simplicity for trading applications. With the exception of the meta models, the average PnL_3 is in fact monotonically decreasing in the number of coefficients. That is, as the regularization parameter is increased, the average PnL_3 decreases, as illustrated in Figure 18.

It is interesting to contrast these findings with the analogous results for the PnL_1 values of all models where execution cost is ignored, as shown in Figure 19. One is immediately struck by the fact that the meta models perform considerably worse than, say, the baseline model according to PnL_1 . The meta-models now have the *smallest* average value, 13,402, whereas the baseline model has an average of 17,695. For the LASSO models, the average PnL_1 decreases monotonically in the regularization parameter λ , from 17,602 for $\lambda = 0.001$ to 15,128 for $\lambda = 0.256$.

How can we reconcile the fact that the meta-models yield the best results in terms of PnL_3 , but the worst results for PnL_1 ? Since $PnL_3^i = PnL_1^i - fee_i \cdot n_i$ where fee_i is the taker

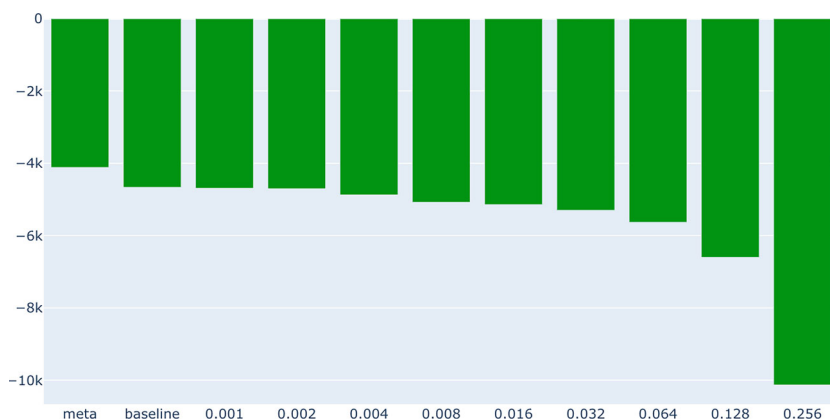


Figure 18. PnL₃ values averaged across markets for each type of model. The numeric values denote the LASSO regularization parameters.

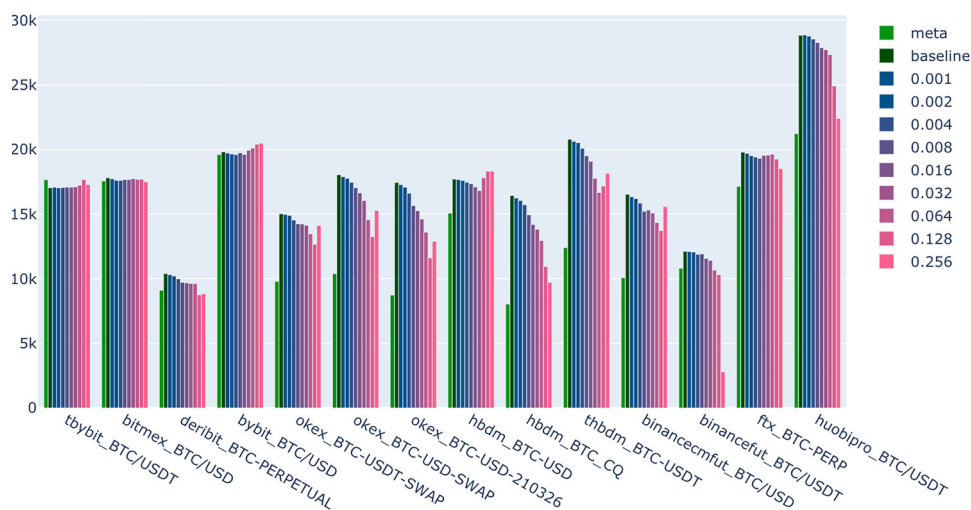


Figure 19. PnL₁ values for each model.

fee on market i and n_i is the number of trades on market i , we must have a smaller number of trades for the meta-models compared with the baseline models. That is, the baseline model strategy trades more often than that of the meta-model, hence accumulating basis points in profit when fees are ignored, but accumulating losses when fees are accounted for. The meta-model strategy trades an average of 25.3% less often relative to its baseline model counterpart. The discrepancy is especially large on the Huobi spot market and quarterly futures contracts, where the number of trades is less than half that of the baseline models. The difference in the number of trades per model is shown in Figure 20.

6. Market-Making Experiments

The strategies described above are *taker* strategies. That is, they rely on limit orders that would lead to immediate execution: buy orders use the top ask as their limit price, while

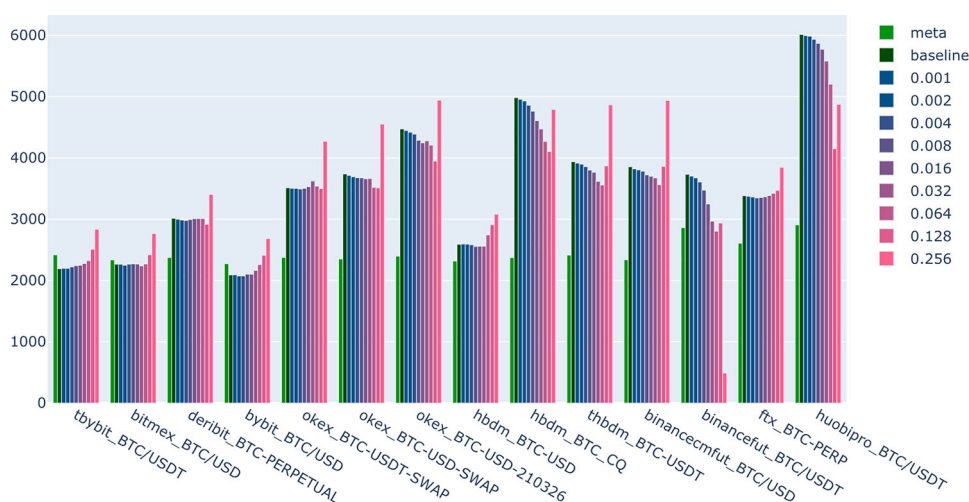


Figure 20. Number of trades for each model.

sell orders use the top bid. We found positive PnL values, even after taking execution cost into account, in a number of cases when the taker fee was particularly low. Some markets (such as the Bybit BTC perpetual) which have among the highest taker fees gave low PnL values, despite their strong tendency to lag.

