Fill Probability Is Not Enough: Empirical Fill Modeling and Trade Selection in BTC/USDT Market Making*

Michel Bassil Chandrasekhara Devarakonda

April 2026

Abstract

This paper examines whether short-horizon fill prediction is sufficient for profitable passive BTC/USDT market making. Using Binance 100 ms L2 order book updates and tick-by-tick trades from 21 trading days, we first estimate empirical fill curves and then evaluate several quoting objectives in a simulator with fees, slippage, adverse selection, inventory constraints, and end-of-day liquidation. Fill prediction is statistically strong: an exponential specification explains more than 99.9% of the cross-depth variation in fill rates, and feature-based classifiers achieve out-of-sample AUC close to 0.87. However, strategies that rank quotes primarily by predicted fill probability remain unprofitable after costs, even when supplemented with volatility gating or a separate adverse-selection model. Within the set of strategies considered here, positive aggregate performance appears only when the model is trained directly on realized order-level PnL, with non-filled orders assigned zero outcome. Adding longer-context features further improves selectivity, yielding +\$21.8 over 21 days with an annualized Sharpe of approximately 1.03 in a conservative single-level configuration, while multi-level quoting remains negative. The main implication is that fill prediction is a useful component of execution modeling, but not an adequate standalone objective for passive market making in this setting.

1. Setup

Passive market making is an execution problem under uncertainty. Quotes placed closer to the midprice execute more frequently but capture less spread, while deeper quotes earn more conditional on execution but fill much less often. The central question is whether accurate execution modeling is sufficient to support profitable quoting once fees, slippage, adverse selection, and inventory liquidation are taken into account.

The analysis uses Binance BTC/USDT spot data consisting of 100 ms L2 order book updates and tick-by-tick trades over 21 trading days. Synthetic bid and ask orders are evaluated every second on the depth grid

$$\mathcal{D} = \{0.1, 0.25, 0.5, 1, 2, 3, 5, 7, 10, 15, 20, 30\} \text{ bps}$$

and horizons $T \in \{60, 300, 600\}$ seconds. Because the data are L2 rather than L3, fills are approximated using a queue-ahead rule: an order is classified as filled if opposing trade flow is sufficient to consume the displayed queue ahead and trade through the simulated price within horizon T.

Trading performance is evaluated in a simulator with a 1.0 bps maker fee, 0.5 bps slippage, 0.01 BTC order size, inventory cap of ± 0.10 BTC, 8 bps inventory skew at the position limit, and 2 bps end-of-day liquidation slippage. PnL is computed from realized cash flows.

2. Fill Modeling and Economic Objectives

The parametric baseline is an exponential fill model,

$$P_{\text{fill}}(d \mid T, s) = A(T, s)e^{-\kappa(T,s)d},$$

where d is depth in bps and s denotes market state. Empirically, this form fits extremely well across horizons and volatility regimes.

To capture local variation, we train feature-based fill classifiers using depth, log-depth, side, queue-ahead, top-of-book size, imbalance, relative spread, and short-term return/volatility features. The strongest short-context models achieve out-of-sample AUC near 0.87.

Adverse selection is the post-fill midprice move against the quote. In the two-stage setup it enters as a separately predicted cost term $\widehat{AS}_i$:

$$S_i^{\text{fill}} = \hat{p}_i(d_i - c), \qquad S_i^{\text{two}} = \hat{p}_i(d_i - \widehat{AS}_i - c),$$

*Work in progress, Results are preliminary

where $\hat{p}_i$ is predicted fill probability and c is explicit trading cost. The unified model instead predicts realized order-level PnL directly:

$$S_i^{\text{uni}} = \widehat{\mathbb{E}[y_i \mid x_i]},$$

with $y_i = 0$ for non-fills. This forces the model to rank quotes by expected economic outcome rather than execution likelihood alone.

