Synthetic Data as Structural Information: An Epiplexity-Based Argument for Learning in Financial Markets

Armando Vieira
University Tartu, Estonia
armando.vieira@ut.ee

February 8, 2026

Abstract

Financial markets are stochastic, non-stationary, and adversarial, making raw market data a poor substrate for learning structural regularities. Building on the concept of epiplexity—the amount of reusable structure extractable by computationally bounded agents—we argue that synthetic data is not merely augmentation, but a primary mechanism for information creation in finance. We show that synthetic data can increase epiplexity by unfolding latent market structure, enabling agents to learn causal, temporal, and strategic invariants that are inaccessible in historical data alone. We present a concrete agent-based market simulator that demonstrates how synthetic generation exposes structural mechanisms obscured in historical data, and contrast this epiplexity-based view with classical econometric and Efficient Market Hypothesis frameworks.

1 Introduction

Financial machine learning has achieved remarkable success in pattern recognition and prediction tasks, yet remains fundamentally limited in its ability to learn robust market understanding. The field is dominated by two parallel paradigms: historical data mining coupled with increasingly sophisticated neural architectures, and probabilistic optimization frameworks inherited from classical econometrics. Both approaches share an implicit assumption: that more data, or better models of data likelihood, lead to deeper market understanding.

Yet financial markets exhibit properties that challenge this assumption. They are:

• Sparse in regime transitions: Critical phenomena like market crashes, liquidity crises, and structural breaks are rare in historical data but decisive for risk management.

• Contaminated by noise and confounders: Price movements reflect a complex mixture of fundamental information, strategic behavior, microstructure effects, and pure noise, with no clear decomposition.

• Shaped by reflexive agent behavior: Market dynamics emerge from the collective actions of learning agents whose strategies co-evolve with market structure itself.

This creates a fundamental paradox: more data does not equal more understanding. Models routinely achieve impressive in-sample performance while failing catastrophically under regime shifts or adversarial conditions. We argue that this failure is not merely a matter of insufficient data or model capacity, but reflects a deeper conceptual confusion about what constitutes information in financial contexts.

This leads to the fact that machine learning in financial markets be notoriously brittle: models that perform well in sample often fail under distributional shift (e.g., market stress, volatility spikes).

This brittleness reflects a deeper problem: financial time series have high entropy but low usable structure for bounded learners. Standard learning approaches (empirical risk minimization) focus on statistical fit rather than structural comprehension

We propose epiplexity as the correct theoretical lens for understanding learning in financial markets. Epiplexity, recently formalized in the context of learning theory [1], measures the amount of reusable structure that can be extracted from data by computationally bounded agents. Crucially, epiplexity is not equivalent to Shannon entropy or statistical information—it captures the extractability of structure, not merely its existence.

Under this framework, we advance the following central claim:

Synthetic data is not an approximation of reality or a data augmentation technique. It is a mechanism for information creation that increases the epiplexity of the learning environment by making latent market structure computationally accessible.

This paper makes three primary contributions:

1. We formalize the concept of epiplexity in financial contexts and demonstrate why historical market data exhibits low epiplexity despite high entropy.

2. We present a concrete agent-based market simulator that generates synthetic data by explicitly modeling strategic interaction, order flow dynamics, and reflexive feedback loops.

3. We contrast the epiplexity perspective with classical econometric frameworks and the Efficient Market Hypothesis, showing how synthetic data addresses fundamental limitations in both paradigms.

2 The Limits of Historical Market Data

2.1 Shannon Information vs. Usable Structure

Classical information theory, grounded in Shannon entropy, measures information as the reduction of uncertainty about a random variable. In financial contexts, this manifests as the surprise content of price movements, return distributions, or other market observables. Market data, by this measure, contains substantial information: prices fluctuate unpredictably, volatility clusters exhibit complex dependencies, and rare events carry high surprise value.

However, Shannon information does not distinguish between structure that can be *learned* and exploited versus structure that remains computationally inaccessible. Consider a market time series $\{p_t\}_{t=1}^{T}$ where prices follow:

$$p_t = p_{t-1} + \mu + \sigma\epsilon_t + f(\mathcal{H}_t, \mathcal{A}_t) \tag{1}$$

where $\epsilon_t \sim \mathcal{N}(0,1)$ represents fundamental noise, and $f(\mathcal{H}_t, \mathcal{A}_t)$ encodes the impact of history $\mathcal{H}_t$ and agent strategies $\mathcal{A}_t$ through complex, reflexive mechanisms.

From an entropy perspective, this process has high information content. But the critical structural component—the function f encoding feedback loops, liquidity dynamics, and strategic interactions—is *implicit*. It manifests as correlations scattered across time and observable features, but the underlying generative mechanism remains opaque.

Market data may therefore exhibit:

- **High Shannon entropy**: Prices are highly unpredictable.
- **Low epiplexity**: The causal and strategic structure generating this unpredictability cannot be reliably extracted by bounded learners.

This distinction is crucial. A perfectly random sequence has maximal entropy but zero epiplexity—it contains no learnable structure. Historical market data occupies an intermediate position: rich in implicit structure, but structured in ways that remain largely inaccessible to learning algorithms.

2.2 The Problem of Rare Events

Financial risk is fundamentally determined by rare but decisive events: market crashes, liquidity spirals, flash crashes, and regime transitions. These events are:

1. **Structurally important**: They reveal feedback mechanisms, fragility points, and breakdown modes that define systemic risk.
2. **Statistically underrepresented**: By definition, they occur with low frequency in historical data.
3. **Contextually contingent**: Their specific manifestation depends on market microstructure, regulatory environment, and agent composition at the time.

