Intelligent Inventory Management for Cryptocurrency Brokers

Christopher Felder[*]
University of Tübingen

Johannes Seemüller[†]
Bankhaus Scheich

Abstract

In equity trading, internalization is the predominant execution method for uninformed order flow, allowing retail brokers to realize cost savings and thereby offer price improvements to customers. In cryptocurrency trading, there are doubts as to whether informed and uninformed traders can be distinguished in the same way, leading brokers to seek cost savings through internal order matching instead. Using the historical order flow of the German cryptocurrency broker BISON, we present a prediction-based approach to internal order matching: Upon receiving a customer order, our model forecasts whether future order flow will be sufficient to neutralize the order before the settlement date. With a prediction accuracy of 85%, it enables brokers to match three-quarters of order volume internally, which is three times as much as a traditional static approach, and realize meaningful cost savings, even after accounting for common minimum price improvements.

Keywords: Segmentation, Retail internalization, Cryptocurrency.
JEL classification: C45, C55, C63, D49, G17, G24.

[*] Christopher Felder (`christopher.felder@uni-tuebingen.de`), University of Tübingen, Department of Banking, Nauklerstraße 47, 72074 Tübingen

[†] Johannes Seemüller (`j.seemueller@bankhaus-scheich.de`), Bankhaus Scheich, Roßmarkt 21, 60311 Frankfurt am Main

1 Introduction

Only few retail stock traders are aware that their orders are typically not executed on exchanges (Comerton-Forde et al., 2018). Instead, firms execute them internally against their own book. Internalization represents the major trading modality for retail order flow in equity trading (Fox et al., 2019; Barardehi et al., 2022) and pays off because retail traders are supposed to not have superior information (Chakravarty, 2001). Since order filling at exchanges is not necessary, firms offer price improvements over the quoted spread for internalized trades. Grammig and Theissen (2012) state that, consequently, internalization can be profitable both for brokers and customers.

In cryptocurrency (crypto) trading, given the large retail share in the total trade volume (Bianchi and Dickerson, 2019), brokers may have similar incentives for retail internalization. However, while financial research is confident that retail orders are less informed than institutional orders in equity trading, this is rather unclear in crypto trading. Makarov and Schoar (2020) find that ‘it is less obvious whether there are traders who are more informed than others and what the nature of the information is’. Bianchi and Dickerson (2019) argue that the opaqueness of information flow and participants’ heterogeneous beliefs generate information asymmetry.

As a consequence, trading against retail order flow in crypto trading is not necessarily profitable, and brokers may seek for alternative off-exchange execution methods. One method discussed by Battalio and Loughran (2008) is to match customer orders internally instead. Internal order matching allows brokers to save the bid-offer spread without having to trade against customers for their own account. Hence, the broker stores orders in inventory for which she expects a contrary order in the future and sends all other orders to the exchange. However, this order segmentation is not straightforward as it depends on the future order flow. If a broker had knowledge about the future, she could store exclusively orders covered by the future order flow in inventory, and route all other orders to the exchange. In practice, risk limits serve to manage the market risk of the inventory position: Indeed, if she cannot internally match an order from inventory before its settlement date, she will have to execute it on the exchange at possibly worse prices and pay the full spread.

In this paper, we present a prediction-based model for crypto retail order segmentation. When a broker receives an order that she cannot immediately execute against open orders in inventory, the model predicts whether future order flow will be sufficient to neutralize the order before its settlement date. Using this prediction, the broker can decide whether to hold the order in inventory or route it to an exchange for immediate execution. Based on an order sample from BISON, the crypto broker of Börse Stuttgart with more than half a million active users and a trade volume of €5.6 billion in 2021, we develop the prediction model in two steps:

First, we define the target variable *optimal internal matching rate*, which represents the maximum fraction of an order that a broker could have matched internally given a $t + 2$ settlement cycle, and thus is the true case we want to predict. Second, we derive predictors from an analysis of market and order flow dynamics prior to the submission of internally matched orders and orders routed to the exchange. Our proposed predictors are the buy volume surplus, average order volume, total trade volume, high-low price range, and one-minute price changes. Prior to order submission, each predictor exhibits behavior specific to each of the two outcomes.

We arrive at the following findings: 85% of order volume is followed by contrary order volume within two days. Consequently, the broker can internally match a maximum of 85% of traded volume. Both logistic regression and artificial neural network predict 85% of order flow correctly, allowing brokers to internally match three-quarters of trade volume. Interestingly, our models achieve the best prediction results during highly volatile and trading-intensive periods. Since these periods are typically associated with large spreads, internal matching pays off disproportionately for brokers and customers. Lastly, assumed that customers and brokers share the cost savings equally, prediction-based segmentation leads to a 35% reduction in effective spreads compared to quoted spreads.

Our results thus support the theory of cost savings, consistent with empirical (Battalio, 1997; Hansch et al., 1999; Battalio et al., 2001; Peterson and

Sirri, 2003; Grammig and Theissen, 2012) and theoretical studies (Battalio and Loughran, 2008; Degryse et al., 2022) of execution costs for internalized trades. However, literature reaches mixed conclusions about market quality under internalization. While Easley et al. (1996), Bessembinder and Kaufman (1997), Chakravarty and Sarkar (2002), and Preece and Rosov (2014) argue that off-exchange trade execution can negatively impact market quality, Battalio (1997), Battalio et al. (1997), and Hansch et al. (1999) find no evidence of detriments to traders. Consistent with Larrymore and Murphy (2009) and Comerton-Forde et al. (2018), who report that mandatory price improvements for internalized orders can improve market quality, our results are marginally sensitive to minimum price improvements, suggesting a potential improvement in market quality.

Our work contributes to previous studies of order segmentation (Fleming and Nguyen, 2013; Garriott and Walton, 2018; Brolley and Cimon, 2020) and builds on study results on the predictive power of retail order flow in the stock market (Kelley and Tetlock, 2013; Boehmer et al., 2021) and crypto market (Scaillet et al., 2018; Silantyev, 2019; Ante, 2020) and on the drivers of internalization (Anolli and Petrella, 2007; Kwan et al., 2015; Barardehi et al., 2022). We also contribute to the analysis of crypto order flow, although we do not attempt to measure informed trading (Wang et al., 2021; Feng et al., 2018), but follow the argument of Makarov and Schoar (2020) that it is unclear what the nature of superior information is in crypto trading. Our model eliminates the need to identify informed traders and thus contributes a meaningful improvement to off-exchange trade execution in crypto trading.

