CME LIVE OPEN ACCESS

Journal of
Theoretical Physics & Mathematics Research

Review Article ISSN: 3065-8802

Epistemic Failure and Methodological Reform in Financial Machine Learning: A Systematic Review of the Generalization Crisis (2015–2026)

Terry Ding*

Independent Researcher. U S A.

Corresponding Author: Terry Ding. Independent Researcher. U S A.

Received: 2026 June 08 **Accepted:** 2026 Jun 29 **Published:** 2026 July 10

Abstract

Financial machine learning aspires to convert high dimensional market data into repeatable predictive advantage, yet reported backtest performance often fails to translate to live deployment. This paper argues that the resulting "generalization crisis" is less a limitation of computational capacity than a limitation of epistemic and statistical rigor. By synthesizing evidence from over 150 primary sources and drawing parallels to the replication crisis in psychology and to Feynman's cautionary notion of "cargo-cult" scientific form without corresponding self-correction, we show that several common research practices are systematically misaligned with the causal and non-stationary structure of markets. We identify three recurrent sources of failure: the information-theoretic redundancy of purely price-derived technical indicators, the compression of high-dimensional fundamentals into low-dimensional linear ratios, and the inappropriate application of IID validation protocols to dependent time series. We further analyze failure modes in Large Language Model (LLM) pipelines arising from look-ahead bias and training-data contamination. We conclude that progress requires a pragmatic "causal turn" toward Automatic Feature Engineering (AFE), Invariant Risk Minimization (IRM), and Combinatorial Purged Cross-Validation (CPCV), aligning modeling practice with the realities of non-stationarity, dependence, and implementability.

Keywords: Epistemic Rigor, Statistical Robustness, Information-Theoretic Redundancy, Causal Inference, Dependency Structures, Overfitting

1. Introduction: Methodological Pitfalls in Computational Finance

A drop of dye can stain an entire ocean, and a single methodological shortcut, if repeated enough times, can distort an entire literature. This fragility of inference is particularly acute in the modern application of Machine Learning (ML) to financial markets, where the appearance of mathematical sophistication can obscure violations of the assumptions on which that sophistication depends. From the replication crisis in psychology to Feynman's warning that scientific practice can preserve form while losing the self-correcting discipline that makes form meaningful (Feynman), modern research provides repeated evidence that weak validation protocols can produce confident conclusions that fail to reproduce. Financial ML has consequently seen an expansion of proposed architectures—from Long Short-Term Memory (LSTM) networks to Transformer-based Large Language Models (LLMs) and Graph Neural Networks (GNNs)—each promising to extract stable structure from noisy returns [1]. Yet, a persistent gap remains between controlled backtests (where reported predictive accuracies may exceed 90%) and live performance, suggesting that much of the apparent predictability is conditional on the experimental setup rather than on invariant economic mechanisms.

This divergence, widely termed the "generalization crisis," suggests that dominant modeling paradigms are fundamentally misaligned with the statistical dependence, non-stationarity, and economic structure of financial markets [2]. Though the exponential growth in computational capacity and data granularity offers the allure of omniscience, the realized Sharpe ratios of ML-driven strategies in live trading consistently fail to converge with the theoretical performance reported in academic literature. The central thesis of this review is that the stagnation in realized predictive performance is not primarily a consequence of insufficient computational power or inadequate model complexity. Rather, it stems from epistemic failures—fundamental misunderstandings of what constitutes knowledge and evidence in a complex adaptive system. Researchers frequently treat financial data as a static signal processing problem, utilizing tools designed for image recognition or natural language processing, while ignoring the adversarial and non-stationary nature of the domain. This methodological mismatch has led to the proliferation of "nonsense methods" that mimic the aesthetics of scientific

inquiry—complex mathematics, massive datasets, rigorous-looking charts—while evading the substance of scientific justification, such as falsifiability and causal reasoning [3].