However, as we noted before, a large taker fee usually also implies a large maker rebate. This is because the net fee (the difference between taker fee and maker rebate) is largely uniform across markets, as one would expect in a competitive exchange environment, since a market with comparatively very large net fees would naturally find it difficult to attract traders, while a market which undercuts competing exchanges with exceptionally low net fees would likely struggle to be profitable. Let us consider more closely those markets which are laggards, have a high taker fee and consequently also a high maker rebate. The maker rebate on the Bybit BTC perpetual, for instance, is 2.5 bpts, which is the highest on our 14 markets. It is natural to wonder whether we can leverage the large maker rebate on Bybit in conjunction with its lagging nature to devise a *maker* strategy which achieves positive PnL values.

In contrast with taker strategies, maker strategies employ *passive* limit orders (also called maker orders). That is, for a sell order, the order price may be any price strictly greater than the top bid price (note that the associated orderbook level need not have pre-existing liquidity). Similarly, a passive buy order must have an order price strictly smaller than the top ask. The order thus submitted rests passively in the orderbook, can be observed by others, and can only be executed by being consumed a taker order from another trader.⁶ The execution time, if any, of a passive order is therefore uncertain, in compensation for which a maker order submitted at time t necessarily achieves a better price than a taker order submitted at the same time. In addition, maker orders incur smaller (in some cases negative) execution fees than taker orders. The trade-off between passive and aggressive orders can be concisely summarized as one of immediacy on one hand versus ‘goodness’ of execution price on the other hand.

Market makers (MMs) facilitate trades by quoting passive buy and sell orders and profiting from the price difference of the buy and sell orders. Suppose for a moment the spread

is minimal (i.e., the difference between top ask and top bid is exactly one tick size). A classic MM strategy in this case would be to post a buy order at the best bid, followed by a sell order at the best ask after the buy order is filled. When the sell order fills, the MM's inventory is rebalanced and they have profited from the spread. The procedure is then repeated, as the MM accumulates spread profits. However, the above strategy makes two critical assumptions, one of which turns out to be detrimental to profitability in the real world. The first assumption is a minor one, namely that the spread is minimal. In practice the spread can of course be several tick sizes, meaning there exist multiple 'empty levels' between the top bid and the top ask. In this case, the MM is faced with the additional choice of whether to post their buy order as aggressively as possible (one tick below the top ask), on the top bid, or on one of the other empty levels in between. The second assumption we made is that the midprice does not move in the between the fill time of the MM's buy order and that of their sell order. In this interim period, the midprice could also have moved down such that the new top ask is much lower than the fill price of the buy order. If this happens and the MM posts at the new top ask and receives a fill, they have rebalanced their inventory but have incurred a loss. On the other hand, if they continue posting a sell order on the old top ask level (one tick above their buy fill price), there is more uncertainty associated with this order being filled at all. The fill probability of this order is often modelled as an exponentially decaying function of the distance of its price to the midprice.

The phenomenon described above is dubbed '*adverse selection*'. In fact, in the above scenario, since the filling of the buy order hinges on the arrival of a trader who places an aggressive sell order, it can be argued that, conditioned on the MM's buy order filling, an adverse price move (down) is more likely than a static price or a price move up. The more informed active traders tend to be, the more forcefully this point applies.

As a matter of fact, according to some measures, the MM always experiences an adverse price move of a certain magnitude, almost tautologically by the definition of their initial buy order (in our example) filling. To illustrate this point, we introduce the so-called 'microprice' which can be thought of as a volume-weighted midprice. It is defined by

$$p_m := \frac{p_a \cdot v_a + p_b \cdot v_b}{v_a + v_b},$$

where p_a and p_b denote the top ask and top bid prices, respectively, while v_a and v_b denote the top ask and top bid quantities, respectively. At the instant of the buy order filling, the top bid quantity is reduced by *at least* the amount of the MM's buy order. This means that the microprice at the instant of the MM's fill is smaller than it was immediately before. This can be expressed in an inequality as follows. Let us write $p_{m,\text{before}} := (p_a \cdot v_a + p_b \cdot v_b) / (v_a + v_b)$ for the microprice immediately before the fill and $p_{m,\text{after}}$ for the microprice at the fill time, and let us denote the size of the MM's passive order by v_{fill} . We then have

$$p_{m,\text{before}} > \frac{p_a \cdot v_a + p_b \cdot (v_b - v_{\text{fill}})}{v_a + (v_b - v_{\text{fill}})} \geq p_{m,\text{after}}$$

where equality holds if and only if (i) the filled buy order was at the front of the queue on the top bid level and (ii) the aggressor's order size exactly matched the size of the MM's buy order.

A key ingredient in a successful MM strategy therefore is to minimize the likelihood of adverse selection by decreasing the probability of trading against informed taker flow (and

thereby increasing the probability of trading against uninformed flow). This is typically achieved by anticipating when an adverse price move is likely to occur in the near future, and then cancelling the passive order before its arrival. The crux of the matter is developing a model with these capabilities. This requires monitoring the same trading signals considered by adversarial takers, which usually involves careful observation and processing of market data from correlated liquid markets.

In this section, we describe a maker strategy that builds on the previously trained meta-models. To test these strategies properly, and noting the difficulties in backtesting maker strategies, we carried out a series of real-world trading experiments on Bybit, where the maker rebate is largest. We then compare the results of our maker strategy with those of a naive benchmark strategy, to showcase the efficacy of the trading alpha provided by the meta models.

6.1. Strategy Specification

We now outline a maker strategy based on the meta-models, which we use for two reasons. First, the meta-models are far more parsimonious than the baseline models. Second, while the meta-models have lower explanatory power than some of the other models, their efficiency in the sense of R^2 per feature was by far the largest, and they achieved the largest PnL values when fees were taken into account.

What is the simplest, most natural maker strategy? Let us first consider what data is required in order to fully specify a maker strategy. For concreteness, let us consider in the following the case of a single passive sell order; the case of a passive buy order is analogous. The specification of a strategy requires

- (1) a criterion for when to post a new order;
- (2) a limit price $p >$ top bid and an amount ν for the posted order; and
- (3) a criterion for when to cancel the ask order (if not filled).

