Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The leakage premium in decomposition-based cryptocurrency forecasting

Ozan Nadirgil

onadirgil@proton.me

Independent https://orcid.org/0009-0003-5584-0488

Research Article

Keywords: Cryptocurrency; Forecasting; Signal decomposition; CEEMDAN; VMD; Look-ahead bias; Data leakage

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11529129/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

The leakage premium in decomposition-based cryptocurrency forecasting

Ozan Nadirgil

Abstract

Decomposition-ensemble models, which split a price series into components with methods such as CEEMDAN and VMD and forecast each component with a neural network, report remarkable accuracy for cryptocurrency returns. This paper shows that most of that accuracy is look-ahead. The decompositions are two-sided, so components computed on the full sample carry future prices into every input. A typical pipeline that stacks VMD on top of CEEMDAN is run on daily prices of five major crypto-assets from 2018 to 2026 under three protocols that differ only in the information each decomposition uses, with data, networks and random seeds held fixed. With the full-sample decomposition the model calls the direction of next-day returns correctly on 93 per cent of days and explains 96 per cent of their variance. Computed causally, the same model is right on 52 per cent of days and has an out-of-sample R^2 of −23 per cent, worse than a forecast of zero. Look-ahead accounts for essentially the entire premium in directional accuracy, a premium that grows with each added decomposition layer and hardly moves across market regimes. A second mechanism explains why causal decomposition forecasts are harmful rather than merely uninformative. The revisions of end-point components cancel across components in the target but not in the forecasts. A long-flat rule driven by the leaky forecasts would show phantom profits of about 146 basis points a day net of costs. A simple falsification test, proposed here, detects look-ahead in any such pipeline.

Keywords: Cryptocurrency; Forecasting; Signal decomposition; CEEMDAN; VMD; Look-ahead bias; Data leakage

JEL classification: C45; C53; C58; G14; G17

1. Introduction

Daily cryptocurrency returns are hard to forecast. Their linear autocorrelations are small, their volatility is high, and studies that evaluate machine-learning forecasts of their direction out of sample report accuracies only modestly above a coin toss (Jaquart et al., 2021). A second and much larger literature tells another story. Models that first split the price series into components, with CEEMDAN, VMD or a related method, forecast each component with a neural network and then add the forecasts back together routinely report directional accuracies above 70 or 80 per cent, with errors a fraction of those of a naive no-change forecast. The two bodies of evidence cannot both describe the same market.

This paper argues that, to a large extent, they do not. The decomposition methods at the heart of these models are two-sided. The value of a component on a given day depends on prices observed after that day, so when the whole sample is decomposed before it is split into an estimation and an evaluation period, as it commonly is, every input the forecasting model sees carries a little of the future. Hydrologists noticed this a decade ago (Zhang et al., 2015; Du et al., 2017; Quilty and Adamowski, 2018), and recent work has started to measure it in other settings (Yang et al., 2024). Cryptocurrency forecasting is where the decomposition-ensemble approach has spread furthest. Yet, to the author's knowledge, the size of the effect has not been measured there, and its sources have not been taken apart.

This paper measures it. It builds the kind of multi-layer pipeline that is typical of this literature, CEEMDAN followed by permutation-entropy grouping and a second decomposition with VMD, with a separate neural network per component, and applies it to daily prices of five major crypto-assets from May 2018 to September

2026. The pipeline is then run under three protocols that differ in one respect only, the information used to compute each quantity. P0 decomposes the full sample once, as most published studies do. P2 reads the training set from a decomposition of what is known at each re-estimation date and reads every forecast input from a decomposition that ends on the forecast date, so that no forecast uses the future. P3 goes a step further. Each training input comes from a decomposition that ends on the day it describes, which means the model is trained on exactly the kind of input it is later asked to use. The data, the networks, their settings and their random seeds are the same throughout. A falsification test confirms that only the P0 forecasts move when prices after the forecast date are altered.

The results are stark. Under P0, the two-layer model applied to Bitcoin appears to explain 96 per cent of the variation in next-day returns, and it calls their sign correctly on 92 days out of a hundred. No plausible model of an actively traded asset could do that. Under P3 the same model, on the same days, is right on 53 days out of a hundred, and its out-of-sample R^2 is −26 per cent, so in squared error it does worse than a forecast of zero. The other four assets tell the same story. Across the five, directional accuracy at depth 2 averages 93.2 per cent under P0 and 52.0 per cent under P3. The model confidence set of Hansen et al. (2011) keeps the zero forecast, the historical mean and a one-layer linear benchmark for every asset, and it excludes the causal decomposition models in 38 of 40 cases.

Several further findings refine this picture. The premium behaves like a dose-response relationship in the depth of decomposition. It is zero by construction when nothing is decomposed, which provides a placebo check on the bookkeeping. With one layer the premium is 15.6 percentage points of directional accuracy, with two it is 41.2. The second layer, which the literature adds because it raises accuracy, owes about 96 per cent of that gain to the extra leakage it brings in. The three-protocol design also separates the premium into two parts. In directional accuracy look-ahead accounts for essentially all of it, and the effect of training on interior rather than end-point values is too small to detect. In squared error the picture changes, because a model trained on interior values and fed end-point values at forecast time produces forecasts that are far too large for the returns they are meant to predict. Across five market regimes between the COVID-19 crash and 2026 the premium in directional accuracy hardly moves, and volatility and liquidity explain almost none of its variation once the depth of decomposition is accounted for. That points to a property of the method, not of the market.

The causal forecasts are worse than a zero forecast for a reason that is easy to miss. Under P3 each component network is trained on the change between two successive end-points of its component, and most of that change is a revision of yesterday's value. The component targets are therefore partly predictable, and component by component the forecasts look good. But the revisions cancel across components, because every decomposition adds up to the same price, and so must whatever is predictable about them. The errors of separately trained networks do not cancel. Their sum is, to a close approximation, noise.

The economic stakes are easy to state. A trader who followed the P0 forecasts with a simple long/flat rule would have expected to earn about 146 basis points a day net of a ten-basis-point transaction cost, against 8 for the same rule driven by P3 forecasts. That difference is phantom alpha. It exists on paper and could never have been earned. The P0 rule keeps beating buy-and-hold up to one-way costs of 200 to 350 basis points, far above

any fee an exchange charges. The P3 rule stops beating it at 2.4 basis points at most, well inside the range that a rule without any skill reaches by luck.

The paper speaks to three literatures. To the large and still growing work on decomposition-based forecasting of cryptocurrency prices it offers a causal benchmark for reading existing results, together with a measure of how much of the reported accuracy survives it. For the methodological work on leakage in machine-learning research (Kaufman et al., 2012; López de Prado, 2018; Kapoor and Narayanan, 2023), it adds a channel that runs through pre-processing rather than through the sample split. It also shows how to separate the information effect from the train-test consistency effect that comes bundled with it whenever leakage is removed. And it bears on the evidence about how predictable cryptocurrency returns are (Urquhart, 2016; Liu and Tsyvinski, 2021), because it shows that the extraordinary accuracy of decomposition-based models is no evidence against the near-unpredictability of daily returns found elsewhere.

The argument is not that decomposition has no value in forecasting. It is narrower. Whatever value decomposition has must be established under a protocol that a forecaster could actually have followed. To make that easy, the code for every step is released with the paper, the falsification test included. The test takes minutes to run and is worth adding to any study that puts a two-sided transformation in front of a forecasting model.

Section 2 reviews related work. Section 3 introduces the sample, and Section 4 the decomposition, the three protocols and the forecasting models. Section 5 explains how the forecasts are evaluated. Section 6 reports the findings, Section 7 discusses what they imply and where they stop, and Section 8 concludes.

2. Related literature

2.1. Decomposition-ensemble forecasting

The idea of forecasting a difficult series by first splitting it into simpler parts is old in signal processing. Its modern form in economics goes back to Yu et al. (2008), who split crude oil prices into intrinsic modes with EMD, the empirical mode decomposition of Huang et al. (1998), forecast each mode with a neural network and summed the forecasts. They called this a decomposition-ensemble learning paradigm. The template has changed remarkably little since then. What has changed is the decomposition. EMD gave way to its ensemble variants, EEMD (Wu and Huang, 2009) and CEEMDAN (Torres et al., 2011), which reduce mode mixing by averaging over noise-perturbed copies of the signal, and to variational mode decomposition (Dragomiretskiy and Zosso, 2014), which extracts band-limited modes by solving an optimisation problem instead of sifting. Multi-layer designs combine the two, typically by applying VMD a second time to the noisiest CEEMDAN components, and entropy measures, most often permutation entropy (Bandt and Pompe, 2002), decide which components are grouped or decomposed again.

Cryptocurrency prices have become a popular testing ground, and Table 1 summarises five recent studies. Between them they use daily, hourly and minute data and pair VMD, CEEMDAN, ensemble patch transforms (EPT) and SSA (singular spectrum analysis) with recurrent, convolutional and tree models. All conclude that decomposition, alone or combined with other devices, improves accuracy. Wang et al. (2026) make the

broadest claim, reporting that VMD improves all seven classes of model they try. The reported accuracy is often striking. Li et al. (2022) call the direction of the next price move correctly 81.7 per cent of the time, and on the price level Jin and Li (2023) and Tsai et al. (2026) report average absolute errors below one per cent.

Table 1 records, for each of these studies, the asset and data frequency, the decomposition, the forecasting model, whether accuracy is measured on prices or on returns, the accuracy reported, and whether the decomposition is applied before or after the evaluation sample is set aside. The last of these matters most for the argument developed here, and it is also the one that methods sections most often leave implicit.

Table 1. Decomposition-based forecasting studies of crypto-asset prices.

Study	Data	Decomposition	Forecasting model	Target	Accuracy reported	Decomposition window
Li et al. (2022)	BTC, daily, 2013-2021; minute data in a robustness check	VMD, 11 modes	BiLSTM	Normalised price	MSE, RMSE, MAE, MAPE, DA; best DA 81.7%	Not stated; whole series implied
Jin and Li (2023)	BTC and ETH, daily, 2017-2020	VMD, residual decomposed again	Attention GRU, LSTM combiner	Price	RMSE, MAE, MAPE; MAPE 0.39% for BTC	Not stated
Koo and Kim (2024)	BTC	SSA	LSTM with input filtering	Price	RMSE	Not stated
Tsai et al. (2026)	BTC, hourly, 2019-2023	CEEMDAN, VMD, EPT and combinations	LSTM, BiLSTM, TCN	Price	MAPE, 0.07% to 0.76% by year	Not stated; whole series implied
Wang et al. (2026)	BTC	VMD	Seven model classes, linear regression to CNN	Price	RMSE, MAE, MAPE; DA	Not stated

BiLSTM, bidirectional LSTM; TCN, temporal convolutional network; DA, directional accuracy; MAPE, mean of absolute percentage errors. The last column records what each paper says about the window that is decomposed, that is, whether the decomposition is computed on the full sample before the evaluation period is set aside or only on data available at each forecast date.

2.2. Look-ahead in two-sided transformations

Leakage, the use at estimation time of information that would not be available when the model is used, is among the most common reasons why machine-learning results fail to replicate. Kaufman et al. (2012) give a formal definition and a taxonomy. Kapoor and Narayanan (2023), surveying machine-learning-based research across many scientific fields, find leakage to be a pervasive source of over-optimistic results. Finance knows the problem under other names. López de Prado (2018) shows how overlapping labels leak information across cross-validation folds and proposes purging and embargoing, and Bailey et al. (2014) show how easily a strategy selected on a backtest can be made to look profitable.

