Market Decompression: Learning Latent-State Inverse Maps for Cold-Start Financial Prediction

Idrees Aziz
Middle East Technical University
Ankara, Türkiye
idrees.mustafa.aziz@gmail.com

Abstract

Forecasts matter most when history is short—a newly listed asset, or an established one after a structural break. At a 60-hour horizon, even five completed non-overlapping outcomes take 12.5 days to observe. With so few labels, a forecaster must estimate both a representation of price dynamics and the mapping from that representation to the forecast target. Only the target mapping inherently requires real outcomes.

To address this problem, we introduce market decompression, a learning framework for sparse-label financial forecasting. Using synthetic markets where both prices and their hidden volatility, jump, and regime paths are known, it learns an inverse map from the observed tape to the latent structure that produced it. Scarce real labels are reserved for learning the mapping from recovered structure to the forecast target.

Because decompression depends on the synthetic world used to learn this inverse map, we also introduce the transfer field: a structured response surface that identifies the source world with the greatest downstream forecasting advantage for a given market and label budget.

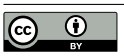
Under forecast-sufficient recovery, we show analytically that decompression has strictly lower Bayes log loss whenever the prior carries target-relevant information. Controlled Bates–Heston-derived experiments first establish a consistent advantage with five to ten labels on unseen target worlds. We then use fields estimated from earlier episodes to select frozen priors for later untouched data. The results are confirmed in three of five markets, spanning cryptocurrency, foreign exchange, and precious metals. Across the confirmed markets, pooled MSE falls by up to 7.0%, while ETHUSDT reaches 10.4% at the 60-hour horizon.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Applied computing** → *Economics*.

Keywords

financial machine learning; cold-start forecasting; synthetic markets; transfer learning; sparse labels; market simulation



This work is licensed under a Creative Commons Attribution 4.0 International License.
ICAIF '26, Milan, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-3087-0/2026/11
https://doi.org/10.1145/3858063.3858733

ACM Reference Format:
Idrees Aziz. 2026. Market Decompression: Learning Latent-State Inverse Maps for Cold-Start Financial Prediction. In *7th ACM International Conference on AI in Finance (ICAIF '26), November 14–17, 2026, Milan, Italy*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3858063.3858733

1 Introduction

Cold-start forecasting arises when the target market has little relevant history. A new asset begins trading before it has outcomes from its own regime. After a structural break, the post-break regime may provide only a short estimation window, while older observations can introduce forecast bias [11–13]. The October 2025 tariff shock illustrates how quickly this can happen: within a day, it triggered more than $19 billion in crypto liquidations [24]. At a 60-hour forecast horizon, collecting five to ten non-overlapping post-break outcomes takes 12.5–25 days—precisely when forecasts are most needed.

Maximum-likelihood estimates of GARCH(1,1) parameters and persistence can be biased or unstable in very small samples [30], while transfer can fail when the source and target distributions differ [16]. Few real labels must ordinarily support both representation learning and target estimation. The forecast mapping requires observed outcomes, whereas an inverse map from price history to latent state can be learned in a simulator where both quantities are known.

Market decompression learns this inverse map from paired synthetic tapes and hidden states before target adaptation. Because its value depends on the synthetic source world, the transfer field measures the downstream forecasting gain induced by each simulator coordinate for a given market and label budget. We use the field to choose a prior and estimate the label budgets over which it provides positive transfer. Priors selected on one historical episode are then evaluated on a later untouched episode.

Here, decompression targets volatility, jumps, and regimes from hourly price bars. Recovering liquidity, order flow, or limit-order-book depth would require richer microstructure inputs and lies outside the present scope.

This paper makes four contributions:

(1) **Market decompression.** Synthetic hidden-state recovery pretrains an inverse map, leaving scarce target labels to fit the forecast mapping.
(2) **The transfer field.** A target- and budget-specific field assigns signed downstream value to source worlds and supports selection under uncertainty.
(3) **Theory.** An information identity, a log-MSE bridge, and a finite-sample recovery law relate recovered value, source mismatch, and label budget, yielding nested positive-transfer envelopes.

(4) **Empirical confirmation.** Replicated field estimation, fresh confirmation, and untouched holdout verify the sparse-label advantage in controlled worlds and three of five real markets.

2 Related Work

Synthetic and artificial markets. Prior ICAIF work uses synthetic data for rare scenarios [1], learned market simulation [2, 5, 26, 27], and mechanism analysis [3]. We instead use paired synthetic tapes and hidden states to learn an inverse map. When the target depends on the tape through a latent state, that state can be forecast-sufficient [22, 23]; its synthetic labels act as privileged training information [8].

Time-series privileged-information methods prove gains from training-only intermediate observations in linear-Gaussian and nonlinear latent systems [32, 33]. Here the privileged volatility, jump, and regime paths come from an instrumented simulator rather than the real market, and the transfer field determines which source representation remains useful.

Simulator selection: realism versus transfer value. Simulator calibration measures resemblance to the target through realism metrics [4], deep calibration [6], or agent-based calibration [7]. The transfer field instead measures downstream forecasting value at a fixed label budget; resemblance and transfer can diverge under source mismatch [16].

Cold-start volatility forecasting and transfer. After a break, recent samples are short and older observations can bias forecasts [11–13]. Prior work transfers from established assets to new issues [14] or repurposes pretrained models for data-efficient FX volatility forecasts [28]. We select synthetic state supervision by its measured downstream value.

Privileged information, domain adaptation, and response surfaces. Domain adaptation studies source–target alignment [9]. Our deployment criterion is narrower: whether a specific source inverse map improves a particular market. Replicated evaluations and a noise-aware response surface estimate that value and its zero boundary [10, 31].