To simplify the problem, we assume the following:

- $p :=$ top bid + tick size,
- $\nu :=$ 2000USD, and
- Post a new order : \iff cancel criterion returns ‘do not cancel’ and we currently have no ask order posted.

Here the *tick size* refers to the minimum price movement of the market in question. That is, we use a fixed price, namely the lowest possible ask level and a fixed (somewhat arbitrarily chosen) small order amount for the sake of testing.⁷ It is left for future work to investigate the scalability to larger order amounts. With these simplifications in mind, to complete the maker strategy we need only to specify a condition for when to cancel an order posted at the top ask. The cancel criterion for a passive order at the top ask level should be capable of anticipating when an adverse price move in excess of the maker rebate is likely, and in that case return a ‘cancel’ decision in time to successfully pull the order. As mentioned earlier, adverse selection is unavoidable – the question is just how much of it our fills are subject to it. If this is less than the maker rebate received per fill, we achieve profitability.

How can we leverage the meta-models to define a natural cancel criterion? Let us fix a market $i \in \{1, \dots, 14\}$ on which to trade. (As previously noted, we conduct our experiments on the Bybit BTC perpetual.) The meta-model for market i is the fitted linear model

$$\begin{aligned} \text{fret}_t^{\delta,i} = & \mu_i + \beta_{i,1} \text{mIMB}_t^{a,i} + \beta_{i,2} \text{mIMB}_t^{b,i} + \beta_{i,3} \text{mTFI}_t^i \\ & + \beta_{i,4} \text{mPRET}_t^i + \beta_{i,5} \text{mDIV}_t^i + \epsilon_{i,t}, \end{aligned} \quad (27)$$

and we fix a future returns horizon $\delta = 1000$ ms for the remainder of this section. Let

$$F_i : \mathbb{R}^5 \rightarrow \mathbb{R}, \quad (28)$$

denote the market-specific function that maps an observation of meta-features to the prediction of the fitted meta-model according to Equation (27). Suppose we have a sell order posted on the top ask level at time t , and we observe a new sample $x_t \in \mathbb{R}^5$. We can then define the following cancel criterion parametrized by a constant T

$$\text{cancel top ask order} : \iff F_i(x_t) > T. \quad (29)$$

That is, when the meta-model predicts a rise price of a certain magnitude (parametrized by T) based on the most recent feature observation, we cancel any order posted at the top ask.

Now consider the choice of the threshold T . If T is too large, we will end up cancelling very rarely and hence experiencing significant adverse selection. In the limit case $T = \infty$, we in fact never cancel (which will serve as our benchmark for comparison, as we will describe later). If, however, T is too low, we will cancel frequently and thus receive very few fills (and consequently collect very few maker rebates). As a matter of fact, if we were to choose $T = -\infty$ the cancel condition would always return ‘cancel’ and we would receive no fills whatsoever, since we would never post a new order to begin with. Given the trade-offs highlighted by the limiting cases, how can we arrive at a sensible parameter value where we maximize the number of fills we receive, while simultaneously minimizing the amount of adverse selection our fills are subject to?

We develop a heuristic answer to this question by learning from cancellation behaviour of other market participants. That is, we attempt to characterize, by means of meta-model predictions, times at which a significant amount of cancellations from the top ask level is imminent. In particular, we search for meta-feature observations in the union of training and test set (over 22 February–1 March) where, at times t the following conditions hold:

- $v_t > M$,
- $p_{t+\delta} = p_t$ and $v_{t+\delta} < m$ for some $\delta \in (0, 500 \text{ ms})$.

Here v_t and p_t denote the top ask liquidity and price at time t , respectively; $M > 0$ represents a moderately large amount of liquidity, while $m > 0$ represents a small amount of liquidity. Empirically, choosing m and M as the first and second quartile (respectively) gives reasonable results. The quartiles are computed over all observations of the top ask liquidity spanning the dates 22 February to 1 March 2021 midnight UTC time (this is the union of training and test sets used in previous section). We denote the subset of meta-feature samples defined by the above two condition by \mathcal{F} .

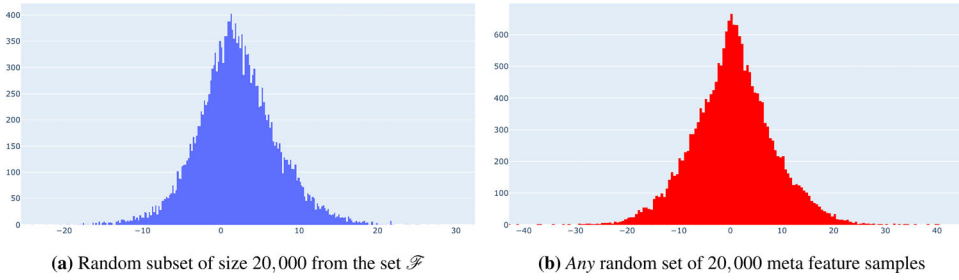


Figure 21. Distribution of meta model predictions: (a) random subset of size 20,000 from the set \mathcal{F} and (b) any random set of 20,000 meta feature samples.

The intuition here is to scan our data set for examples where moderately high liquidity (on the top ask level) is followed by low liquidity. These transitions represent the cutoff point at which the market (i.e., most market participants) ‘agrees’ that it is no longer profitable to post an order at the top ask level. We investigate such cases via our meta models. Consider Figure 21(a), where we display a histogram of values from the set

$$\mathcal{P} := \{F_i(x) : x \in \mathcal{F}\}. \quad (30)$$

That is, we show the distribution of predictions made by the meta-model on samples from the set \mathcal{F} defined above.⁸ The average and median predictions are 2.04 and 1.83 respectively, while the standard deviation is 5.14. We compare this with the histogram shown in Figure 21(b), where we display a uniformly randomly chosen subset of size 20,000 of the set $\{F_i(x) : x \text{ is any training or test sample}\}$. Here we find an average prediction value of 0.47, a median of 0.44 and a standard deviation of 7.88.