2 Market structure

2.1 Internal order matching of a retail broker

We assume a retail broker who deploys a Request for Quote system for crypto trading. When receiving a Request for Quote from a customer, the broker displays the price at which order execution is possible. The displayed price thus depends on how the broker executes the order: When routing the order to an exchange, the broker determines the public price at which immediate trade execution would be possible, i. e., the current volume-weighted offer (bid) price for buy (sell) orders. When executing the order internally against other customer orders, the broker may offer a better price. We assume that the brokerage market is competitive and brokers offer price improvements.

The price improvement determines how customer and broker share the cost savings. Battalio and Loughran (2008) argue that brokers should pass the full monetary benefit to the customers, which, however, implies that brokers have no monetary incentive to internalize. We assume that brokers and customers equally share the benefit and trade at the mid point between the mid price and quoted price, both saving one-fourth of the quoted spread. When BTC trades at €30,000/€30,020, the broker internally matches a buy (sell) order at €30,015 (€30,005).

By accepting the displayed quote, the customer submits a request for trade. If confirmed by the broker, the trade is legally binding for both sides and no subsequent changes to the execution price are possible. Next, when executing the order at an exchange, the broker places a market order. When internally matching the order instead, the broker either executes the order immediately against open complementary positions or stores it in inventory. We assume $t + 2$ (48 hours) settlement cycle, i. e., the broker has to transfer the traded amounts between the customer's and the own wallet 48 hours after trade confirmation at the latest. When she cannot fully execute the order against incoming orders within two days, she has to close the position at an exchange two days after, which is detrimental to the broker as she pays the quoted spread and thus makes a loss of one-fourth of the quoted spread on that trade. Hence, when internally matching customer orders, older orders in inventory have a higher priority.

2.2 Order segmentation

In size-based segmentation, brokers segment orders by order size, which is the standard approach for filtering uninformed trades (Kim and Verrecchia, 1991; Grundy and McNichols, 2015; Shen et al., 2017). Similar to Anolli and Petrella (2007), our proxy for the critical order size is the European Union regulation of Systematic Internalizers that are allowed to internalize orders up to the standard market size (Commission Delegated Regulation 2017/587), amounting to €$10,000$ for most liquid stocks in Europe.

In prediction-based segmentation, brokers segment orders based on a prediction. When receiving a customer order that the broker cannot immediately execute against inventory positions, a prediction model processes private and public information and forecasts whether the future order flow will neutralize the requested order before the settlement. If predicted so, the broker offers an improved price and stores the order in inventory. Otherwise, the broker offers the public quotes and routes the order to an exchange.

Besides, brokers limit their exposure to market risk by setting a maximum open position. As this limit typically depends on undisclosed firm-specific risk metrics, we assume a discretionary inventory limit of €100,000, i. e., if a broker is short €98,000 and receives another €5,000 buy order, she has to route it to an exchange.

3 Data

We gratefully acknowledge the complete order history of the BISON app from Boerse Stuttgart for Bitcoin Cash (BCH), Bitcoin (BTC), Ethereum (ETH), Litecoin (LTC), and Ripple (XRP). We analyze market orders and triggered limit orders between July 01, 2020 and March 30, 2022. In addition, we download historical open, high, low, close price and trade volume at minute granularity from the crypto exchanges Bitstamp (2022), Coinbase (2022), and Kraken (2022), which serve as the broker's reference markets in our model. As historical order book data are not publicly available, we follow Brauneis, Mestel, Riordan and Theissen (2021) and Brauneis, Mestel and Theissen (2021) and estimate spreads from high, low, and close prices based on Abdi and Ranaldo (2017).

Lastly, we divide the data set into training and test periods. Ji et al. (2019) show that for classification problems in crypto trading, n-fold cross-validation is the most robust sequencing method. n-fold cross validation divides the period into n equal-sized sub-intervals and trains n independent models for each of the sub-intervals. We divide the entire period into sub-intervals of five weeks in length, of which the first three weeks are for training (60%), the fourth week is for validation (20%), and the last week is for test (20%). The validation sets serve to design the architecture of the prediction model and to regularize the training (see Section 4.3).

4 A prediction model for internal order matching

4.1 Optimal internal matching rate

A suitable target variable for our prediction model is the optimal internal matching rate of an order, where the internal matching rate measures the share of the internally matched order volume in the total order volume (Anolli and Petrella, 2007). The optimal internal matching rate represents the maximum achievable internal matching rate under $t + 2$ settlement that a broker with complete knowledge of the future would achieve by holding in inventory all orders for which she will receive a contrary order and routing to the exchange only those orders that cannot be neutralized by future order flow.

We determine the optimal internal matching rate by chronologically routing all orders to inventory, where we either execute immediately against open positions, or store and check whether future orders are sufficient to neutralize the position or not. An internal matching rate of 1 occurs when either open inventory positions at the time of order receipt or subsequent order flow have neutralized the order. A rate of 0 indicates that we either have the order stored in inventory but future orders could not neutralize it, or we had to pass it to an exchange due to the inventory limit.

Figure 1 illustrates the daily volume-weighted optimal internal matching rate across all currencies, which reaches 85% on average. The blue (orange) bars represent internally matched buy (sell) order volume, while the green (red) bars represent buy (sell) volume routed to the exchange.