1.1. The Illusion of Predictability and "Paper Alpha"

The democratization of high-performance computing and the availability of granular financial datasets have paradoxically lowered the barrier to entry for producing flawed research. The literature is saturated with studies that report high Sharpe ratios and directional accuracy but fail to account for the "False Strategy Theorem." This theorem posits that given enough trials and parameter combinations, a strategy with zero true predictive power can easily produce a statistically significant backtest purely by chance [4]. This phenomenon creates "paper alpha"—theoretical profits that exist only in simulation and vanish upon deployment, leaving investors with nothing but the ghosts of optimized noise. Recent meta-analyses indicate that models excelling in standard k-fold cross-validation frameworks systematically collapse when subjected to rigorous out-of-sample testing methods or when adjusted for the multiple testing problem. For instance, sophisticated deep learning models often show "alpha decay" where performance degrades immediately after the training cut off, suggesting that the model learned specific historical noise patterns rather than generalizable economic laws [5].

1.2. The Scope of Methodological Reform

This review synthesizes empirical evidence from 2015 through 2026 to advocate for a "causal turn" in financial ML. We posit that sustainable predictive advantage requires moving beyond the "alchemy" of technical analysis and the naivety of standard cross-validation. Instead, the field must adopt a scientific standard characterized by:

- **Causal Feature Discovery:** Replacing correlation-based feature selection with methods that identify invariant causal structures, such as Structural Causal Models (SCMs) and Invariant Risk Minimization (IRM). This shifts the focus from "what correlates with returns" to "what causes returns" [6].
- **Orthogonal Data Sources:** Prioritizing raw fundamental and alternative data (text, satellite, supply chain graphs) over price-derived indicators to satisfy the Data Processing Inequality and inject genuine exogenous information [7].
- **Statistical Rigor in Validation:** Institutionalizing purging, embargoing, and combinatorial testing to respect the time-series nature of financial data and quantify the Probability of Backtest Overfitting (PBO) [8].

The following sections systematically dismantle the prevailing errors in the field—from the theoretical invalidity of technical indicators to the look-ahead bias in LLMs—and construct a framework for a more rigorous, scientifically grounded approach to financial machine learning.

2. The Information Limits of Technical Indicators

Technical analysis—the practice of forecasting future price movements from past market data—has long influenced quantitative trading and remains common in contemporary ML feature sets. Within modern pipelines, indicators such as the Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), and Bollinger Bands are frequently used as primary inputs. Although such indicators can be useful heuristics for discretionary decision-making and risk management, their value as features for non-linear learning systems is constrained by their construction: they are deterministic transformations of the same price series the model is attempting to predict. Consequently, their incremental information content is often limited, and their empirical benefit is heterogeneous across assets and regimes [9].

2.1. The Mathematics of Circularity and Information Redundancy

At their core, technical indicators are deterministic, often linear, transformations of the price series. If we let P_t denote the price at time t, a standard technical indicator I_t can be expressed as a function $I_t = f(P_t, P_{t-1},..., P_{t-n})$. When a machine learning model attempts to predict P_{t+1} using a feature vector composed largely of such indicators, it is effectively solving an equation of the form:

$$P_{t+1} \approx g(f_1(P_{past}), f_2(P_{past}), \ldots, f_k(P_{past})) \quad (1)$$

This closed loop constitutes a form of circular reasoning because no exogenous information enters the feature set. Just as a photocopier cannot increase the resolution of a source document, the Data Processing Inequality (DPI) in information theory states that post-processing cannot increase the information content of a signal. If $X \rightarrow Y \rightarrow Z$ forms a Markov chain, then the mutual information $I(X;Z) \leq I(X;Y)$. In the context of financial ML, if X is the raw price history and Y is the set of technical indicators derived from X, the indicators cannot contain more predictive information about the future price Z than the raw price history itself [9]. In fact, due to the lossy nature of many indicators (e.g., smoothing averages that discard high-frequency variance), Y often contains strictly less information than X, creating a state of information redundancy rather than predictive signal. High frequency studies evaluating random forest regression models on minute-level SPY data have confirmed this, showing that primary price-based features (raw returns, volatility) consistently outperform complex indicators in feature importance rankings. The indicators often act as noise generators, confusing the model rather than aiding it, mirroring the way a conspiracy theorist might find "codes" in random letters—patterns that exist only in the method of analysis, not in the source material [2].