This analysis suggests that $T := 1.83$, the median prediction on the set \mathcal{F} , is a reasonable candidate for the cancel threshold. We now examine this choice from another perspective and compare with other possible values for T . In so doing, it is useful to measure the expected adverse selection when the prediction exceeds T . Consider Figure 22 which shows the average price (top ask) increase (in basis points) after time δ , for $\delta \leq 500$ s, conditioned on a meta-model prediction greater than T_q , defined as the q th quantile of the set \mathcal{P} from Equation (30), for $q \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. (Note that $T_{0.5} = 1.83$ is our candidate threshold.) We observe that our candidate threshold (corresponding to $q = 0.5$) reaches 2.5 bpts after a few seconds; this is precisely the breakeven point of tolerable adverse selection, since the maker rebate is 2.5 bpts. This provides us with further evidence of the suitability of the threshold $T = 1.83$.

Impossibility of backtesting. Thus far, we have a candidate maker strategy (or a set of candidate strategies) to assess for its capability to generate positive PnL. The next question is how to do this. A number of difficulties make it virtually impossible to do so with any reasonable degree of precision on the basis of a backtesting approach. Apart from the general point that any backtest necessitates an alternative version of history to accommodate the backtest orders, there is a specific challenge arising from the nature of our data, which is snapshot data as opposed to order-by-order, so that it is impossible to see the sequence of events between snapshots, and thereby deduce the outcome of the backtest strategy. For an example of the insurmountable ambiguities that can arise in backtesting, see Section D in the appendix.

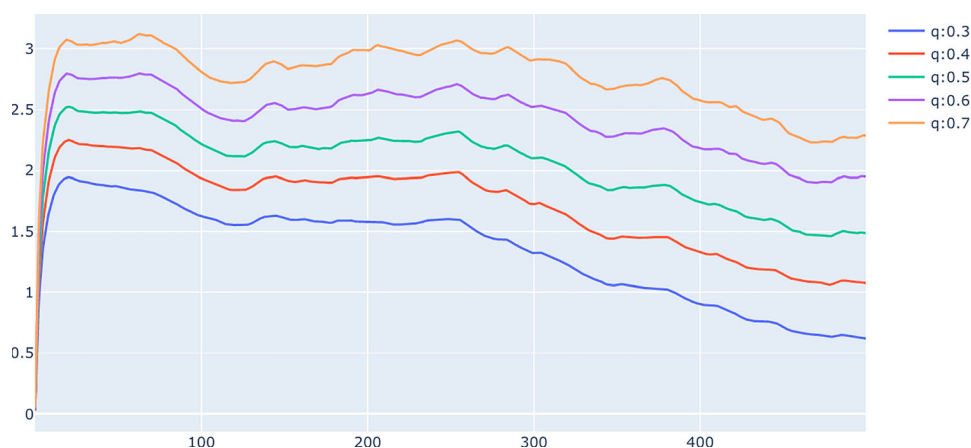


Figure 22. Average top ask price returns after δ second conditioned on meta model prediction greater than T_q .

6.2. Methodology and Experimental Setup

As just noted, backtesting of maker strategies with any reasonable degree of precision is not possible in crypto markets; however, it is relatively simple for individuals to trade in an automated fashion on crypto markets, as it is easy to create exchange accounts and access market data and trade infrastructure through public API endpoints. This is in contrast to traditional financial markets with a high barrier to entry, where one must usually interface with brokers, undergo a number of elaborate verification procedures, set up a prohibitively elaborate trading infrastructure to interface with relevant markets, and so on. We exploit the comparatively easy access to crypto markets to carry out ‘real-money’ tests of the performance of our previously defined candidate maker strategy. We now describe the framework and scope of our trading experiments, along with details of how they are conducted.

Some practical matters. We begin by describing a number of practical circumstances involving the real-time data collection and feature calculation process. The data feeds we use here are those also used to gather historical data, as described in Section 2. That is, we use websocket API feeds for each of our 14 markets whose data. This is generally the quickest way to obtain real-time data in crypto markets, since the exchange automatically pushes orderbook or trades updates to the subscriber, who does not have to explicitly prompt the exchange for the most recent market data (which is the case for the alternative type of data feed, the so-called REST API requests). Orderbook and trades data are received via different websocket feeds, with order book data typically a short time later than trade data. On rare occasions, mostly when volatility is high, this can lead to contradictory information, which we are able to detect and pause trading momentarily while updates resolve the contradiction.

We mine this real-time data in a server which is cloud-located to the server of the exchange on which we are trading, namely Bybit. To explain the term ‘cloud-located’, we first remark that Bybit, like most other crypto exchanges, is hosted in a cloud computing facility. In the case of Bybit this is the Singapore region of amazon Web Services (AWS);

colocation can then be easily achieved by renting server space in the same AWS region, and we thereby achieve a sub-1ms latency with Bybit.

We process the market data in real time, to compute first the values of the features (or transformed features, where applicable), then the meta-features, and finally the output of the meta-model and the corresponding output decision. This entire computation pipeline takes no more than two milliseconds. The main bottlenecks are the computations of the average prices $p_{a,t}^i(N_i)$ and $p_{b,t}^i(N_i)$ in the orderbook imbalance (3), because their computation can involve iterations over many orderbook levels. We trigger a new feature calculation (on the latest market data) and decision output roughly every 2–10 ms, depending on market volatility.

Experimental setup. Our goal is to evaluate the experimentally-realized PnL of the maker strategy described above. Recall the specification of the strategy. Considering first a sell order, it is defined by the following set of rules:

- (a) The sell post price at time t is $p_{b,t} + \Delta$, where $p_{b,t}$ is the top bid price on the Bybit BTC perpetual at time t and Δ represents the tick size.
- (b) The order amount is fixed at 2000 USD.
- (c) Cancel a top ask order : $\iff F_i(x_t) > T := 1.83$ where $F_i(x_t)$ is the prediction of Bybit's meta-model on the latest meta-feature observation x_t .
- (d) Post a new top ask order : $\iff F_i(x_t) \leq T$.

However, due to rate limit constraints imposed by the exchange, we have found it necessary in practice to further restrict condition (d). Specifically, the exchange allows only a limited number of API requests (order placements or cancellations) per minute. If the limit is exceeded, any subsequent requests fail for about a minute. Luckily, the exchange provides information about the number of remaining API requests. We incorporate this information by additionally requiring that there be at least two remaining API requests, as we need at least two requests so we can cancel after posting an order. Thus letting rrl_t denote the remaining rate limit at time t , we restrict (d) by additionally imposing the condition $\text{rrl}_t \geq 2$.