4.2 Predictors of internal matches

The prediction task is a classification problem with the optimal internal matching rate as target variable, labeled as ‘1’ if it is greater or equal to 50%, and labeled as ‘0’ otherwise. Thus, the model predicts the probability

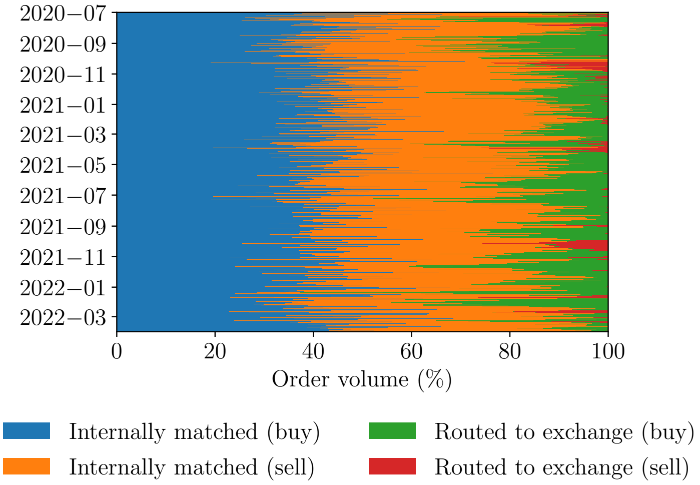


Figure 1: Optimal internal matching rate

that future order flow neutralizes an incoming order. Next, we derive suitable predictors for that prediction task. To this end, Figure 2 analyzes market and order flow dynamics before and after order submission.

The upper left plot illustrates the average hourly *buy volume share*, which is the buy order volume during a particular period divided by the total order volume during that period. The x-axis reports the time difference to the order submission time in hours. For instance, the value reported at -5 hours depicts the average *buy volume share* between 5 and 6 hours prior to receiving an order. The calculation of the value includes two steps: First, for each order, we aggregate all orders between 5 and 6 hours prior to receiving that order, and divide the total buy order volume by the total order volume of the aggregated group. Second, we take the equal-weighted average of the calculated ratios across all orders.

The upper right plot illustrates the *average order volume*, which measures the total order volume during a period divided by the number of orders during that period, normalized in the range $[0, 1]$. The center left plot presents the logarithmic *relative order size* of orders, which is the volume of the submitted order relative to the previous total trade volume. For instance, the value reported at -5 hours depicts the order volume divided by the total

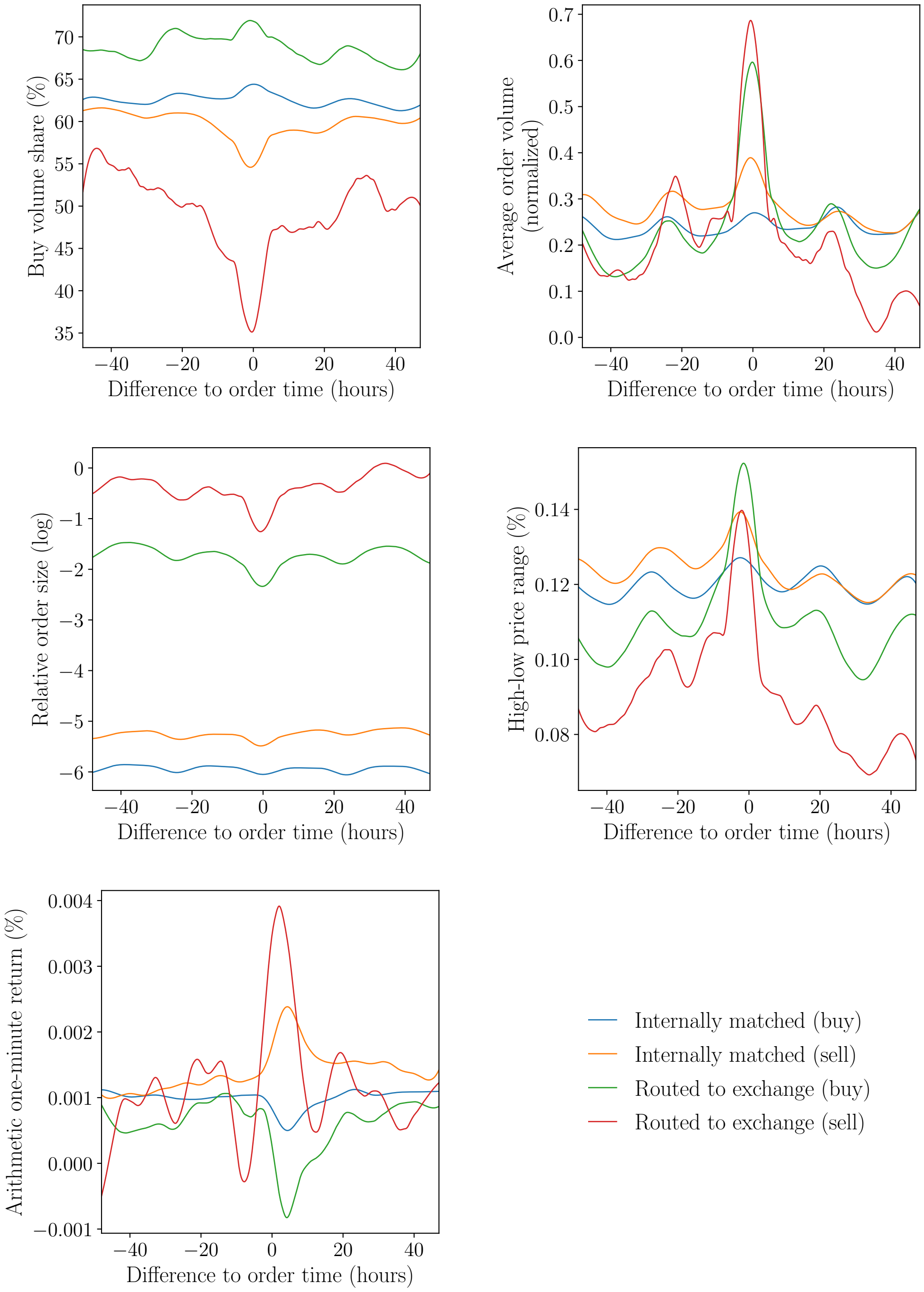


Figure 2: Market and order flow dynamics before and after order submission

order volume between five and six hours ago. The center right chart shows the one-minute percentage *high-low price range*, calculated by dividing the distance between the high and low price in a minute by the high price. The bottom chart shows the average *arithmetic one-minute return*, determined by the one-minute percentage change of volume-weighted mid prices at Kraken, Bitstamp and Coinbase.