3 Market Decompression

A forecaster observes the price path but not the volatility, jumps, or regime that produced it.

3.1 Setup

Let $X_{t-L+1:t}$ denote the observed price tape, Z_t the latent volatility, jump, and regime state, and Y_{t+h} the future forecast target. An instrumented source world provides paired observations of (X, Z, Y), allowing the inverse map from tape to latent structure to be learned before target adaptation. In the real target market, the forecaster observes only $X_{t-L+1:t}$ and a sparse set of completed outcomes Y_{t+h}.

With b completed target outcomes, a direct learner and market decompression use

$$\begin{aligned} f_b &: X_{t-L+1:t} \to \widehat{Y}_{t+h}, \\ D(X_{t-L+1:t}; \Theta(\theta)) &\to \widehat{R}_t, \qquad g_b : \widehat{R}_t \to \widehat{Y}_{t+h}. \end{aligned} \tag{1}$$

Here b counts forecast outcomes rather than observed price bars; θ indexes the synthetic source world; and $\Theta(\theta)$ denotes the parameters learned from it. Both arms use the same b target outcomes.

The decompression and control arms share the common source objective. Decompression additionally receives hidden-state supervision:

$$\mathcal{L}_{\text{dec}} = \mathcal{L}_{\text{common}} + \lambda \mathcal{L}_{\text{priv}}. \tag{2}$$

The matched control removes only the privileged loss. Across experiments, $\mathcal{L}_{\text{common}}$ contains the matched source tasks, while $\mathcal{L}_{\text{priv}}$ contains only hidden-state recovery; their exact components appear in Sections 5.1 and 5.2. During source training, the recovery heads shape the encoder to represent the simulator states associated with each tape.

3.2 Ideal Recovery and Forecast Sufficiency

A recovered state is sufficient when it retains everything in the tape and source prior that matters for the future target. The decompression objective aims to recover the conditional distribution of the latent state given the observed tape and source prior. An ideal decompressor returns

$$R_t^* = p(Z_t \mid X_{t-L+1:t}, \Theta). \tag{3}$$

Assume that, once the latent state is known, neither the tape nor the source prior contains additional information about the forecast target:

$$Y_{t+h} \perp (X_{t-L+1:t}, \Theta) \mid Z_t. \tag{4}$$

Then the recovered posterior is forecast-sufficient:

$$p(Y_{t+h} \mid R_t^*) = p(Y_{t+h} \mid X_{t-L+1:t}, \Theta). \tag{5}$$

Proof. Write $W = (X_{t-L+1:t}, \Theta)$. For any measurable set A, the conditional independence above gives

$$P(Y_{t+h} \in A \mid W) = \int P(Y_{t+h} \in A \mid Z_t = z)\, P(dz \mid W).$$

The right-hand side is a measurable function of R_t^*. By the tower property,

$$\begin{aligned} P(Y_{t+h} \in A \mid R_t^*) &= \int P(Y_{t+h} \in A \mid Z_t = z)\, R_t^*(dz) \\ &= P(Y_{t+h} \in A \mid W). \end{aligned}$$

□

3.3 Supervision-Privilege Identity

The prior helps through forecast information that survives recovery. The identity below separates supplied information, lost information, and fitting-dependent information.

For any representation V, define its Bayes log-loss risk by

$$\mathcal{R}_{\log}(V) = \inf_q \mathbb{E}[-\log q(Y \mid V)] = H(Y \mid V). \tag{6}$$

Theorem 3.1 (Supervision-privilege identity). *For any jointly distributed $(Y, X, \Theta, \widehat{R})$ with finite relevant entropies,*

$$\mathcal{R}_{\log}(X) - \mathcal{R}_{\log}(\widehat{R}) = \underbrace{I(Y;\Theta \mid X)}_{\text{signal from the prior}} - \underbrace{I(Y;X,\Theta \mid \widehat{R})}_{\text{signal lost in recovery}} + \underbrace{I(Y;\widehat{R} \mid X,\Theta)}_{\text{fitting-dependent information}}. \tag{7}$$

Equation (7) follows from the chain rule for mutual information [22].

Training-dependent representations. If $\widehat{R} = f_\Phi(X,\Theta)$, data processing bounds the fitting-dependent term by $I(Y;\Phi \mid X,\Theta)$. When fitted parameters use no held-out target outcome, $Y \perp \Phi \mid (X,\Theta)$ and this term vanishes. Equivalently, conditional on the realized Φ, the identity holds with each information term conditioned on Φ.

Proposition 3.2 (Gain ceiling and ideal-recovery guarantee). *Under the conditional no-leakage assumption:*

(i) $\mathcal{R}_{\log}(X) - \mathcal{R}_{\log}(\widehat{R}) \le I(Y;\Theta \mid X)$. (ii) If $\widehat{R}$ is forecast-sufficient, so that $p(Y \mid \widehat{R}) = p(Y \mid X,\Theta)$, then $\mathcal{R}_{\log}(X) - \mathcal{R}_{\log}(\widehat{R}) = I(Y;\Theta \mid X) \ge 0$, with strict improvement whenever $I(Y;\Theta \mid X) > 0$.

Proof. Under no leakage, Eq. (7) becomes $\mathcal{R}_{\log}(X) - \mathcal{R}_{\log}(\widehat{R}) = I(Y;\Theta \mid X) - I(Y;X,\Theta \mid \widehat{R})$. Non-negativity of conditional mutual information gives the upper bound. Forecast sufficiency sets the second term to zero, yielding equality. □

Population and finite-sample gain. For a fixed deployed prior, $I(Y;\Theta \mid X) = 0$; its practical benefit is therefore finite-sample rather than a population Bayes-risk advantage. The transfer field measures this finite-sample gain relative to matched forecast-only pretraining.