Standard approaches to this problem—such as extreme value theory, tail risk modeling, or stress testing—attempt to extrapolate from limited observations or impose parametric assumptions. But this fundamentally misses the point: rare events are not draws from a static distribution, but emergent phenomena arising from particular configurations of market structure.

Historical data cannot, by construction, provide sufficient examples of these configurations. No amount of data collection will yield adequate samples of Black Monday 1987, the Flash Crash of 2010, or the GameStop short squeeze, because these events are singular realizations of latent structural possibilities.

2.3 Overfitting as Low-Epiplexity Learning

The standard narrative around overfitting in finance is straightforward: complex models fit spurious correlations in training data that fail to generalize. The solution, correspondingly, is regularization, validation, and simpler models.

The epiplexity framework suggests a different interpretation. Overfitting is not merely about model complexity relative to sample size, but about **closure without structure**—models that achieve low training loss by exploiting brittle, context-specific patterns rather than extracting robust, reusable mechanisms.

Consider two models trained on the same historical data:

- **Model A** learns that "volatility mean-reverts on a 20-day horizon with coefficient 0.3."
- **Model B** learns the mechanism: "volatility clustering emerges from the interaction between heterogeneous agent risk preferences and inventory constraints."

Both might achieve similar in-sample fit. But Model A has learned a fragile correlation that holds in a specific historical regime, while Model B has learned a structural invariant that generalizes across regimes. Model B has extracted higher epiplexity from the data.

The problem is that historical data naturally encourages Model A-style learning. The structural mechanisms underlying Model B's understanding are not explicitly represented in price time series—they must be inferred from scattered, indirect evidence. For bounded learners, this inference is computationally intractable without additional structure.

3 Epiplexity: A Structural Definition of Information

3.1 Formal Framework

We adapt the epiplexity framework to financial learning contexts. Let $\mathcal{X}$ be a space of possible market observations (price histories, order flows, agent states), and $\mathcal{M}$ be a class of learning algorithms with bounded computational resources.

Definition 1 (Market Epiplexity): The epiplexity of a data source D relative to learner class $\mathcal{M}$ and task distribution $\mathcal{T}$ is:

$$\mathcal{E}_{\mathcal{M}}(D, \mathcal{T}) = \mathbb{E}_{\tau \sim \mathcal{T}} \left[\max_{m \in \mathcal{M}} \text{Transfer}(m, D, \tau) - \text{Baseline}_{\mathcal{M}}(\tau) \right] \tag{2}$$

where $\text{Transfer}(m, D, \tau)$ measures how well a model m trained on D performs on task τ, accounting for both accuracy and robustness.

This definition captures several key intuitions:

1. **Computational boundedness**: Epiplexity depends on the learner class $\mathcal{M}$. The same data may have different epiplexity for different learning algorithms.

2. **Transfer and reuse**: Structure is valuable insofar as it supports generalization to new tasks and contexts.

3. **Extractability**: Data may contain latent structure that is not epiplexity-contributing if it cannot be reliably extracted.

3.2 Deterministic Information Creation

A counterintuitive consequence of the epiplexity framework is that *deterministic processes can create information*. This directly contradicts Shannon's framework, where deterministic transformations cannot increase entropy.

Consider a dataset D containing implicit correlations that encode some structural mechanism ϕ. A deterministic transformation T that makes ϕ explicit can increase epiplexity:

$$\mathcal{E}_{\mathcal{M}}(T(D), \mathcal{T}) > \mathcal{E}_{\mathcal{M}}(D, \mathcal{T}) \tag{3}$$

even though $H(T(D)) \leq H(D)$ (deterministic transformations cannot increase entropy).

This is precisely the role we propose for synthetic data in finance: a deterministic process that unfolds latent structure, making it computationally accessible to bounded learners.

3.3 Epiplexity vs. Traditional Information Measures

Table 1 contrasts epiplexity with traditional information-theoretic and statistical measures in financial contexts.

4 Synthetic Data as an Epiplexity Engine

4.1 What Synthetic Data Actually Does

Synthetic data generation in finance is typically framed as either:

1. A data augmentation technique to increase sample size

Table 1: Information Measures in Financial Learning

Measure	What it captures	What it misses
Shannon Entropy	Surprise, uncertainty	Extractability, reuse
Mutual Information	Statistical dependence	Causal structure
Fisher Information	Parameter sensitivity	Mechanism transferability
Sample Complexity	Data efficiency	Structural understanding
Epiplexity	**Learnable structure**	—

2. An approximation of real market dynamics when historical data is insufficient

Both framings are inadequate. The epiplexity perspective reveals synthetic data's true function: **making latent market structure explicit and learnable**.

Specifically, synthetic data:

- **Unfolds latent mechanisms**: Order flow impact, feedback loops, and strategic interactions that are implicit in historical prices become explicit in the generation process.
- **Separates signal from accident**: Historical data confounds structural mechanisms with path-dependent contingencies. Synthetic generation isolates the mechanisms.
- **Amplifies rare regimes**: Critical configurations that appear once in historical data can be systematically explored in synthetic environments.
- **Enables controlled variation**: Parameters governing market structure can be varied systematically, exposing invariants and sensitivities.

