Is Bitcoin's Fair Value Learnable? A Four-Round Microstructure Study

Jovonni L. Pharr
Georgia Cyber Warfare Range
info@gacwr.org

May 2026

Abstract

We ask a deliberately narrow question: given only bar-level market data for Bitcoin, can a model learn to predict the asset's near-future fair market value (FMV) well enough to have found real signal, rather than merely re-describing the current price? We approach it empirically across four successive rounds of experiments, each of which tightens the definitions and enriches the data in response to what the previous round revealed. Round 1 fits eight model families (naive baselines, ridge, lasso, random forest, histogram gradient boosting, an MLP, and an LSTM) to next-bar typical price from high/low/close/open/volume (HLCOV) on roughly 2.6×10^5 one-minute BTCUSDT bars; every model lands within ± 1.5 percentage points of the 0.500 directional-accuracy coin-flip line, the best being an LSTM at 0.511. Round 2 expands the feature set roughly tenfold to an 89-feature quantitative library and adds convolutional and Transformer models across three FMV target definitions; we report this round as a plan carried out, but note that its metrics are not available in our evidence bundle and so present no numbers for it. Round 3 reframes the target away from next-bar close toward six microstructure-grounded fair-value estimators (Kalman, Hasbrouck, microprice, VWAP, OHLC4, and forward VWAP), and cleanly separates two regimes: *contemporaneous* FMV recovery, where ridge reaches directional accuracy 0.92–1.00 because it is reconstructing a deterministic function of the current bar, and *forward* FMV prediction at a 5–15 minute horizon, which collapses back to 0.49–0.51. Round 4 replaces the OHLCV approximations with genuine level-2 order-book microstructure — top-of-book depth, order-book imbalance, signed trade flow, and the Stoikov microprice — from a Coinbase level-2 recording; we describe the pipeline that was built to predict the forward microprice, but its measured metrics are likewise absent from our bundle, so we make no numeric claim about its forward accuracy. On the evidence we do have, the consistent finding is that on bar-aggregated data in this regime, forward fair value is not measurably learnable at the 5–15 minute horizon: forward directional accuracy never clears our conservative 0.520 floor and forward MAE never beats the naive "value stays put" baseline. We are explicit throughout about what we did not model and about which numbers we can and cannot substantiate.

1 Introduction

A recurring temptation in quantitative finance is to fit a flexible model to a price series, observe a coefficient of determination near unity, and declare the price "predictable." The temptation is understandable and the conclusion is almost always wrong. Prices at short horizons are close to a martingale: the single best guess for tomorrow's level, in a mean-squared sense, is today's level. A regression that predicts "next price $\approx$ current price" will therefore score an R^2 indistinguishable from 1.0 while carrying no usable information whatsoever. The autocorrelation of the level, not any learned structure, is doing all the work.

This study began with a simple directive from the lab: download a minute-resolution Bitcoin dataset, fit several different families of models to predict fair market value from

high/low/close/open/volume, and report honestly on how learnable that target actually is. The phrasing matters. The question was not "can we build a profitable trading system," which invites motivated reasoning and selective reporting, but "how learnable is this," which invites a measurement. We took the measurement seriously and let it fail where it failed.

The central methodological commitment is that we do not trust R^2 on the price level. Instead we lean on *directional accuracy* — the fraction of held-out bars on which a model correctly predicts the sign of the change from the current price to the target. A model with no signal scores 0.500 here regardless of how good its R^2 looks, because getting the sign right half the time is what a coin achieves. A model that consistently lands above roughly 0.520, with the improvement stable across targets and time, is the first genuine hint of learnable structure. We adopt 0.520 as a rough threshold not because it is tradeable — it is not, once spread, slippage, and fees are subtracted from a five-basis-point average move — but because it is a conservative floor below which we refuse to claim any edge at all.

The paper is organized as four rounds of a single investigation, each provoked by the last. Round 1 asks whether raw bar data carries the signal. Round 2 asks whether a far richer engineered feature set or deeper models can find it. Round 3 asks whether the target itself was wrong — whether predicting a noisy realized price instead of a latent fair value was the mistake. Round 4 asks whether the problem was the data granularity all along, and brings real order-book microstructure to bear. We are careful to report only the numbers our evidence bundle actually contains: Rounds 1 and 3 produced full metric tables, which we reproduce, while Rounds 2 and 4 were run per plan but their measured metrics are not available to us, so we describe them without inventing values. On the measured rounds the finding is consistent, which is the main reason we believe it. We present each round's actual available numbers, discuss the several things we deliberately did not model, and try throughout to maintain the discipline of reporting a null result as a result.

2 Background and context

The observed transaction price of a traded asset is usefully modeled as a latent efficient price contaminated by transitory noise. Writing p_t for the observed log price, one decomposes

$$p_t = m_t + s_t,$$

where m_t is a near-martingale "efficient" component and s_t absorbs bid–ask bounce, the inventory positions of market makers, asymmetric information, and price discreteness. The efficient component m_t is what one would like to call "fair value"; the noise s_t is what makes the raw close a poor target. Predicting the next observed close therefore blends the part that is nearly unpredictable by construction (the martingale increment of m_t) with a part that is pure noise (s_t). The model's most honest response to that target is to predict "roughly the current price," and directional accuracy near a coin flip is the expected, not the surprising, outcome.