Even with this additional restriction, however, we ran into some issues. Whenever $F_i(x_t) \approx T$, there are often many up- and down-crossings of T in a short time, leading to rapid depletion of the number of permitted API requests, after which we must pause for about a minute before a new order can be submitted. The outcome is a significant number of periods of enforced inactivity. We address this problem by further restricting condition (d). Specifically, we additionally require certain trade flow imbalance in favour of the potential sell order. More explicitly, this condition takes the form

$$\text{mTFI}_t^i < T',$$

for some threshold T' . Although this choice is heuristic, it works well in practice; we leave further calibration and a more rigorous approach to future work. In the meantime, the net result of our restriction measures is to reduce the number of new submissions and to mitigate extended inactivity periods due to quickly depleted rate limits. The updated condition (d) now takes the form

$$\text{Post a new top ask order} : \iff F_i(x_t) \leq T \quad \wedge \quad \text{mTFI}_t^i < T' \quad \wedge \quad \text{rrl}_t > 1.$$

We have decided to avoid subsequent same-side fills (here, sells) to simplify post-analysis of the results. The net effect is that the magnitude of the inventory never exceeds one unit (long or short), which removes any additional volatility in the PnL calculation that might be induced by strategies where the inventory fluctuates more widely.

We address this issue by additionally tracking a variable $\Pi_t \in \{-1, 1\}$ which tells us the net position (or inventory) at time t . The inventory takes only values -1 and 1 and it is initialized by setting $\Pi_0 = 1$. If the inventory is negative (i.e., we are short), we rule out submissions of new sell orders. By the same token, if we have long inventory, we do not permit the posting of a new buy order. With this new restriction, the final form of condition (d) from above is as follows:

Post a new top ask order

$$: \iff F_i(x_t) \leq T \quad \wedge \quad \text{mTFI}_t^i < T' \quad \wedge \quad \text{rrl}_t > 1 \quad \wedge \quad \Pi_t > 0.$$

As previously mentioned, we also define a companion maker *buy* strategy, specified completely analogously

- (a) The buy post price at time t is $p_{a,t} - \Delta$ where $p_{a,t}$ is the top ask price on the Bybit BTC perpetual at time t and Δ represents the tick size.
- (b) The order amount is fixed at 2000 USD.
- (c) Cancel a top bid order : $\iff F_i(x_t) < -T$
- (d) Post a new top bid order : $\iff F_i(x_t) \leq T \quad \wedge \quad \text{mTFI}_t^i < T' \quad \wedge \quad \text{rrl}_t > 1 \quad \wedge \quad \Pi_t < 0.$

For each of the maker buy and sell strategies, we launch an independent ‘bot’ that operates according to the above specified rules. That is, we launch a *sell bot* with access to real time data (including the remaining rate limit rrl_t and net position Π_t) which acts according to the set of rules for the sell strategy. Likewise, we launch a *buy bot* which uses the same data feeds, trades on the same exchange, and acts according to the rules of the buy strategy. Note that the data feeds rrl_t and Π_t are shared. This means, for instance, that if the buy bot is barred from posting a new buy order due to $\Pi_t > 0$ (see condition (d) above) and the sell bot receives a fill at time $t + \delta$ resulting in $\Pi_{t+\delta} < 0$, then the buy bot will receive the updated value $\Pi_{t+\delta}$ and become active again. Both bots operate in perpetuity according to these guidelines. We refer this pair of bots as the *metaMM* bots (or strategy).

In order to better interpret our results and draw comparisons, we launch another pair of buy and sell bots operating the same way as we outlined above, with the exception that they use $T = \infty$ and fixed order amount $v := 100$ (we reduced the amount since this strategy is expected to be extremely unprofitable). The modification of the threshold T has large ramifications: the resulting pair of bots never cancel after posting an order, hence cannot avoid adverse selection. We call this the *naive benchmark* or just *benchmark* strategy.

6.3. Results and Comparison with Benchmark

We ran the strategy (i.e., the bots were active) nonstop from 2021-07-08 13:44:02 until 2021-07-13 23:54:02+00:00, using training and test data from February 22nd until March 1st (midnight UTC), 2021. The total filled amount for the metaMM strategy was 1.537 million USD, implying an average hourly fill volume of 11.810k; this

corresponds to approximately six ‘units’ since the order size is fixed at 2000 USD. Over the entire sample, the difference between the average sell and buy price is -4.4749 bpts. More specifically, let us define the average sell price by

$$\bar{p}_s := \frac{1}{V_s} \sum_{(p_s, v_s) \in \mathcal{G}^{(s)}} p_s \cdot v_s,$$

where $\mathcal{G}^{(s)}$ is the set of all sell fills over the sample period, and $V_s := \sum_{(p_s, v_s) \in \mathcal{G}^{(s)}} v_s$, and the average buy price by

$$\bar{p}_b := \frac{1}{V_b} \sum_{(p_b, v_b) \in \mathcal{G}^{(b)}} p_b \cdot v_b$$

over the set $\mathcal{G}^{(b)}$ of buy fills. Then we have

$$\text{bpts}_{\text{metaMM}} := (\bar{p}_s / \bar{p}_b - 1) \cdot 10,000 = -4.4749,$$

In other words, a roundtrip trade loses on average -4.4749 bpts. However, recall that each leg (buy or sell) of the trade receives a 2.5 bpts maker rebate so that a roundtrip trade nets a total of 5 bpts in rebates. This implies an average profit per roundtrip trade of $5 - 4.4749 = 0.5251$ bpts. The median interarrival time between opposite-side fills is 328 s.

Note that our real money experiments were carried out some four months after the paper experiments from previous sections, including the calibration of the meta models. Better results might be achieved were we to repeat the experiment using more recent data for the model calibration. Indeed, we performed preliminary real-money trading experiments at the end of February and March 2021 which exhibited slightly better performance than reported here. These experiments were conducted over a longer period, used larger order sizes (up to 20k USD), and employed a classification approach with labels heavily reliant on human input. In total, the notional USD turnover in the preliminary trading experiments was more than 100 million USD. In the interim four month period, it was necessary to deal with technical challenges related to changes in API configurations, and we adapted the methodology of the earlier experiments to better fit in with earlier work exhibited in this paper, which explains the delay.