Since all variables show individual behavior for orders internally matched and orders routed to the exchange, we propose as predictors the differences between the recent one-hour average and the recent two-hours, three-hours, six-hours, twelve-hours, one-day and two-days averages. For example, if the *buy volume share* has been 0.4 in the past hour and 0.5 in the past six hours, the six hours change is -0.1. We standardize each predictor separately for each currency by subtracting the mean and dividing by the standard deviation prior to splitting into training, validation and test periods. With controlling for the order side, order volume, order amount, open inventory position and currency, all combinations yield a total of $5 \cdot 6 + 5 = 35$ predictors.

4.3 Prediction methodology

Elastic Net As our predictors set is large and predictors are partly correlated or irrelevant, which causes the simple logistic regression to overfit noise rather than extract signal (Gu et al., 2020), we consider penalized logistic regression. Ridge regression (Hoerl and Kennard, 1970) penalizes the sum of squared coefficients in order to shrink them towards zero (L_2 regularization). LASSO regression (Tibshirani, 1996) penalizes the sum of the coefficients' absolute values and performs variable selection by shrinking coefficients to exactly zero (L_1 regularization). Elastic Net (Zou and Hastie, 2005) combines both L_1 and L_2 regularization based on a mixing parameter α that determines each penalty's share. We follow Gu et al. (2020) and optimize α and λ, the level of coefficient shrinkage, based on prediction accuracy for the validation samples. The grid includes each 100 values for α ranging from 0.01 to 1.00 and for λ on a logarithmic scale ranging from 0.0001 to 10,000.

Artificial Neural Network The Artificial Neural Network (ANN) is an 'universal approximator' for any predictive association (Hornik et al., 1989). Through multiple layers equipped with activation functions, ANNs extract different levels of characteristics from input information and generally can approximate any complex, nonlinear function. We consider feed-forward neural networks, where the number of units in the first and last layer equals the dimensions of predictors and target variable. We use the validation sets to test fully connected architectures of up to five hidden layers with dimensions according to the geometric pyramid rule (Masters, 1993). Based on prediction accuracy, our final selection is the architecture with two hidden layers.

Each training step minimizes the cross-entropy loss. The optimization uses stochastic gradient descent and the Adam algorithm (Kingma and Ba, 2015) with an initial learning rate of 0.001 and a batch size of 16. The activation function of all hidden nodes is the rectified linear unit. We regularize training in two ways. First, 10% dropout randomly ignores every tenth node in both hidden and visible layers, approximating model averaging. Second, an early stopping rule serves to halt training when the training does not contribute to the learning for the validation set anymore and since more than two epochs.

5 Prediction-based internal order matching

5.1 Prediction quality

A classification is true-positive (TP) if the broker holds the order in inventory and subsequent order flow actually neutralizes it. If the position remains uncovered until $t+2$ and the broker has to close it on the exchange, the classification is false positive (FP). If the broker routed it to an exchange and subsequent order flow would actually not have neutralized it, the classification is true negative (TN). Otherwise, if the order flow had been sufficient to neutralize it, the classification would have been false negative (FN).

Traditional quality measures are accuracy (ACC), true positive rate (TPR), and true negative rate (TPN). As these measures disregard the financial impact of classifications, we consider volume-weighted variants instead. The

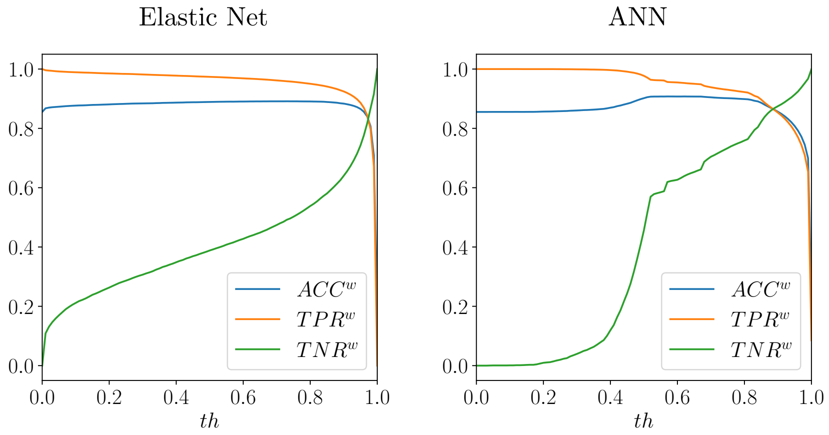


Figure 3: Prediction quality by probability threshold

volume-weighted accuracy is $ACC^w = (V_{TP} + V_{TN}) / (V_{TP} + V_{TN} + V_{FP} + V_{FN})$, where V_{TP} (V_{TN}) is the total order volume of TP (TN) classifications. Correspondingly, $TPR^w = V_{TP}/ (V_{TP} + V_{FN})$ and $TNR^w = V_{TN}/ (V_{TN} + V_{FP})$.

The broker segments orders based on the prediction probability p and the probability threshold th. If $p > th$, she stores the order in inventory, and if $p \leq th$, she routes it to the exchange. As our data set is imbalanced with relatively more internally matched orders, we should choose th carefully. Figure 3 analyzes the prediction quality for different th. A question that arises is whether brokers or customers have individual preferences for a particular prediction quality outcome. A high TNR^w combined with a low TPR^w implies that the broker has to close more inventory positions on the exchange than she erroneously sends to the exchange. Consequently, she prefers a high TPR^w. On the other hand, customers save more cost when the broker sends more orders to inventory by accepting a lower TPR^w and thus prefer a high TNR^w. In order to consider both sides' preferences, we propose th for which $TNR^w \approx TPR^w$, which is 97% in case of Elastic Net and 89% for ANN. In other words, the broker routes orders to inventory if the predicted probability of an internal match for this order is more than 97% or 89%, respectively.