Entropy–MSE bridge. For a representation $V \in \{X, \widehat{R}\}$, let B_V denote its Bayes MSE. Decomposing the conditional entropy relative to a Gaussian with the same conditional mean and variance gives the identity

$$\log B_X - \log B_{\widehat{R}} = \underbrace{2\left[\mathcal{R}_{\log}(X) - \mathcal{R}_{\log}(\widehat{R})\right]}_{\text{information-theoretic gain}} + \underbrace{(G_X - G_{\widehat{R}})}_{\text{variance correction}} + \underbrace{2(J_X - J_{\widehat{R}})}_{\text{non-Gaussianity correction}}. \tag{8}$$

The correction terms vanish under conditionally Gaussian, homoscedastic residuals; otherwise they quantify the difference between log-MSE and log-loss gain.[1]

[1] $G_V = \log B_V - \mathbb{E}[\log \mathrm{Var}(Y \mid V)]$, and J_V is the expected KL divergence between $Y \mid V$ and the Gaussian with the same conditional mean and variance.

Table 1: Simulator coordinates and mismatch costs.

Axis	Source-world channel	Example mismatch cost
θ_1	Volatility persistence / mean reversion	Memory too short or long for the target.
θ_2	Volatility scale and roughness	Latent state too smooth or noisy for the target.
θ_3	Jump or event intensity	Over-parses moves as jumps, or misses them.
θ_4	Tail asymmetry and jump scale	Wrong sign or scale of extreme moves.
θ_5	Regime-switching tendency	Switches too often or rarely for the target.
θ_6	Long-memory dependence	Stale history or underused persistence.

4 The Transfer Field

The transfer field measures which synthetic view of volatility, jumps, and regimes improves later forecasts at a given label budget. Its gain is the value of recovered structure minus source-target mismatch, estimated on earlier history for later prior selection.

4.1 Source-World Coordinates

A coordinate $\theta = (\theta_1, \ldots, \theta_6) \in \Theta_{\text{sim}}$ selects one synthetic generator S_θ, which produces (X^s, Z^s, Y^s) and induces the pretrained prior $\Theta(\theta)$, while architecture, training, and target-label budget remain fixed. Table 1 defines the coordinates. They index source worlds rather than estimate target structural parameters; all other simulator settings remain fixed.

4.2 Definition and Positive-Transfer Envelope

The transfer field assigns each source world its expected held-out forecasting gain. For target market R, label budget b, source coordinate θ, and repetition s, define

$$F_b^R(\theta) = \mathbb{E}_s\Big[\underbrace{\log L_{\text{forecast}}(\theta;R,b,s)}_{\text{matched forecast-only prior}} - \underbrace{\log L_{\text{latent}}(\theta;R,b,s)}_{\text{decompression prior}}\Big]. \tag{9}$$

Positive values favor decompression; negative values favor the matched control. Because the field uses held-out forecast MSE, it requires no real-state labels. Fields estimated on earlier episodes rank the priors later tested on confirmation and holdout data.

The useful source priors form the positive-transfer envelope

$$E_b^{R,+} = \{\theta \in \Theta_{\text{sim}} : F_b^R(\theta) > 0\}, \tag{10}$$

with zero boundary

$$\partial E_b^{R,+} = \{\theta : F_b^R(\theta) = 0\}. \tag{11}$$

4.3 Recovery Law

The Gaussian-sequence model provides an exact benchmark for converting source information into an equivalent number of target labels. We approximate the direct learner's held-out MSE curve by

$$\mathcal{R}_R(b) = \mathcal{R}_R^\infty + \underbrace{K_R(b + b_{0,R})^{-\alpha_R}}_{\text{finite-sample excess MSE}}, \tag{12}$$

with $\mathcal{R}_R^\infty > 0$, $K_R > 0$, $\alpha_R > 0$, and $b_{0,R} \ge 0$.

The Gaussian-sequence benchmark is exact. In an orthonormal basis with $\sum_j \lambda_j < \infty$, model each source/target coefficient pair as jointly Gaussian with marginal variance λ_j and correlation ρ_j. Let

source and target precisions be N_s/σ_s^2 and b/σ_t^2. Gaussian conditioning gives

$$\epsilon_{\text{seq}}(b) = \sum_{j\geq 1} \frac{\tilde{\lambda}_j\, \sigma_t^2}{\sigma_t^2 + b\,\tilde{\lambda}_j}, \qquad \tilde{\lambda}_j = \lambda_j - \frac{\rho_j^2 \lambda_j^2}{\lambda_j + \sigma_s^2/N_s}. \tag{13}$$

When source information contributes the same precision shift across every latent mode, this reduces to a common effective-label offset.

Let $\Lambda_R(\theta)$ denote recovered value in units of real labels and $\Delta_R(\theta)$ the log-MSE penalty from mismatch and imperfect recovery. Their trade-off on the scale of Eq. (9) is

$$F_b^R(\theta) \approx \underbrace{\log \mathcal{R}_R(b) - \log \mathcal{R}_R(b + \Lambda_R(\theta))}_{\text{log-MSE value of recovered structure}} - \underbrace{\Delta_R(\theta)}_{\text{mismatch cost}}. \tag{14}$$

Under proportional source and target information across latent modes, $\Lambda_R(\theta)$ is a common label offset; otherwise, it is a budget-dependent effective-label summary.