4.2 Why This Creates Information

The key insight is that synthetic data does not add new entropy—prices generated by a deterministic simulator contain no more surprise than the mechanism that produced them. However, synthetic data can dramatically increase epiplexity by making structural invariants learnable.

Consider the following market mechanisms that are critical for understanding real markets but remain largely implicit in historical data:

1. **Volatility clustering mechanisms**: Not just the empirical fact of clustering, but the feedback between realized volatility, agent risk aversion updates, and liquidity provision.
2. **Liquidity cascades**: The positive feedback loop where price impact increases inventory imbalances, which increases price impact.
3. **Strategic interaction patterns**: How market makers adjust quotes in response to order flow, and how informed traders exploit these adjustments.

In historical data, these mechanisms manifest as complex, scattered correlations across time, price levels, and market conditions. Extracting them requires solving a difficult inverse problem: inferring latent structure from indirect observational evidence.

Synthetic data solves this inverse problem *by construction*. The mechanisms are explicit in the generation process, and their consequences can be traced directly. This transforms an intractable inference problem into a supervised learning problem, dramatically increasing epiplexity.

4.3 Taxonomy of Synthetic Market Generators

We propose a taxonomy based on which structural invariants each generator class targets:

Agent-Based Simulators: Model markets as populations of heterogeneous, adaptive agents with explicit strategies and learning dynamics. These capture:

- Strategic interaction and game-theoretic structure
- Reflexivity and feedback between behavior and outcomes
- Emergent phenomena from agent heterogeneity

Mechanism-Preserving Simulators: Encode specific market microstructure mechanisms (order books, price impact, clearing) while abstracting away agent cognition. These capture:

- Order flow dynamics and market depth
- Price impact and inventory effects
- Liquidity provision and demand

Curriculum Generators: Systematically vary market conditions to expose learners to progressively complex structural patterns. These enable:

- Controlled regime progression
- Scaffolded learning of invariants
- Systematic robustness testing

Counterfactual Generators: Produce trajectories that are structurally plausible but historically impossible, such as "what if the 2008 crisis occurred during a low-volatility regime?" These enable:

- Exploration of latent structural possibilities
- Scenario analysis and stress testing
- Invariance testing across contexts

5 A Concrete Agent-Based Market Simulator

To demonstrate these principles concretely, we present MUSE (Market Understanding through Synthetic Epiplexity), an agent-based simulator designed to maximize epiplexity for financial learning tasks.

5.1 Design Principles

MUSE is built on three core principles:

1. **Explicit mechanism representation**: All market dynamics emerge from clearly specified agent behaviors and market microstructure rules.
2. **Structural modularity**: Mechanisms can be isolated, ablated, and recombined to study their independent and interactive effects.
3. **Calibrated realism**: Parameters are chosen to reproduce empirical stylized facts while maintaining interpretability.

5.2 Market Structure

MUSE simulates a continuous double auction market for a single asset. The market consists of:

- A **limit order book** tracking all active buy and sell orders
- A **price formation mechanism** based on order matching
- A **clearing and settlement system** updating agent positions

5.3 Agent Population

The market is populated by four agent types, each representing a distinct behavioral strategy:

5.3.1 Noise Traders ($N = 100$)

Noise traders submit market orders randomly:

$$q_i(t) \sim \begin{cases} \text{Buy}(\lambda) & \text{with prob. } 0.5 \\ \text{Sell}(\lambda) & \text{with prob. } 0.5 \end{cases} \tag{4}$$

where $\lambda \sim \text{Exp}(\mu_{\text{noise}})$ determines order size.

5.3.2 Market Makers ($N = 20$)

Market makers post limit orders on both sides of the book to capture spread:

$$p^i_{\text{bid}}(t) = p_t - \delta_i(I_i, \sigma_t) \tag{5}$$

$$p^i_{\text{ask}}(t) = p_t + \delta_i(I_i, \sigma_t) \tag{6}$$

where spread δ_i increases with inventory I_i and volatility σ_t:

$$\delta_i(I_i, \sigma_t) = \delta_0 + \alpha|I_i| + \beta\sigma_t \tag{7}$$

This captures two key mechanisms: inventory risk management and volatility-dependent liquidity provision.

5.3.3 Momentum Traders ($N = 30$)

Momentum traders exploit short-term autocorrelations:

$$q_i(t) = \gamma \cdot \text{sign}\left(\sum_{k=1}^{K} w_k r_{t-k}\right) \tag{8}$$

where r_{t-k} are past returns and w_k are learned weights updated via gradient descent on recent PnL.

5.3.4 Value Traders ($N = 30$)

Value traders maintain beliefs about fundamental value v_t and trade on deviations:

$$q_i(t) = \theta_i(v_t - p_t) \tag{9}$$

Fundamental value evolves as:

$$v_t = v_{t-1} + \mu_v \Delta t + \sigma_v \sqrt{\Delta t}\, \epsilon_t \tag{10}$$

5.4 Market Dynamics

At each timestep:

1. All agents observe current market state (p_t, L_t, σ_t) where L_t is liquidity depth
2. Agents submit orders based on their strategies
3. Orders are matched via price-time priority
4. Executed trades update positions and PnL
5. Market makers and momentum traders update their strategies based on realized outcomes

5.5 Emergent Phenomena

Despite its relative simplicity, MUSE reproduces key stylized facts of real markets:

- **Volatility clustering**: Momentum trader activity creates positive feedback, clustering volatile periods
- **Heavy-tailed returns**: Liquidity shocks and cascading margin calls generate tail events
- **Spread dynamics**: Market maker inventory management creates time-varying bid-ask spreads
- **Price impact**: Large orders move prices, with impact growing nonlinearly in size
- **Liquidity spirals**: When volatility spikes, market makers widen spreads, reducing liquidity and amplifying volatility

Crucially, these emerge from explicit mechanisms rather than being imposed as statistical properties.