Signal decomposition opens a less obvious channel. CEEMDAN, VMD, wavelet transforms and SSA are all two-sided, in the sense that the value of a component at time t depends on observations after t. A component computed on the full sample therefore contains a little of the future at every date, and a model trained and evaluated on such components is partly forecasting something it has already been shown. Hydrologists were the first to spell this out. Zhang et al. (2015) asked whether hybrid models with data pre-processing are suitable for monthly streamflow forecasting at all. Du et al. (2017) showed that singular-spectrum and wavelet hybrids had commonly been built in a way that lets future observations enter the inputs, and that their apparent advantage depends on it. Quilty and Adamowski (2018) then set out best practices for wavelet-based forecasting in real-world settings. More recently, Yang et al. (2024) quantified the effect for EMD-based models and found that

forecast errors rise substantially once the leakage is removed, and Feng et al. (2026) proposed a sample-wise VMD that is free of leakage by construction.

Three questions remain open, and this paper takes them up. The first is how large the premium is in cryptocurrency markets, where the decomposition-ensemble literature is now large and the reported accuracy unusually high. The second is what the premium consists of. Removing leakage changes two things at once. The forecast no longer sees the future, and the model is no longer trained on the same kind of input it is evaluated on. Existing comparisons set a leaky design against a leakage-free one and attribute the whole difference to leakage, whereas the three-protocol design used here keeps the two effects apart. The third question is whether the premium is a fixed property of the method or moves with market conditions and with the depth of decomposition, and what it would have been worth to an investor who believed it.

2.3. How predictable are cryptocurrency returns?

Any claim about forecast accuracy has to be read against how much predictability there is to be found. The evidence on the informational efficiency of Bitcoin is mixed. It points to modest, time-varying departures from random-walk behaviour that have shrunk as the market has matured (Urquhart, 2016; Nadarajah and Chu, 2017; Bariviera, 2017). In the cross-section, Liu and Tsyvinski (2021) find that cryptocurrency returns are predicted by time-series momentum and by investor attention but not by the factors that price equities, and Lee and Wang (2025) find that components of return variance help to predict cryptocurrency returns. Studies that forecast the direction of daily or intraday returns by machine learning, and evaluate the forecasts out of sample on returns, report accuracies only modestly above 50 per cent (Akyildirim et al., 2021; Jaquart et al., 2021).

The forecasting literature supplies a second set of reference points. Campbell and Thompson (2008) show that an out-of-sample R^2 of a fraction of one per cent against the historical mean can matter economically to an investor, and Gu et al. (2020) report monthly out-of-sample R^2 values below one per cent for the best machine-learning models of equity returns. Makridakis et al. (2018) find that standard statistical methods outperform a range of machine-learning methods on series from the M3 competition. Zeng et al. (2023) show that a one-layer linear model matches or beats far more elaborate Transformer architectures on standard benchmarks. And when many models are compared on the same data, the best among them will look good by chance. White (2000), Hansen (2005) and Hansen et al. (2011) develop tests that allow for this, and Harvey et al. (2016) argue that the bar for significance in empirical finance should be raised accordingly.

Set against these benchmarks, a daily directional accuracy above 80 per cent, or an error variance a tenth of that of a zero forecast, implies a degree of predictability that is very hard to reconcile with everything else that is known about these markets. That tension is where this paper starts.

3. Data

3.1. Sample

The sample covers Bitcoin (BTC), Ether (ETH), BNB, XRP and Cardano (ADA), five large and liquid crypto-assets quoted against Tether (USDT) on Binance. Prices come from Binance's public market-data archive (data.binance.vision), which publishes the exchange's own candlestick files at fixed addresses. The files were

downloaded in October 2026. Working with a single venue keeps the five series on one clock and under one set of trading rules. The archive is also static, so anyone can rebuild the sample.

The daily sample runs from 5 May 2018 to 30 September 2026, which gives 3,071 observations per asset. The start date is dictated by the data. XRP/USDT began trading partway through 4 May 2018, so 5 May is the first complete day on which all five pairs exist. Solana was considered as well, but its USDT pair starts only in August 2020, and including it would have shortened the common sample by 829 days, or 27 per cent. A longer history was judged worth more than a sixth asset. Daily bars close at 00:00 UTC. Because these markets trade around the clock, every calendar day is a trading day, and the daily series contain no gaps.

The robustness exercise in Section 6.8 also uses four-hour BTC/USDT bars from the same archive, 18,426 in all. Nine of them are absent. Each of the nine is an exchange-wide maintenance window between June 2018 and February 2020, missing for every pair at the same timestamp, and each is filled with the previous close and zero volume, the usual way of recording an interval without trades. None falls inside the window used by the four-hour experiment. As a check on the data, the four-hour bars were aggregated to days, and the resulting open, high, low, close and volume reproduce Binance's daily file exactly on all 3,071 days.

The out-of-sample period begins on 1 January 2020 and contains 2,465 daily forecast targets per asset. The 606 days before it, from 5 May 2018 to 31 December 2019, serve as the initial estimation window and are the only data used to choose the tuning constants of the decomposition (Section 4.2). Starting the evaluation in January 2020 puts the March 2020 crash, the bull market of 2020 and 2021, the collapse of Terra and FTX in 2022, the spot-ETF approvals of 2024 and everything since inside the evaluation sample. That spread of market conditions is what makes the regime analysis of Section 6.4 possible.

3.2. Descriptive statistics

Table 2 summarises daily log returns. Three of their features matter for what follows. Returns are volatile and heavy-tailed. Annualised volatility ranges from 62 per cent for BTC to 98 per cent for ADA, excess kurtosis lies between 10 and 24, and the Jarque-Bera test rejects normality for every asset at any conventional level. Each series has a single day on which the log price fell by more than 0.50, a loss of roughly 40 per cent. For four of the five assets that day is 12 March 2020. For XRP it is 23 December 2020, the day after the US Securities and Exchange Commission sued Ripple Labs.

There is little linear structure in returns and a great deal in their squares. First-order autocorrelations are small and negative, between −0.04 and −0.07. The Ljung-Box statistic on ten lags rejects white noise for four of the five assets, XRP being the exception, but the predictability it implies is small. Squared returns, in contrast, are strongly autocorrelated, and ARCH-LM tests reject conditional homoskedasticity everywhere.

Returns are stationary, log prices are not. Augmented Dickey-Fuller tests fail to reject a unit root in the log price, with p-values between 0.46 and 0.78, and reject it decisively in returns.

None of this is new, and it is reported mainly to set expectations. On a price this close to random-walk behaviour, with returns that carry little linear memory, an honest one-day-ahead forecast should find it hard to improve on a forecast of zero. That benchmark, and the margin by which published decomposition models claim to beat it, is what this paper is about.

Table 2. Descriptive statistics of daily log returns, 5 May 2018 to 30 September 2026.

Statistic	BTC	ETH	BNB	XRP	ADA
Panel A. Moments and quantiles					
Mean (%)	0.070	0.039	0.129	0.016	−0.013
Standard deviation (%)	3.27	4.37	4.38	5.11	5.16
Annualised volatility (%)	62.4	83.5	83.6	97.7	98.5
Skewness	−1.20	−0.97	−0.37	0.59	0.13
Excess kurtosis	21.2	14.5	24.0	18.7	10.3
Minimum (%)	−50.3	−59.1	−58.2	−53.9	−53.3
1st percentile (%)	−9.3	−12.8	−11.5	−13.2	−13.0
Median (%)	0.04	0.06	0.12	−0.04	−0.09
99th percentile (%)	9.4	11.7	11.8	16.2	15.1
Maximum (%)	17.8	23.4	53.2	54.9	54.3
First-order autocorrelation	−0.066	−0.073	−0.073	−0.041	−0.056
Panel B. Tests (p-values in parentheses)					
Jarque-Bera	58,246 (< 0.001)	27,405 (< 0.001)	73,907 (< 0.001)	45,093 (< 0.001)	13,640 (< 0.001)
Ljung-Box Q(10), returns	33.3 (< 0.001)	44.3 (< 0.001)	52.7 (< 0.001)	14.0 (0.172)	25.5 (0.004)
Ljung-Box Q(10), squared returns	57.3 (< 0.001)	108.3 (< 0.001)	350.9 (< 0.001)	162.3 (< 0.001)	155.4 (< 0.001)
ARCH-LM(10)	51.4 (< 0.001)	89.0 (< 0.001)	237.0 (< 0.001)	118.7 (< 0.001)	126.2 (< 0.001)
ADF, log price (p-value)	0.780	0.754	0.623	0.462	0.651
ADF, log return (p-value)	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Mean daily dollar volume (USD million)	1863	956	237	269	128

Daily log returns in per cent, 3,070 per asset. Q(10) is the Ljung-Box statistic with ten lags and ARCH-LM(10) the Lagrange multiplier test for conditional heteroskedasticity with ten lags. The augmented Dickey-Fuller regressions include a constant, with the lag length chosen by the Akaike criterion. Dollar volume is the daily close times the daily base-asset volume.

4. Methodology

4.1. Forecasting problem

Let $x_t = \ln P_t$ denote the log closing price on day $t = 0,1,\ldots,T$, with $T = 3{,}070$. The object of interest is the h-day log return forecast with the information available at the close of day t. Every forecast in the paper targets this quantity, and every accuracy measure is computed on it. The point deserves emphasis because much applied work reports accuracy on the price level, where a forecast that simply repeats yesterday's price achieves an R^2 close to one. On returns, that same forecast is the benchmark to beat.

$$y_t^{(h)} = x_{t+h} - x_t, \qquad h \in \{1,5,10\},$$

Forecasts are produced in an expanding window. The first forecast origin is $t_0 = 31$ December 2019, and the models are re-estimated every $R = 20$ days, at refit dates $t_k = t_0 + 20k$. At refit date t_k the training set consists of all samples $s \in \{s_{\min},\ldots,t_k - h\}$, so every training target has been observed by t_k. The index $s_{\min} = 127$ marks the end of the shortest window considered long enough to decompose (128 observations). The estimated models then forecast every origin from t_k to $t_k + 19$ before the next refit. For $h = 1$ this gives 124 refits and 2,465 forecasts per asset and model.

4.2. Decomposition

CEEMDAN: The first layer is the complete ensemble empirical mode decomposition with adaptive noise of Torres et al. (2011), which extends EMD (Huang et al., 1998) and its ensemble version, EEMD (Wu and Huang, 2009). Applied to a window $x_a,\ldots,x_b$, it returns IMFs (intrinsic mode functions) ordered from the highest frequency to the lowest, the last one being the residual trend, and these sum exactly to the input. The

ensemble comprises 50 noise realisations, and the noise amplitude is set to $\varepsilon_0 = 0.2$ times the standard deviation of the current residual, the amplitude that Wu and Huang (2009) recommend for the ensemble method and that Torres et al. (2011) use. Sifting relies on cubic splines and on the default stopping rule of the PyEMD implementation. The noise seed is fixed, so the decomposition of any given window is deterministic.

Grouping by permutation entropy: How many IMFs the decomposition returns depends on the length of the window, from five for 128 observations to nine or ten for the full sample. That is awkward for a forecasting model that must see the same inputs at every refit, and, following common practice, IMFs of similar complexity are merged into a few co-IMFs. Complexity is measured by the normalised permutation entropy of Bandt and Pompe (2002), computed for every IMF with embedding dimension $m = 4$ and delay $\tau = 1$, so that the 24 possible ordinal patterns stay few compared with the shortest window (see Riedl et al., 2013, on this trade-off). The scalar entropies are clustered with k-means, and the clusters are ordered by decreasing mean entropy, so that co-IMF 0 collects the most irregular IMFs and co-IMF $k-1$ the smoothest.