For the benchmark strategy we employed the 20-fold smaller order amount 100 USD. The total filled amount was 153.26k USD representing 1532.6 units. The activity period of the benchmark bots was from 2021-07-17 12:11:27 until 2021-07-18 06:59:26 (thus, considerably shorter than for the metaMM strategy), giving an average fill volume of 8.152k USD per hour. The benchmark strategy achieved the average value (defined similarly to the metaMM case)

$$\text{bpts}_{\text{benchmark}} = -6.8184,$$

on a roundtrip trade which, in line with our expectation, did not yield a profit despite the 5 bpts maker rebate per roundtrip trade. However, the median interarrival time between opposite-side fills is far lower for the benchmark bots, namely only 27 s, which is less than 10% of the corresponding metaMM value. This is plausible, since the metaMM strategy cancels much more often after it posts, and therefore accumulates fills much more slowly.

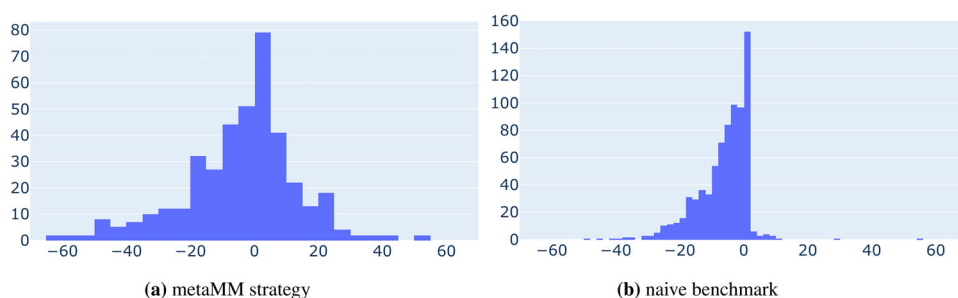


Figure 23. Distribution of the basis point difference between buy and subsequent sell fill: (a) metaMM strategy and (b) naive benchmark.



Figure 24. Cumulative PnL in USD, across time. To give an indication of the price volatility, we have included the bitcoin price in gray in the left plot. Note that the bet size for the benchmark strategy is 1/20th that of the metaMM strategy, and that the horizon is less than one fifth of that for metaMM; for a direct comparison one can rescale the vertical axis of the right-hand plot accordingly: (a) MetaMM strategy and (b) Naive benchmark strategy.

In Figures 23(a) and 23(b), we show histograms, one for the metaMM strategy and one for the benchmark bots, of the differences in basis points between pairs of sell and subsequent buy orders (maker rebates not yet taken into account). It is interesting to note here that the benchmark strategy appears to have a much smaller proportion of roundtrip trades with positive basis-point difference.

In Figure 24(a), we show PnL over time⁹ for the metaMM bots, as well as an overlaid bitcoin price chart, which we included to give an indication of price volatility over the sample period.¹⁰ We strongly suspect that the strategy's expected performance is worse during periods of high volatility; we also conjecture that the variation in the strategy's PnL is far greater during highly volatile times. This is made plausible by the observation that high volatility implies more frequent price jumps and hence larger PnL jumps due to inventory held during a large price rise or fall (even if the inventory is small). Both of these are interesting topics for future investigation. Note that the activity period of our bots was a period of relatively low volatility.

For comparison, we show a similar plot for the benchmark strategy in Figure 24(b). It is striking how steadily the PnL decreases in this case. The contrast between the two PnL graphs provides clear visual evidence of the additional 'alpha' that the meta model provides for this maker strategy. To obtain quantitative comparison of risk-adjusted profitability, we compute an annualized Sharpe ratio by dividing the average hourly returns by their

standard deviation and multiplying the result by $\sqrt{365 \cdot 24}$ (since bitcoin trading occurs 365 days a year and 24 h a day). The metaMM strategy yields a value of 8.68 compared with -137.63 for the benchmark strategy – that is, the former is medium sized and positive while the latter is large and negative.

For our last performance evaluation metric, we sought a measure for the adverse selection to which our fills were subject. More precisely, we want to compare the fill price with another price some time into the future and examine the difference between those two prices. This requires that one specify a time horizon and a notion of price. For our comparisons, we use the last traded price on the Bybit BTC perpetual as the reference price and employ a number of different time horizons, namely $\delta \in \{0.5 \text{ s}, 5 \text{ s}, 10 \text{ s}, 30 \text{ s}, 60 \text{ s}, 150 \text{ s}, 300 \text{ s}, 600 \text{ s}, 1200 \text{ s}, 2400 \text{ s}, 4800 \text{ s}\}$. For each of these horizons, we measure the mean, standard deviation, maximum, minimum, and the 25th, 50th (median) and 75th percentiles of the price moved against (adverse to) our fill price over the set of all our trades. The results are shown in Table 11 for the metaMM bots, and in Table 12 for the naive benchmark. In this table, negative numbers indicate an adverse price move. Note that while most of the measured statistics (except the standard deviations) are negative they must be compared with the maker rebate of 2.5 bpts received for each fill.

When comparing the adverse selection that the benchmark fills are subject to with that of the metaMM fills, we note that after a ‘settling period’ or around 10 s, the difference between the two stabilizes at around one basis point. That is, the benchmark fills experience an average adverse price move which is one basis point worse than that experienced by the metaMM fills. While a mere basis point may not sound particularly substantial, we see in the PnL graphs what ramifications this has. The 500 ms time horizon is anomalous in that the benchmark sees less adverse selection over this time period (but note that the average holding period is much longer than 500 ms for both strategies). The standard deviation in adverse selection is smaller for the metaMM strategy, with exception of the 500 ms time window. The difference becomes more pronounced the larger the time horizon.