Table 1 reports the prediction quality of four segmentation scenarios: In the scenario 'None', the broker does not segment order flow at all and routes

Segmentation	None	Order size	Elastic Net	ANN
ACC^w	0.14	0.40	0.84	0.86
TPR^w	0.00	0.30	0.84	0.86
TNR^w	0.14	0.99	0.82	0.87
Correlation	–	0.13	0.38	0.51

Table 1: Prediction quality

all orders to the exchange. The scenario ‘Order size’ represents size-based order segmentation, whereas ‘Elastic Net’ and ‘ANN’ represent prediction-based order segmentation following Elastic Net and ANN predictions, respectively. The bottom line focuses on the Pearson correlation of true and predicted labels as a measure of model fit. Prediction-based order segmentation correctly sends around 85 out of 100 Euros to either exchange or inventory, which is more than double the amount of size-based segmentation, with ANN being slightly superior to Elastic Net.

Lastly, we implement a Diebold and Mariano (2002) test for the null that the difference between size- and prediction-based classifications is zero. We can reject the null at the 0.0001 significance level for both Elastic Net and ANN, and conclude that the forecasts are significantly different.

5.2 Internal matching rate

Precisely, evaluating our model by statistical measures is insufficient as for both brokers and traders only the share of successfully internally matched order volume is crucial. To this end, we determine the internal matching following order segmentation based on Elastic Net or ANN predictions. Figure 4 illustrates the proportion of internally matched volume by currency across all test periods. The bottom row represents the average of all currencies and shows an internal matching rate of one quarter for size-based segmentation and three quarters for forecast-based segmentation, which is about 10 percentage points below the optimal internal matching rate indicated by red diamonds. Thus, by using our forecast model instead of order size limits, a broker can triple internal matching.

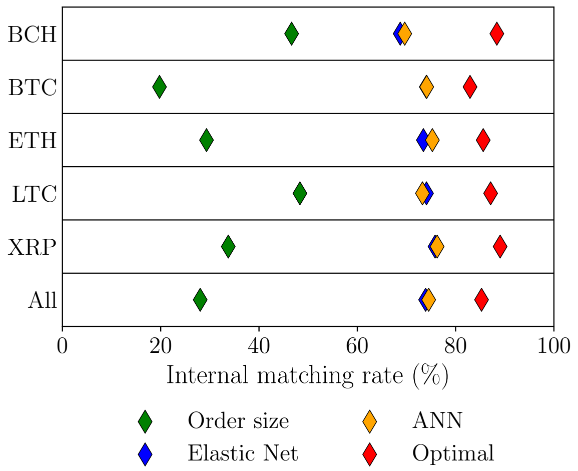
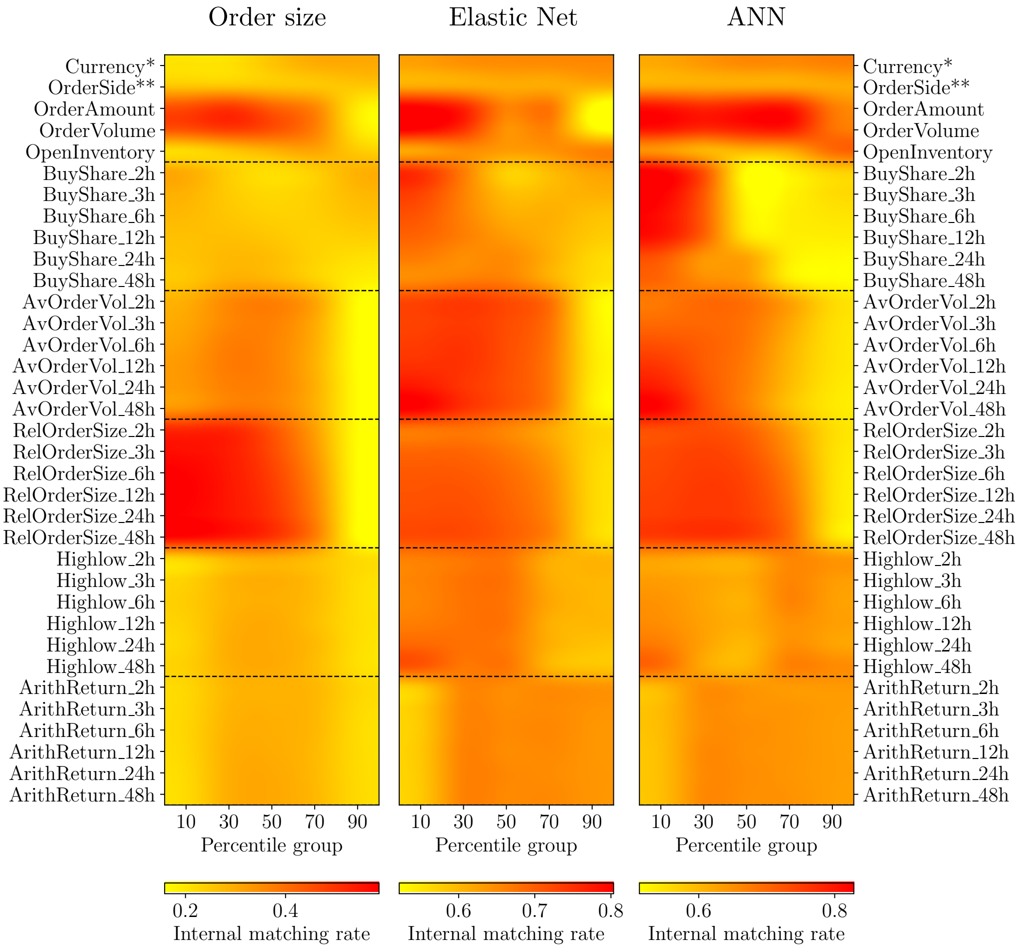


Figure 4: Internal matching rate by currency

Next, we analyze the relationship between internal matching rate and predictors. Figure 5 illustrates the volume-weighted internal matching rate for percentile groups of predictors. For instance, percentile group 1 (100) contains the orders in the lowest (highest) 1% of a predictor across all test periods. Size-based segmentation shows characteristic behavior for small orders and small changes in the average order volume. Prediction-based segmentation achieves high (low) internal matching when changes in *buy volume share* and *average order volume* are small (large), which is consistent with our findings from Figure 2 that suggests lower probability of an internal match given strong increases in the two variables. However, the models adapt less to the changes in *high-low price range* and *one-minute arithmetic return*. Overall, the size of an order is critical to its internal matching rate. Orders executed on the exchange refer to relatively large orders in an above-average volatile market environment accompanied by a buy or sell surplus. Comparing ANN with Elastic Net, we notice that ANN internally matches more large orders than Elastic Net.