How many groups to form is not fixed in advance. Instead k is picked from the values 2 to 4 by the silhouette criterion (Rousseeuw, 1987), on the pre-sample window alone, 5 May 2018 to 31 December 2019. The criterion selects $k = 2$ for all five assets, with silhouette values between 0.70 and 0.81 against 0.51 to 0.64 for $k = 3$ (Table 3). The structure is the same everywhere. The first two IMFs have entropies of about 0.98, close to the value of one that white noise would give, and form co-IMF 0, while the remaining IMFs, the trend included, form co-IMF 1. Three groups, a common choice in applied work, are not supported by the data on any asset. Once k is fixed, the entropies and the grouping are recomputed inside every window that is decomposed, so the assignment of IMFs to co-IMFs never uses data from outside the window.

VMD of the irregular group. The second layer applies VMD (Dragomiretskiy and Zosso, 2014) to co-IMF 0, the component that multi-layer designs treat as too noisy to forecast directly. VMD needs a number of modes K and a bandwidth penalty α, and both are again chosen on the pre-sample window only. K follows a centre-frequency rule. It is raised step by step until the centre frequency of the highest mode changes by less than 2 per cent from K to $K + 1$, the point at which additional modes start to split an existing band rather than resolve a new one. Given K, the penalty α is chosen from $\{100,250,500,1000,2000,4000\}$ to minimise the sum of the residual energy ratio and the index of orthogonality (Huang et al., 1998), which trades reconstruction fidelity against mode mixing. Table 3 reports the selections. K lies between 3 and 5 and α between 100 and 500.

Two implementation details turn out to matter for a causal design.[1] VMD does not reconstruct its input exactly, and the reconstruction error is largest at the edges of the window, which is precisely where a real-time forecaster reads its inputs. The VMD residual is therefore carried as an explicit component, so that the components still sum exactly to the price. The second detail is less visible. The widely used Python port of VMD (vmdpy) silently drops the last observation of any input of odd length, and in an expanding window this would make every second forecast rest on a decomposition that ends one day before the forecast origin. The first observation is dropped instead, which leaves the end-point intact.

[1] Neither detail is visible in aggregate results, and both are easy to miss. The replication package contains a unit test for each.

Depth: Three levels of decomposition are compared. At depth 0 nothing is decomposed and the model sees lagged returns. At depth 1 it sees the two co-IMFs. At depth 2 it sees the K VMD modes of co-IMF 0, their residual and co-IMF 1, between five and seven components depending on the asset.

Table 3. Tuning constants of the decomposition, selected on the pre-sample window.

Asset	IMFs	Permutation entropy of IMF 1 to IMF n	Silh. k = 2	k = 3	k = 4	k^*	K^*	α^*	RER (%)	OI
BTC	7	0.98 0.98 0.54 0.38 0.29 0.25 0.22	0.81	0.64	0.53	2	4	250	3.25	0.129
ETH	6	0.98 0.98 0.54 0.38 0.30 0.20	0.79	0.57	0.44	2	5	500	4.16	0.137
BNB	7	0.99 0.98 0.56 0.40 0.30 0.22 0.00	0.70	0.52	0.46	2	4	250	3.30	0.152
XRP	7	0.98 0.96 0.55 0.37 0.27 0.22 0.00	0.71	0.51	0.49	2	3	100	1.54	0.131
ADA	6	0.98 0.98 0.56 0.38 0.30 0.18	0.76	0.57	0.51	2	4	250	3.40	0.135
BTC (4h)	7	0.99 0.99 0.59 0.41 0.32 0.26 0.18	0.75	0.59	0.46	2	5	500	4.58	0.154

Pre-sample window 5 May 2018 to 31 December 2019 (606 daily observations); for the four-hour BTC series, the 1,024 bars that end at the first four-hour forecast origin (Section 6.8). The permutation entropy (order 4, delay 1) is normalised to the unit interval; the final IMF is the residual trend. Silh. is the silhouette value of the k-means grouping of the entropies. K* is the count of VMD modes chosen by the centre-frequency rule and α* the bandwidth penalty that minimises RER + OI, where RER is the residual energy ratio and OI the index of orthogonality.

4.3. Three decomposition protocols

Write $c_j^{[a,b]}(s)$ for the value at time s of component j in the decomposition of the window $x_a,\dots,x_b$, and $\Delta c_j^{[a,b]}(s) = c_j^{[a,b]}(s) - c_j^{[a,b]}(s-1)$ for its first difference inside that decomposition. The input for component j at time s is the vector of its last $L = 10$ differences,

$$\mathbf{z}_j^{[a,b]}(s) = \left(\Delta c_j^{[a,b]}(s-L+1),\dots,\Delta c_j^{[a,b]}(s)\right),$$

and each component gets a network of its own. The forecast of the return is the sum of the component forecasts, as in the decomposition-ensemble paradigm of Yu et al. (2008). The three protocols differ only in the decomposition from which each quantity is read, and Fig. 1 sets them side by side.

P0 (full sample): Training inputs $\mathbf{z}_j^{[0,T]}(s)$, training targets $c_j^{[0,T]}(s+h) - c_j^{[0,T]}(s)$ and forecast inputs $\mathbf{z}_j^{[0,T]}(t)$ all come from a single decomposition of the full sample $x_0,\dots,x_T$. Most published hybrid models are built this way. CEEMDAN and VMD are two-sided operations, so the component values at t depend on prices after t, and so does every forecast.

P2 (periodic re-decomposition): At each refit date the estimation window $x_0,\dots,x_{t_k}$ is decomposed once, and the training inputs $\mathbf{z}_j^{[0,t_k]}(s)$ and targets $c_j^{[0,t_k]}(s+h) - c_j^{[0,t_k]}(s)$ are read from that decomposition. On each forecast day t the series up to that day is decomposed, and the network is fed the end-point inputs $\mathbf{z}_j^{[0,t]}(t)$. No forecast uses a price observed after its origin. The training inputs, however, are interior values of a two-sided decomposition, whereas the inputs seen at forecast time are end-point values, which are noisier and are revised as new data arrive.

P3 (fully causal): Every quantity is read from a decomposition that ends on the date it describes. The training input for sample s is $\mathbf{z}_j^{[0,s]}(s)$, its target is the change between successive end-points, $c_j^{[0,s+h]}(s+h) - c_j^{[0,s]}(s)$, and the forecast input is $\mathbf{z}_j^{[0,t]}(t)$, as in P2. P3 needs one decomposition for every day in the sample, 2,944 per asset, all of which are computed once and stored.

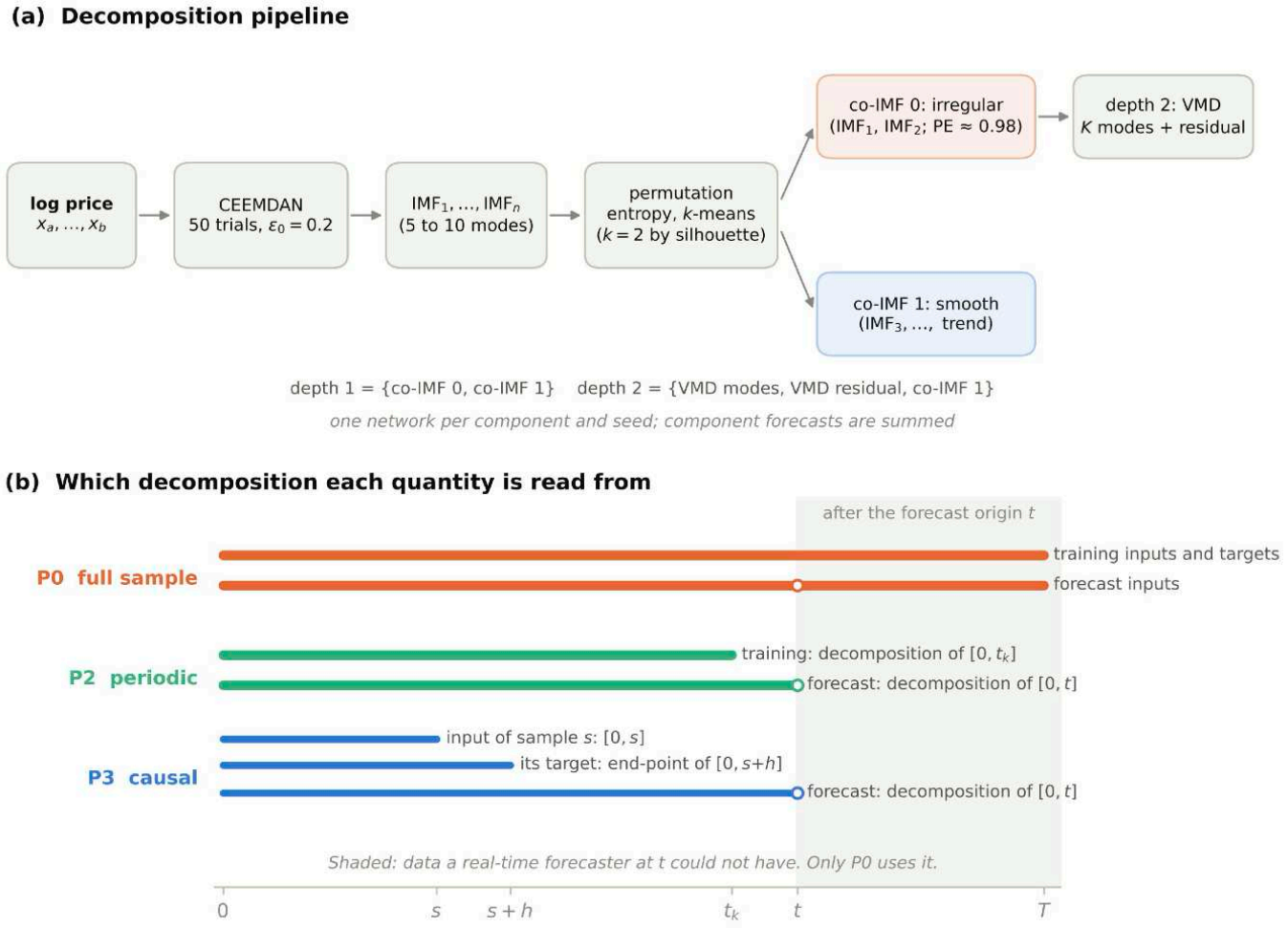


Fig. 1. The three decomposition protocols. Each row shows which observations enter the decomposition from which a training input, a training target and a forecast input are read. Under P0 a single decomposition of the full sample is used throughout. Under P2 the training set comes from a decomposition that ends at the refit date and each forecast input from a decomposition that ends on the forecast date. Under P3 every quantity comes from a decomposition that ends on the date it refers to.

The components of every decomposition sum to the price at every date, so the component targets sum to the realised return $y_s^{(h)}$ under all three protocols. The networks always forecast the same quantity, and only the information set changes. Depth 0 involves no decomposition and is therefore identical under the three protocols.

This design makes it possible to split the gain that the literature attributes to decomposition. For any accuracy measure A oriented so that larger is better, define the *leakage premium* $\Pi = A(\mathrm{P0}) - A(\mathrm{P3})$ and write

$$\Pi = \underbrace{A(\mathrm{P0}) - A(\mathrm{P2})}_{\text{look-ahead}} + \underbrace{A(\mathrm{P2}) - A(\mathrm{P3})}_{\text{train-test consistency}} .$$

The first term is the value of information that no forecaster could have had, the leakage in the narrow sense. The second term measures what happens when the training set is built from the same kind of end-point values that the model will see in use. Its sign is not obvious in advance. Consistent training may help the network learn the distribution it will face, or the noisier end-point inputs may simply be harder to learn from.