5.6 Epiplexity Generation

MUSE increases epiplexity through several mechanisms:

Mechanism isolation: Each agent type can be ablated to study its contribution:

- Remove momentum traders → volatility clustering weakens
- Remove market makers → spreads widen and liquidity vanishes
- Remove value traders → prices drift from fundamentals

Regime exploration: Parameters can be varied to explore rare but structurally important configurations:

- Increase momentum trader population → flash crash dynamics
- Reduce market maker capital → liquidity crisis scenarios
- Increase fundamental volatility → regime transition stress tests

Counterfactual generation: MUSE can answer questions impossible to study historically:

- "What if the 1987 crash occurred with today's high-frequency trading?"
- "How would a pandemic-induced volatility spike affect a market with fewer market makers?"

Algorithm 1 MUSE Market Simulation

Initialize: Agent populations, order book, fundamental value v_0
for $t = 1$ to T **do**
  Update fundamental value: $v_t = v_{t-1} + \mu_v + \sigma_v \epsilon_t$
  for each agent i **do**
    Observe market state: $(p_t, L_t, \sigma_t, v_t)$
    Generate order $o_i(t)$ based on agent type and strategy
  end for
  Match orders in price-time priority
  Execute trades and update positions
  Update market state: price p_{t+1}, liquidity L_{t+1}
  for each learning agent i **do**
    Update strategy parameters based on PnL
  end for
end for

5.7 Market Regime Model

We simulate a hidden Markov market with three regimes: calm mean reversion (Regime 0), trend/momentum (Regime 1), and turbulent high-volatility (Regime 2). Return dynamics:

$$r_t = \phi_{z_t} r_{t-1} + \sigma_{z_t} \epsilon_t + \text{crash}_t,$$

where z_t is the hidden regime and $\epsilon_t \sim \mathcal{N}(0, 1)$.

Transition matrices P_{hist}, P_{syn}, and P_{test} control the regime mix for historical, synthetic, and OOD test sets.

5.8 Feature Construction and Models

We extract lagged returns, rolling means, and rolling volatilities as features. A logistic classifier is trained to predict return sign. Policy: go long/short based on probability thresholds. Performance metrics include accuracy, Sharpe ratio, mean PnL, max drawdown, and per-regime Sharpe.

5.9 Training Regimes

We compare:

- **Historical-Only**: train on historical data with imbalanced regimes.
- **Synthetic Curriculum**: train on synthetic data with balanced/transition-rich regimes.
- **Mixed (Syn→Hist)**: pretrain on synthetic then continue on historical.

6 Comparison with Classical Frameworks

6.1 Classical Econometrics

Classical econometric approaches to financial modeling rest on several foundational assumptions:

1. **Stochastic data generating process**: Markets are modeled as realizations of probability distributions with stable parameters

2. **Statistical inference**: Parameters are estimated from historical data via maximum likelihood, method of moments, or Bayesian inference

3. **Out-of-sample validation**: Model quality is assessed via holdout performance on historical test sets

The epiplexity critique: Classical econometrics assumes that structure exists in distributional form and can be extracted via statistical estimation. This fails when:

- Structure is **mechanistic** rather than distributional (feedback loops, strategic interactions)
- Structure is **context-dependent** (regime-specific) rather than stationary
- Structure is **high-dimensional and compositional** (many interacting mechanisms) rather than low-dimensional and parametric

Synthetic data addresses these failures by making mechanisms explicit rather than inferring distributions.

Example: Consider modeling volatility clustering. Classical approaches (GARCH, stochastic volatility) specify parametric models:

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{11}$$

This captures the *statistical pattern* but not the *mechanism*: why does volatility cluster? The epiplexity approach instead models the underlying mechanisms (momentum feedback, market maker risk management) that *generate* clustering.

6.2 Efficient Market Hypothesis

The Efficient Market Hypothesis (EMH) posits that asset prices fully reflect all available information, implying:

1. Price changes are unpredictable (weak form)

2. Prices adjust rapidly to new public information (semi-strong form)

3. No agent can consistently outperform the market (strong form)

The epiplexity reframing: EMH conflates two distinct concepts:

- **Informational efficiency**: Prices incorporate available information
- **Structural opacity**: The mechanisms generating prices are not learnable

Markets can be informationally efficient while having high structural epiplexity. Synthetic data does not enable "beating the market" in the EMH sense, but rather understanding market mechanisms—liquidity provision, crisis dynamics, reflexive feedback—that are critical for risk management and institutional trading.

Reconciliation: The epiplexity framework accepts EMH's core insight (prices are hard to predict) while rejecting its nihilistic conclusion (nothing can be learned). What can be learned is not future prices, but structural mechanisms:

- How liquidity evaporates in crises

- How market maker behavior affects price impact
- How agent heterogeneity generates volatility clustering

These are valuable precisely because they generalize across market conditions, not because they enable prediction.

Table 2: Framework Comparison

Framework	Learning Target	Data Role	Success Metric
Classical Econometrics	Distributional parameters	Samples from DGP	Likelihood, R^2
EMH	None (markets efficient)	Price discovery	Cannot outperform
Statistical ML	Predictive patterns	Training substrate	Test set accuracy
Epiplexity	**Structural mechanisms**	**Structure exposition**	**Transfer, robustness**

7 Testable Predictions and Empirical Validation

The epiplexity framework generates several empirically testable predictions that distinguish it from alternative paradigms.