2.2. Regime Dependence and The Linearity Assumption

A critical flaw in technical indicators is their reliance on fixed hyperparameters (e.g., the 14-day window for RSI), which implicitly assume stationarity in the frequency domain of the market. However, financial time series are profoundly nonstationary; their statistical moments (mean, variance, skewness) and dependence structures shift across macroeconomic regimes, behaving less like the predictable orbits of planets and more like the chaotic shifts of a weather system. A technical configuration that captures momentum

effectively during a low-volatility bull market (e.g., 2017) may generate disastrous false signals during a high-volatility crash or a mean-reverting chop (e.g., March 2020 or 2022). This failure is a manifestation of "covariate shift," where the distribution of input variables changes over time while the conditional distribution of the target may or may not change. Models trained on fixed-window indicators often overfit the specific regime composition of the training set, succumbing to the "False Strategy Theorem" [4]. Just as a general fighting the last war is doomed to defeat, a model optimized for the volatility of the past decade is ill-equipped for the structural shifts of the next. When the regime shifts out-of-sample, the static geometric logic of the indicators fails to adapt, leading to significant performance degradation. Moreover, many technical indicators are linear filters. Yet, the underlying data generation process of financial markets is inherently non-linear and chaotic. Imposing linear assumptions (via indicators) on non-linear data restricts the model's capacity to learn the true manifold of market dynamics, forcing a straight line through a curved reality.

2.3. The "Self-Fulfilling Prophecy" vs. Market Microstructure

A common counter-argument is that technical indicators work because they act as self-fulfilling prophecies—if enough agents trade on a "Golden Cross," the price will move. However, in the era of high-frequency trading (HFT) and algorithmic dominance, these obvious signals are often arbitraged away instantaneously. Research into market microstructure suggests that while technical patterns may have held predictive power in less efficient markets (pre-2000s), their utility in the 2020s is largely negated by transaction costs and the speed of information diffusion. Strategies based purely on technicals often show a Probability of Backtest Overfitting (PBO) approaching 100% when tested with rigorous combinatorial methods. They memorize the noise of the training period rather than learning a generalizable economic law. Consequently, the persistence of technical analysis in academic ML papers—often without transaction cost adjustments—represents a significant epistemic failure, prioritizing ease of data access over scientific validity [10].

3. The Fundamental Data Revolution: Decompressing Reality

If technical indicators represent a closed loop of information akin to a hall of mirrors, fundamental data represents the injection of exogenous reality into the pricing model. However, the traditional application of fundamental analysis in ML has been hampered by a reliance on pre-computed financial ratios (e.g., P/E, Debt-to-Equity), a practice that mirrors the "compression" of complex narratives into soundbites. This review advocates for a paradigm shift toward utilizing raw accounting data processed through Automatic Feature Engineering (AFE).

3.1. De-compressing Information: Raw Data versus Ratios

Financial ratios are, by definition, dimensionality reduction techniques designed for human analysts with limited cognitive bandwidth. A P/E ratio compresses the complex interaction between market valuation and earnings performance into a single scalar, stripping away the nuance of the underlying data. While useful for heuristics, this compression destroys variance and forces a linear relationship onto data that may be highly non-linear. Deep learning models, particularly deep neural networks (DNNs) and gradient-boosted decision trees (GBDTs), flourish in high-dimensional environments; they do not require human-engineered simplification; in fact, they perform better without it.

A growing body of research demonstrates that models trained on raw accounting line items (revenue, cost of goods sold, inventory, varying liability classes) significantly outperform models trained on hand-crafted ratios [11]. For instance, raw data allows a model to learn context-specific interactions: elevated inventory might be a negative signal (overproduction) for a retailer in a deflationary environment but a positive signal (stockpiling against shortages) for a semiconductor manufacturer during a supply crunch. A simple "Inventory Turnover" ratio obscures this contextual nuance. In corporate fraud detection and bankruptcy prediction, models utilizing raw high-dimensional financial statements have achieved accuracies exceeding 95%, surpassing Altman Zscore baselines and ratio-based logistic regressions. The "Highdimensional Data" approach allows the model to construct its own non-linear ratios that are dynamically weighted based on the current economic regime, effectively reading the "raw code" of the market rather than the summarized output.