4.4. Forecasting models and benchmarks

Networks: The main model is a feed-forward network (DNN) with two hidden layers of 64 and 32 units, ReLU activations and dropout of 0.2 after the first layer. To check that the results do not hinge on the

architecture, an LSTM (Hochreiter and Schmidhuber, 1997) with one hidden layer of 32 units and a linear read-out is estimated as well. Inputs and targets are standardised component by component, with moments computed on the fitting part of the training window. The networks minimise squared error with Adam (Kingma and Ba, 2015), a learning rate of 10^{-3} and mini-batches of 32. Training stops once the loss on the last 10 per cent of the training window, held out in chronological order, has failed to improve for ten epochs (with a ceiling of 100), and the weights from the best epoch are restored.

These settings were fixed before any data were seen and do not vary across assets, depths or protocols. Hyperparameters are deliberately left untuned for two reasons. Tuning would introduce a second source of variation across protocols, and any difference in accuracy could then reflect the tuning budget instead of the information set. And a search over network configurations on the evaluation sample would itself be a form of data snooping (White, 2000; Hansen et al., 2011), precisely the problem the evaluation design is meant to avoid.

Every configuration is estimated with ten random seeds, which control the initial weights, the order of the mini-batches and the dropout masks. The same seeds are used under P0, P2 and P3, so comparisons between protocols are paired. Accuracy is reported for the ensemble forecast, the average over seeds, together with its spread across individual seeds. The DNN is re-estimated every 20 days as described above. The LSTM costs about seven times as much per training epoch, so it is re-estimated every 120 days and used for $h = 1$ only.[2]

Horizons: For $h = 5$ and $h = 10$ two strategies are compared. The *direct* strategy trains separate networks on h-day component targets. The *recursive* strategy iterates the one-day networks, appending each predicted difference to the input window and summing the predicted changes over h steps.

Benchmarks: The simplest benchmark, RW, sets every forecast to zero, which is what random-walk log prices imply, and it serves as the denominator of Theil's inequality coefficient (Section 5.1). The historical mean, the expanding average of realised h-day returns, is the benchmark of the out-of-sample R^2 of Campbell and Thompson (2008). Two learned benchmarks use the same ten lagged returns, refit schedule and seeds as the networks at depth 0. DLinear (Zeng et al., 2023) splits the input window into a moving-average trend (kernel 5) and a remainder and maps each to the forecast with a linear layer. It is included because one-layer linear models of this kind have been shown to match or beat far more elaborate architectures on standard forecasting benchmarks. LightGBM (Ke et al., 2017) is a gradient-boosted tree ensemble, included to separate gains from non-linearity from gains specific to neural networks.

4.5. A falsification test

A protocol is causal if its forecasts cannot change when the data after the forecast origin change. This property can be tested directly. A price series is simulated, every decomposition and every forecasting protocol is run on it, the series is altered after a date t^*, and the whole pipeline is run again. For all causal configurations the forecasts at every origin $t \leq t^*$ are identical to the last bit in the two runs. That holds at depth 0, under P2 and

---

[2] All seed-by-component networks of a configuration are trained together as one batched model in which every member has its own parameters, initialisation, mini-batch order and early-stopping counter. Batched training was checked against separate training of each member and reproduces its forecasts to within 3×10^{-7}, with identical stopping epochs.

P3, for both networks, for both benchmarks and for the recursive forecasts. Under P0 the same forecasts change, by up to 0.025 in log-return units at the one-day horizon. The test also confirms that the component targets sum to the realised return to within 10^{-14} under every protocol. It is cheap, and any study that places a two-sided transformation in front of a forecasting model would gain from running it.

5. Evaluation

5.1. Accuracy

All measures are computed on the h-day log return over the 2,465 out-of-sample days, or slightly fewer for $h > 1$. Absolute accuracy is summarised by the root mean squared error (RMSE) and the mean absolute error, complemented by three relative measures that are easier to compare across assets. Theil's inequality coefficient U divides the model's RMSE by that of RW, the zero forecast, so values below one favour the model. The out-of-sample R^2 of Campbell and Thompson (2008),

$$R^2_{OS} = 1 - \frac{\sum_t (y_t^{(h)} - \hat{y}_t^{(h)})^2}{\sum_t (y_t^{(h)} - \bar{y}_t^{(h)})^2},$$

measures the gain over the historical mean $\bar{y}_t^{(h)}$, the average of all h-day returns realised by day t. DA, the directional accuracy, is the share of days on which the forecast has the sign of the realised return. It is tested against the null of no directional skill with the statistic of Pesaran and Timmermann (1992), which allows for the fact that a forecast that is always positive will be right more often than half the time in a rising market.

Each measure is reported for the ensemble forecast, the average over the ten seeds, and also as the mean and standard deviation across seeds. In practice a user would trade on the ensemble, and the spread across seeds shows how much of an apparent difference between two configurations could come from initialisation alone.

5.2. Model confidence set

When dozens of configurations are evaluated on the same days, the best among them will look good by chance. The comparisons are therefore summarised with the model confidence set (MCS) of Hansen et al. (2011), which contains the best model with a given probability while allowing for the multiplicity of comparisons. The loss is the squared forecast error. The test uses the range statistic and the stationary bootstrap (Politis and Romano, 1994) with 5,000 replications, with a block length equal to the median of the optimal lengths of Politis and White (2004), as corrected by Patton et al. (2009), across models and never shorter than the forecast horizon. The tables report the 90 per cent set.

The set is built from the forecasts that a user could actually have produced. These are RW, the historical mean, DLinear, LightGBM, the networks at depth 0, and the networks at depths one and two under P2 and P3. P0 forecasts were never feasible and are left out. The exercise is then repeated with them included, to show which conclusions a reader of the leaky results would have drawn.

5.3. The leakage premium

The premium of Section 4.3 is computed for the out-of-sample R^2, for DA and for $1-U$, three measures on which larger values are better. For each asset, depth, network and horizon the tables report the total premium $\Pi = A\,(\mathrm{P0}) - A\,(\mathrm{P3})$ and its two parts, the look-ahead component $A\,(\mathrm{P0}) - A\,(\mathrm{P2})$ and the consistency component $A\,(\mathrm{P2}) - A\,(\mathrm{P3})$. At depth 0 the three protocols coincide and the premium is zero by construction, which is a useful check on the bookkeeping.

5.4. Regimes

The evaluation period spans very different markets. Rather than by calendar year, it is split at turning points of the Bitcoin price that can be read off the data. The five regimes are listed in Table 4. A forecast belongs to the regime of its target date. A Markov-switching classification was considered, but turning points are preferred because they are transparent, involve no estimated object and correspond to episodes that readers will recognise.

Table 4. Market regimes in the evaluation period.

Regime	Period	Description
R1	1 Jan 2020 to 30 Apr 2020	COVID-19 crash (BTC low on 12 March 2020) and rebound
R2	1 May 2020 to 8 Nov 2021	Bull market up to the BTC closing high of 8 November 2021
R3	9 Nov 2021 to 21 Nov 2022	Bear market, including the Terra/Luna and FTX collapses, down to the BTC closing low of 21 November 2022
R4	22 Nov 2022 to 31 Dec 2024	Recovery, the US spot-ETF approvals and the April 2024 halving
R5	1 Jan 2025 to 30 Sep 2026	Most recent period

Turning points are BTC/USDT daily closes. A forecast belongs to the regime of its target date.

5.5. What drives the premium?

If the premium is a mechanical property of the decomposition, it should be much the same everywhere. If it depends on the data, it should vary with the properties of each series. To distinguish the two, the premium is regressed on market characteristics,

$$\Pi_{a,r,d} = \beta_0 + \beta_1\,\sigma_{a,r} + \beta_2 \ln V_{a,r} + \beta_3\,1\,[d = 2] + \varepsilon_{a,r,d},$$

where $\Pi_{a,r,d}$ is the premium for asset a in regime r at depth $d \in \{1,2\}$, $\sigma_{a,r}$ is the annualised volatility of daily returns on asset a in regime r, and $V_{a,r}$ is its average daily dollar volume, a proxy for liquidity. Five assets, five regimes and two depths give 50 observations. OLS estimates are reported with heteroskedasticity-robust (HC3) standard errors, with and without asset fixed effects.

5.6. Economic value

Accuracy is translated into money with the simplest rule a forecast supports. The investor holds the asset on day $t + 1$ if the one-day forecast made on day t is positive and holds cash otherwise. Each change of position costs c per unit traded. Results are reported for c between 0 and 20 basis points, a range that spans the standard spot fees of large exchanges over the sample period, with 10 basis points as the central case. Performance is measured by the annualised return, the Sharpe ratio (with a zero risk-free rate) and the maximum drawdown, and compared with buy-and-hold.

Two quantities summarise the comparison. The *break-even cost* is the one-way cost at which the rule's Sharpe ratio falls to that of buy-and-hold. A rule without any timing skill will beat buy-and-hold by luck in some samples, so each break-even cost is compared with a permutation null. The rule's positions are shuffled across days, which preserves the share of time spent in the market but breaks any link with future returns, and the break-even cost is recomputed 500 times. The *phantom alpha* is the difference in mean daily net return between the rule driven by P0 forecasts and the same rule driven by P3 forecasts, at the central cost. It is stated per day, because the annualised figures it implies are too large to be meaningful. It is the return a reader of the published accuracy would have expected, and it could not have been earned.

6. Results

6.1. What the inputs know

Before turning to forecasts, it is instructive to examine the inputs themselves. Panel (a) of Fig. 2 asks how stable the real-time value of a component is. For every day t, the end-point value of co-IMF 0 in the decomposition of $x_0,\ldots,x_t$ is compared with the value that the same day receives once the window has grown by k more days. Because the two co-IMFs sum to the known price, any revision of one is exactly offset by the other, and a single curve per asset describes how the split between the smooth and the irregular part of today's price changes as the future arrives. For BTC, a single additional day revises the split by 0.72 standard deviations of daily returns on average, five days by 1.08, and after that the revision hardly grows. The full-sample value, the one that P0 uses, lies 1.07 standard deviations away from the real-time value. The other assets look much the same, with five-day revisions between 0.90 and 1.09. In other words, the real-time value of the irregular component is uncertain by about as much as a typical daily return, and the uncertainty is resolved only by data that a forecaster does not yet have.

Panel (b) shows what this means for prediction. For every depth-2 component of every asset it plots the correlation between the latest input, the last first difference of the component, and the next day's return. Under P0 the correlations reach about 0.5 in absolute value for every asset and average about 0.3. Under P3 the largest of them is 0.09, and the averages range from 0.03 to 0.05. The full-sample inputs are, in effect, partly made of tomorrow's price. The causal inputs carry far less, though not nothing; the small amount they do carry is taken up in Section 7.3.