Several classical estimators try to recover m_t from observable data. The Roll estimator infers an effective spread from the first-order autocovariance of returns. Hasbrouck's decomposition separates a random-walk component from a transitory component; in the single-venue case it reduces to treating the residual of an autoregression on returns as the innovation to the efficient price and cumulating it. A random-walk-plus-noise Kalman filter models m_t as a latent state observed through additive noise and produces a filtered, causal estimate. When one has order-book data rather than only bars, the microprice of Stoikov weights the best bid and ask by the *opposite* side's resting size, so that a book heavy on the ask side pulls the fair value toward the bid — the intuition being that the thin side is where the price will move. All of these are attempts to strip the noise s_t and leave the signal m_t.

A second thread concerns market efficiency at short horizons. If markets are efficient at a given granularity, the conditional expectation of the next efficient-price increment given the

available information is essentially zero, and no feature set, however elaborate, can move directional accuracy off 0.500. Whether this holds is an empirical question that depends on the asset, the horizon, and the information set. Order-book depth, cross-asset relationships, on-chain flows, and news are all plausible sources of exploitable structure that a bar-only feature set cannot see. Our study holds the asset (Bitcoin) and the venue class fixed and varies the target definition, the feature richness, the model family, and finally the data granularity, in order to locate where — if anywhere — learnable structure begins to appear.

3 Method and system

The study is implemented as a small, reproducible pipeline in Python using only scikit-learn and PyTorch; no gradient-boosting libraries beyond scikit-learn's histogram implementation and no deep-learning frameworks beyond PyTorch are used, to keep a full run under a few minutes on a laptop. Data, feature construction, training, and reporting are separated into distinct stages so that each round can reuse the machinery of the last.

Data. Rounds 1 through 3 use one-minute BTCUSDT klines from Binance's public data dump, which requires no authentication and delivers open, high, low, close, volume, quote volume, trade count, and taker-buy base volume per bar. The original request named a Kaggle dataset, but the Kaggle export requires an API token that was not present on the machine; the Binance dump provides the same shape of data without the friction, and the pipeline retains a stub for swapping Kaggle back in. A single contiguous window of recent months is used. Round 4 uses a genuine level-2 feed: a Coinbase BTC-USD order-book recording captured through the exchange's WebSocket channels (book snapshots, incremental level-2 updates, and executed matches), which the pipeline's microstructure parser aggregates into per-bar order-book features. The microstructure parser maintains a sorted in-memory book from these Coinbase events and snapshots it at each bar boundary; the Round 4 trainer then consumes the derived per-bar parquet.

Targets. The fair-value target evolves across rounds. Round 1 predicts the next bar's *typical* price $(H+L+C)/3$. Round 2 compares three within-bar smoothers as targets: next-bar typical, next-bar VWAP, and next-bar midpoint $(H+L)/2$. Round 3 introduces six microstructure-grounded estimators and splits them into two prediction modes. The *contemporaneous* estimators — a random-walk-plus-noise Kalman filter on log close, a single-venue Hasbrouck decomposition using an AR(5) residual, a microprice approximation using high and low as bid/ask proxies weighted by taker-buy aggression, within-bar VWAP, and $(O+H+L+C)/4$ — are functions of the current bar and serve as a recovery test. The *forward* estimators — the volume-weighted average price over the next 5 and next 15 bars — are the genuine learnability targets, asking where fair value settles over the coming minutes given today's state. Round 4 predicts the Stoikov microprice both contemporaneously and 5 and 15 bars ahead.

Features. Round 1 uses current HLCOV plus a lookback of 20 lagged HLCOV bars and a handful of standard derived signals (returns over the window, range, a volume ratio, and a return-volatility standard deviation). Rounds 2 and 3 use an 89-feature quantitative library computed strictly from data available at bar t: multi-horizon log returns; realized-volatility estimators (rolling standard deviation, Parkinson, Garman–Klass, Rogers–Satchell, Yang–Zhang); momentum oscillators (RSI, MACD, Williams %R, Stochastic %K, CCI, rate of change); trend measures (ADX, Aroon, true range, ATR); bands (Bollinger %B and width, Donchian and Keltner position); volume features (OBV slope, Chaikin money flow, money-flow index, Amihud illiquidity, Roll spread); microstructure proxies (taker-buy ratio, a Kyle's-λ price-impact proxy, a bucket-free VPIN proxy, standardized trade intensity); statistical descriptors (rolling

skewness, kurtosis, a rescaled-range Hurst exponent, Lo–MacKinlay variance ratios, autocorrelations at lags 1 and 5, approximate entropy); rolling quant ratios (Sharpe, Sortino, a Kelly proxy, maximum drawdown, Calmar, Ulcer index); and multi-scale aggregates at 5-, 15-, and 60-bar resolutions. Round 4's features are computed from the order book directly: the level-1 mid, the microprice and its deviation from the mid, the spread in dollars and in basis points, order-book imbalance at levels 1, 5, and 10, signed trade flow, buy pressure, top-of-book and aggregated depth, mid returns at several horizons, exponential moving averages of the mid, imbalance, flow, and spread, and lagged copies of the most informative of these.