Table 11. Adverse selection (in bpts per trade) over different time horizons for the metaMM strategy.

| | 0.5 s | 5 s | 10 s | 30 s | 60 s | 150 s | 300 s | 600 s | 1200 s | 2400 s | 4800 s |
|-----|--------|--------|--------|--------|--------|--------|--------|---------|---------|---------|---------|
| avg | −0.51 | −0.97 | −1.24 | −1.69 | −1.99 | −2.27 | −2.44 | −2.19 | −2.88 | −2.91 | −1.73 |
| std | 1.61 | 2.74 | 3.57 | 5.41 | 7.54 | 12.52 | 16.21 | 23.12 | 30.04 | 40.94 | 57.06 |
| min | −12.66 | −24.18 | −31.75 | −23.75 | −37.99 | −53.18 | −69.18 | −139.03 | −129.19 | −215.28 | −217.02 |
| 25% | −0.15 | −1.07 | −1.94 | −3.74 | −5.71 | −9.48 | −12.45 | −15.26 | −20.32 | −25.09 | −34.52 |
| 50% | 0.00 | 0.00 | −0.15 | −0.74 | −1.22 | −2.32 | −2.56 | −3.14 | −3.69 | −1.64 | −1.80 |
| 75% | 0.00 | 0.00 | 0.00 | 0.15 | 0.30 | 4.73 | 6.82 | 11.52 | 14.24 | 18.27 | 33.22 |
| max | 3.87 | 13.72 | 17.41 | 28.14 | 30.46 | 42.77 | 61.28 | 110.43 | 108.80 | 192.34 | 199.02 |

Table 12. Adverse selection (in bpts per trade) over different time horizons for the naive benchmark strategy.

| | 0.5 s | 5 s | 10 s | 30 s | 60 s | 150 s | 300 s | 600 s | 1200 s | 2400 s | 4800 s |
|-----|--------|--------|--------|--------|--------|--------|---------|---------|---------|---------|---------|
| avg | −0.49 | −1.65 | −2.28 | −2.78 | −2.94 | −3.32 | −3.26 | −3.00 | −3.16 | −3.24 | −3.06 |
| std | 1.25 | 2.96 | 4.12 | 7.10 | 10.40 | 16.33 | 23.63 | 30.26 | 38.49 | 51.81 | 89.72 |
| min | −12.05 | −19.97 | −26.90 | −38.16 | −61.74 | −91.39 | −122.95 | −164.41 | −164.73 | −184.83 | −368.74 |
| 25% | −0.34 | −2.78 | −4.15 | −6.31 | −8.58 | −13.23 | −17.64 | −22.91 | −28.00 | −36.19 | −45.80 |
| 50% | −0.00 | −0.43 | −1.19 | −2.16 | −2.41 | −3.47 | −3.08 | −2.99 | −3.33 | −3.31 | −3.64 |
| 75% | 0.08 | 0.00 | 0.00 | 0.09 | 2.04 | 6.51 | 11.70 | 16.21 | 21.17 | 30.10 | 40.48 |
| max | 3.53 | 16.77 | 31.85 | 54.40 | 50.37 | 73.54 | 119.77 | 132.37 | 159.60 | 182.54 | 365.70 |

7. Conclusion

In this study, by performing a quantitative analysis on a highly granular data set comprising market data from the most liquid bitcoin markets, we have established that variables derived from short-term order flow and relative price differences explain a high proportion of variance in short-term price movements. There is a degree of forecastability of future price movements, which varies across markets.

However, markets differ in their predictability and their predictive power over other markets. Part of these differences can be attributed to differences in the fee structure: a market with a high taker fee and high maker rebate is prone to being a laggard, while a market with a low taker fee and low maker rebate tends to be a leader. The fee structure itself thus induces an apparent forecastability, which is not arbitrageable because of the way the fees work. However, there are residual differences in predictability which suggests a prevailing ranking of markets in terms of their importance to the price formation process. In this ranking, markets on Binance and Huobi feature particularly dominantly as leaders.

When we attempted to leverage these differences in predictability to devise a trading strategy based on taker orders, we found only marginally profitable strategies on a small subset of markets and under the extreme assumption of a heavily reduced taker fee, typically granted only to high-volume traders. On one hand, these findings can be interpreted as evidence of the efficiency of bitcoin markets at sub-1s and 1s time scales, despite their reputation as being highly inefficient. On the other hand, with some effort and a number of refinements, we were able to demonstrate, by means of a fully automated real-money trading experiment, a profitable maker strategy (over a sample period spanning roughly 5 days in July 2021) which leveraged our previously developed models and a large maker rebate. The dramatic contrast between its performance and that of a benchmark strategy which we deployed for comparative purposes, furnished clear evidence in support of our methodological approach.

Future work. Our framework opens up a variety of avenues for future research. For one, a natural extension would be to incorporate alternative cryptocurrencies, such as Ethereum, into our analysis. Daily Ethereum trade volumes now routinely represent around half of those for bitcoin (and occasionally even exceed them). Given the high correlation between these two assets, we would certainly expect significant cross-impact to take place, even on small time scales and potentially as a useful predictive signal. Similarly, one could also explore the interplay between equities and bitcoin markets, particularly around times of market open and close where equities trade volumes tend to be largest.

A second avenue for potential future work is to consider extensions of the taker and maker strategies we presented in this paper. One could, for instance, allow for greater inventory risk in the maker strategy, and possibly vary the post (i.e., submission) price depending on the size of the inventory, thus trading off the strength of the alpha with the inventory risk. For the taker strategy, one could explore how profitability is affected by different execution styles, which go beyond the simple top-of-the-book execution that we assume.

Further related to the market making strategy, it would be interesting to investigate its scalability to larger order amounts, along with the interplay of its profitability with volatility conditions. Building models for estimating queue position and incorporating this into the trading behaviour would likely lead to more realistic backtesting, and potential additional profitability when deployed in a live market experiment.

Lastly, at a more foundational level, our feature set can be refined and extended in a number of ways, or simply processed differently. For instance, instead of applying our nonlinear feature transform to the base features, one might experiment with nonlinear methods, such as tree ensembles or neural networks, applied directly to the set of base features.