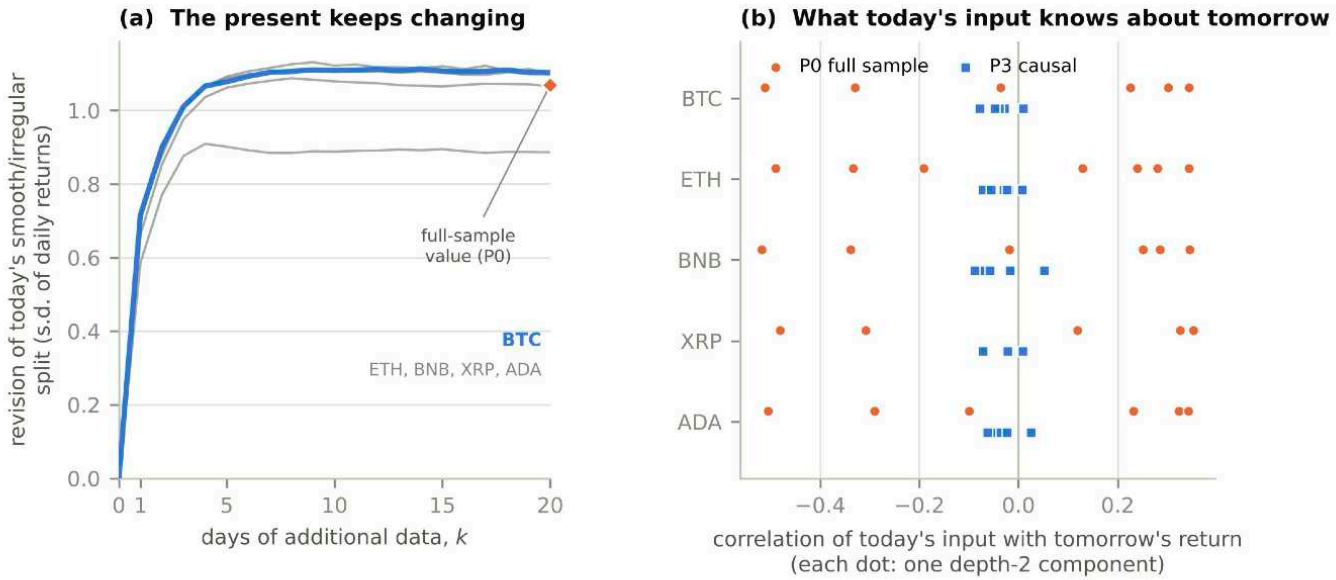


Fig. 2. What the inputs know. (a) Root mean square revision of the real-time split between the two co-IMFs of day t, as the window is extended by k further days, in units of the standard deviation of daily returns; the diamond marks the distance between the real-time value and the full-sample value for BTC. (b) Correlation between the latest input of each depth-2 component and the next day's return, for the full-sample (P0) and the causal (P3) decomposition.

6.2. Accuracy under the three protocols

Table 5 reports one-day-ahead accuracy averaged over the five assets, and Fig. 3 shows the asset-level values behind each average. Table S1 in the Supplementary Material gives every measure for every asset. Consider the benchmarks first. DLinear, LightGBM and the DNN without decomposition all have a U within 0.007 of one and an out-of-sample R^2 within 1.4 percentage points of zero, and none of them calls the direction of the next day's return correctly on more than 52.3 per cent of days. For none of these fifteen forecasts does the Pesaran-Timmermann test reject the absence of directional skill. This is what the descriptive statistics of Section 3 would lead one to expect, and it is the yardstick for everything that follows.

Table 5. One-day-ahead forecast accuracy.

Model	U	R^2_{OS} (%)	DA (%)	PT rejections at 5%
Random walk	1.000 [1.000, 1.000]	0.11 [−0.02, 0.22]		
Historical mean	1.001 [1.000, 1.001]	0.00 [0.00, 0.00]	49.5 [47.5, 52.3]	0/5
DLinear	1.005 [1.003, 1.007]	−0.81 [−1.40, −0.39]	51.2 [49.9, 52.3]	0/5
LightGBM	1.001 [0.999, 1.004]	−0.16 [−0.63, 0.46]	50.6 [49.0, 52.3]	0/5
DNN, depth 0	1.003 [0.999, 1.006]	−0.47 [−0.97, 0.31]	50.5 [48.9, 51.9]	0/5
DNN, depth 1, P0	0.851 [0.845, 0.858]	27.61 [26.31, 28.75]	66.5 [65.5, 67.2]	5/5
DNN, depth 1, P2	1.374 [1.324, 1.413]	−88.62 [−99.35, −75.33]	51.3 [49.8, 52.6]	1/5
DNN, depth 1, P3	1.066 [1.039, 1.089]	−13.49 [−18.65, −7.84]	50.8 [49.3, 51.7]	1/5
DNN, depth 2, P0	0.201 [0.193, 0.205]	95.95 [95.82, 96.28]	93.2 [91.7, 95.0]	5/5
DNN, depth 2, P2	1.085 [1.062, 1.124]	−17.66 [−26.13, −12.51]	50.9 [50.0, 51.9]	2/5
DNN, depth 2, P3	1.108 [1.069, 1.136]	−22.73 [−28.70, −14.12]	52.0 [51.4, 53.2]	3/5

Out-of-sample period 1 January 2020 to 30 September 2026, 2,465 forecasts per asset. Each cell gives the mean over the five assets and, in brackets, the lowest and highest asset value. U is Theil's inequality coefficient relative to RW, R^2_{OS} the out-of-sample R^2 against the historical mean, and DA the directional accuracy on days with a non-zero return. The last column counts the assets for which the one-sided Pesaran-Timmermann statistic is significant at the 5 per cent level. All learned models are ensembles of ten seeds. The sign of a zero forecast is undefined, so no DA is reported for RW.

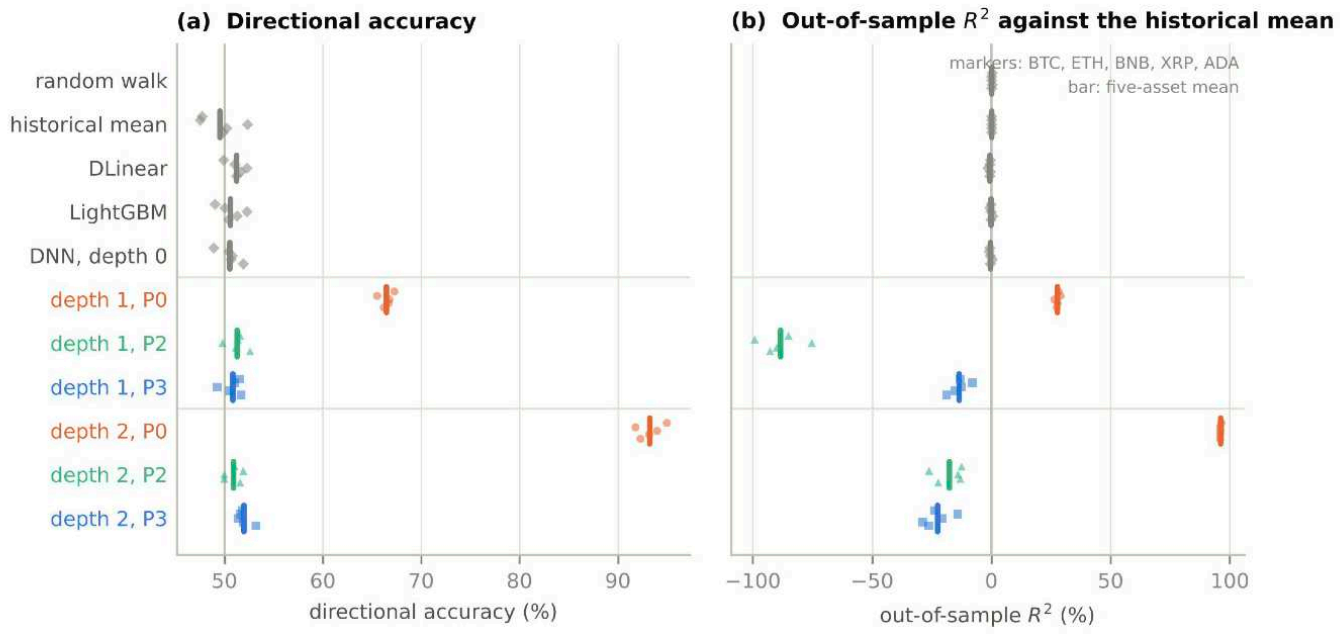


Fig. 3. One-day-ahead accuracy under the three protocols. Each small marker is one asset and each bar the five-asset mean. (a) Directional accuracy. (b) Out-of-sample R² against the historical mean.

Under P0 the decomposition models reach a level that no benchmark comes close to. At depth 1 the DNN reaches an average U of 0.85, an R^2 of 27.6 per cent and 66.5 per cent directional accuracy. These figures would not look out of place in the published literature, and that is part of the problem, because nothing about them signals that anything is wrong. At depth 2 they become 0.20, 96.0 per cent and 93.2 per cent. The five assets behave almost identically. At depth 2 directional accuracy lies between 91.7 and 95.0 per cent and R^2 between 95.8 and 96.3 per cent, a uniformity across markets as different as Bitcoin and Cardano that is itself a hint that the accuracy comes from the method and not from the markets.

None of this survives the move to a protocol that a forecaster could actually have run. Under P2 and P3 the R^2 of the decomposition models is negative for every asset and at both depths, between −99.4 and −7.8 per cent, and U exceeds one in every case, so the causal forecasts are worse than a forecast of zero. Directional accuracy falls back to 51.3 per cent on average. It is significantly above one half in 7 of the 20 causal configurations, and its best value, 53.2 per cent for BTC at depth 2 under P3, is about what careful machine-learning studies of daily returns report. In direction, then, the causal models are at best a little ahead of the benchmarks; the depth-2 model under P3 averages 52.0 per cent, about one point above DLinear. In squared error they are clearly worse.

Table 6 turns these numbers into a premium, and Table S2 gives it asset by asset. In directional accuracy the premium is 15.6 percentage points at depth 1 and 41.2 at depth 2, against zero by construction at depth 0, so it grows with each decomposition layer. The second layer raises directional accuracy by 26.7 points under P0 and by 1.2 points under P3, which means that about 96 per cent of what it appears to add is leakage. The look-ahead component accounts for essentially the whole premium. The consistency component is small, 0.5 points at depth 1 and −1.1 at depth 2, and for no asset is the difference in directional hits between P2 and P3 statistically significant, the smallest McNemar p-value being 0.13 (Table S6). In direction it therefore makes no difference whether the networks are trained on interior or on end-point values.

Table 6. The leakage premium and its two components, one-day horizon.

Measure	P0	P2	P3	Premium	Look-ahead	Consistency
Directional accuracy, depth 1	66.5	51.3	50.8	15.6	15.2	0.4
Directional accuracy, depth 2	93.2	50.9	52.0	41.2	42.3	−1.1
R^2_{OS}, depth 1	27.6	−88.6	−13.5	41.1	116.2	−75.1
R^2_{OS}, depth 2	96.0	−17.7	−22.7	118.7	113.6	5.1
$1-U$, depth 1	14.9	−37.4	−6.6	21.5	52.3	−30.8
$1-U$, depth 2	79.9	−8.5	−10.8	90.7	88.4	2.3

Accuracy under each protocol and the premium P0 minus P3, split into look-ahead (P0 minus P2) and train-test consistency (P2 minus P3), in percentage points, averaged over the five assets (asset-level values in Table S2 of the Supplementary Material). At depth 0 the three protocols coincide and the premium is zero by construction.

In squared error the split looks different. P2, which trains on interior values and then forecasts from end-point values, is heavily penalised for the mismatch at depth 1. Its average R^2 there is −88.6 per cent, against −13.5 per cent under P3, and the Diebold-Mariano test rejects equal accuracy in favour of P3 for all five assets. Hence the look-ahead component of the R^2 premium, 116 points, is larger than the premium itself, 41 points, and the consistency component is large and negative. At depth 2 the two causal protocols are closer, with R^2 of −17.7 and −22.7 per cent. Training on inputs unlike those seen at forecast time is a problem in its own right, separate from look-ahead, and a protocol that merely stops using future prices does not resolve it.