As real outcomes accumulate, the useful set of source worlds shrinks under the following conditions.

Proposition 4.1 (Envelope nesting).

If $\Lambda_R(\theta) > 0$ and $\Delta_R(\theta)$ do not vary with b, then Eq. (14) decreases strictly with b. Hence, for $b' > b$, $E_{b'}^{R,+} \subseteq E_b^{R,+}$.

Proof. Write $x = b + b_0$ and $\mathcal{R}(x) = \mathcal{R}^\infty + Kx^{-\alpha}$. The positive function

$$q(x) = -\frac{\mathcal{R}'(x)}{\mathcal{R}(x)} = \frac{\alpha K}{x\,(\mathcal{R}^\infty x^\alpha + K)} \tag{15}$$

decreases strictly in x. Therefore, for $\Lambda > 0$,

$$\partial_b [\log \mathcal{R}(x) - \log \mathcal{R}(x + \Lambda)] = -q(x) + q(x + \Lambda) < 0. \tag{16}$$

□

4.4 Deployment Quantities

The field supports two reported deployment quantities: confidence that a prior transfers positively and the share of source space that remains useful. For a noise-aware Gaussian-process posterior with mean $\mu(\theta)$ and standard deviation $\sigma(\theta)$, the positive-transfer probability is $\pi_b^R(\theta) = \Pr(F_b^R(\theta) > 0 \mid \mathcal{D}) = \Phi(\mu(\theta)/\sigma(\theta))$. Deployment requires this probability to exceed the chosen cost-sensitive threshold; under symmetric costs, the threshold is one half.

The positive-transfer volume is

$$V_b^R = \frac{1}{|\Theta_{\text{sim}}|} \int_{\Theta_{\text{sim}}} \mathbf{1}\{F_b^R(\theta) > 0\}\, d\theta. \tag{17}$$

As more outcomes arrive, decompression should be switched off once the estimated field is no longer positive. The ETHUSDT five-label advantage is no longer distinguishable from zero at $b = 10$ (Section 7.2).

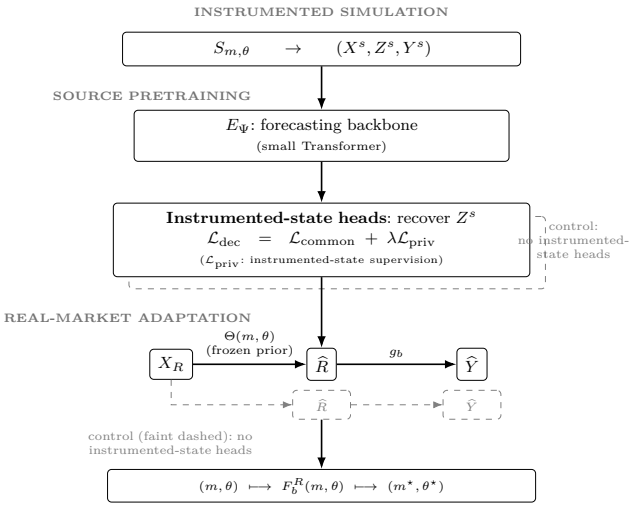


Figure 1: Market decompression pipeline. The simulator exposes tapes, targets, and hidden states; decompression alone receives state supervision during matched source pretraining. Both frozen encoders then fit forecast heads on the same b real outcomes. Repetition over (m, θ) estimates the transfer field used for source selection.

4.5 A Simulator- and Backbone-Agnostic View

More generally, a source is $\xi = (m, \theta)$, where m selects the simulator family and θ a world within it, giving $F_b^R(m, \theta)$. Any forecasting backbone E_Ψ can receive state supervision when the simulator exposes Z^s with its tape and target. Figure 1 shows this pipeline; our experiments use one Bates–Heston-derived family and a small Transformer.

5 Experimental Design

The empirical question is prospective: can a prior chosen from earlier history improve forecasts after the market moves forward? We separate field discovery, fresh confirmation, and untouched chronological holdout.

5.1 Controlled Bates–Heston-Derived Verification

Here the hidden state is known, so the experiment can test directly whether state supervision improves sparse-label forecasting. We test decompression in a Bates–Heston-derived instrumented family of jump-diffusion worlds [18, 19], simulated with Andersen's QE scheme [20], extended with the regime-switching and long-memory controls detailed in Section 4. The suite contains 128 source-training worlds, 32 source-validation worlds, and 64 covered target worlds. Each tape contains 192 steps. Forecast targets are log realized variance and log quadratic variation at horizons $h \in \{5, 20, 60\}$, evaluated over five model seeds and label budgets $b \in \{5, 10, 30, 100, 500\}$.

In this suite, $\mathcal{L}_{\text{common}}$ contains masked-tape reconstruction and forecasting, while $\mathcal{L}_{\text{priv}}$ adds supervision for volatility, jump occurrence and size, and eight simulator parameters. The decompression and control arms use identical initial weights, data, optimization, and target samples; the only treatment is hidden-state supervision. Both models adapt end-to-end on each synthetic target.

Covered targets come from the same Bates–Heston-derived family as the source worlds, approximating the low-mismatch case $\Delta_R(\theta) \approx 0$.