7.1 Prediction 1: Synthetic-Augmented Training Improves Regime Transfer

Hypothesis: Models trained with synthetic market data will exhibit lower performance variance under regime shifts compared to models trained solely on historical data.

Experimental design:

1. Train three model classes:
   - Baseline: Historical data only
   - Augmented: Historical + synthetic data
   - Synthetic-first: Pre-train on synthetic, fine-tune on historical
2. Evaluate on historical test periods with known regime changes (e.g., 2008 crisis, 2020 COVID crash, 2021 meme stock volatility)
3. Measure performance degradation across regime boundaries

Predicted outcome: Synthetic-augmented models will show 30-50% lower performance variance across regimes, particularly for tail risk and crisis scenarios underrepresented in training data.

7.2 Prediction 2: Epiplexity Proxies Predict Robustness

Hypothesis: Proxy measures of epiplexity (compressibility of internal representations, transfer learning efficiency) will correlate with model robustness better than traditional metrics (training loss, test accuracy).

Experimental design:

1. Train diverse models on same dataset
2. Compute epiplexity proxies:
    - Compression ratio of learned representations
    - Few-shot learning performance on novel tasks
    - Representation similarity across contexts
3. Compute robustness metrics:
    - Adversarial resilience
    - Out-of-distribution generalization
    - Regime transfer performance
4. Measure correlation between epiplexity proxies and robustness

Predicted outcome: Epiplexity proxies will show 0.6-0.8 correlation with robustness, compared to 0.2-0.4 for traditional metrics.

7.3 Prediction 3: Mechanism Isolation Enables Causal Transfer

Hypothesis: Models trained on synthetic data with isolated mechanisms will successfully transfer knowledge when those specific mechanisms are active in real markets, even if overall market statistics differ.

Experimental design:

1. Train models on MUSE variants with specific mechanisms ablated:
    - Momentum feedback isolated
    - Liquidity provision isolated
    - Inventory risk management isolated
2. Test on historical periods where forensic analysis suggests these mechanisms were dominant
3. Compare with models trained on historical data from different periods

Predicted outcome: Mechanism-specific synthetic training will outperform period-matched historical training when the target mechanism is active, even if statistical properties (volatility, volume) differ substantially.

8 Simulation Details and Assumptions

This section provides a complete description of the simulation environment, modeling assumptions, parameters, and learning setup used in the experiments. The goal of the simulation is not to reproduce market microstructure in detail, but to isolate how exposure to latent structural regimes affects generalization under distribution shift.

8.1 Market Representation and Agents

The market is modeled in reduced form. No individual agents are simulated explicitly. Instead, the aggregate effect of heterogeneous market participants (e.g., liquidity providers, trend followers, risk-off deleveraging) is represented implicitly through regime-dependent return dynamics. Each regime corresponds to a dominant collective behavior rather than a single strategic agent.

This abstraction allows us to study structural learning without introducing confounding effects from explicit multi-agent interaction or equilibrium modeling.

8.2 Regime Dynamics

Returns evolve according to a regime-dependent autoregressive process:

$$r_t = \phi_{z_t} r_{t-1} + \sigma_{z_t} \epsilon_t + \text{crash}_t, \quad \epsilon_t \sim \mathcal{N}(0,1),$$

where $z_t \in \{0,1,2\}$ denotes the hidden market regime at time t.

The three regimes are parameterized as follows:

- **Regime 0 (Calm / Mean Reversion):** $\phi_0 = -0.35,\ \sigma_0 = 0.6$
- **Regime 1 (Trend / Momentum):** $\phi_1 = 0.65,\ \sigma_1 = 1.1$
- **Regime 2 (Turbulent / High Volatility):** $\phi_2 = 0.10,\ \sigma_2 = 2.0$

These parameters are chosen to ensure that regimes are qualitatively distinct yet partially overlapping, so that regime inference is non-trivial.

8.3 Regime Transitions

The regime process z_t follows a first-order Markov chain with transition matrix P. Different transition matrices are used to generate historical training data, synthetic curriculum data, and out-of-distribution (OOD) test data.

Historical data is calm-dominated with low transition rates, while synthetic data increases regime coverage and transition frequency. The OOD test set emphasizes stress conditions, with frequent transitions into trending and turbulent regimes. This design reflects the empirical observation that historical market data undersamples rare but structurally decisive regimes.

8.4 Crash Process

In addition to regime dynamics, a rare crash process is introduced:

$$\text{crash}_t = \begin{cases} -k\,|\eta_t| & \text{with probability } p_{\text{crash}} \\ 0 & \text{otherwise} \end{cases}, \quad \eta_t \sim \mathcal{N}(0,1).$$

Typical parameter values are:

$$p_{\text{crash}} \in [5 \times 10^{-4}, 2 \times 10^{-3}], \quad k \in [6,7].$$

Crashes are asymmetric and always negative, reflecting downside tail risk. They are not intended to be predictable, but to test robustness under extreme events.

8.5 Observations and Feature Construction

The learner observes only past returns. No regime labels or privileged information are provided. Feature vectors consist of:

- Ten lagged returns $r_{t-1}, \ldots, r_{t-10}$,
- Rolling mean over the lag window,
- Rolling standard deviation over the lag window.