6.3. Accurate components, noisy sums

Why are the causal forecasts not merely uninformative but worse than a forecast of zero? Table 7 looks inside them, with the asset-level values in Table S3. Under P3 the target for component j is the change between two successive end-points, $c_j^{[0,t+1]}(t+1) - c_j^{[0,t]}(t)$, and this change has two parts. One is the change inside the new window, $c_j^{[0,t+1]}(t+1) - c_j^{[0,t+1]}(t)$. The other is the revision of yesterday's end-point value, $c_j^{[0,t+1]}(t) - c_j^{[0,t]}(t)$. The revisions are large. They account for 66 per cent of the sum of squared component targets at depth 1 and for 87 per cent at depth 2. At depth 2 the networks fit these targets well. On average they explain 36 per cent of the out-of-sample variation of their own targets, and for the highest-frequency VMD mode, a nearly regular oscillation with a period of about two and a half days, they explain more than 80 per cent for every asset. A study that reported accuracy component by component would conclude that the causal pipeline works.

It does not, because the revisions sum to zero across components. Every decomposition reproduces the price at every date, so whatever is revised in one component is revised in the opposite direction in the others. Much of what is predictable about the component targets is therefore predictable only component by component, and it has to cancel when the forecasts are added up. Separately trained networks do not deliver forecasts that cancel in this way. Their sum has a standard deviation of 37 per cent of that of the return at depth 1 and 51 per cent at depth 2, while its correlation with the return is 0.00 and 0.04. For a forecast whose standard deviation is a fraction s of that of the return and whose correlation with it is ρ, the out-of-sample R^2 is approximately $2\rho s - s^2$. With the values in Table 7 this gives −13.5 per cent at depth 1 and −22.6 per cent at depth 2, almost exactly the R^2 of Table 5. The causal forecasts are, to a first approximation, noise of a known size.

A simple check confirms that the trouble lies in the scale of the forecasts and that little information hides behind it. Each forecast is rescaled by the slope of a regression of past returns on past forecasts, estimated at

every origin from data available at the time. The R^2 of every causal configuration then rises to between −2.4 and −0.1 per cent. Rescaled, the causal decomposition forecasts are about as good as the historical mean and no better. The P0 forecasts, by contrast, hardly change when rescaled, because they were never noisy to begin with.

Table 7. Component forecasts and return forecasts, one-day horizon.

Model	Revision share of component targets (%)	Component R^2 (%)	sd(forecast) / sd(return)	corr(forecast, return)	R^2_{OS} (%)	R^2_{OS} after rescaling (%)
Depth 1, P0	0.0	58.2	0.49	0.525	28.2	28.01
Depth 1, P2	66.4	−20.3	0.93	−0.010	−90.9	−0.58
Depth 1, P3	66.4	7.5	0.37	−0.002	−14.9	−0.57
Depth 2, P0	0.0	92.3	0.89	0.985	96.5	97.16
Depth 2, P2	86.7	10.9	0.42	0.000	−17.7	−0.66
Depth 2, P3	86.7	36.1	0.51	0.038	−26.6	−0.49

The revision share is the part of the sum of squared component targets that is due to revising the preceding end-point value (zero under P0, where targets come from one decomposition). Component R² is the out-of-sample R² of each component forecast against its own realised target, averaged over components. The next two columns compare the sum of the component forecasts, the return forecast, with the realised return. The last two columns give the out-of-sample R² of the return forecast before and after a causal rescaling by the slope of a regression of past returns on past forecasts, both on the origins after a burn-in of 250 days. Means over the five assets; asset-level values in Table S3 of the Supplementary Material.

6.4. Market regimes and the drivers of the premium

If the premium came from some property of the markets, it would rise and fall with them. In direction it does not. Fig. 4(a) shows the premium in each of the five regimes, and Table S4 lists it by asset. Averaged over the assets it ranges from 13.6 to 16.9 percentage points at depth 1 and from 38.3 to 42.5 at depth 2. The COVID-19 crash, by far the most volatile episode, is the regime with the smallest premium at both depths, not the largest. Panel (b) sets the 50 asset-regime premia against each asset's annualised volatility within the regime, which ranges from 44 to 148 per cent, and the two depths form two flat bands.

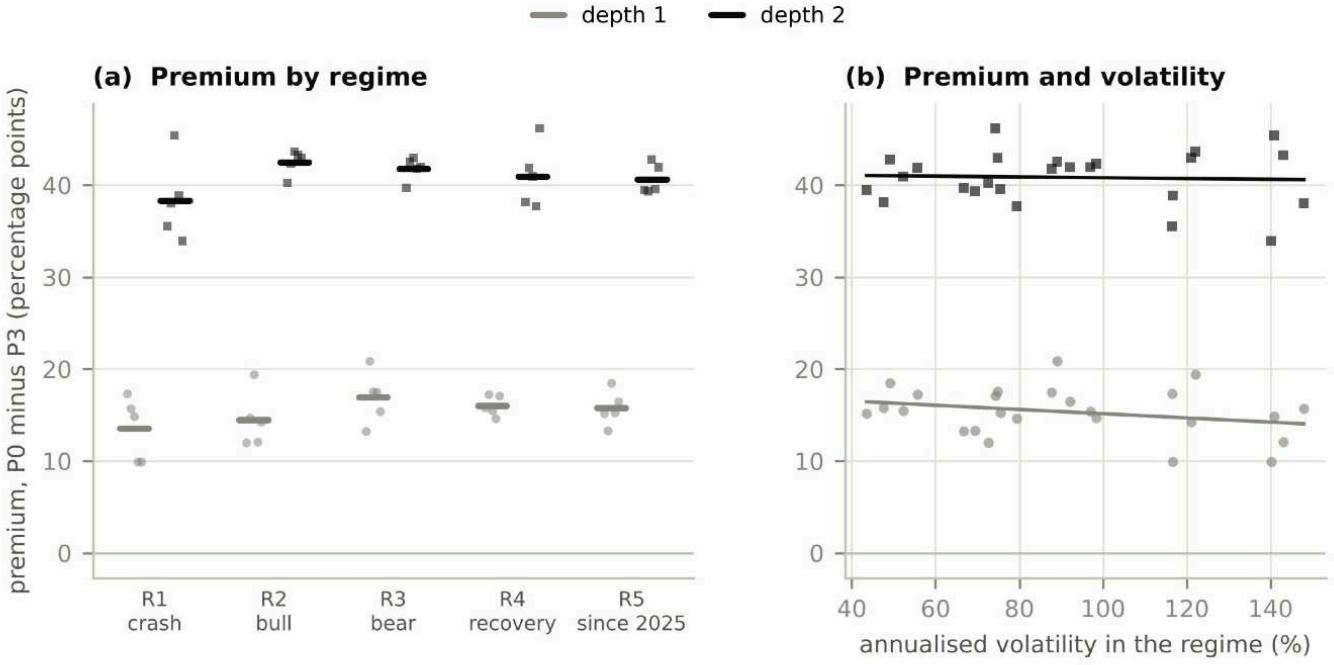


Fig. 4. The leakage premium in directional accuracy across market regimes. (a) Premium of P0 over P3 in each regime, by depth; each small marker is one asset and each bar the five-asset mean. (b) The same premia against each asset's annualised volatility within the regime, with a least-squares line for each depth.

Table 8 confirms the impression. In the regression for the directional premium the coefficient on volatility is −0.01, with a standard error of 0.015, so even a doubling of volatility from 60 to 120 per cent would move the premium by less than one point. Dollar volume is insignificant without asset fixed effects. Within assets its coefficient is 1.1 points per unit of log volume and significant at the 5 per cent level, but that is small next to the depth-2 dummy of 25.5 points. The regression explains 96 per cent of the variation in the premium, and almost all of that through depth. The directional premium is a property of the method, set by how many decomposition layers are stacked, not a property of the market.

Table 8. What drives the premium? Panel regressions across assets, regimes and depths.

	DA premium	DA premium	R^2_{OS} premium	R^2_{OS} premium
Annualised volatility (per cent)	−0.012 (0.015)	−0.005 (0.015)	−0.286*** (0.058)	−0.294*** (0.064)
Log dollar volume	0.076 (0.442)	1.147** (0.556)	0.977 (1.760)	2.925 (2.335)
Depth 2	25.491*** (0.839)	25.491*** (0.747)	78.441*** (3.633)	78.441*** (3.639)
Asset fixed effects	No	Yes	No	Yes
Observations	50	50	50	50
R^2	0.96	0.97	0.92	0.93

OLS regressions of the premium (P0 minus P3, percentage points) for asset a in regime r at depth d on the asset's annualised daily-return volatility within the regime, the log of its mean daily dollar volume and a dummy for depth 2. Five assets, five regimes and two depths. Heteroskedasticity-robust (HC3) standard errors in parentheses; *, ** and *** denote significance at the 10, 5 and 1 per cent levels.

In squared error volatility does matter, for a reason that Section 6.3 makes easy to see. The R^2 premium falls by 0.29 points for every point of volatility. The standard deviation of the causal forecasts varies far less across regimes than that of returns, for BTC at depth 2 between 1.3 and 2.5 per cent a day against 2.3 to 6.1 per cent for returns, so the noise they add costs less R^2 when markets are turbulent. The leaky forecasts, whose R^2 is close to 96 per cent everywhere, play no part in this. Table S5 repeats both regressions for the two components of the premium. In squared error the effect of volatility runs through the consistency component, P2 minus P3, while the look-ahead component is not significantly related to volatility at the 5 per cent level.

6.5. The model confidence set

Table 9 asks which forecasts survive once the multiplicity of comparisons is allowed for. Among the forecasts a user could have produced, the 90 per cent model confidence set contains RW, the historical mean, DLinear, LightGBM and both networks without decomposition for every asset. The causal decomposition models, counting both networks, are excluded in 38 of the 40 cases. The two exceptions are BNB forecasts under P2, from the depth-1 DNN and the depth-2 LSTM, whose mean squared errors are 75 and 16 per cent above that of RW. They survive only because the heavy tails of BNB returns make the bootstrap distribution of the loss differences very wide.

Table 9. Model confidence sets, one-day horizon.

Model	BTC	ETH	BNB	XRP	ADA
Random walk	**0.458**	**0.719**	**0.614**	**1.000**	**0.811**
Historical mean	**0.458**	**0.295**	**0.614**	**0.135**	**0.205**
DLinear	**0.207**	**0.295**	**0.247**	**0.372**	**0.537**
LightGBM	**0.425**	**0.719**	**0.614**	**0.372**	**0.537**
DNN, depth 0	**0.252**	**0.719**	**0.508**	**0.372**	**0.204**
DNN, depth 1, P2	0.000	0.000	**0.178**	0.000	0.000
DNN, depth 1, P3	0.000	0.000	0.001	0.000	0.000

Model	BTC	ETH	BNB	XRP	ADA
DNN, depth 2, P2	0.000	0.000	0.097	0.001	0.000
DNN, depth 2, P3	0.000	0.000	0.002	0.000	0.000
LSTM, depth 0	**1.000**	**1.000**	**1.000**	**0.535**	**1.000**
LSTM, depth 1, P2	0.000	0.000	0.018	0.000	0.000
LSTM, depth 1, P3	0.000	0.000	0.000	0.001	0.000
LSTM, depth 2, P2	0.000	0.000	**0.147**	0.000	0.000
LSTM, depth 2, P3	0.000	0.000	0.000	0.001	0.000
Set when P0 forecasts are added	DNN, depth 2, P0; LSTM, depth 2, P0	DNN, depth 2, P0	DNN, depth 2, P0	DNN, depth 2, P0	DNN, depth 2, P0

MCS p-values (Hansen et al., 2011) with squared-error loss, range statistic, stationary bootstrap with 5,000 replications. Bold entries belong to the 90 per cent model confidence set. Only forecasts a user could have produced enter the main comparison. The LSTM is re-estimated every 120 days. The last row reports the set obtained when the P0 forecasts of both networks, at depths one and two, join the candidates.