5.3 What drives segmentation performance?

This section aims to understand what circumstances make which segmentation approach perform well. For this purpose, we compare the one-day



* Currency represents BCH (1-20), BTC (21-40), ETH (41-60), LTC (61-80) and XRP (81-100)
** Order side represents sell (1-50) and buy (51-100) side

Figure 5: Internal matching rate for percentile groups of predictors

TPR^w and TNR^w with the daily means of *buy volume share*, *average order volume*, *relative order size*, *high-low price range* and *arithmetic one-minute return*. We group the test periods into ten groups, where decile group 1 covers the 10% days with the lowest TPR^w (TNR^w) and decile group 10 covers the 10% days with the highest TPR^w (TNR^w). Prior to the grouping, we standardize each variable to a mean of zero and a standard deviation of 1.

The upper (lower) panel in Figure 6 illustrates the daily means by TPR^w (TNR^w) decile groups. Size-based segmentation delivers the best (worst) inventory classifications in a low-volatile (high-volatile) market environment

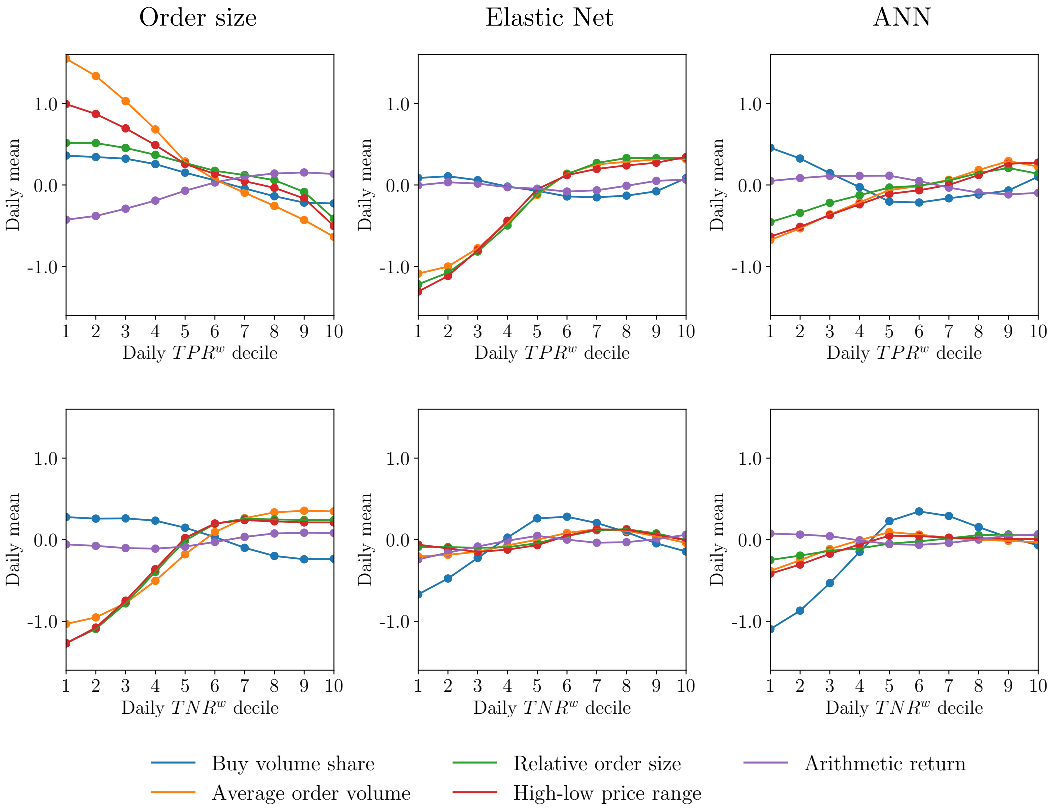


Figure 6: Market and order flow dynamics of daily TNR^w (TPR^w) decile groups.

with small (large) daily means of *average order volume*, *relative order size*, *high-low price range*, and *buy volume share*, associated with negative (positive) returns. Exchange classifications are poor when *average order volume* is small and volatility is low while prices are rising. In contrast, surprisingly, prediction-based segmentation provides precise inventory decisions in an above-average volatile environment, whereas quality declines in relatively low-volatile market environments. Exchange classifications are poor when prices fall sharply.

In relating these results to our previous results from Figure 2 and Figure 5, it is first important to distinguish between statistical measures and internal matching and to keep in mind that, e.g., a large TPR^w does not automatically imply high internal matching. One might wonder why in volatile markets the chance of internal matching decreases, while at the same time the prediction quality for internal matches increases. One possible reason is that we chose a probability threshold th optimized for volatile environments. Since trading volume is highest in volatile market environments, and we parameterized th based on volume-weighted quality measures, our parameterization is possibly less suited for low-volatile periods. We support this argument by testing a prediction model with th parameterized based on equally weighted quality measures, which reports inverse behavior of TPR^w and TNR^w.

5.4 Analysis of cost savings

In the absence of internal matching, investors always trade at the quoted price. In the presence of internal matching, investors trade at either the quoted or the improved price. We measure investors' cost savings by the distance between the actual trade and the quoted price based on the effective spread defined by $(2 \times |\text{trade price} - \text{mid price}|)/\text{mid price}$ (Bollen et al., 2004). The mid price is the average of the offer and bid price, and the trade price of an internally matched order is the mid point between quoted and mid price (see Section 2.1).

Table 2 reports volume-weighted effective spreads by segmentation scenario. We follow Bollen et al. (2004) and weight spreads once across exchanges according to the trade volume share of each exchange, and once according to the volume of each individual trade. The average quoted spread across all currencies is 0.17%. Internal matching by size-based segmentation leads to a 10% decrease, whereas Elastic Net and ANN reduce effective spreads by around 35%.