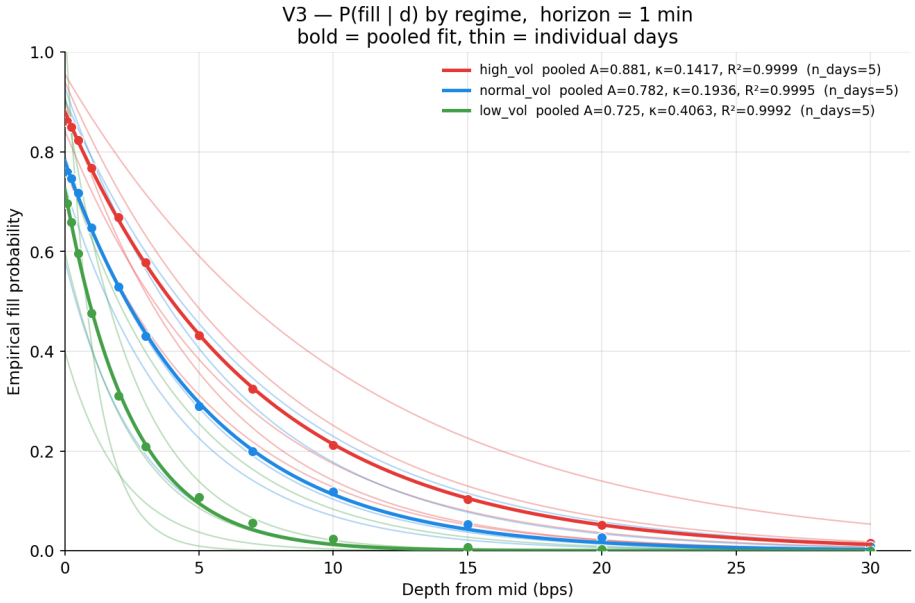

Figure 1: Empirical and fitted fill curves at 60-second horizon by volatility regime. Fill probability is very well described by an exponential decay in depth.

3. Main Results

The execution modeling results are strong. Fill probability is well described by an exponential decay in depth, with pooled regime fits achieving $R^2 > 0.999$. The differences across volatility regimes are also economically meaningful: high-volatility days exhibit flatter fill curves than low-volatility days at the same horizon, implying slower decay of execution likelihood with depth. In the feature-based models, depth, queue-ahead, imbalance, and short-horizon price dynamics are the most informative variables. Imbalance displays a clear side-dependent effect: bid-heavy books reduce bid fill rates and increase ask fill rates, with the opposite pattern under ask-heavy conditions.

The trading results are materially less favorable. Once explicit trading costs and end-of-day inventory liquidation are incorporated, strategies based primarily on predicted fill probability are unprofitable across all realized intraday volatility buckets. Volatility gating and cancellation overlays reduce losses, but they do so mainly by lowering trading activity rather than by producing consistently better trade selection. A separate adverse-selection model improves the ranking margin on some configurations, but its performance remains unstable and economically sparse.

Within the configurations tested here, positive aggregate performance appears only under the unified objective trained directly on realized order-level PnL. Adding longer-context features improves selectivity further, leading to fewer trades but better average trade quality. The strongest configuration is a conservative single-level strategy at the 300-second horizon. By contrast, simultaneous multi-depth quoting remains negative, consistent with the view that correlated exposure across depth levels is not well handled by independent per-quote scoring.

Configuration	Total PnL	Fills/day	Annualized Sharpe
Fill-only, ungated	-$827.7	208.5	–
Best adaptive volatility gate	-$300.6	38.7	–
Best cancellation overlay	-$25.4	2.5	-1.31
Two-stage fill + AS, gate only	-$60.4	9.3	-4.72
Unified single-level, short context	+$8.0	4.9	+0.39
Unified single-level, long context	+$21.8	3.2	+1.03
Long-context multi-level	-$86.2	1.8	-0.76

Table 1: Ablation path from fill-centric quoting to direct order-level PnL prediction. Positive aggregate performance appears only under the unified objective.

4. Interpretation and Limitations

The main empirical finding is that execution prediction and economic trade selection are not equivalent objectives. Fill probability provides useful information about the likelihood of execution, but it does not by itself constitute an adequate quoting objective once adverse selection, explicit trading costs, and inventory effects are incorporated. For the class of strategies examined here, objectives based primarily on execution likelihood tend to concentrate exposure in quotes that are more vulnerable to toxic flow, whereas direct prediction of order-level PnL yields more selective and economically stronger rankings.

The results should nevertheless be interpreted with caution. Since the analysis relies on public L2 data, the backtest cannot fully represent exact FIFO queue priority, cancellations ahead of the simulated order, latency competition, or detailed partial-fill behavior. The positive aggregate PnL obtained in the strongest single-level configuration remains modest relative to these execution-model uncertainties. The contribution of the study is therefore methodological rather than deployable: it provides evidence that direct economic supervision is better aligned with the market-making objective than fill-centric proxy formulations in this medium-frequency setting.

5. Conclusion

The evidence in this study indicates that, for BTC/USDT passive market making in the present medium-frequency setting, accurate fill prediction is not sufficient for profitable quote selection. Fill probabilities exhibit stable empirical structure and can be predicted with high out-of-sample accuracy, but objectives based primarily on execution likelihood remain economically inadequate once costs and liquidation effects are included. The strongest results are obtained only when the model is trained directly on realized order-level PnL, with additional gains from longer-context features in the single-level strategy.

The practical implication is that execution probability is better viewed as an input to trade selection than as a standalone optimization target.