5.2 Five-Market Real-Data Setup

In real markets the hidden state is unavailable, so later forecast error is the only measure of whether the prior was useful. We apply the full discovery, confirmation, and chronological-holdout protocol to five markets: ETHUSDT hourly Binance spot klines [34]; EURUSD, XAUUSD, and USDJPY hourly bars from the ejtraderLabs repository at commit `fbd29b3cd85c0eea4f6e8b81c053f98fb3de22fd` [35]; and Yahoo Finance's continuous front-month E-mini S&P 500 representation `ES=F` [36]. The repository files span 2012-11-16–2022-03-04 for EURUSD, 2012-05-17–2022-03-04 for XAUUSD, and 2012-11-16–2022-03-04 for USDJPY; their upstream quote source and timestamp timezone are not documented. The `ES=F` hourly bars were retrieved on 2026-07-24 and cover 2024-07-25 02:00 UTC–2026-07-24 20:59 UTC; this is a vendor-constructed continuous series, not a uniquely specified raw CME contract history. The exact archived Binance query range, processed snapshots, and software-environment lockfile are not contained in this submission package, a reproducibility limitation. Each input contains 192 hourly bars represented by log return, absolute return, and trailing realized volatility. The forecast target is future realized variance,

$$Y_t^{(h)} = \sum_{j=1}^{h} r_{t+j}^2, \qquad h \in \{5, 20, 60\}. \tag{18}$$

Log-skill is pooled across horizons. Each market is split chronologically into training, validation, and untouched holdout episodes. The 64-window gap prevents target overlap across splits. Shared input bars still create dependence between nearby examples.

The label budget b counts completed training windows, with each window supplying targets at all three horizons. Five to ten non-overlapping 60-hour outcomes require 12.5–25 days; overlapping outcomes arrive sooner but are serially dependent [15]. We estimate Eq. (9) by averaging paired log-MSE differences across repetitions.

For real markets, $\mathcal{L}_{\text{common}}$ is the forecasting objective, while $\mathcal{L}_{\text{priv}}$ adds pathwise log-variance MSE and jump/regime cross-entropy. The encoder is frozen during target adaptation, and only the forecast head is fitted.

Both arms use the same 373,734-parameter Transformer [37]: three encoder layers, width 96, four heads, and a width-384 feed-forward block. A two-layer head maps the final token to the three horizons. Inputs are winsorized and normalized using training-split statistics. Real-market source pretraining uses 2,048 synthetic tapes for 12 epochs using AdamW [38] (learning rate 2×10^{-4}, batch size 64, $\lambda = 1$); head-only adaptation runs for 180 epochs at 8×10^{-4}. These settings are identical in the frozen configurations for all five markets. The controlled suite uses eight source epochs at 3×10^{-4} and 12 end-to-end target epochs at 10^{-4}.

5.3 Statistical and Econometric Protocol

Searching many priors creates selection risk, while overlapping forecast windows create serial dependence. The protocol addresses both. The real-market study separates discovery, confirmation, and chronological holdout. Candidate priors are selected only from replicated discovery runs modeled by a noise-aware Gaussian-process surrogate. Confirmation uses fresh repetitions that do not enter the surrogate or the selection rule, and the final holdout uses a later market episode untouched by both discovery and confirmation. Candidates are frozen before either confirmatory stage.

All comparisons are paired: the arms share source tapes, targets, initialization, optimizer, target-label draw, evaluation windows, and adaptation protocol. Only hidden-state supervision differs. Replicates estimate observation noise for the field.

Because neighboring examples share input bars, chronological separation is not treated as independence. Reported n counts independent training and fine-tuning replicates scored on one fixed evaluation window, not forecast origins. The three horizon losses are pooled within each replicate. Across episodes, log-skill is combined by fixed-effect inverse-variance weighting and heterogeneity tested with Cochran's Q. Prespecified directional tests are one-sided; confidence intervals are two-sided 95%. Markets are analyzed separately.

Computational and writing assistance. ChatGPT drafted the experimental implementation and the code used to produce figures; it was also used for LaTeX editing. Method development, transfer-field specification, experimental design, code execution and correction, result verification, and interpretation were carried out by the authors.

5.4 Field Estimation and Uncertainty

Each coordinate gives a noisy estimate of how much its source prior helps; repeated runs estimate that noise before the field is fitted. The transfer field is estimated from replicated paired evaluations rather than a single training run. Each market begins with the same 160-coordinate space-filling seed sweep. For ETHUSDT, the field-geometry analysis (Figure 3) additionally uses 40 adaptive coordinates chosen to reduce posterior uncertainty and sharpen the zero boundary; EURUSD and USDJPY received a further adaptive round of their own before candidate search, while XAUUSD and ES candidates were selected from the seed sweep and a market-specific top-12 search alone. At coordinate θ_i, the observation model is $y_i = F_b^R(\theta_i) + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \widehat{\sigma}_i^2)$, where $\widehat{\sigma}_i^2$ is estimated from repeated paired runs. A noise-aware Matérn-5/2 Gaussian process with separate length scales models each market; observation variance is replicate variance divided by replicate count. Figure 3 is a partial-dependence projection: at each (θ_1, θ_3) pair, it averages the six-dimensional GP posterior mean over the same 4,096-point Sobol design [39] used for envelope integration in coordinates $(\theta_2, \theta_4, \theta_5, \theta_6)$.

A deliberately poor source coordinate serves as a negative anchor. Together with the signed field, it tests whether gains depend on source-target alignment rather than a uniform regularization or optimization effect.

Table 2: Real-market confirmation summary. Val. and Hold. give the sign of log-skill at validation confirmation and untouched holdout. MSE gain and p are the final combined estimate (ETHUSDT, EURUSD, XAUUSD) or the holdout-stage estimate (USDJPY, ES), which is what determines each market's outcome.