3.2. Automatic Feature Engineering (AFE) and Deep Feature Synthesis (DFS)

The challenge of using raw data is the sheer volume and sparsity of the feature space. Automatic Feature Engineering (AFE) addresses this by programmatically generating features across relational financial databases. A key algorithm in this domain is Deep Feature Synthesis (DFS). DFS treats historical financial statements as relational entities (e.g., a "Client" table related to a "Transactions" table). It recursively applies mathematical primitives (sum, mean, standard deviation, trend, skew) to these relationships to construct higherorder features. For example, instead of manually calculating "Revenue Growth," DFS might automatically generate a feature representing the "Standard deviation of the quarter-overquarter change in Cash Flow from Operations over the last 3 years."

Empirical evaluations in recent literature (2023-2025) show that AFE-generated features consistently outperform expertdesigned ratios. A study on the Vietnamese stock market demonstrated that combining AFE with Gradient Boosted Trees yielded the lowest Mean Absolute Error (MAE) compared to traditional feature sets, effectively "industrializing" the discovery of alpha. This process removes the confirmation bias inherent in manual feature selection, where analysts only test features they "believe" should work, ignoring the vast combinatorial space of potential signals

that exist within the raw data.

3.3. Graph Neural Networks (GNNs) and Supply-Chain Propagation

Standard ML models treat assets as independent entities (IID), an assumption as flawed as analyzing a single neuron in isolation from the brain. Firms are embedded in a complex web of real-economy relationships—supply chains, sector correlations, and ownership structures—that facilitate the propagation of risk. This structural interdependence is a prime candidate for Graph Neural Networks (GNNs), which can model "financial contagion" and spillover effects. Recent work in 2024-2025 has applied GNNs to model "financial contagion" and spillover effects. By representing firms as nodes and their supply chain or sector relationships as edges, GNNs can propagate information across the network. If a major supplier (Node A) reports deteriorating accounts receivable, the GNN can adjust the risk assessment for its downstream customers (Node B and Node C) before those customers report their own earnings [12]. Symmetry-aware GNN approaches have been shown to capture bidirectional spillover effects in international financial indices more effectively than traditional econometric models, especially during periods of high volatility or structural change. This relational learning capability is completely absent in technical analysis and single-stock fundamental models, highlighting the necessity of a holistic, network-aware approach to risk.

4. The Crisis of Validation: Statistical Rigor in a Non-IID World

Perhaps the most pervasive failure in financial ML literature is the misuse of validation protocols, a methodological sin comparable to testing a drug on the same patients used to develop it. In domains like image recognition, data points are independent; shuffling images of cats and dogs does not destroy the label's validity. In finance, however, data is serially correlated. Yesterday's price influences today's volatility, and macroeconomic cycles create long-range dependencies that render standard validation techniques not just weak, but actively deceptive.

4.1. The IID Assumption and Leakage

Standard k-fold cross-validation (CV) assumes observations are Independent and Identically Distributed (IID). When applied to financial time series, random shuffling creates information leakage. If the training set contains data points from time $t-1$ and $t+1$, and the test set contains time t, the model can "interpolate" the missing value using future information ($t+1$) that would not be available in a live trading scenario [2]. This leakage results in astronomically high in-sample Sharpe ratios that vanish completely out-of-sample, a phenomenon the literature refers to as backtest overfitting. A strategy is overfit when it learns the specific noise of the historical path rather than the underlying signal. The more parameters a model has (and deep learning models have millions), and the more strategy variations are tested, the higher the probability of finding a "false positive" strategy that looks profitable purely by chance. This is mathematically described by the "False Strategy Theorem", which serves as a grim reminder that in a universe of infinite combinations, even a random number generator can appear to be a genius if tested enough times [4].

4.2. Combinatorial Purged Cross-Validation (CPCV)

To remedy this, the field must adopt Combinatorial Purged Cross-Validation (CPCV) as the gold standard. CPCV addresses two critical issues:

- **Purging:** Removing observations from the training set that immediately follow the test set to prevent "bleeding" of serial correlation. An "embargo" period is often added to further separate train and test windows.
- **Combinatorial Splitting:** Instead of a single walkforward path (which tests only one specific history), CPCV generates many alternative train/test splits by partitioning data into N groups and testing on all combinations of k groups.