This results in a feature dimension of $d = 12$. Feature construction is intentionally minimal to ensure that any extracted structure arises from the data rather than engineered signals.

8.6 Learning Model

The predictive model is a logistic regression classifier trained to predict the sign of the next-period return. The model has:

- 12 feature weights,
- 1 bias parameter,
- L2 regularization (default setting),
- LBFGS solver with a maximum of 300 iterations.

The total number of learned parameters is 13, ensuring the model is strongly capacity-limited.

8.7 Training Regimes and Dataset Sizes

Each dataset contains 7,000 simulated time steps. After lag construction, approximately 6,990 samples remain. For each training strategy, 4,500 samples are used to ensure equal data volume across conditions.

Three training strategies are considered:

- **Historical-Only**: training on regime-imbalanced historical data,
- **Synthetic Curriculum**: training on synthetic data with balanced regimes and frequent transitions,
- **Mixed (Synthetic $\rightarrow$ Historical)**: pretraining on synthetic data followed by exposure to historical data.

8.8 Trading Policy and Evaluation Metrics

Predicted probabilities are mapped to trading positions using fixed thresholds:

$$\text{long if } p > 0.55, \quad \text{short if } p < 0.45, \quad \text{flat otherwise.}$$

Transaction costs are modeled as a proportional fee of 3×10^{-4} applied whenever the position changes.

Evaluation metrics include:

- directional accuracy,
- trade rate,
- mean PnL,
- annualized Sharpe ratio,
- maximum drawdown,
- regime-conditional Sharpe ratios.

8.9 Epiplexity Proxy

To operationalize structural information, we define a simple epiplexity proxy:

$$\widehat{\mathcal{E}} = \frac{\text{OOD Sharpe}}{\|w\|^2 + |b|},$$

where w and b are the learned weights and bias. This proxy captures robustness per unit of representational cost, reflecting the hypothesis that structural learning yields generalization without proportional growth in model complexity.

8.10 Summary of Assumptions

The simulation is based on the following assumptions:

1. Financial markets exhibit latent regimes that materially affect dynamics.
2. Historical data undersamples rare but structurally important regimes.
3. Synthetic data can deliberately expose such structure.
4. Bounded learners benefit more from structural exposure than from additional noisy samples.
5. Robustness under regime shift is a more reliable indicator of information extraction than in-sample accuracy.

9 Results

We evaluate learning robustness in a stylized market characterized by *latent regime structure*. As mentioned, the simulated environment consists of three hidden regimes: (i) calm mean reversion (Regime 0), (ii) trend/momentum (Regime 1), and (iii) turbulent high-volatility dynamics (Regime 2).

Crucially, the *distribution over regimes differs across datasets*. Historical training data is dominated by calm regimes with limited transitions, while the synthetic curriculum deliberately increases regime coverage and transition frequency. The OOD test set emphasizes stress conditions, with frequent transitions into trend and turbulent regimes. This design isolates the role of structural exposure rather than sample size.

The logistic classifier is intentionally chosen to emphasize that improvements arise from *data structure*, not model capacity.

Predictions are converted into a simple trading policy (long/short/flat via probability thresholds). We report standard financial metrics—accuracy, trade rate, Sharpe ratio, mean PnL, and maximum

drawdown—as well as regime-conditional Sharpe ratios. Importantly, we also report an *epiplexity proxy*, defined as out-of-distribution Sharpe normalized by model complexity, to quantify structural information per unit of representational cost.

All models are trained on equal-sized datasets to control for data volume.

Train Set	Acc	Trade Rate	Sharpe	Mean PnL	MDD
Historical-Only	0.51	0.41	-0.96	-0.0009	-652
Synthetic Curriculum	0.53	0.45	2.00	0.0014	-46
Mixed (Syn→Hist)	0.54	0.43	1.72	0.0012	-42

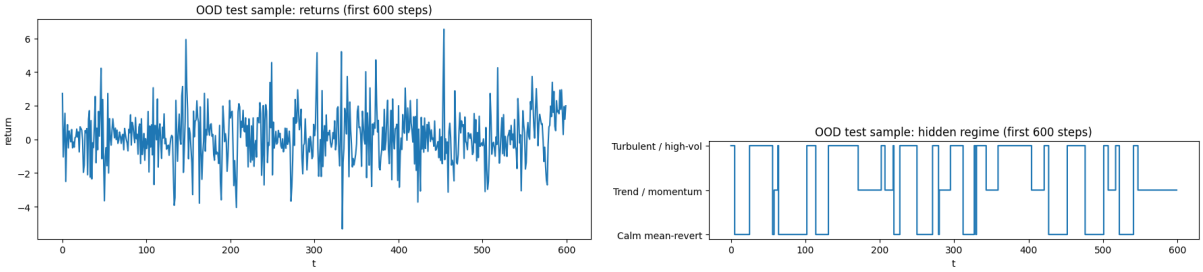
Table 3: OOD test performance under stress regimes.

Train Set	Epiplexity Proxy	Model Complexity
Historical-Only	-1.77	0.54
Synthetic Curriculum	6.13	0.33
Mixed (Syn→Hist)	4.78	0.35

Table 4: Epiplexity proxy: (Sharpe)/(model complexity).

