

# Interpretable Hypothesis-Driven Trading: A Rigorous Walk-Forward Validation Framework for Market Microstructure Signals

Gagan Deep\*      Akash Deep†      William Lamptey‡

December 16, 2025

## Abstract

We develop a rigorous walk-forward validation framework for algorithmic trading designed to mitigate overfitting and lookahead bias. Our methodology combines interpretable hypothesis-driven signal generation with reinforcement learning and strict out-of-sample testing. The framework enforces strict information set discipline, employs rolling window validation across 34 independent test periods, maintains complete interpretability through natural language hypothesis explanations, and incorporates realistic transaction costs and position constraints. Validating five market microstructure patterns across 100 US equities from 2015 to 2024, the system yields modest annualized returns (0.55%, Sharpe ratio 0.33) with exceptional downside protection (maximum drawdown  $-2.76\%$ ) and market-neutral characteristics ( $\beta = 0.058$ ). Performance exhibits strong regime dependence, generating positive returns during high-volatility periods (0.60% quarterly, 2020–2024) while underperforming in stable markets ( $-0.16\%$ , 2015–2019). We report statistically insignificant aggregate results (p-value 0.34) to demonstrate a reproducible, honest validation protocol that prioritizes interpretability and extends naturally to advanced hypothesis generators, including large language models. The key empirical finding reveals that daily OHLCV-based microstructure signals require elevated information arrival and trading activity to function effectively.

---

\*Department of Mathematics & Statistics, Texas Tech University. Email: gdeep@ttu.edu. Corresponding author.

†Department of Mathematics & Statistics, Texas Tech University. Email: akash.deep@ttu.edu.

‡Department of Mathematics & Statistics, Texas Tech University. Email: wilampte@ttu.edu.

The framework provides complete mathematical specifications and open-source implementation, establishing a template for rigorous trading system evaluation that addresses the reproducibility crisis in quantitative finance research. For researchers, practitioners, and regulators, this work demonstrates that interpretable algorithmic trading strategies can be rigorously validated without sacrificing transparency or regulatory compliance.

**Keywords:** Algorithmic Trading, Walk-Forward Validation, Market Microstructure, Interpretable Machine Learning, Reinforcement Learning, Backtesting Methodology

**JEL Classification:** G11, G12, C53, C63

## 1 Introduction

Quantitative trading research faces a reproducibility crisis. Studies consistently document trading strategies generating double-digit annual returns through backtesting, yet institutional investors report that over 90% of academic strategies fail when implemented with real capital [Harvey et al., 2016]. This credibility gap threatens the practical relevance of finance research and has generated increasing skepticism toward published trading anomalies. The fundamental problem is methodological: standard backtesting procedures suffer from overfitting through in-sample parameter optimization, lookahead bias through the use of information unavailable in real-time, and lack of interpretability through reliance on black-box machine learning models.

This paper develops a rigorous validation framework that addresses these deficiencies while maintaining generality across hypothesis generation approaches. Our framework makes four key methodological innovations. First, it enforces strict information set discipline where features, signals, and execution decisions use only data available up to that point in time, preventing lookahead bias that pervades much backtesting research. Second, it employs walk-forward validation with rolling windows, where the system must prove itself repeatedly across 34 independent out-of-sample test periods spanning multiple market regimes rather than succeeding in one fortunate backtest. Third, it maintains complete interpretability by requiring every trade to originate from a human-interpretable hypothesis expressed in natural language, enabling regulatory compliance and post-hoc auditing. Fourth, it incorporates realistic execution assumptions including commission costs, slippage, position limits, and stop-loss rules that reflect actual trading constraints.

Critically, the validation framework is *agnostic to hypothesis source*. While our proof-of-concept implementation uses five hand-crafted rule-based hypothesis types, the framework readily extends to more sophisticated generation methods including genetic programming for

symbolic pattern discovery, neural networks with post-hoc interpretation, and large language models that generate hypotheses in natural language. This generality distinguishes our approach from prior work: we provide validation infrastructure rather than a specific trading strategy, enabling researchers to test their own hypotheses under rigorous conditions while maintaining interpretability and preventing overfitting.