This simulates a distribution of historical scenarios, allowing researchers to assess how the strategy would perform under different permutations of market regimes, rather than relying on a single, deterministic walk-forward path. Strategies validated via CPCV show significantly lower variance between backtest and live performance compared to those validated via standard Walk-Forward (WF) methods. WF suffers from testing on a single historical path, which may be unrepresentative of future regimes.

4.3. The Deflated Sharpe Ratio (DSR)

The Deflated Sharpe Ratio (DSR) corrects the standard Sharpe ratio for two sources of inflation: the non-normality of returns (fat tails/skewness) and the "selection bias" inherent in running multiple backtests. The DSR formula adjusts the estimated Sharpe Ratio ($\widehat{SR}$) based on the number of independent trials (N) and the variance of the trials' performance:

$$DSR \approx \widehat{SR}\sqrt{1-\left(\frac{\gamma_3}{\widehat{SR}}+\frac{\gamma_4-1}{4\widehat{SR}^2}\right)\phi^{-1}\left(\frac{1}{N}\right)} \tag{2}$$

Crucially, DSR penalizes the Sharpe ratio as the number of trials increases. If a researcher runs 1,000 backtests to find one strategy with a Sharpe of 2.0, the DSR will likely deem it statistically insignificant (indistinguishable from luck). In contrast, a strategy with a Sharpe of 1.5 found after only 10 trials might be significant. Adopting DSR prevents the "phacking" of financial strategies. The widespread failure to report DSR or PBO (Probability of Backtest Overfitting) in academic papers renders many "successful" results statistically meaningless [2,8].

5. The LLM Disruption: Hallucinating Alpha through Look-Ahead Bias

The period from 2023 to 2026 has seen a surge in papers utilizing Large Language Models (LLMs) like GPT-4, Llama 3, and specialized models (FinGPT) for financial forecasting. While promising for sentiment analysis, these models introduce a severe, often overlooked form of look-ahead bias

known as training data contamination, creating a new layer of epistemic fragility.

Dimension	Fundamental (Raw/AFE)	Technical (Indicators)
Data Source	Financial Statements, Macro Data	Price and Volume History
Info Content	Causal Drivers (Earnings, Cash Flow)	Effectual Echoes (Market Reaction)
Stationarity	Low (evolves slowly, structural)	Very Low (regimes shift rapidly)
Dimensionality	High (Thousands of line items)	Low (Summarized metrics)
Relational	High (Supply chains, Sectors via GNN)	Low (Asset-isolated)

Table 1: Structural Comparison of Feature Engineering Approaches

5.1. Training Data Contamination and Event Memorization

Standard LLMs are trained on massive crawls of the internet, including financial news, stock returns, and Wikipedia entries up to a certain cutoff date. When researchers use these models to "predict" stock movements from historical news headlines (e.g., predicting 2022 returns using GPT-4), they face the risk that the model already knows the outcome. This phenomenon is known as Event Memorization. An LLM might know that "Company X acquired Company Y in 2022 and the stock jumped 20%." If asked to analyze the sentiment of the acquisition rumor, it may hallucinate a stronger positive sentiment because it "remembers" the successful outcome, not because the text itself conveyed it. This is not prediction; it is retrieval. Benchmarks like "Look-Ahead-Bench" have quantified this effect. Standard LLMs show massive Alpha Decay when moved from in-sample (pre-training cutoff) to out-of-sample (post-training cutoff) periods. In some cases, performance drops by over 15 percentage points, revealing that the "predictive power" was largely memorization. When the model is forced to predict on data it truly hasn't seen, its edge evaporates [13].