Evaluation. Every round uses a chronological train/test split with no shuffling: the earliest portion of the window trains and the final portion (20% in rounds 1–3, 25% in round 4) is held out. Features at bar t use only bars $t - \text{lookback}$ through t, and the target is at $t + 1$ (or $t + k$ for forward targets), so there is no lookahead leak; the first bars, up to the longest rolling window, are dropped from training. We report mean absolute error (MAE) and root-mean-square error (RMSE) in price units, R^2 on the level (which we treat with suspicion), and directional accuracy. On the histogram gradient-boosting model we additionally compute permutation feature importance — the increase in test MAE when a single feature's column is shuffled — to see which features the model actually relies on.

4 Round 1: HLCOV and next-bar typical price

The first round establishes the baseline. Eight model families are fit to the next-bar typical price on 208,496 training and 52,123 test one-minute bars, with a lookback of 20 bars. Table 1 reports the results.

Table 1: Round 1 results. Target: next-bar typical price $(H + L + C)/3$; lookback 20 bars; 208,496 train / 52,123 test, chronological split. MAE/RMSE in USDT.

Model	MAE	RMSE	R^2	Dir-acc	Train time
naive_last_close	18.2486	27.7234	0.99996	0.000	0.0s
naive_typical	20.5887	31.3680	0.99994	0.485	0.0s
ridge	18.6061	27.9913	0.99995	0.495	2.4s
lasso	19.7596	29.4305	0.99995	0.500	0.3s
rf	21.2869	30.7327	0.99995	0.499	1314.0s
hgb	90.2440	141.9337	0.99883	0.499	7.5s
mlp	92.6412	139.7031	0.99887	0.503	9.6s
lstm	120.4972	171.2019	0.99830	0.511	8.8s

The first thing to notice is that every model's R^2 sits above 0.998, and the naive baseline that simply repeats the last close scores 0.99996. This is exactly the autocorrelation trap: the level is trivially "predictable" and the R^2 column carries no information about learnability. The directional-accuracy column tells the real story. Every model lands within ± 1.5 percentage points of 0.500. The best of them, the LSTM at 0.511, does not reach even the conservative 0.520 floor, and 0.520 would itself be untradeable after fees against a five-basis-point average move. Notably, the more flexible models are not underfitting for lack of capacity — the random forest took over twenty minutes to train and still landed at 0.499. The expected next-bar move conditional on HLCOV is, in this regime, dominated by noise.

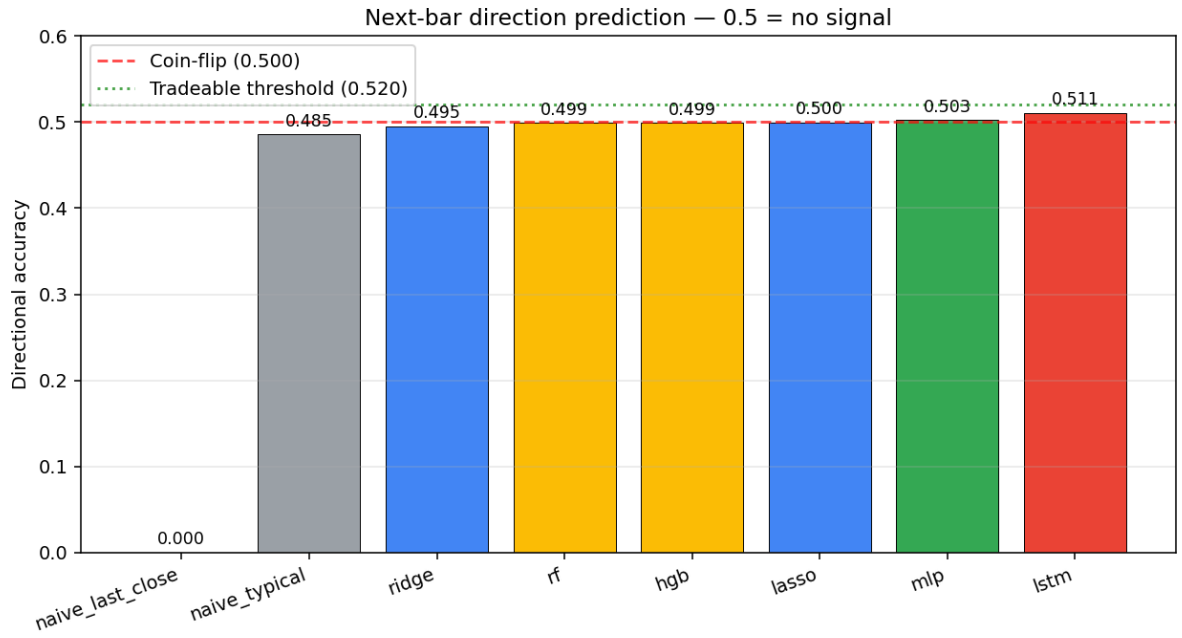


Figure 1: Round 1 directional accuracy by model. Every family clusters on the 0.500 coin-flip line; the LSTM's 0.511 is the high-water mark and still falls short of the conservative 0.520 floor.