To demonstrate the framework’s capabilities, we implement five illustrative hypothesis types encoding common market microstructure patterns: institutional accumulation, flow momentum, mean reversion, breakouts, and range-bound value signals. These patterns span diverse trading concepts and generate sufficient activity for statistical analysis but make no claim to exhaustiveness—they serve to validate the methodology rather than represent comprehensive strategy development. We test these hypotheses across 100 US equities from 2015-2024 using a reinforcement learning agent that learns which hypothesis types to execute based on historical performance, all within the walk-forward validation structure.

Our empirical results serve two purposes: demonstrating realistic out-of-sample performance and illustrating methodological principles. The system generates modest overall returns of 0.55% annualized with 41% win rate at the fold level, substantially below typical published claims but representative of genuine out-of-sample performance after rigorous validation. Critically, aggregate returns are not statistically significant ( $p$ -value 0.34), and we report this honestly rather than p-hacking or selectively presenting results—transparency essential for correcting publication bias. Performance exhibits strong regime dependence: positive returns during high-volatility periods (2020-2024: +2.4% annualized average across relevant folds) versus negative returns during stable markets (2015-2019: -0.16% annualized), revealing that market microstructure signals derived from daily data work primarily during elevated volatility. Risk management is exceptional, with maximum drawdown of only -2.76% compared to -23.8% for SPY, and the strategy exhibits market-neutral characteristics ( $\beta = 0.058$ , correlation 0.53).

These modest results reflect methodological rigor rather than deficiency. By reporting realistic returns that survive strict testing alongside non-significant  $p$ -values and regime-dependent failures, we demonstrate what honest validation looks like—providing a reference point against which other studies can be evaluated. The primary contribution is not a profitable trading system but a validation template that other researchers can apply to their own hypotheses, confident that results will be reproducible and free from lookahead bias.

This work makes several contributions to trading system validation methodology. We provide a complete, reproducible framework with mathematical specifications and open-source implementation that researchers can directly apply. We demonstrate that rigorous walk-forward validation dramatically tempers conclusions, with our modest 0.55% return

contrasting sharply with typical published claims of 15-30% annual returns. We show that aggregate statistics mask important regime-dependent heterogeneity, with testing across multiple market conditions revealing when and why strategies succeed or fail. We contribute to correcting publication bias by reporting non-significant results alongside methodological transparency. Finally, we establish that interpretability and adaptive learning can be successfully combined without sacrificing either dimension.

The remainder of this paper proceeds as follows. Section 2 reviews related literature on algorithmic trading, machine learning in finance, and validation methodologies. Section 3 describes our walk-forward framework in detail with complete mathematical formulations, emphasizing framework generality and extensibility. Section 4 presents comprehensive empirical results across 34 out-of-sample tests for our illustrative hypothesis implementation. Section 5 analyzes regime-dependent performance and discusses implications for research and practice. Section 6 concludes with limitations, extensions to more sophisticated hypothesis generation methods, and future research directions.

## 2 Literature Review

### 2.1 The Replication Crisis and the Need for Rigorous Validation

Financial economics faces a reproducibility crisis. [Harvey et al. \[2016\]](#) document that at least 316 factors have been proposed to explain cross-sectional returns, arguing that most claimed research findings in financial economics are likely false when properly accounting for multiple testing. They demonstrate that newly discovered factors require t-statistics exceeding 3.0 (not the traditional 2.0) to be considered genuinely significant given extensive data mining across the field. This echoes [Ioannidis \[2005\]](#)’s foundational argument that most published research findings are false when study power is low, tested hypotheses outnumber true relationships, and flexibility in designs enables p-hacking.

[McLean and Pontiff \[2016\]](#) quantify the consequences of this crisis: examining 97 return predictors, they find portfolio returns decline 26% out-of-sample and 58% post-publication, with roughly half attributable to data-mining bias and half to publication-informed arbitrage. [Hou et al. \[2020\]](#) confirm the severity through systematic replication of 452 anomalies—65% fail single-test significance hurdles using value-weighted returns and NYSE breakpoints, rising to 82% under multiple-testing adjustments. More recently, [Jensen et al. \[2023\]](#) examine whether there is a replication crisis in finance, finding that while replication rates are higher than in other social sciences, significant concerns remain about the robustness of many documented anomalies. Despite these warnings, many studies still omit formal multiple-testing

adjustments or ignore regime shifts, producing strategies that decay sharply once implemented [Lo and MacKinlay, 1990, Sullivan et al., 1999]. These findings establish both the problem our paper addresses and the imperative for methodologies that distinguish genuine alpha from statistical artefact.