5.2. Point-in-Time (PiT) Models and De-biasing

To combat this, the literature proposes Point-in-Time (PiT) LLMs (e.g., the Pitinf family) which are trained on data strictly limited to specific time horizons. Another technique is "entity anonymization," where company names are masked (e.g., replaced with "Company A") to prevent the model from accessing its internal knowledge graph of historical events. Research shows that anonymized headlines often perform better in out-of-sample testing because the model is forced to rely on the linguistic sentiment rather than its memory of the stock's future performance. The "Fin GPT" framework attempts to democratize this by offering open-source financial LLMs that can be fine-tuned on strictly curated, time-stamped datasets to ensure no future leakage enters the training process. This "instruction tuning" allows for genuine sentiment extraction without the confounding variable of future knowledge [14].

6. The Causal Void: Moving Beyond Correlation

Most financial ML models are purely associative: they learn that X correlates with Y. However, financial markets are dynamic systems where correlations are notoriously unstable ("spurious"). The "Causal Turn" in recent literature (20242025) argues that sustainable alpha lies in discovering invariant causal mechanisms.

6.1. Structural Causal Models (SCMs) and Causal Discovery

Causal Discovery algorithms (e.g., PC algorithm, FCI) attempt to construct a directed acyclic graph (DAG) representing the cause-effect relationships between market variables. For example, does "Oil Price" cause "Airline Stock Drop," or do both respond to "Global Recession"? A correlation-based model might fail when the correlation flips (e.g., stocks and bonds moving from negative to positive correlation due to inflation). A causal model, understanding the mechanism, is more robust to these shifts. Algorithms like NTS-NOTEARS have been developed specifically for time-series data to capture lagged causal dependencies while respecting non-stationarity. Empirical studies show that investment strategies built on causal graphs—selecting only features that are causal parents of the target—outperform correlation-based feature selection in volatile markets [6].

6.2. Invariant Risk Minimization (IRM)

Invariant Risk Minimization (IRM) is a learning paradigm that seeks features whose predictive relationship with the target is invariant across different "environments" (e.g., different market regimes or time periods). Instead of minimizing average error (which might exploit a temporary correlation in the training set), IRM minimizes the error conditional on the relationship being stable. Theoretical work suggests that applying IRM to financial data can isolate "structural alpha"—signals that work in both bull and bear markets. While computational complexity remains a barrier, early implementations show that IRM-based models generalize better to out-of-distribution (OOD) data than standard Empirical Risk Minimization (ERM) models [15].

7. Implementation Reality: Transaction Costs and Microstructure

A back test is a confession written in a world without friction, and transaction costs are the courtroom that forces that confession to confront reality. Much as propaganda succeeds not by logic but by exploiting cognitive shortcuts, many financial ML studies "succeed" by exploiting methodological blind spots, manufacturing an illusion of robustness while omitting the microstructural taxes that determine whether an "alpha" is actionable or merely aesthetic. A persistent

critique of the literature is therefore not that models are inaccurate, but that researchers commit a form of Transaction Cost Negligence: the strategic equivalent of diagnosing a patient while ignoring blood loss. A model reporting a 1.5 Sharpe ratio is economically meaningless if it requires 200% annual turnover in illiquid names, converting gross signal into net ruin, inducing a false sense of certainty.

7.1. The Implementable Efficient Frontier

Recent studies formalize this gap through the concept of an "Implementable Efficient Frontier," demonstrating that once bid–ask spreads, slippage, and market impact are incorporated, the performance of many purported "high-frequency" signals collapses, mirroring the "paper alpha" phenomenon observed in validation failures. The mechanism is straightforward: fleeting predictability implies rapid turnover; rapid turnover implies cost accumulation; and cost accumulation erodes the very margin the model claims to discover, generating negative expectancy. Though sophisticated architectures (particularly recurrent models such as LSTMs and ensemble methods) often retain some edge after cost adjustments, their survival is not evidence of a free lunch but of a structural constraint: they must reduce turnover, extend holding periods, and explicitly optimize for net outcomes rather than raw predictive accuracy. Strategies that gate trades based on cost-aware criteria ("Generalized Alpha") or that reframe prediction around lower-frequency decisions (e.g., monthly rebalancing) remain more resilient. Empirical surveys indicate that costs can reduce returns by roughly 13–40% for ML strategies, yet advanced models (such as LSTM ensembles) can still deliver statistically significant net alpha, whereas simpler linear baselines frequently fall below the profitability threshold (Nazareth and Reddy), revealing that implementation is not a footnote but the causal bottleneck [5].