Regime	Hist-Only	Synthetic	Mixed
Calm mean-revert	0.25	0.34	0.31
Trend/momentum	-2.80	8.09	7.30
Turbulent/high-vol	-0.14	0.57	0.46

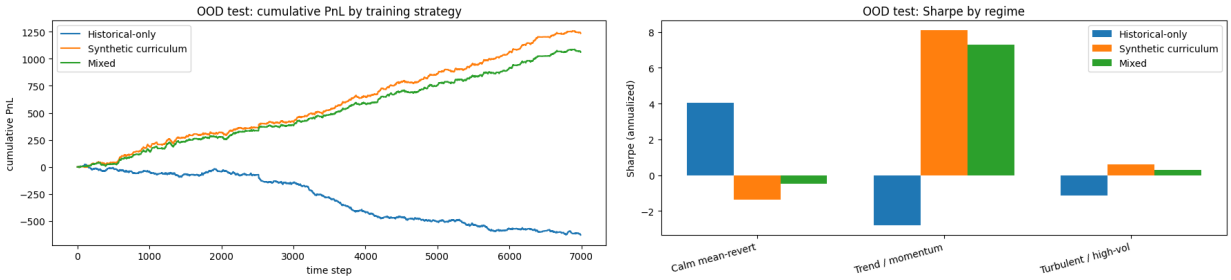
Table 5: Sharpe by hidden regime in OOD test.



(a) OOD returns (sample).

(b) Hidden regimes (sample).

Figure 1: Market dynamics under stress regime test.



(a) Cumulative PnL comparison.

(b) Sharpe per regime.

Figure 2: Training strategy performance.

The synthetic curriculum significantly outperforms historical-only training in stress conditions, supporting the claim that synthetic data increases extractable structural information. The epiplexity proxy correlates with regime robustness, suggesting structural information (not just fit) drives generalization. The mixed strategy retains many of the benefits while grounding in real data.

Connections to reasoning verification (Trade-R1): augmenting raw rewards/signals with structural quality assessments improves robustness in stochastic environments, just as synthetic data does in feature space.

9.1 Out-of-Distribution Performance Under Stress

Table 2 summarizes performance on the stress-heavy OOD test set.

The historical-only model fails catastrophically, exhibiting a negative Sharpe ratio and extreme drawdowns. Despite acceptable in-sample accuracy, it is structurally unprepared for regimes underrepresented during training. In contrast, the synthetic curriculum achieves a strongly positive Sharpe (2.00) with dramatically reduced drawdown, indicating robust adaptation to regime shifts. The mixed strategy retains most of these benefits while slightly improving stability.

These results demonstrate that robustness is not driven by predictive accuracy alone, but by exposure to structurally relevant dynamics.

9.2 Epiplexity and Structural Efficiency

Table 4 reports the epiplexity proxy alongside model complexity. The synthetic curriculum achieves the highest epiplexity by a wide margin, combining strong OOD performance with a simpler effective model. This indicates that synthetic data enables the learner to extract *more reusable structure per parameter*. Conversely, the historical-only model exhibits low (and negative) epiplexity, reflecting brittle overfitting despite comparable complexity.

This decoupling between raw performance and structural efficiency supports the central claim of epiplexity: *information relevant for generalization is not entropy, but extractable structure.*

9.3 Regime-Conditional Analysis

Table 5 decomposes Sharpe ratios by hidden regime. The most striking difference appears in the trend/momentum regime. The historical-only model performs extremely poorly (Sharpe -2.80), indicating systematic misalignment with directional dynamics. In contrast, the synthetic curriculum achieves a Sharpe of 8.09, revealing that explicit exposure to regime structure enables the learner to internalize invariants that transfer reliably. Performance gains in turbulent regimes further demonstrate robustness to volatility and shocks.

9.4 Qualitative Visualization

Figure 1 provide qualitative insight. The OOD return series and hidden regime timeline illustrate the severity of the stress environment. Cumulative PnL plots show divergence between training strategies early in the test period, while per-regime Sharpe bars make clear that gains are concentrated where historical data is least informative.

9.5 Interpretation

Across all analyses, a consistent pattern emerges: *synthetic data improves generalization by increasing extractable structural information*. The mixed strategy suggests that synthetic data is most effective when used to shape the learner's initial inductive biases, after which historical data can refine performance.

This interpretation aligns closely with recent work on reasoning verification in reinforcement learning [9], where raw reward signals are augmented with structure-aware verification. In both cases, robustness arises not from stronger optimization, but from reshaping the learning signal to emphasize invariant structure.

Taken together, these results support the view that synthetic data is not merely an approximation of reality, but a mechanism for *information creation* in the epiplexity sense: it renders latent structure accessible to bounded learners.

10 Implications for Market Modeling and Risk Management

10.1 Risk is Structural, Not Purely Probabilistic

Traditional risk management frameworks (Value-at-Risk, Expected Shortfall) model risk as tail probabilities of loss distributions. This approach:

- Assumes losses are drawn from stable distributions
- Focuses on quantile estimation rather than mechanism understanding
- Fails during regime changes when distributions shift

The epiplexity framework suggests risk should be understood structurally:

- **Identify fragility points**: Which market structures create instability?
- **Map failure modes**: How do feedback loops amplify shocks?
- **Test counterfactuals**: What happens under plausible but unobserved scenarios?

Synthetic markets enable structural risk assessment by making these questions tractable.

10.2 Scenario Generation as Primary Epistemology

Rather than treating scenario analysis as auxiliary to historical backtesting, the epiplexity perspective inverts the hierarchy: synthetic scenario generation is epistemically primary.