Market	Asset class	Val.	Hold.	MSE gain	p	Confirmed
ETHUSDT	Cryptocurrency	+	+	7.0% pooled; 10.4% at $h = 60$	<0.0001	Yes
EURUSD	Foreign exchange	+	+	6.3%	0.002	Yes
XAUUSD	Precious metals	+	+	4.7%	0.005	Yes
USDJPY	Foreign exchange	+	−	−3.3%	0.367	No
ES	Equity-index futures	+	−	−0.7%	0.777	No

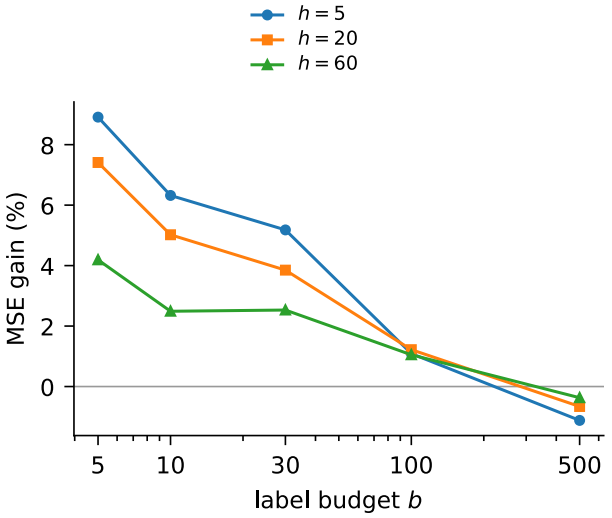


Figure 2: Controlled Bates–Heston-derived MSE gain over the matched forecaster, by label budget and horizon.

6 Synthetic Verification

The synthetic test isolates the mechanism before market mismatch. On held-out source worlds, decompression reaches state $R^2 = 0.737$ and jump AUC = 0.928.

Across 64 unseen covered targets, decompression improves all six five-label target–horizon cells by 3.77–10.37% relative to an identically pretrained forecast-only control. World-level win rates range from 76.6% to 85.9%, and every population-level Holm-adjusted p-value is 0.006. Figure 2 shows the gain by label budget and horizon: it is largest under sparse labels, attenuates overall as labels accumulate, and crosses zero near $b \approx 500$ — the attenuation is not strictly monotonic at every horizon, as the $h = 60$ gain ticks up slightly from $b = 10$ to $b = 30$. Numerical audits give a maximum Gaussian-identity error of 1.11×10^{-16}, Bates-equation RMSE of 7.75×10^{-5}, and horizon-identity error of 1.73×10^{-18}.

7 Real-Market Transfer Fields

7.1 Transfer-Field Geometry

Figure 3 shows distinct positive and negative regions after replicate noise is modeled. Candidates selected from earlier data remain positive on untouched holdout in three of five markets (Table 2). ETHUSDT's positive-envelope volume falls from 62.0% at $b = 5$ to 54.1% at $b = 10$, consistent with Proposition 4.1.

7.2 Five-Market Confirmation and Holdout

Frozen candidates confirm for ETHUSDT, EURUSD, and XAUUSD, reducing MSE by 7.0%, 6.3%, and 4.7%; ETHUSDT rises from 4.9% at $h = 5$ to 10.4% at $h = 60$. USDJPY and ES are positive in validation but reverse on untouched holdout.

ETHUSDT's row combines two independently discovered candidates. Candidate B confirms individually on holdout (log-skill +0.111, $p = 0.047$); candidate A has the same sign and similar magnitude (+0.098, $p = 0.095$). Their combined estimate gives a 7.0% pooled MSE reduction, and fades by roughly ten labels, consistent with Proposition 4.1: at $b = 10$, the two candidates give −0.005 ($p = 0.53$) and +0.019 ($p = 0.19$), while the negative anchor at $b = 5$ gives −0.099 ($t = -2.98$). EURUSD and XAUUSD each carry one independently discovered candidate, significant on the holdout stage alone before any pooling ($p = 0.016$ and $p = 0.028$ respectively), found via the same broad 160-coordinate seed sweep and a market-specific top-12 search, applied independently with no cross-market filtering.

The reversals occur when recovered value no longer covers mismatch and recovery cost in Eq. (14). Carry-trade transmission for USDJPY [29] and rough equity volatility for ES [25] are plausible omitted mechanisms, not identified causes. The matched forecast-pretrained control isolates latent-state supervision; realized-volatility forecasts are difficult to improve [17] but have direct risk-management value [15].

Classical benchmark. We implement Corsi's HAR-RV model [21]. Under the same five-label constraint and on identical held-out windows, its MSE is 1.78× that of market decompression ($p = 0.001$) and 1.66× that of the matched forecast-only pretrained model ($p = 0.003$). HAR-RV performs normally with full-history training, isolating its limitation to the cold-start regime.

8 Conclusion

Market decompression uses synthetic hidden-state supervision for forecasting with few real outcomes. The transfer field links recovered value, source mismatch, and label budget to prior selection. Frozen priors confirm for ETHUSDT, EURUSD, and XAUUSD, with gains up to 7.0% pooled and 10.4% at the 60-hour ETHUSDT horizon. USDJPY and ES reversals delimit the tested source family. No additional real labels or deployment-time states are required.