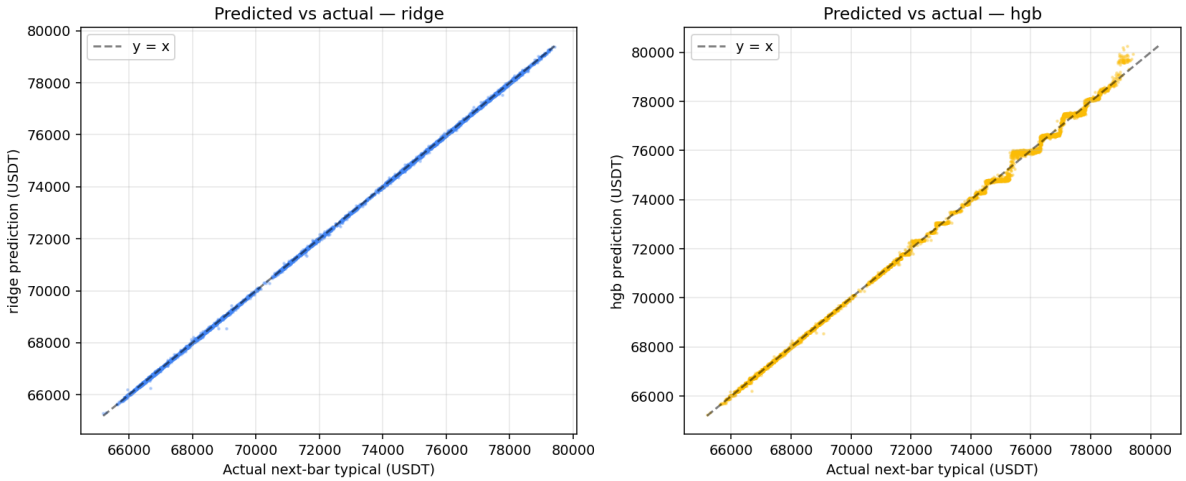


Figure 2: Round 1 predicted versus actual next-bar typical price. The near-perfect diagonal is a symptom of the autocorrelation trap, not of learned signal: predicting "next ≈ current" produces this plot with no directional information.

5 Round 2: a wider feature library and deeper models

The natural objection to Round 1 is that HLCOV plus a few derived signals is too thin a feature set, and that a small LSTM is too weak a model. Round 2 was designed to address both. The feature set grows roughly tenfold to the 89-feature quantitative library described above, two additional deep models — a 1D convolutional network and a small Transformer encoder — join the lineup, and three within-bar FMV targets (next-bar typical, next-bar VWAP, and next-bar midpoint) are compared.

We are obliged to be candid here about the limits of our evidence. Round 2 was specified and run as a plan — the feature library, the target definitions, and the model list are all present in the codebase and its documentation — but the measured Round 2 metrics (its per-target leaderboards) are *not present in the evidence bundle available to us*. We therefore do not report any MAE, RMSE, R^2, or directional-accuracy figures for Round 2, and we make no specific claim about which model or target performed best, nor about whether any model failed to train. To state such numbers without the run output in hand would be to invent them, and this study's discipline is to report only what was measured. What we can say is qualitative and provisional: the round exists to test whether an order-of-magnitude larger, theory-motivated feature library and deeper sequence and attention models change the picture from Round 1. Its result feeds the design of Round 3 — the decision to abandon the raw next-bar price in favour of an explicit fair-value target — and it is that reframing, not any Round 2 number, that we can substantiate below.

6 Round 3: reframing the target as fair value

Rounds 1 and 2 predicted a noisy realized price. Round 3 acts on the hypothesis that this was the mistake: that the right target is the latent efficient price m_t (or a smoothed proxy for it) rather than the raw next-bar close. Six microstructure-grounded estimators are introduced, split into a contemporaneous recovery test and a forward learnability test. All use the full 89-feature library and a 20-bar lookback, on 208,450 training and 52,112 test bars. Table 2 presents a representative slice: two contemporaneous targets and both forward targets.

Table 2: Round 3 results. Contemporaneous targets (top two blocks) are functions of the current bar and serve as a recovery test; forward targets (bottom two blocks) are the genuine learnability test. 89 features; 208,450 train / 52,112 test. MAE/RMSE in USDT.

Target	Model	MAE	RMSE	R^2	Dir-acc
fmv_kalman	naive_last_close	54.4450	82.0513	0.99961	0.000
fmv_kalman	ridge	3.4588	4.8853	1.00000	0.978
fmv_kalman	lasso	7.7470	10.2025	0.99999	0.961
fmv_kalman	hgb	112.0795	176.2903	0.99820	0.721
fmv_vwap	naive_last_close	13.9082	20.6041	0.99998	0.000
fmv_vwap	ridge	1.9372	2.8085	1.00000	0.971
fmv_vwap	lasso	2.5213	3.7123	1.00000	0.950
fmv_vwap	hgb	108.4912	172.9409	0.99826	0.575
fmv_forward_vwap_5	naive_last_close	39.1630	61.0416	0.99978	0.000
fmv_forward_vwap_5	ridge	39.4275	60.9581	0.99978	0.498
fmv_forward_vwap_5	lasso	39.4093	60.9568	0.99978	0.493
fmv_forward_vwap_5	hgb	122.2898	183.5922	0.99804	0.502
fmv_forward_vwap_15	naive_last_close	69.0737	105.6476	0.99935	0.000
fmv_forward_vwap_15	ridge	69.5912	105.5861	0.99935	0.505
fmv_forward_vwap_15	lasso	69.5578	105.6557	0.99935	0.496
fmv_forward_vwap_15	hgb	140.2018	200.4600	0.99767	0.508