## 2.2 Walk-Forward Validation: The Gold Standard

Pardo [1992, 2008] pioneered walk-forward analysis as the gold standard for trading-strategy validation, introducing continuous re-optimization on rolling windows where strategies must prove themselves repeatedly across different market conditions rather than succeed in one fortunate backtest. Yet early implementations lacked the statistical rigor demanded by modern multiple-testing standards.

Bailey and López de Prado [2014] provided the mathematical foundation, proving that high simulated performance is easily achievable after backtesting relatively few strategy configurations, with memory effects in financial series causing over-fitted strategies to systematically under-perform (not merely fail to outperform) out-of-sample. Bailey and López de Prado [2014] introduced the *Deflated Sharpe Ratio* to correct for selection bias under multiple testing and non-normal returns, while Bailey et al. [2017] developed *Combinatorially Symmetric Cross-Validation (CSCV)* to compute the Probability of Backtest Overfitting. Recent work by Arian et al. [2024] compares validation methods for machine learning in finance, finding that *Combinatorial Purged Cross-Validation* shows superiority in mitigating over-fitting risks, though walk-forward remains the industry standard for realistic trading simulation. Recent extensions introduce regime-aware segmentation, in which training and testing windows are conditioned on volatility or macroeconomic regimes to enhance robustness in non-stationary environments [Kirschenmann et al., 2022].

Our paper advances this literature by integrating Pardo’s walk-forward methodology with modern statistical adjustments and extending it to machine-learning settings, yielding a unified validation framework that simultaneously addresses the overfitting concerns highlighted by Bailey et al. [2014] and the multiple-testing requirements emphasized by Harvey et al. [2016].

## 2.3 Machine Learning in Finance: Power vs. Interpretability

Gu et al. [2020] demonstrate in their landmark *Review of Financial Studies* paper that machine-learning methods, particularly trees and neural networks, substantially outperform traditional linear models in measuring equity risk premia—in some cases doubling Sharpe ratios of regression-based strategies. Their long-short decile strategies achieve Sharpe ratios

of 1.35 (value-weighted) and 2.45 (equal-weighted), with out-of-sample  $R^2$  of 0.33%–0.40% for stock-level predictions. Fischer and Krauss [2018] show LSTM networks achieve daily returns of 0.46% and Sharpe ratios of 5.8 before transaction costs for S&P 500 constituent prediction, though performance declined notably post-2010.

Ensemble methods including XGBoost and LightGBM have proliferated in finance applications, with recent work increasingly employing SHAP (SHapley Additive exPlanations) for feature-importance analysis [Lundberg and Lee, 2017]. Freyberger et al. [2020] develop sparse methods that achieve both good predictive performance and interpretability through factor selection. Yet these advances come with a profound interpretability deficit.

Rudin [2019] argues forcefully that for high-stakes decisions—including financial trading—the field should prioritize inherently interpretable models rather than explaining black-box models post-hoc, demonstrating that the perceived accuracy-interpretability trade-off is often a myth for structured data. She shows rule-based models like CORELS achieve comparable accuracy to complex systems while maintaining transparency. Chen et al. [2021] demonstrate this principle in credit risk with globally interpretable two-layer additive models that match neural-network accuracy. While SHAP and LIME are widely used for post-hoc explainability, they do not guarantee that the underlying model is interpretable or aligned with economic theory [Ribeiro et al., 2016, Molnar, 2020]. Contemporary symbolic-regression and hybrid-AI techniques explicitly embed domain knowledge, delivering accuracy comparable to black-box models while preserving full transparency [La Cava et al., 2021, Kronberger et al., 2022].

Our hypothesis-driven approach addresses this interpretability gap by building strategies on explicit, testable hypotheses about market microstructure rather than opaque learned representations, while our walk-forward framework rigorously validates whether these interpretable strategies maintain performance out-of-sample.

## 2.4 Market Microstructure Theory and Daily Data Applications

Kyle [1985] established the foundational framework for understanding price impact, liquidity, and information asymmetry, showing how informed traders camouflage trades within noise. Easley et al. [1996] introduced the *Probability of Informed Trading (PIN)* model to quantify adverse-selection risk, later demonstrating that information risk is priced in cross-sectional returns [Easley et al., 2002]. Hasbrouck [1995] developed VAR frameworks for measuring price discovery and information shares that became standard methodology. While these seminal papers focused on high-frequency data, recent work demonstrates microstructure signals can be extracted from daily OHLCV data.