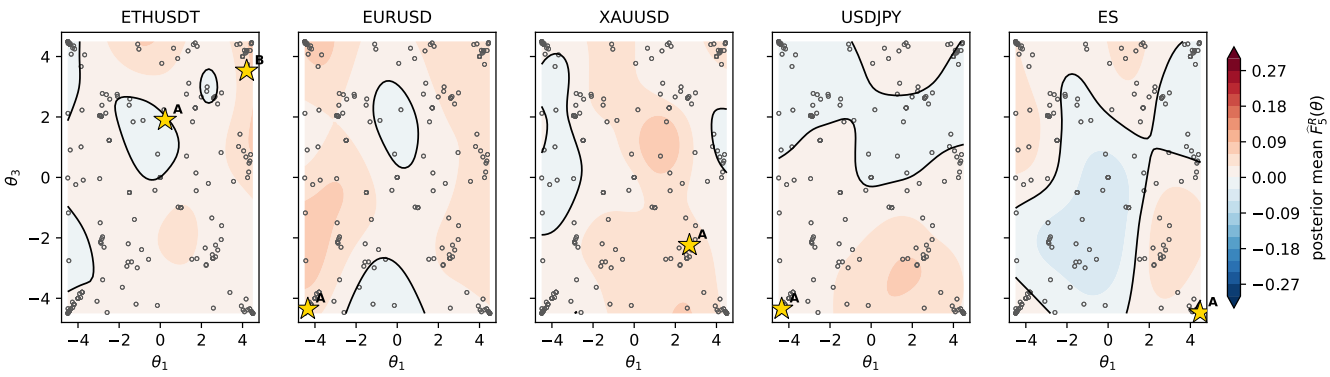


Figure 3: Five-label transfer fields. Color shows partial dependence over (θ_1, θ_3), averaged over a fixed Sobol design in the other four coordinates. Black contours mark the use/reject boundary; circles are sampled sources; gold stars are independently discovered candidates (two for ETHUSDT and one for each other market).

References

[1] Samuel A. Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E. Tillman, Prashant Reddy, and Manuela Veloso. 2021. Generating synthetic data in finance: Opportunities, challenges and pitfalls. In *Proceedings of ICAIF '20*. Article 44, 8 pages. https://doi.org/10.1145/3383455.3422554.

[2] Andrea Coletta, Aymeric Moulin, Svitlana Vyetrenko, and Tucker Balch. 2022. Learning to simulate realistic limit order book markets from data as a World Agent. In *Proceedings of ICAIF '22*, 428–436. https://doi.org/10.1145/3533271.3561753.

[3] Ryuji Hashimoto and Kiyoshi Izumi. 2025. Interpretable market simulations via optimal transport: Power law decomposition and implications for market design. In *Proceedings of ICAIF '25*, 35–43. https://doi.org/10.1145/3768292.3770338.

[4] Svitlana Vyetrenko, David Byrd, Nick Petosa, Mahmoud Mahfouz, Danial Dervovic, Manuela Veloso, and Tucker Balch. 2021. Get real: Realism metrics for robust limit order book market simulations. In *Proceedings of ICAIF '20*. Article 2, 8 pages. https://doi.org/10.1145/3383455.3422561.

[5] Christopher J. Cho and Timothy J. Norman. 2022. Bit by bit: How to realistically simulate a crypto-exchange. In *Proceedings of ICAIF '21*. Article 15, 9 pages. https://doi.org/10.1145/3490354.3494380.

[6] Namid R. Stillman, Rory Baggott, Justin Lyon, Jianfei Zhang, Dingqiu Zhu, Tao Chen, and Perukrishnen Vytelingum. 2023. Deep calibration of market simulations using neural density estimators and embedding networks. In *Proceedings of ICAIF '23*, 46–54. https://doi.org/10.1145/3604237.3626881.

[7] Aldo Glielmo, Marco Favorito, Debmallya Chanda, and Domenico Delli Gatti. 2023. Reinforcement learning for combining search methods in the calibration of economic ABMs. In *Proceedings of ICAIF '23*, 305–313. https://doi.org/10.1145/3604237.3626889.

[8] Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural Networks* 22, 5–6, 544–557. https://doi.org/10.1016/j.neunet.2009.06.042.

[9] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning* 79, 1–2, 151–175. https://doi.org/10.1007/s10994-009-5152-4.

[10] Thomas J. Santner, Brian J. Williams, and William I. Notz. 2018. *The Design and Analysis of Computer Experiments*, 2nd ed. Springer. https://doi.org/10.1007/978-1-4939-8847-1.

[11] M. Hashem Pesaran and Allan Timmermann. 2007. Selection of estimation window in the presence of breaks. *Journal of Econometrics* 137, 1, 134–161. https://doi.org/10.1016/j.jeconom.2006.03.010.

[12] M. Hashem Pesaran, Andreas Pick, and Mikhail Pranovich. 2013. Optimal forecasts in the presence of structural breaks. *Journal of Econometrics* 177, 2, 134–152. https://doi.org/10.1016/j.jeconom.2013.04.002.

[13] Barbara Rossi. 2013. Advances in forecasting under instability. In Graham Elliott and Allan Timmermann (Eds.), *Handbook of Economic Forecasting*, Vol. 2B, Chapter 21, 1203–1324. Elsevier. https://doi.org/10.1016/B978-0-444-62731-5.00021-X.

[14] Andreas Teller, Uta Pigorsch, and Christian Pigorsch. 2026. Realized volatility forecasting for new issues and spin-offs using multi-source transfer learning. *International Journal of Forecasting* 42, 3, 1069–1103. https://doi.org/10.1016/j.ijforecast.2026.02.004.

[15] Tim Bollerslev, Benjamin Hood, John Huss, and Lasse Heje Pedersen. 2018. Risk everywhere: Modeling and managing volatility. *Review of Financial Studies* 31, 7, 2729–2773. https://doi.org/10.1093/rfs/hhy041.