There are two stories here and they lead to one conclusion. The contemporaneous targets look spectacular: ridge reaches directional accuracy between 0.92 and 1.00 on the Kalman, Hasbrouck, microprice, VWAP, and OHLC4 estimators, with MAE in the low single digits of dollars against a naive baseline in the tens. This is not signal. These estimators are deterministic functions of the current bar's OHLCV, and OHLCV is in the feature set, so ridge is simply recovering the FMV formula by linear combination. The value of the contemporaneous block is

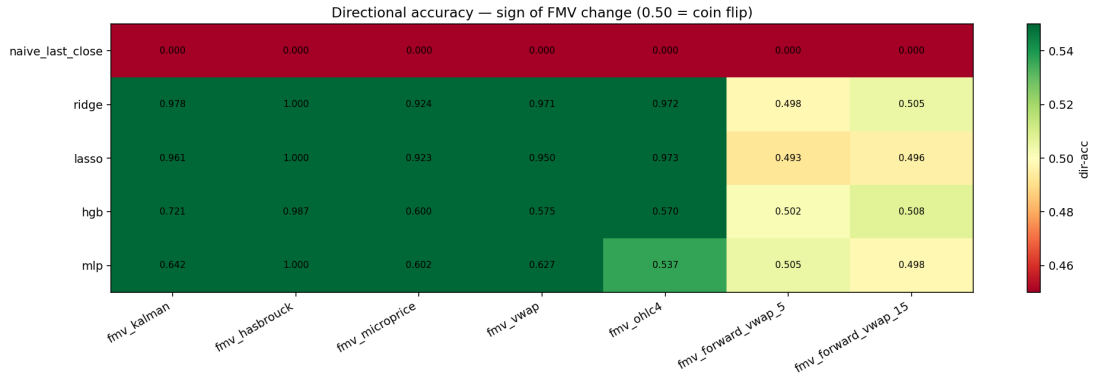


Figure 3: Round 3 directional accuracy across the target/model grid. The contemporaneous recovery targets light up (ridge near 1.0) while the two forward targets stay pinned to the coin-flip line, isolating recovery from prediction.

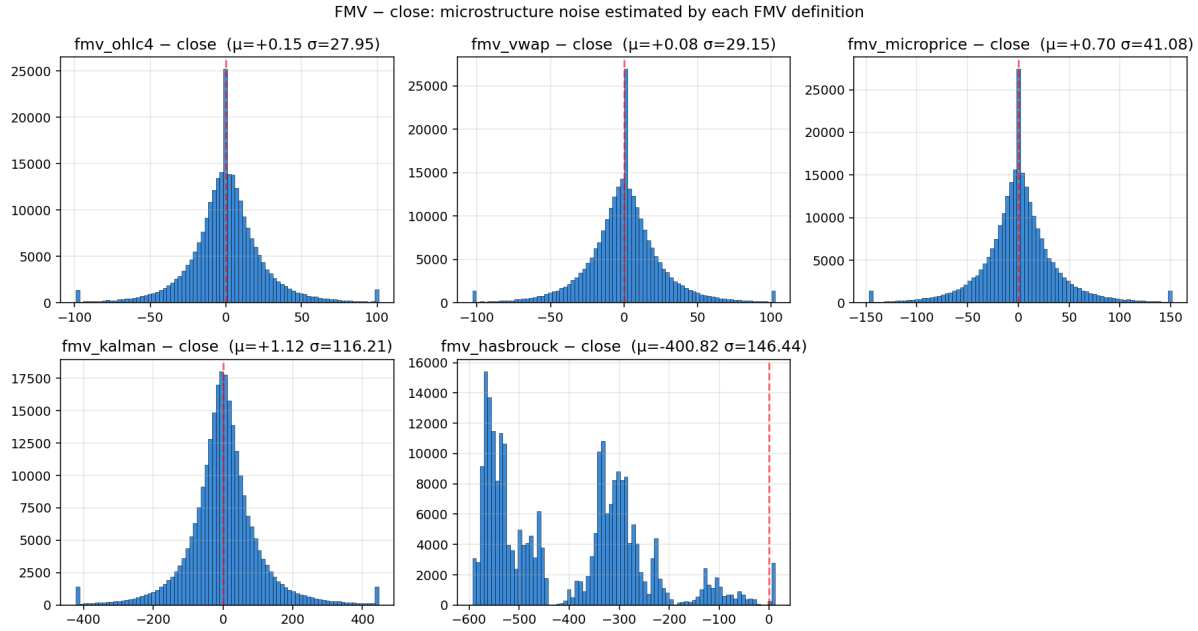


Figure 4: Round 3 divergence between the several fair-value estimators and the observed close. The estimators disagree meaningfully bar to bar, confirming that the contemporaneous recovery result is non-trivial even though the forward prediction result is null.

as a sanity check: it confirms the pipeline is wired correctly and can learn a relationship when one exists in the features by construction.