Finally, we evaluate whether these results are robust to potential regulatory changes in crypto trading. To date, crypto trading is neither subject to the Markets in Financial Instruments Directive (MiFID 2), nor does the European Commission's draft of the Crypto-Asset Regulation MiCA address internalization. However, it is possible that in the future the European Union will regulate crypto trading similar to equity trading. MiFID 2 allows internalization only if a certain minimal price improvement is met. Hence, firms have to either internalize retail order flow at a minimal price improvement or route it to an exchange.

Currency	None	Order size	Elastic Net	ANN
BCH	0.25%	0.20%	0.16%	0.15%
BTC	0.13%	0.12%	0.08%	0.08%
ETH	0.17%	0.15%	0.11%	0.10%
LTC	0.23%	0.18%	0.14%	0.14%
XRP	0.23%	0.20%	0.14%	0.14%
All	0.17%	0.15%	0.11%	0.11%

Table 2: Average volume-weighted effective spreads

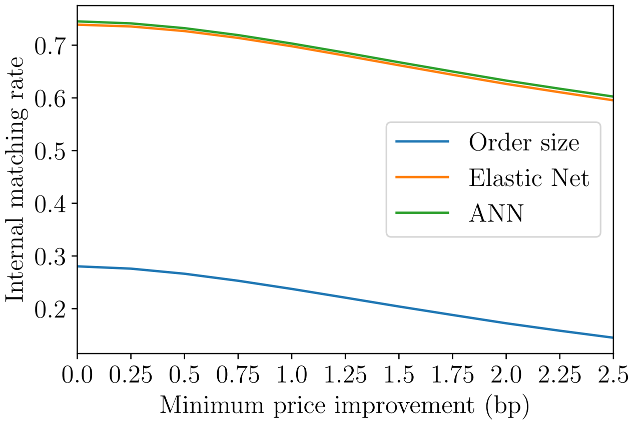


Figure 7: Internal matching rate under minimum price improvement

We measure price improvement in basis points (bp) of the volume-weighted mid price, i.e., given a minimum price improvement of 1 bp, an order about 1 BTC trading at a mid price of €30,000 must receive a price improvement of at least €3. Since our market model assumes that customers and brokers share the cost savings equally, 1 bp price improvement also implies 1 bp additional gain for the broker. Figure 7 illustrates internal matching by minimum price improvements. At 1 bp, the broker could internally match 70% of the order volume, but would save relatively more in costs than the customer on most of these orders. The broker could also internally match more than 70% if she accepts that the customer saves relatively more costs. More than 60% of total order volume could receive a price improvement of more than 2.5 bp.

6 Conclusion

In this paper, we introduce a new approach to internal order matching in crypto trading. To this end, we present an order segmentation methodology based on predicting whether or not an incoming order that the broker cannot immediately execute against orders in inventory will be neutralized by future order flow. We show that a broker using our model can triple internally matched volume compared to traditional segmentation, allowing both brokers and traders to realize cost savings of 35% of the quoted spread.

A limitation of our work is that we cannot assess the actual impact of increased internal matching on market quality. Our analyses assume a positive linear relationship between internal matching and cost savings. However, in equity trading, empirical studies question this relationship (Easley et al., 1996; Bessembinder and Kaufman, 1997; Chakravarty and Sarkar, 2002; Preece and Rosov, 2014). Future work should examine whether increased internal matching in crypto markets affects market quality and, if so, whether customers' cost savings could be eroded by widening spreads in the main market.

Overall, our model is robust to market dynamics and provides a data-driven solution for preferenced trading with crypto order flow, that realizes meaningful cost savings both for brokers and traders, even after accounting for common minimum price improvements.

References

Abdi, F. and Ranaldo, A. (2017), 'A Simple Estimation of Bid-Ask Spreads from Daily Close, High, and Low Prices', *The Review of Financial Studies* **30**(12), 4437–4480.

Anolli, M. and Petrella, G. (2007), 'Internalization in European Equity Markets Following the Adoption of the EU MiFID Directive', *The Journal of Trading* **2**(2), 77–88.

Ante, L. (2020), 'Bitcoin transactions, information asymmetry and trading volume', *Quantitative Finance and Economics* **4**(3), 365–381.

Barardehi, Y. H., Bernhardt, D., Da, Z. and Warachka, M. (2022), Institutional Liquidity Demand and the Internalization of Retail Order Flow: The Tail Does Not Wag the Dog, The Warwick Economics Research Paper Series (TWERPS), University of Warwick, Department of Economics.

Battalio, R., Greene, J. and Jennings, R. (1997), 'Do Competing Specialists and Preferencing Dealers Affect Market Quality?', *The Review of Financial Studies* **10**(4), 969–993.

Battalio, R. H. (1997), 'Third Market Broker-Dealers: Cost Competitors or Cream Skimmers?', *The Journal of Finance* **52**(1), 341–352.

Battalio, R. H. and Loughran, T. (2008), 'Does payment for order flow to your broker help or hurt you?', *Journal of Business Ethics* **80**(1), 37–44.

Battalio, R., Jennings, R. and Selway, J. (2001), 'The Relationship Among Market-Making Revenue, Payment for Order Flow, and Trading Costs for Market Orders', *Journal of Financial Services Research* **19**(1), 39–56.

Bessembinder, H. and Kaufman, H. M. (1997), 'A cross-exchange comparison of execution costs and information flow for NYSE-listed stocks', *Journal of Financial Economics* **46**(3), 293–319.

Bianchi, D. and Dickerson, A. (2019), Trading volume in cryptocurrency markets, Working paper, Available at SSRN 3239670. unpublished.

Bitstamp (2022), 'Bitstamp http api', `https://www.bitstamp.net/api/`.

Boehmer, E., Jones, C. M., Zhang, X. and Zhang, X. (2021), 'Tracking Retail Investor Activity', *The Journal of Finance* **76**(5), 2249–2305.

Bollen, N. P., Smith, T. and Whaley, R. E. (2004), 'Modeling the bid/ask spread: measuring the inventory-holding premium', *Journal of Financial Economics* **72**(1), 97–141.

Brauneis, A., Mestel, R., Riordan, R. and Theissen, E. (2021), 'How to measure the liquidity of cryptocurrency markets?', *Journal of Banking & Finance* **124**, 106041.