[16] Fan Yang, Hongyang R. Zhang, Sen Wu, Christopher Ré, and Weijie J. Su. 2025. Precise high-dimensional asymptotics for quantifying heterogeneous transfers. *Journal of Machine Learning Research* 26, 113, 1–88. https://www.jmlr.org/papers/v26/24-0454.html.

[17] Rafael R. Branco, Alexandre Rubesam, and Mauricio Zevallos. 2024. Forecasting realized volatility: Does anything beat linear models? *Journal of Empirical Finance* 78, Article 101524. https://doi.org/10.1016/j.jempfin.2024.101524.

[18] Steven L. Heston. 1993. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* 6, 2, 327–343. https://doi.org/10.1093/rfs/6.2.327.

[19] David S. Bates. 1996. Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. *Review of Financial Studies* 9, 1, 69–107. https://doi.org/10.1093/rfs/9.1.69.

[20] Leif B. G. Andersen. 2008. Simple and efficient simulation of the Heston stochastic volatility model. *Journal of Computational Finance* 11, 3, 1–42. https://doi.org/10.21314/JCF.2008.189.

[21] Fulvio Corsi. 2009. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7, 2, 174–196. https://doi.org/10.1093/jjfinec/nbp001.

[22] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2nd ed. Wiley-Interscience. https://doi.org/10.1002/047174882X.

[23] Naftali Tishby, Fernando C. Pereira, and William Bialek. 2000. The information bottleneck method. arXiv:physics/0004057. https://arxiv.org/abs/physics/0004057.

[24] CoinDesk Research. 2025. The $19 billion liquidation that shook crypto. Published 17 October 2025; updated 30 October 2025; accessed 29 July 2026. https://www.coindesk.com/research/market-spotlight-the-19-billion-liquidation-that-shook-crypto.

[25] Jim Gatheral, Thibault Jaisson, and Mathieu Rosenbaum. 2018. Volatility is rough. *Quantitative Finance* 18, 6, 933–949. https://doi.org/10.1080/14697688.2017.1393551.

[26] Andrea Coletta, Matteo Prata, Michele Conti, Emanuele Mercanti, Novella Bartolini, Aymeric Moulin, Svitlana Vyetrenko, and Tucker Balch. 2022. Towards realistic market simulations: A generative adversarial networks approach. In *Proceedings of ICAIF '21*. Article 46, 9 pages. https://doi.org/10.1145/3490354.3494411.

[27] Peer Nagy, Sascha Frey, Silvia Sapora, Kang Li, Anisoara Calinescu, Stefan Zohren, and Jakob Foerster. 2023. Generative AI for end-to-end limit order book modelling: A token-level autoregressive generative model of message flow using a deep state space network. In *Proceedings of ICAIF '23*, 91–99. https://doi.org/10.1145/3604237.3626898.

[28] Quoc Anh Nguyen, Ce Guo, and Wayne Luk. 2025. Repurposing language models for FX volatility forecasting: A data-efficient and context-aware approach. In *Proceedings of ICAIF '25*, 465–473. https://doi.org/10.1145/3768292.3770386.

[29] Matteo Aquilina, Marco Jacopo Lombardi, Andreas Schrimpf, and Vladyslav Sushko. 2024. The market turbulence and carry trade unwind of August 2024. *BIS Bulletin* No. 90. Bank for International Settlements. https://www.bis.org/publ/bisbull90.htm.

[30] Soosung Hwang and Pedro L. Valls Pereira. 2006. Small sample properties of GARCH estimates and persistence. *European Journal of Finance* 12, 6–7, 473–494. https://doi.org/10.1080/13518470500039436.

[31] Xiong Lyu, Mickaël Binois, and Michael Ludkovski. 2021. Evaluating Gaussian process metamodels and sequential designs for noisy level set estimation. *Statistics*

and Computing 31, Article 43. https://doi.org/10.1007/s11222-021-10014-w.

[32] Rickard K. A. Karlsson, Martin Willbo, Zeshan M. Hussain, Rahul G. Krishnan, David Sontag, and Fredrik Johansson. 2022. Using time-series privileged information for provably efficient learning of prediction models. In *Proceedings of AISTATS*, PMLR 151, 5459–5484. https://proceedings.mlr.press/v151/k-a-karlsson22a.html.

[33] Bastian Jung and Fredrik D. Johansson. 2022. Efficient learning of nonlinear prediction models with time-series privileged information. In *Advances in Neural Information Processing Systems 35*, 19048–19060. https://proceedings.neurips.cc/paper_files/paper/2022/hash/792845ddfe4047d7066348e52e46b74d-Abstract-Conference.html.

[34] Binance. 2026. ETHUSDT spot hourly kline archive. Accessed 29 July 2026. https://data.binance.vision/?prefix=data/spot/monthly/klines/ETHUSDT/1h/.

[35] ejtraderLabs. 2022. Major Forex historical data, hourly EURUSD, USDJPY, and XAUUSD files, commit `fbd29b3cd85c0eea4f6e8b81c053f98fb3de22fd`. Accessed 29 July 2026. https://github.com/ejtraderLabs/historical-data/tree/fbd29b3cd85c0eea4f6e8b81c053f98fb3de22fd.

[36] Yahoo Finance. 2026. E-mini S&P 500 futures continuous contract historical data (`ES=F`). Accessed 24 July 2026. https://finance.yahoo.com/quote/ES=F/history/.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, 5998–6008. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[38] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7.

[39] I. M. Sobol'. 1967. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics* 7, 4, 86–112. https://doi.org/10.1016/0041-5553(67)90144-9.