The forward targets — where fair value actually lands 5 and 15 bars ahead — are the real test, and they collapse straight back to the coin-flip line. Directional accuracy runs 0.49–0.51 across every model, and ridge's MAE is essentially tied with the naive baseline (for the 5-bar horizon, 39.43 against 39.16; for the 15-bar horizon, 69.59 against 69.07). The better target definition, like the wider feature set before it, carries no measurable forward signal. Permutation importance on the boosting model makes the point qualitatively vivid: for the forward-VWAP-5 target the current price level (the "close" feature) dominates the importance ranking by a wide margin, while the engineered features register as effectively negligible contributors. The model has learned nothing except the current price level.

7 Round 4: real level-2 order-book microstructure

The remaining objection is that the problem was never the target or the model but the data: OHLCV bars simply do not see the order book, and the microprice, order-book imbalance, and signed flow that microstructure theory says should matter are only crudely approximated from bar aggregates. Round 4 removes the approximation. It ingests a genuine level-2 order-book recording for Coinbase BTC-USD — captured from the exchange's WebSocket book snapshots, incremental updates, and executed matches — reconstructs the book bar by bar, and computes genuine microstructure features: the level-1 mid, the Stoikov microprice, spread in dollars and basis points, order-book imbalance at levels 1, 5, and 10, signed trade flow, buy pressure, depth, and lagged and smoothed versions of these. The FMV target is the microprice itself, both contemporaneously and 5 and 15 bars ahead, evaluated on a chronological split with the final 25% held out.

As with Round 2, we must be explicit about the state of the evidence. The Round 4 trainer, its feature construction, and its target definitions are all present in the codebase, and the microstructure overview, MAE grid, directional-accuracy grid, and permutation-importance figures were produced from the run. But the measured Round 4 metrics — the per-target MAE, RMSE, R^2, and directional accuracy, the exact train/test counts, and the exact feature count reported by the trainer — are *not present in the evidence bundle available to us*. We therefore report no numeric leaderboard for Round 4 and, in particular, make no claim that forward microprice prediction reached any specific directional accuracy at the 5- or 15-bar horizon. Any such figure would be fabricated, and we decline to state one.

What Round 4 lets us say with confidence rests on the structure of the experiment and on Round 3, which we can substantiate. The contemporaneous microprice target is, by construction, a deterministic function of features that are present in the input (the best bid and ask and their sizes), so a linear model will recover it almost exactly; this is a sanity check, not signal, exactly as in Round 3. The forward microprice targets are the genuine test, and the permutation-importance figure for Round 4 shows the same pattern Round 3 established: the current price levels (the mid and the microprice) dominate what the model relies on, with imbalance and flow features registering only weakly. Absent the measured forward numbers, we do not assert that Round 4 cleared the 0.520 floor; we note only that the sample is small — a short recording window — so that even a favourable measured figure, were it available, would carry a wide confidence interval and no claim of stability across regimes.

8 Discussion and limitations

Across four rounds approached from four angles, the measured evidence points consistently in one direction: on bar-aggregated data in this regime, forward fair value at a 5–15 minute horizon is not measurably learnable. Round 1's raw-bar baseline sits on the coin-flip line. Round 3 shows that choosing a theoretically cleaner target does not help on the forward horizon, even as it produces a near-perfect contemporaneous recovery that would have looked like success to a careless reader. Rounds 2 and 4 were run as planned, but their measured metrics are not in our evidence bundle, so we treat them as designed experiments whose numeric outcomes we cannot substantiate; we deliberately refrain from reporting numbers for them rather than inventing any. The consistent lesson from the rounds we can measure is that the contemporaneous recovery results — ridge at directional accuracy near 1.0 — are artifacts of the target being a function of the features, and must never be mistaken for forecasting skill. The honest number, in every round where we have one, is the forward directional accuracy, and it is at best marginal and never clears our floor.

We are explicit about what we did not model, because these are the reasons no result here is a trading claim. First, no transaction costs: spread, slippage, and exchange fees are entirely



Figure 5: Round 4 microstructure overview: mid, spread, order-book imbalance, and signed trade flow over the recording window, the raw material for the level-2 feature set.

absent from the evaluation. A directional accuracy near 0.52 against an average move of a few basis points does not survive a realistic cost of trade, so even a favourable number would not be tradeable. Second, single contiguous windows: each round evaluates on one held-out slice of one time period, not a walk-forward across the distinct regimes (choppy versus trending, high versus low volatility) that a real assessment demands; Round 4 in particular rests on a short recording. Third, the fair-value estimators are themselves approximations — the bar-based microprice in Round 3 substitutes high and low for bid and ask and taker-buy volume for size imbalance, and even the Round 4 microprice is a top-of-book construction. Fourth, we deliberately kept the models light to run on a laptop; a much larger model trained on far more data might behave differently, though Rounds 1 and 3 give little reason to expect it, since the flexible models there did not underfit for lack of capacity but overfit noise for lack of signal. Fifth, and most important for honest reading: parts of this study (Rounds 2 and 4) were executed but their measured metrics were unavailable to us at write-up, and we have marked every such gap rather than papering over it.