Notes

1. Bitcoin derivatives can often be traded with $100\times$ leverage or more [Bitmex Perpetual Guide](#) while regulated marketplaces of traditional assets typically do not allow more than $5\text{--}10\times$ leverage, with some prime brokers allowing leverage up to $20\times$ on some exchanges (as in London and Tokyo Stock Exchanges).
2. An example of the latter would be a limit sell order posted at the best bid price with amount less than the amount available at the top bid level.
3. Differences in tick sizes are too small to materially impact our analysis, though it is interesting to note their impact on the distributions of liquidity across levels. More information can be found in Appendix A.3.
4. By this we mean *buyer-initiated* trades where the taker placed a buy order.
5. To this end, intraday LOBSTER data could be employed, which provides detailed tick-by-tick limit order book data for NASDAQ stocks. LOBSTER contains all limit order submissions, cancellations and executions on each trading day. All events are time-stamped with millisecond precision (Bibinger et al. 2019), and are reconstructed from NASDAQ's historical TotalView-ITCH data (Huang and Polak 2011).
6. An exception is the iceberg order type supported by some exchanges which allows one to submit passive orders that are not visible to other market participants. These orders are typically subject to different fees.
7. It should be noted that on Bybit, the vast majority of the time the spread is minimal and our chosen post price coincides with the top ask level. Generally, the larger the maker rebate is, the more MMs are incentivized to quote minimal spreads – a more systematic study of spread sizes and their relation to fee structure is deferred to future study.
8. In practice, we uniformly select a random subset of size 20,000 from these predictions in order to avoid many successive samples and to keep the histogram more legible.
9. In this plot, the PnL value used at time t is the sum of (i) the cash value of realized PnL up to time t and (ii) the unrealized PnL of any open inventory, calculated as if it was immediately liquidated so as to return the inventory to zero.
10. The proceeds from these and earlier trading experiments amounted to slightly more than 500 USD and were donated to provide basic supplies to people in need in Mozambique.

Disclosure Statement

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

References

- “Bitmex Perpetual Guide.” <https://www.bitmex.com/app/perpetualContractsGuide>
- “Bybit Websocket API Documentation.” <https://bybit-exchange.github.io/docs/linear/#t-websocket>
- “Coin Metrics Market Manipulation Report.” <https://coinmetrics.substack.com/p/coin-metrics-state-of-the-network>
- “Cryptocompare Spot vs Derivatives Volumes.” https://www.cryptocompare.com/media/37746011/cryptocompare_exchange_review_2021_01.pdf
- “Deribit Index Documentation.” <https://test.deribit.com/pages/docs/general>
- “Kimchi Premium Investopedia.” <https://www.investopedia.com/terms/k/kimchi-premium.asp>
- “Pandas Resample Function.” <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.core.resample.Resampler.last.html>

- “Spot vs Leveraged Trading.” <https://www.binance.com/en/support/faq/360033162052>
- “Volume Monitor.” <https://ftx.com/volume-monitor>
- “Websocket Wikipedia Article.” <https://en.wikipedia.org/wiki/WebSocket>
- Aleti Saketh, and Mizrach Bruce. 2020. “Bitcoin Spot and Futures Market Microstructure.” *Journal of Futures Markets* 41 (2): 194–225.
- Alexander Carol, Choi Jaehyuk, Park Heungju, and Sohn Sungbin. 2019. “Bitmex Bitcoin Derivatives: Price Discovery, Informational Efficiency, and Hedging Effectiveness.” *Journal of Futures Markets* 40 (1): 23–43.
- Alexander Carol, and Heck Daniel F. 2020. “Price Discovery in Bitcoin: The Impact of Unregulated Markets.” *Journal of Financial Stability* 50: 100776.
- Alfonsi Aurélien, Fruth Antje, and Schied Alexander. 2009. “Optimal Execution Strategies in Limit Order Books with General Shape Functions.” *Quantitative Finance* 10 (2): 143–157.
- Bibinger Markus, Hautsch Nikolaus, Malec Peter, and Reiss Markus. 2019. “Estimating the Spot Covariation of Asset Prices – Statistical Theory and Empirical Evidence.” *Journal of Business & Economic Statistics* 37 (3): 419–435.
- Bouchaud Jean-Philippe, Farmer J., and Lillo F. 2008. “How Markets Slowly Digest Changes in Supply and Demand.” *Capital Markets: Market Microstructure*.
- Bouchaud Jean-Philippe, Mézard Marc, and Potters Marc. 2002. “Statistical Properties of Stock Order Books: Empirical Results and Models.” *Quantitative Finance* 2 (4): 251–256.
- Cartea Álvaro, Jaimungal Sebastian, and Penalva José. 2015. *Algorithmic and High-frequency Trading*. Cambridge: Cambridge University Press.
- Cont Rama, Cucuringu Mihai, and Zhang Chao. 2022. “Price Impact of Order Flow Imbalance: Multi-level, Cross-asset and Forecasting.” [arXiv:2112.13213](https://arxiv.org/abs/2112.13213)
- Cont Rama, Kukanov Arseniy, and Stoikov Sasha. 2014. “The Price Impact of Order Book Events.” *Journal of Financial Econometrics* 12 (1): 47–88.
- Griffin John M., and Shams Amin. 2020. “Is Bitcoin Really Untethered?.” *The Journal of Finance* 75 (4): 1913–1964.
- Hastie Trevor, Tibshirani Robert, and Friedman Jerome. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York Inc.
- Huang R., and Polak T. 2011. “LOBSTER: Limit Order Book Reconstruction System.” Tech. rep.
- Hung Jui-Cheng, Liu Hung-Chun, and Yang J. Jimmy. 2021. “Trading Activity and Price Discovery in Bitcoin Futures Markets.” *Journal of Empirical Finance* 62: 107–120.
- Malinova Katya, and Park Andreas. 2015. “Subsidizing Liquidity: The Impact of Make/take Fees on Market Quality.” *The Journal of Finance* 70 (2): 509–536.
- Plerou Vasiliki, Gopikrishnan Parameswaran, Gabaix Xavier, and Stanley H. Eugene. 2002. “Quantifying Stock-price Response to Demand Fluctuations.” *Physical Review E* 66 (2): 0–0.
- Potters Marc, and Bouchaud Jean-Philippe. 2003. “More Statistical Properties of Order Books and Price Impact.” *Physica A: Statistical Mechanics and its Applications* 324 (1): 133–140. *Proceedings of the International Econophysics Conference*.
- Soska Kyle, Dong Jin-Dong, Khodaverdian Alex, Zetlin-Jones Ariel, Routledge Bryan, and Christin Nicolas. 2021. “Towards Understanding Cryptocurrency Derivatives: A Case Study of Bitmex.” *Proceedings of the Web Conference 2021*.
- Tkacik Daniel. 2021. “Crypto Derivatives Markets are Booming.” <https://engineering.cmu.edu/news-events/news/2021/04/19-bitmex.html>
- Xu Ke, Gould Martin D., and Howison Sam D. 2018. “Multi-level Order-flow Imbalance in a Limit Order Book.” *Market Microstructure and Liquidity* 4 (3–4): 1950011.