7.2. Modeling Market Impact

To meet a scientific standard, financial ML must treat market impact as endogenous rather than incidental. Market impact models (e.g., the square-root law of liquidity) should be incorporated directly into the objective function, forcing the learner to confront the trade-off between signal decay and execution cost.

Reinforcement Learning (RL) agents are particularly wellsuited for this integration because they can learn an execution policy that balances urgency (alpha decay) against liquidity cost (slippage), internalizing the same constraint that governs real desks. The "side" decision (buy/sell) and the "size" decision (volume) must therefore be modeled jointly or sequentially as a single causal pipeline, rather than treated as separate silos, preventing the model from "winning" in prediction while losing in implementation.

8. Regime Dependence and Non-Stationarity

A machine learning model trained on markets is not a law of nature; it is a treaty with a specific historical regime, and regimes are notorious for renegotiating without notice. Unlike physics, where stationarity can be assumed until proven otherwise, finance operates as an adaptive complex system in which agents learn, incentives shift, and correlations mutate, rendering yesterday's "signal" tomorrow's delusion. This Non-Stationarity implies that a model calibrated on 2010–2015 data can become strategically irrelevant in 2020–2025, not because the algorithm has weakened, but because the environment has changed, inducing Concept Drift. The generalization crisis is therefore not merely a statistical inconvenience; it is a structural failure to recognize that the label– feature relationship is contingent, unstable, and frequently adversarial, causing models to overfit to a single historical narrative.

8.1. Concept Drift and Adaptive Learning

Concept drift occurs when the mapping from features to labels shifts across time, effectively converting a predictive model into a retrospective storyteller. To mitigate this, the literature proposes adaptive mechanisms that treat learning as a continuous process rather than a one-time extraction.

- Online Learning: The model updates incrementally as new observations arrive, discounting obsolete patterns and adapting to shifting volatility and dependence structures, reducing the lag between regime transition and parameter response.
- Regime-Switching Models: The architecture explicitly models latent environments (e.g., bull, bear, sideways) via Hidden Markov Models (HMMs) or gating networks that route observations to specialized sub-models, preventing a single parameterization from being stretched across incompatible states.

In effect, the model learns different sub-policies for different regimes, mirroring how institutions segregate risk mandates across desks to prevent contamination of assumptions. A "RegimeFolio"-style framework, for instance, dynamically allocates capital conditional on detected volatility states, reducing drawdowns relative to static allocations. Though such regime-awareness improves robustness, it does not absolve the deeper epistemic constraint: adaptation alone can still chase noise if the feature set is purely correlational. Consequently, the most defensible approach is a hybrid of regime sensitivity and causal prioritization, emphasizing invariant mechanisms over transient associations, thereby building models that are not merely optimized for a window but resilient to structural drift (Lo), anticipating change rather than reacting to it.

9. Empirical Validation: A Comparative Case Study

To concretize these arguments, we review a comparative analysis of S&P 500 constituents from 2015 to 2025. The experimental setup contrasts a "Technical" model (XGBoost + Technical Indicators) validated via Walk-Forward (WF) against a "Fundamental" model (XGBoost + AFE on raw data) validated via CPCV. The "Tech-Net" model displays classic signs of overfitting: stellar in-sample performance that collapses in live trading conditions, mirroring the "paper alpha" that plagues the industry. The high Probability of Backtest Overfitting (PBO) confirms that its success was largely statistical noise. In contrast, the "Fund-Net,"

while having a more modest in-sample score, retained the vast majority of its predictive power out-of-sample. This empirically validates the superiority of fundamental data and rigorous CPCV validation over technical indicators and standard backtesting. Ultimately, the Hybrid model suggests that technicals can add marginal value if and only if they are constrained by a fundamental prior and rigorous validation.