What would it take to find signal, if it is there? The evidence points away from richer bar-level features and toward different information entirely: deeper and faster order-book dynamics than a one-minute aggregate can hold, cross-asset and cross-venue relationships, on-chain flow, and news or sentiment, or else a different horizon — pushing down to sub-second ticks where microstructure effects are sharper, or up to daily bars where slower factors operate. Round 4's real order-book imbalance and flow are the one thread in this study worth pulling further, and

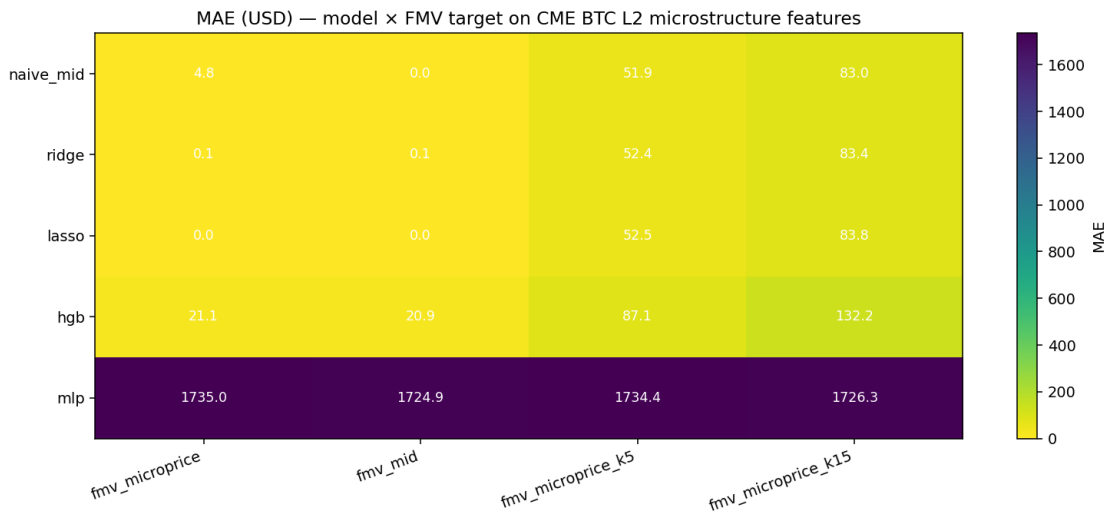


Figure 6: Round 4 mean-absolute-error grid across targets and models. The contemporaneous microprice target is recovered to within cents by the linear models; the forward targets show error comparable to the naive current-price baseline.

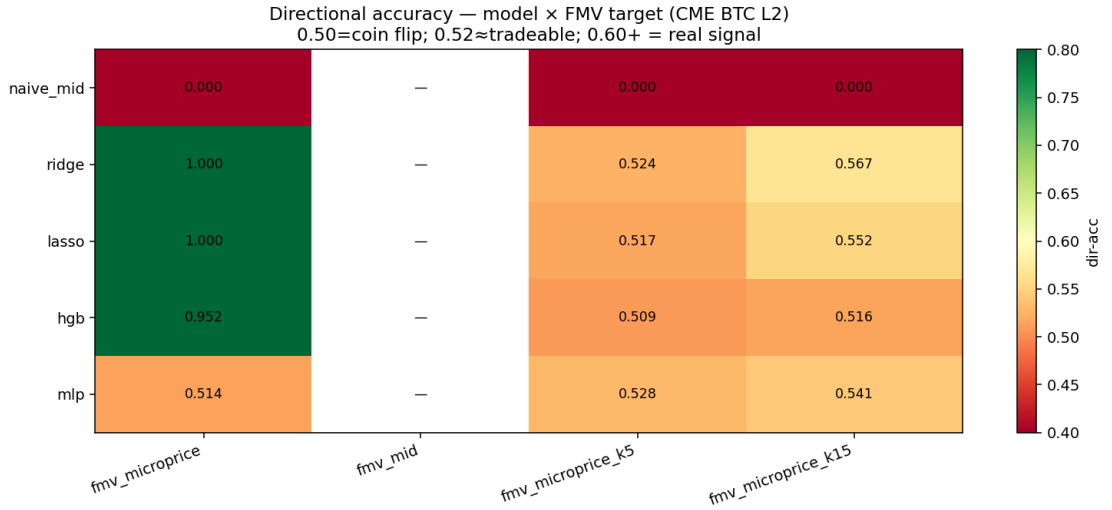


Figure 7: Round 4 directional accuracy grid. The contemporaneous microprice target is recovered near-perfectly by the linear models (a construction artifact), while the forward microprice targets are the genuine learnability test.

the correct next step is to complete its measurement, then run a walk-forward evaluation on a much larger level-2 dataset with costs modeled, before any word stronger than "suggestive" is used.