Brauneis, A., Mestel, R. and Theissen, E. (2021), 'What drives the liquidity of cryptocurrencies? A long-term analysis', *Finance Research Letters* **39**, 101537.

Brolley, M. and Cimon, D. A. (2020), 'Order-flow segmentation, liquidity, and price discovery: The role of latency delays', *Journal of Financial and Quantitative Analysis* **55**(8), 2555–2587.

Chakravarty, S. (2001), 'Stealth-trading: Which traders' trades move stock prices?', *Journal of Financial Economics* **61**(2), 289–307.

Chakravarty, S. and Sarkar, A. (2002), 'A model of broker's trading, with applications to order flow internalization', *Review of Financial Economics* **11**(1), 19–36.

Coinbase (2022), 'Coinbase pro api', `https://docs.cloud.coinbase.com/exchange/docs`.

Comerton-Forde, C., Malinova, K. and Park, A. (2018), 'Regulating dark trading: Order flow segmentation and market quality', *Journal of Financial Economics* **130**(2), 347–366.

Degryse, H., Van Achter, M. and Wuyts, G. (2022), 'Plumbing of securities markets: The impact of post-trade fees on trading and welfare', *Management Science* **68**(1), 635–653.

Diebold, F. X. and Mariano, R. S. (2002), 'Comparing predictive accuracy', *Journal of Business & Economic Statistics* **20**(1), 134–144.

Easley, D., Kiefer, N. M. and O'Hara, M. (1996), 'Cream-Skimming or Profit-Sharing? The Curious Role of Purchased Order Flow', *The Journal of Finance* **51**(3), 811–833.

Feng, W., Wang, Y. and Zhang, Z. (2018), 'Informed trading in the Bitcoin market', *Finance Research Letters* **26**, 63–70.

Fleming, M. and Nguyen, G. (2013), Order flow segmentation and the role of dark trading in the price discovery of u.s. treasury securities, Staff Report 624, Federal Reserve Bank of New York, New York, NY.

Fox, M. B., Glosten, L. and Rauterberg, G. (2019), *The New Stock Market: Law, Economics, and Policy*, Columbia University Press, New York Chichester, West Sussex.

Garriott, C. and Walton, A. (2018), 'Retail order flow segmentation', *The Journal of Trading* **13**(3), 13–23.

Grammig, J. and Theissen, E. (2012), 'Is Best Really BETTER? Internalization of Orders in an Open Limit Order Book', *Schmalenbach Business Review (sbr)* **64**(2), 82–100.

Grundy, B. D. and McNichols, M. (2015), 'Trade and the Revelation of Information through Prices and Direct Disclosure', *The Review of Financial Studies* **2**(4), 495–526.

Gu, S., Kelly, B. and Xiu, D. (2020), 'Empirical Asset Pricing via Machine Learning', *The Review of Financial Studies* **33**(5), 2223–2273.

Hansch, O., Naik, N. Y. and Viswanathan, S. (1999), 'Preferencing, Internalization, Best Execution, and Dealer Profits', *The Journal of Finance* **54**(5), 1799–1828.

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

Hornik, K., Stinchcombe, M. and White, H. (1989), 'Multilayer feedforward networks are universal approximators', *Neural Networks* **2**(5), 359–366.

Ji, S., Kim, J. and Im, H. (2019), 'A Comparative Study of Bitcoin Price Prediction Using Deep Learning', *Mathematics* **7**(10), 1–20.

Kelley, E. K. and Tetlock, P. C. (2013), 'How Wise Are Crowds? Insights from Retail Orders and Stock Returns', *The Journal of Finance* **68**(3), 1229–1265.

Kim, O. and Verrecchia, R. E. (1991), 'Market reaction to anticipated announcements', *Journal of Financial Economics* **30**(2), 273–309.

Kingma, D. P. and Ba, J. (2015), Adam: A method for stochastic optimization, *in* '3rd International Conference on Learning Representations, ICLR May 7-9, 2015', ICLR, San Diego, CA, USA.

Kraken (2022), 'Downloadable historical ohlcvt data', `https://support.kraken.com/hc/en-us/articles/360047124832-Downloadable-historical-OHLCVT-Open-High-Low-Close-Volume-Trades-`

Kwan, A., Masulis, R. and McInish, T. H. (2015), 'Trading rules, competition for order flow and market fragmentation', *Journal of Financial Economics* **115**(2), 330–348.

Larrymore, N. and Murphy, A. (2009), 'Internalization and market quality: An empirical investigation', *Journal of Financial Research* **32**(3), 337–363.

Makarov, I. and Schoar, A. (2020), 'Trading and arbitrage in cryptocurrency markets', *Journal of Financial Economics* **135**(2), 293–319.

Masters, T. (1993), Designing Feedforward Network Architectures, *in* T. Masters, ed., 'Practical Neural Network Recipies in C++', Morgan Kaufmann, San Francisco, pp. 173–185.

Peterson, M. A. and Sirri, E. R. (2003), 'Order Preferencing and Market Quality on U.S. Equity Exchanges', *Review of Financial Studies* **16**(2), 385–415.

Preece, R. and Rosov, S. (2014), 'Dark Trading and Equity Market Quality', *Financial Analysts Journal* **70**(6), 33–48.

Scaillet, O., Treccani, A. and Trevisan, C. (2018), 'High-Frequency Jump Analysis of the Bitcoin Market', *Journal of Financial Econometrics* **18**(2), 209–232.

Shen, D., Li, X. and Zhang, W. (2017), 'Baidu news coverage and its impacts on order imbalance and large-size trade of chinese stocks', *Finance Research Letters* **23**, 210–216.

Silantyev, E. (2019), 'Order flow analysis of cryptocurrency markets', *Digital Finance* **1**(1), 191–218.

Tibshirani, R. (1996), 'Regression Shrinkage and Selection via the Lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Wang, J.-N., Liu, H.-C., Zhang, S. and Hsu, Y.-T. (2021), 'How does the informed trading impact bitcoin returns and volatility?', *Applied Economics* **53**(28), 3223–3233.

Zou, H. and Hastie, T. (2005), ‘Regularization and variable selection via the elastic net’, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67**(2), 301–320.