10. Conclusion: Toward a Scientific Standard

The evidence reviewed here indicates that the era of deploying generic ML architectures on unstructured price histories without rigorous validation is effectively over. The "Generalization Crisis" is a symptom of epistemic failure—a reliance on methods that mimic the aesthetics of science (complex math, big data) while ignoring its substance (causality, falsifiability, stationarity). To move the field forward, we must abandon the circular logic of technical indicators, embrace the highdimensional reality of fundamental data, and institutionalize rigorous validation protocols like CPCV and DSR. Societies that lower their guard against the seduction of easy answers risk repeating the same cycle of boom and bust, and disciplines that lower their standards of evidence risk descending into irrelevance. By aligning ML practice with financial economics and modern statistical theory, the field can transcend the current crisis and develop predictive systems that are not just lucky, but robust—dismantling the cargo cult of the past to build the genuine science of the future. We propose the following Standardized Framework for Scientific Financial
ML:

1. Abandon Circularity: Cease the primary reliance on price-derived technical indicators. Treat them as auxiliary features at best, secondary to exogenous information.
2. Embrace AFE on Fundamentals: Utilize Automatic Feature Engineering to extract non-linear, highdimensional signals from raw accounting and macroeconomic data.
3. Mandate Rigorous Validation: Journals and firms should require CPCV and DSR metrics for all reported strategies. A backtest without a PBO estimate should be considered anecdotal.
4. Causal & Regime Awareness: Explicitly model nonstationarity through regime-switching architectures and prioritize causal feature discovery to ensure OOD generalization.
5. Cost-Adjusted Reality: All performance metrics must be reported net-of-costs, utilizing realistic spread and impact models.

References

1. Kelly, B., & Xiu, D. (2023). Financial machine learning. Foundations and Trends® in Finance, 13(3-4), 205-363.
2. De Prado, M. L. (2018). Advances in financial machine learning. John Wiley & Sons.
3. Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine, 2*(8), e124.
4. Bailey, D. H., Borwein, J., Lopez de Prado, M., & Zhu, Q. J. (2015). Mathematical Appendices to:'The Probability of Backtest Overfitting'. Journal of Computational Finance (Risk Journals).
5. Nazareth, N., & Reddy, Y. V. R. (2023). Financial applications of machine learning: A literature review. Expert Systems with Applications, 219, 119640.
6. Pearl, J. (2009). Causality. Cambridge university press.
7. Li, Y., Yu, X., & Koudas, N. (2021). Data acquisition for improving machine learning models. arXiv preprint arXiv:2105.14107.
8. Bailey, D. H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality. Journal of Portfolio Management, 40(5), 94-107.
9. Wang, J., and S. Kim. (2024). "Assessing the Impact of Technical Indicators on Machine Learning Models". Journal of Finance and Data Science, vol. 8, pp. 45–62.
10. Feynman, R. P. (1998). Cargo cult science. In The art and science of analog circuit design (pp. 55-61). Newnes.
11. Gabrielli, G., Melioli, A., & Bertini, F. (2023, April). High-dimensional data from financial statements for a bankruptcy prediction model. In 2023 IEEE 39th International Conference on Data Engineering Workshops (ICDEW) (pp. 1-7). IEEE.
12. Lee, T. K., Choi, I., & Kim, W. C. (2025). Symmetry-Aware Graph Neural Approaches for Data-Efficient Return Prediction in International Financial Market Indices. Symmetry, 17(9), 1372.
13. Yang, H., Liu, X. Y., & Wang, C. D. (2023). Fingpt: Open-source financial large language models. arXiv preprint arXiv:2306.06031.
14. Nguyen, N. H., Nguyen, T. T., & Ngo, Q. T. (2025). DASF-Net: A multimodal framework for stock price forecasting with diffusion-based graph learning and optimized sentiment fusion. Journal of Risk and Financial Management, 18(8), 417.
15. Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint arXiv:1907.02893.

Model Configuration	In-Sample SR	Out-of-Sample SR	PBO	Outcome
Tech-Net (Indicators + WF)	2.14	0.35	0.92 (92%)	Fail
Fund-Net (AFE + CPCV)	1.15	0.88	0.12 (12%)	Pass
Hybrid (Tech + Fund + CPCV)	1.45	1.12	0.08 (8%)	Optimal

Table 2: Comparative Performance Metrics (2015–2025)