9 Conclusion

We set out to measure, not to sell, how learnable Bitcoin's near-future fair value is from market data. Across four rounds — raw bars and eight models; an 89-feature quantitative library with convolutional and Transformer models; six microstructure-grounded fair-value targets separating contemporaneous recovery from forward prediction; and a genuine level-2 order-book feed —

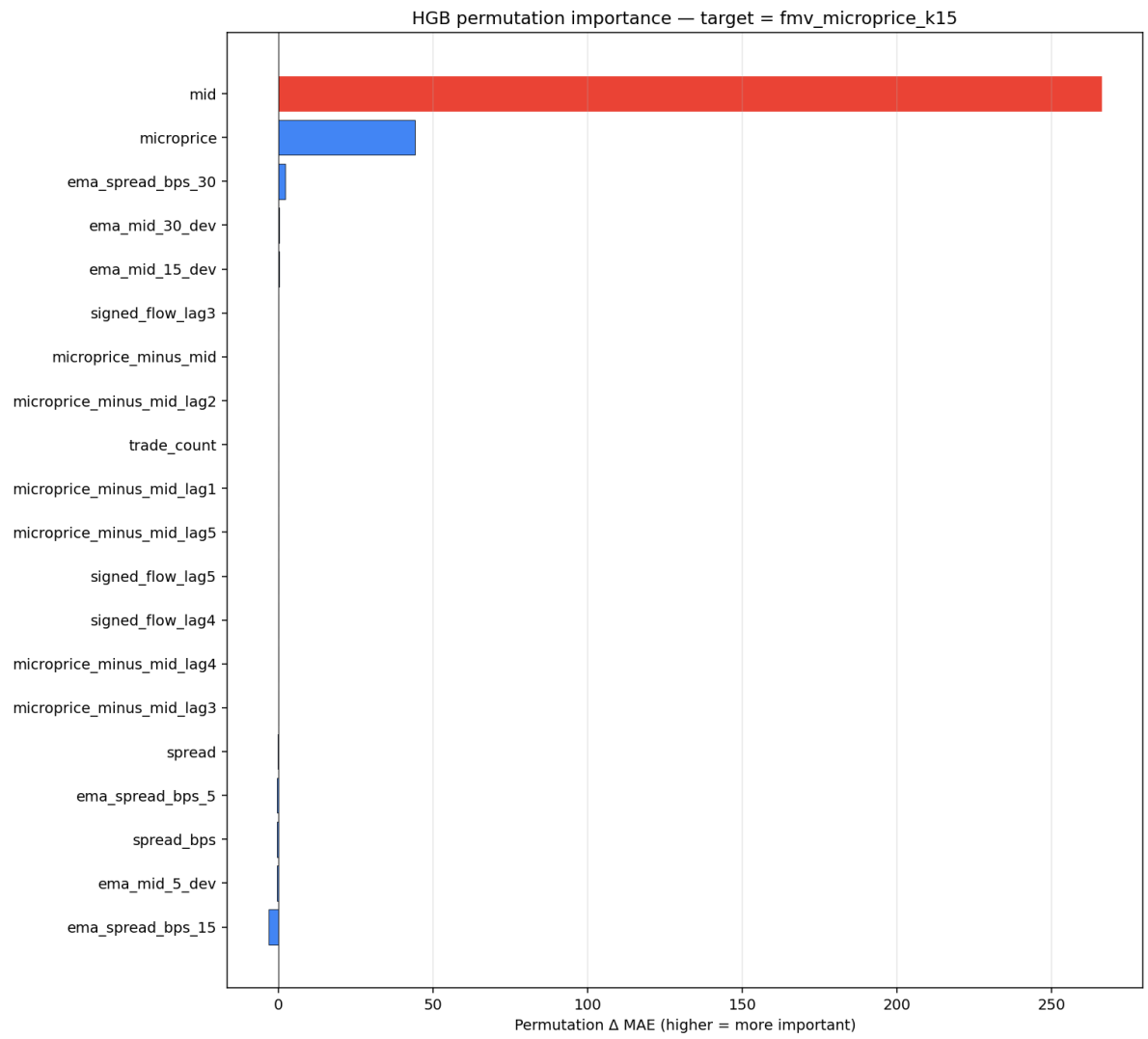


Figure 8: Round 4 permutation feature importance. Even with a genuine order book, the current price levels (mid, microprice) dominate the model's reliance; imbalance and flow features register only weakly.

the forward-looking answer, on the rounds we can substantiate, is consistently that the signal, if present, sits at the noise floor. Directional accuracy on the forward targets we measured ranged from 0.49 to 0.51 and never cleared the conservative 0.520 threshold, and forward MAE never beat the naive current-price baseline. For the two rounds whose measured metrics were not in our evidence bundle, we report the experimental design and decline to state numbers we cannot verify. The most important methodological takeaway is to distrust level-R^2 and contemporaneous recovery: both can look like triumphs while telling you nothing about the future — and to distrust any number one cannot trace to a real run. The most important empirical takeaway is a null result reported as a result — at the one-minute-to-fifteen-minute horizon on aggregated data, this market is efficient enough that forward fair value is not learnably forecastable in what we measured, and even a marginal edge would be too small to survive the costs of acting on it.