Adding the P0 forecasts to the candidates reverses the picture, as the last row of Table 9 and the p-values in Table S7 show. For every asset the set then shrinks to the depth-2 DNN under P0, joined for BTC by the depth-2 LSTM under P0, and every benchmark is rejected. A reader of the leaky results would infer that two layers of decomposition are indispensable, which is exactly the wrong conclusion.

6.6. Economic value and phantom alpha

Table 10 and Fig. 5 translate the forecasts into returns from the long/flat rule at a one-way cost of 10 basis points, and Table S8 repeats the exercise at costs of 0, 5, and 20 basis points. Over the evaluation period buy-and-hold had Sharpe ratios between 0.78 and 1.13. Driven by the P0 forecasts, the rule has Sharpe ratios between 7.6 and 10.2 and earns 146 basis points a day after costs, which compounds into the implausible sums of Fig. 5. It would go on beating buy-and-hold at one-way costs of 203 to 346 basis points, more than ten times the standard fee of any large exchange.

Table 10. Economic value of the one-day forecasts of the depth-2 DNN.

Asset	Sharpe, buy and hold	Sharpe, depth 0	Sharpe, P0	Sharpe, P2	Sharpe, P3	Break-even, P0 (bp)	Break-even, P3 (bp)	Break-even, null, 95th pct. (bp)	Phantom alpha (bp per day)
BTC	0.91	0.57	9.97	0.26	0.43	203	0	3.8	100.2
ETH	0.97	0.68	10.23	0.29	0.78	274	2	4.5	134.3
BNB	1.13	0.70	8.53	0.32	0.35	246	0	3.0	130.1
XRP	0.78	0.44	7.64	0.13	0.22	288	0	8.6	155.9
ADA	0.78	0.03	9.34	0.41	0.52	346	0	6.0	168.4

The long/flat rule holds the asset on day t+1 if the forecast made on day t is positive and holds cash otherwise. Sharpe ratios are annualised, with a zero risk-free rate, at a one-way cost of 10 basis points (none for buy-and-hold). The break-even cost is the one-way cost at which the rule's Sharpe ratio falls to that of buy-and-hold (zero if it is below even at zero cost). The null column gives the 95th percentile of the break-even cost of the P3 rule over 500 random permutations of its positions. Phantom alpha is the difference in mean daily net return between the P0 and the P3 rule at 10 basis points.

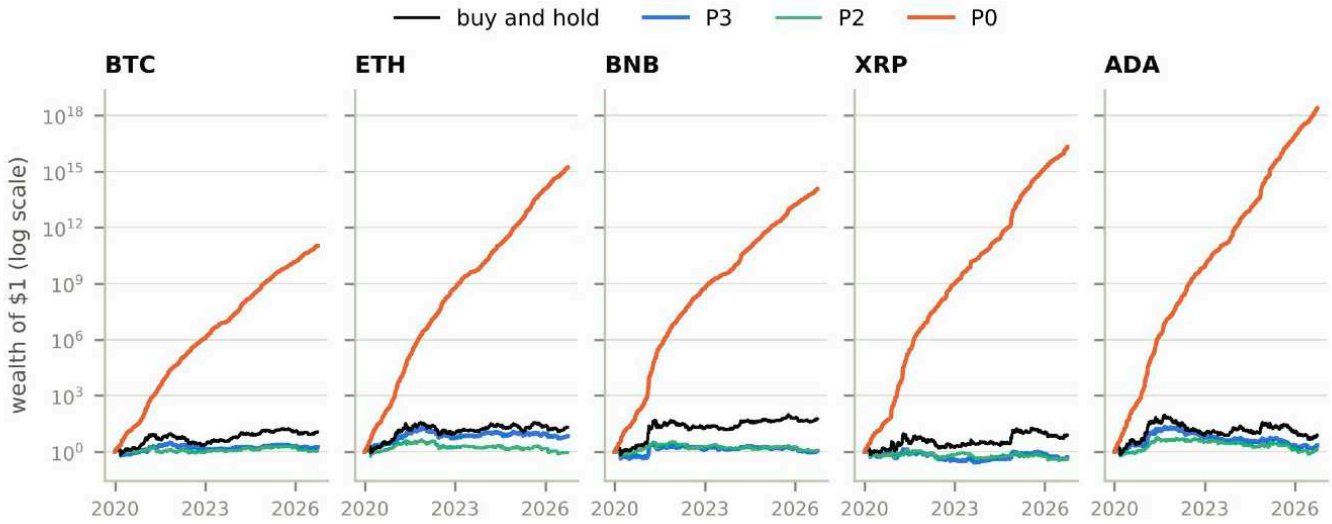


Fig. 5. Phantom alpha. Wealth of one dollar invested on 1 January 2020 in the long/flat rule driven by the one-day forecasts of the depth-2 DNN under each protocol, net of a one-way cost of 10 basis points, and of buy-and-hold. Logarithmic scale, common across panels.

Every feasible rule falls short of buy-and-hold. At 10 basis points the rules driven by the P3, P2 and depth-0 forecasts have average Sharpe ratios of 0.46, 0.28 and 0.48. Their break-even costs are zero for P2 and depth 0 and at most 2.4 basis points for P3, inside the range that randomly shuffled positions reach by luck, whose 95th percentiles lie between 3 and 9 basis points. The phantom alpha, the difference in mean daily net return between the P0 and the P3 rule, averages 138 basis points a day and ranges from 100 to 168 across assets. That is the return a reader of the published accuracy would have expected, and it was never there.

6.7. Longer horizons

Table 11 repeats the comparison at longer horizons. At five days the full-sample protocol is, if anything, more flattering than at one day. The depth-1 model now calls the direction of five-day returns correctly 84.5 per cent of the time, against 66.5 per cent at one day, and its R^2 rises to 81.5 per cent, while the depth-2 model reaches 89.9 and 91.5 per cent. The directional premium therefore grows with the horizon at depth 1, from 15.6 to 34.2 points, and stays at about 40 points at depth 2. The smooth co-IMF accounts for most of the variation of multi-day price changes, and it is also the component whose full-sample values reach furthest into the future. At longer horizons it does most of the leaking.

Under P3 nothing improves. Directional accuracy is 50.3 per cent at both depths and R^2 is −23.8 per cent. The two causal forecasts are in fact almost the same forecast, with correlations between 0.94 and 0.97, because at five days both are dominated by the network for the smooth co-IMF, which the two depths share. Look-ahead again accounts for the whole directional premium, and P2 is again far worse than P3 in squared error, with R^2 of −94 per cent at depth 1 and −111 per cent at depth 2.

At ten days the direct forecasts tell the same story. Under P0 the two depths now look alike, with directional accuracies of 82.4 and 83.0 per cent and R^2 of 71.2 and 73.1 per cent, because the smooth co-IMF, shared by both depths, does almost all of the work. Under P3 directional accuracy is 49.2 and 49.5 per cent and R^2 is −12.6 and −11.6 per cent. At five and at ten days alike, the model confidence set of the feasible forecasts contains benchmarks only, and no causal decomposition model enters it for any asset (Table S9).

The recursive strategy, which iterates the one-day networks instead of training new ones, points the same way. Under P0 it is nearly as accurate as the direct strategy, because every step inherits the leak. Its five-day directional accuracy is 83.9 per cent at depth 1 and 89.5 per cent at depth 2, and at ten days it is still 81.7 and 82.2 per cent. Under P3 the iterated forecasts are no better than chance, between 49.7 and 50.4 per cent, and at depth 2 their errors compound from step to step, until the ten-day forecast has an R^2 of −1,490 per cent. Table S10 reports both strategies for each asset.

6.8. Other networks and a higher frequency

The LSTM experiment asks whether the premium is a peculiarity of the feed-forward network, and the answer is no. Under P0 the LSTM calls the direction of the next day's return correctly on 66.9 per cent of days at depth 1 and on 92.1 per cent at depth 2, with R^2 of 27.8 and 92.5 per cent. Under P3 it is right on 50.6 and 51.5 per cent of days, its R^2 is −15.0 and −23.3 per cent, and U exceeds one for every asset at both depths. The directional premium, 16.3 points at depth 1 and 40.6 at depth 2, lies within a point of what the DNN shows, and look-ahead again accounts for nearly all of it. The one place where the architecture makes a difference is the model without decomposition. There the LSTM has an out-of-sample R^2 between −0.2 and 1.0 per cent and the smallest mean squared error of any feasible forecast for four of the five assets (Table 9). Its margin over RW, less than one per cent of the loss, is too small for the model confidence set to tell the two apart. Table S11 gives the LSTM results by asset.

The four-hour experiment asks whether the premium depends on the sampling frequency. Over the last twelve months of the sample, 2,190 four-hour bars, the BTC results look almost exactly like the daily ones. Under P0 the depth-1 model calls the direction of the next bar correctly 63.9 per cent of the time and the depth-2 model 91.7 per cent, with R^2 of 27.1 and 95.5 per cent. Under P3 the same models are right 49.7 and 50.3 per cent of the time, with R^2 of −8.6 and −26.0 per cent. The directional premium is 14.2 points at depth 1 and 41.4 at depth 2, against 15.6 and 41.2 for daily data. A premium that does not depend on the asset, the network, the market regime or the sampling frequency is hard to attribute to anything but the method.

Table 11. The premium at longer horizons, with another network and at a higher frequency.

Design	DA, depth 1, P0	DA, depth 1, P3	R^2_{OS}, depth 1, P0	R^2_{OS}, depth 1, P3	DA, depth 2, P0	DA, depth 2, P3	R^2_{OS}, depth 2, P0	R^2_{OS}, depth 2, P3
DNN, h = 1 (baseline)	66.5	50.8	27.6	−13.5	93.2	52.0	96.0	−22.7
DNN, h = 5, direct	84.5	50.3	81.5	−23.8	89.9	50.3	91.5	−23.8
DNN, h = 5, recursive	83.9	49.9	79.8	−18.0	89.5	50.4	90.9	−192.7
DNN, h = 10, direct	82.4	49.2	71.2	−12.6	83.0	49.5	73.1	−11.6
DNN, h = 10, recursive	81.7	50.2	70.6	−4.3	82.2	49.7	71.7	−1489.5
LSTM, h = 1	66.9	50.6	27.8	−15.0	92.1	51.5	92.5	−23.3
DNN, BTC, four-hour bars, h = 1	63.9	49.7	27.1	−8.6	91.7	50.3	95.5	−26.0

DA and out-of-sample R^2 (per cent) under P0 and P3, averaged over the five assets (BTC only for the four-hour bars). Direct forecasts come from networks trained on h-day targets, recursive forecasts from iterating the one-day networks. The LSTM is re-estimated every 120 days. The four-hour experiment covers the last twelve months of the sample, with a rolling two-year training window and causal decompositions over a trailing window of 1,024 bars.