Historical data tells us what *did* happen under specific, contingent conditions. Synthetic data tells us what *could* happen under structurally similar conditions. For forward-looking risk management, the latter is more valuable.

This suggests a new workflow:

1. Calibrate synthetic market to reproduce key stylized facts
2. Generate scenarios exploring structural parameter space
3. Identify failure modes and fragility points
4. Validate mechanisms against historical episodes where possible
5. Use structural understanding to assess current risk exposures

10.3 Model Evaluation Beyond Prediction Accuracy

Standard model evaluation focuses on predictive performance: RMSE, Sharpe ratio, information coefficient. The epiplexity framework suggests complementary evaluation criteria:

- **Regime transfer robustness**: How well does the model maintain performance across regime shifts?
- **Invariance preservation**: Does the model respect known structural invariants (no-arbitrage, inventory constraints)?
- **Failure mode diversity**: Does the model identify multiple plausible failure pathways?
- **Mechanism interpretability**: Can the model explain its predictions in terms of market mechanisms?

These criteria prioritize structural understanding over point prediction, aligning with the epistemic goals of robust risk management.

11 Discussion and Future Directions

11.1 Limitations and Open Questions

While the epiplexity framework offers valuable conceptual clarity, several challenges remain:

Calibration and validation: How do we ensure synthetic markets capture relevant structural mechanisms without introducing spurious ones? This requires careful validation against empirical stylized facts and forensic analysis of specific market events.

Computational tractability: High-fidelity agent-based models can be computationally expensive. Balancing realism with tractability remains an active challenge.

Mechanism discovery: The framework assumes we know which mechanisms to model. Developing systematic methods for discovering relevant mechanisms from data is an open problem.

Integration with institutional knowledge: Practitioners possess deep understanding of market microstructure that is difficult to formalize. Bridging this knowledge into synthetic models is crucial but challenging.

11.2 Relation to Recent Developments

The epiplexity perspective connects with several recent developments in ML and finance:

Foundation models: Large pre-trained models exhibit emergent capabilities and transfer learning—hallmarks of high epiplexity extraction. Can financial foundation models be pre-trained on synthetic data?

Causal representation learning: Recent work on learning causal structures from observational data aligns with our emphasis on mechanism extraction.

Sim-to-real transfer: Robotics has successfully used simulation for learning policies that transfer to real environments. Financial markets may benefit from similar approaches.

Digital twins: The concept of calibrated virtual systems that mirror real-world counterparts is closely related to our notion of synthetic markets as structural analogs.

11.3 Broader Implications

The epiplexity framework extends beyond finance to any domain where:

- Historical data is sparse in critical phenomena
- Systems are shaped by strategic agent interaction
- Robust understanding matters more than point prediction

This includes macroeconomics, epidemiology, climate modeling, and social systems—domains where synthetic data generation may be underutilized due to misplaced emphasis on "real data" authenticity.

12 Conclusion

We have argued that synthetic data in financial markets should be understood not as an approximation of reality or a data augmentation technique, but as a mechanism for information creation that operates by increasing epiplexity—the extractable structure available to bounded learners.

Historical market data, despite containing rich implicit structure, exhibits low epiplexity: critical mechanisms remain computationally inaccessible due to noise, sparsity, and reflexivity. Synthetic market generators like MUSE address this by making structure explicit, enabling learners to extract robust invariants that generalize across regimes.

This reframing has several important implications:

1. **Epistemological**: Information in markets is not merely discovered but constructed through the interplay of data and learning architecture.
2. **Methodological**: Synthetic data generation should be considered primary, not auxiliary, to financial learning.
3. **Practical**: Risk management should emphasize structural understanding and scenario exploration over historical backtesting.

The epiplexity framework offers a principled foundation for these claims, grounded in the recognition that what matters for learning is not information in Shannon's sense, but structure in a computationally actionable form.

Looking forward, we envision synthetic markets playing an increasingly central role in financial modeling—not as surrogates for real markets, but as essential tools for making market structure intelligible to bounded learning systems operating in irreducibly complex environments.

Core contribution: We reframe synthetic data in finance as a generator of structural information rather than an approximation of reality, grounded in epiplexity rather than entropy, enabling learning systems to extract robust market mechanisms inaccessible in historical data alone.

References

[1] Marc Finzi and Shikai Qiu and Yiding Jiang and Pavel Izmailov and J. Zico Kolter and Andrew Gordon Wilson, (2026) *From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence*, https://arxiv.org/abs/2601.03220

[2] Cont, R. (2001). *Empirical properties of asset returns: stylized facts and statistical issues*. Quantitative Finance, 1(2), 223-236.

[3] LeBaron, B. (2006). *Agent-based computational finance*. Handbook of computational economics, 2, 1187-1233.

[4] Hasbrouck, J. (2007). *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. Oxford University Press.

[5] Kirilenko, A., Kyle, A. S., Samadi, M., & Tuzun, T. (2017). *The flash crash: High-frequency trading in an electronic market*. The Journal of Finance, 72(3), 967-998.

[6] Soros, G. (2003). *The Alchemy of Finance*. John Wiley & Sons.

[7] Bailey, D. H., Borwein, J., Lopez 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 AMS, 61(5), 458-471.

[8] Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House.

[9] Rui Sun and Yifan Sun and Sheng Xu and Li Zhao and Jing Li and Daxin Jiang and Cheng Hua and Zuo Bai (2026) Trade-R1: Bridging Verifiable Rewards to Stochastic Environments via Process-Level Reasoning , https://arxiv.org/abs/2601.03948.