7. Discussion

7.1. How to read the existing literature

The premium measured here is large enough to explain why the two literatures described in the introduction disagree. A one-layer decomposition evaluated with the full-sample protocol calls the direction of daily returns correctly on about two days in three, which is roughly what many published hybrid models report, and a two-layer decomposition does so on more than nine days in ten. Evaluated causally, the same pipelines are right on about 51 days in a hundred, in line with the careful machine-learning studies of daily returns. Their squared errors exceed those of a forecast of zero. Nothing in these results suggests that the decomposition-based literature has found a kind of predictability that the rest of the literature missed.

This evidence does not imply that every published decomposition-based forecast is leaky. It implies something weaker and more useful, namely that a reported accuracy figure cannot be interpreted until one knows which protocol produced it. Often a reader can tell. Directional accuracies far above the band of 50 to 60 per cent reported by careful studies of daily returns are a warning sign, and so are errors a small fraction of those of a zero forecast, or accuracy measured on the price level instead of on returns. A methods section that decomposes the series first and only then splits it into training and test sets is the clearest sign of all.

Two further traps are less obvious. Accuracy reported component by component says little about the return forecast. Under P3 the depth-2 components are forecast with an average out-of-sample R^2 of 36 per cent against their own targets, while the sum of the component forecasts is worse than a forecast of zero. And a protocol that removes look-ahead but still trains on interior values, P2 in the present design, can be worse in squared error than the fully causal one. A researcher who has taken care to keep future prices out of the forecast might not expect that.

7.2. Recommendations

Most of what follows costs little. A methods section should state explicitly which observations each decomposition uses when a training input, a training target and a forecast input are computed, and the three protocols used here give a vocabulary for doing so. The falsification test should be routine. Changing the data after a date t^* and checking that forecasts up to t^* do not move is an exact test of causality. It needs no theory, and it catches every form of look-ahead, including the kind that software defaults introduce without anyone noticing, like the way vmdpy treats inputs of odd length. It belongs in every replication package.

Evaluation should be done on returns and against simple benchmarks. Accuracy on the price level says little about forecasting skill, whereas U, the out-of-sample R^2 against the historical mean and a one-layer linear model such as DLinear are cheap to add and, in the experiments reported here, hard to beat. A paper that compares tens of configurations should also say how it guards against selecting one of them on noise, with the model confidence set or a test of superior predictive ability. And economic value should be reported net of costs. Directional accuracy becomes money only after transaction costs are paid, and a break-even cost set against a permutation null is a more honest summary than a backtest at zero cost.

For referees and editors the practical point is simple. Asking to see the falsification test, or simply asking for the main result to be re-run with the decomposition computed on the information available at each forecast date, would settle most of the cases in which the question arises.

7.3. What a causal pipeline can still do

On the evidence presented here, not much. The causal inputs do carry a little information about the next day's return, with correlations of up to 0.09, and the causal models' directional accuracy is a point or two above one half, significantly so in more than a third of the configurations. But once their forecasts are rescaled to remove the noise that the component-by-component design creates, the out-of-sample R^2 is slightly negative for every asset, and the trading rules they drive never beat buy-and-hold net of realistic costs. It is telling that for four of the five assets the best feasible forecast comes, by a small margin, from an LSTM that sees nothing but ten lagged returns. The negative result concerns a particular way of using decompositions, however, and not decompositions as such. The revisions that make the end-point components predictable one by one cancel in the sum, and a design that forecasts the components separately throws that cancellation away. Mapping all causal components jointly to the return, or using decompositions that are causal by construction (Feng et al., 2026), would avoid the problem, and the protocols and the falsification test proposed here offer a way to find out whether either helps.

7.4. Limitations

The evidence comes from one exchange and five large crypto-assets quoted against a stablecoin. There is little reason to expect the size of the premium to be specific to these choices, because its source is the two-sided nature of the decomposition and not any property of the market. Still, the evidence for smaller and less liquid assets, for other quote currencies and for other venues has yet to be collected.

The analysis covers one pipeline, CEEMDAN with permutation-entropy grouping and a second VMD layer, chosen because multi-layer designs of this kind dominate the literature. Other two-sided transformations, among them wavelets, SSA and plain EMD, leak through the same channel. How large their premium is depends on how far into the future each of them reaches, a question left for future work.

The networks are deliberately simple, and their hyperparameters were fixed in advance. Tuning them separately under each protocol could change how accurate each configuration is, although it cannot make P0 feasible, and the results should not be read as an estimate of the best accuracy attainable with decomposition-based inputs. The LSTM, which is several times more expensive to train, is re-estimated less often than the DNN, and the GRU that appears in parts of the literature is not reported. The four-hour experiment covers Bitcoin only and, for computational reasons, uses a trailing rather than an expanding decomposition window.

The trading rule is the simplest one the forecasts support. It ignores slippage and market impact, never goes short and trades at the daily close. A more elaborate strategy could extract more from a causal forecast, but the phantom alpha of a leaky forecast would remain phantom.

8. Conclusion

Decomposition-based forecasts of cryptocurrency returns look far better than they are. When a two-sided decomposition is computed on the full sample, every input of the forecasting model carries part of the future, and a standard pipeline that stacks VMD on top of CEEMDAN appears to call the direction of next-day returns correctly on 93 days out of a hundred. Computed from the information a forecaster would actually have had, the same pipeline is right on about 52 days out of a hundred, and its squared error is larger than that of a forecast of zero.

Two lessons follow. The first concerns evaluation. A study that decomposes its series before splitting it reports look-ahead, not forecasting skill, and the error involved is not a matter of decimals. It is large enough to turn a forecast that is worse than a zero forecast into one that explains most of the variation in daily returns. Fortunately it is also easy to detect. Changing the prices after a date and checking that earlier forecasts remain unchanged takes minutes, and it would settle the question for any published model.

The second lesson concerns design. Removing look-ahead is not enough to make decomposition useful. A model trained on interior values of a decomposition and then fed end-point values fails because the two do not look alike, and a model trained consistently on end-point values fails for a subtler reason. The revisions of the end-points cancel across components in the targets but not in the forecasts, so that accuracy at the level of the components hides noise at the level of the return. Component-level accuracy, however good, says little about the forecast that matters.

The evidence concerns one family of pipelines in one market. The mechanism documented here is not specific to either, and it should operate wherever a two-sided transformation sits in front of a forecasting model.

References

Akyildirim, E., Goncu, A., Sensoy, A., 2021. Prediction of cryptocurrency returns using machine learning. Annals of Operations Research 297, 3-36.

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

Bandt, C., Pompe, B., 2002. Permutation entropy: A natural complexity measure for time series. Physical Review Letters 88, 174102.

Bariviera, A.F., 2017. The inefficiency of Bitcoin revisited: A dynamic approach. Economics Letters 161, 1-4.

Campbell, J.Y., Thompson, S.B., 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? Review of Financial Studies 21, 1509-1531.

Dragomiretskiy, K., Zosso, D., 2014. Variational mode decomposition. IEEE Transactions on Signal Processing 62, 531-544.

Du, K., Zhao, Y., Lei, J., 2017. The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series. Journal of Hydrology 552, 44-51.

Feng, W., Tao, R., Cartlidge, J., Zheng, J., 2026. VMDNet: Temporal leakage-free variational mode decomposition for electricity demand forecasting. In: Proceedings of the 34th European Signal Processing Conference (EUSIPCO 2026). arXiv:2509.15394.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Review of Financial Studies 33, 2223-2273.

Hansen, P.R., 2005. A test for superior predictive ability. Journal of Business & Economic Statistics 23, 365-380.

Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79, 453-497.

Harvey, C.R., Liu, Y., Zhu, H., 2016. … and the cross-section of expected returns. Review of Financial Studies 29, 5-68.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9, 1735-1780.

Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.-C., Tung, C.C., Liu, H.H., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A 454, 903-995.

Jaquart, P., Dann, D., Weinhardt, C., 2021. Short-term bitcoin market prediction via machine learning. Journal of Finance and Data Science 7, 45-66.

Jin, C., Li, Y., 2023. Cryptocurrency price prediction using frequency decomposition and deep learning. Fractal and Fractional 7, 708.

Kapoor, S., Narayanan, A., 2023. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4, 100804.

Kaufman, S., Rosset, S., Perlich, C., Stitelman, O., 2012. Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data 6, 15:1-15:21.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30, 3146-3154.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR).

Koo, E., Kim, G., 2024. Centralized decomposition approach in LSTM for Bitcoin price prediction. Expert Systems with Applications 237, 121401.

Lee, S.S., Wang, M., 2025. Variance decomposition and cryptocurrency return prediction. Journal of Financial and Quantitative Analysis 60, 1859-1890.

Li, Y., Jiang, S., Li, X., 2022. Hybrid data decomposition-based deep learning for Bitcoin prediction and algorithm trading. Financial Innovation 8, 31.

Liu, Y., Tsyvinski, A., 2021. Risks and returns of cryptocurrency. Review of Financial Studies 34, 2689-2727.

López de Prado, M., 2018. Advances in Financial Machine Learning. Wiley, Hoboken, NJ.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. Statistical and machine learning forecasting methods: Concerns and ways forward. PLoS ONE 13, e0194889.

Nadarajah, S., Chu, J., 2017. On the inefficiency of Bitcoin. Economics Letters 150, 6-9.

Patton, A., Politis, D.N., White, H., 2009. Correction to “Automatic block-length selection for the dependent bootstrap” by D. Politis and H. White. Econometric Reviews 28, 372-375.

Pesaran, M.H., Timmermann, A., 1992. A simple nonparametric test of predictive performance. Journal of Business & Economic Statistics 10, 461-465.

Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. Journal of the American Statistical Association 89, 1303-1313.

Politis, D.N., White, H., 2004. Automatic block-length selection for the dependent bootstrap. Econometric Reviews 23, 53-70.

Quilty, J., Adamowski, J., 2018. Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models for real-world applications with best practices and a new forecasting framework. Journal of Hydrology 563, 336-353.

Riedl, M., Müller, A., Wessel, N., 2013. Practical considerations of permutation entropy: A tutorial review. The European Physical Journal Special Topics 222, 249-262.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53-65.

Torres, M.E., Colominas, M.A., Schlotthauer, G., Flandrin, P., 2011. A complete ensemble empirical mode decomposition with adaptive noise. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4144-4147.

Tsai, C.-L., Chen, M.-Y., Tsai, T.-Y., Hu, J.-W., Lai, Y.-W., 2026. Using signal decomposition methods and deep learning approaches to forecast Bitcoin price. Computational Economics 67, 1009-1041.

Urquhart, A., 2016. The inefficiency of Bitcoin. Economics Letters 148, 80-82.

Wang, S., Huang, Z., Liao, Z.W., 2026. Multi-scale decomposition for deep learning-based Bitcoin price forecasting. Finance Research Letters 96, 109776.

White, H., 2000. A reality check for data snooping. Econometrica 68, 1097-1126.

Wu, Z., Huang, N.E., 2009. Ensemble empirical mode decomposition: A noise-assisted data analysis method. Advances in Adaptive Data Analysis 1, 1-41.

Yang, X., Li, J., Jiang, X., 2024. Research on information leakage in time series prediction based on empirical mode decomposition. Scientific Reports 14, 28362.

Yu, L., Wang, S., Lai, K.K., 2008. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Economics 30, 2623-2635.

Zeng, A., Chen, M., Zhang, L., Xu, Q., 2023. Are transformers effective for time series forecasting? Proceedings of the AAAI Conference on Artificial Intelligence 37, 11121-11128.

Zhang, X., Peng, Y., Zhang, C., Wang, B., 2015. Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? Some experiment evidences. Journal of Hydrology 530, 137-152.