Critically, Easley et al. [2012] introduced *Volume-Synchronized Probability of Informed Trading (VPIN)* as a real-time order-flow toxicity measure, which predicted the 2010 Flash Crash. Low et al. [2016] adapt VPIN to daily international data, showing that daily BV-VPIN effectively forecasts high volatility across multiple countries—bridging high-frequency microstructure theory with daily-frequency implementation. Chichernea et al. [2024] develop directional option-to-stock trading-volume imbalances using daily option volumes, demonstrating these measures predict future abnormal returns and respond strongly to cash-flow news. Nevertheless, few studies examine regime-dependent performance of microstructure signals. Exceptions include Nagel [2012], who shows that liquidity-based strategies vary with funding liquidity, and Hendershott and Moulton [2011], who link microstructure effects to periods of market stress.

These papers establish that microstructure information persists at daily frequency and can generate tradable signals, providing the empirical foundation for our daily-data approach. Yet the literature lacks systematic examination of how these signals’ effectiveness varies across market regimes—a gap our paper addresses through regime-dependent performance analysis within the walk-forward framework.

## 2.5 Reinforcement Learning for Trading: Adaptability at the Cost of Transparency

Moody and Saffell [2001] pioneered applying reinforcement learning to trading, introducing *Recurrent Reinforcement Learning* that directly optimizes financial objectives without requiring forecasting models. Recent deep-RL applications demonstrate impressive backtest performance: Deng et al. [2017] combine deep learning for feature extraction with RL for decision-making, creating end-to-end learning from raw financial signals; Jiang et al. [2017] present a financial-model-free framework for portfolio management using CNN, RNN, and LSTM architectures that outperformed traditional strategies on cryptocurrency markets. Q-learning, policy-gradient methods (PPO, DDPG, A2C), and multi-armed bandit formulations have all been successfully applied to trading.

However, the literature consistently identifies critical limitations: severe black-box problems making interpretability particularly challenging for sequential decisions, extensive data requirements with pronounced over-fitting risks, vulnerability to non-stationary market dynamics, and difficulty translating simulation success to live trading. Multiple surveys note that many RL approaches fail profitability tests once realistic transaction costs are included, suggesting learned strategies exploit patterns existing only in frictionless environments. In response, hypothesis-driven RL—where agents are constrained by economic priors—has been

proposed as a middle ground between adaptability and transparency [Dixon et al., 2020].

In contrast to RL’s opaque learned policies, our hypothesis-driven approach provides transparent, economically interpretable rules that can be validated against theoretical expectations and audited for regulatory compliance, while walk-forward testing addresses over-fitting concerns that plague RL.

## 2.6 Positioning Our Contributions

This literature review reveals two critical gaps our paper addresses. **Methodologically**, while Pardo established walk-forward analysis and Bailey/Harvey developed statistical corrections for multiple testing and over-fitting, no existing work integrates these approaches into a comprehensive framework applicable to modern machine-learning methods that maintains interpretability. We combine rigorous walk-forward validation with deflated Sharpe ratios, multiple-testing adjustments, and inherently interpretable hypothesis-driven models—providing practitioners with a validation methodology that avoids false discoveries while maintaining the transparency demanded by regulators and risk managers.

**Empirically**, although recent work demonstrates microstructure signals can be extracted from daily data [Low et al., 2016, Chichernea et al., 2024], the literature lacks systematic analysis of how these signals perform across different market regimes. Our regime-dependent performance analysis within the walk-forward framework addresses this gap, testing whether microstructure-based strategies maintain effectiveness or require regime-specific adaptations.

Additionally, we contribute to the growing literature on explainable AI in finance by demonstrating that interpretable, hypothesis-based strategies can be rigorously validated without sacrificing transparency or regulatory compliance [Arrieta et al., 2020, Rudin, 2019]. Together, these contributions provide both a rigorous methodological template for validating algorithmic trading strategies and new empirical evidence on the stability and regime-dependence of daily microstructure signals—advancing both the science of strategy validation and our understanding of information dynamics in equity markets.

## 3 Methodology

### 3.1 Overview and Framework Generality

Before presenting technical details, we emphasize that our framework is designed for *methodological generality*. While this implementation uses five hand-crafted hypothesis types, the validation protocol accommodates any hypothesis generation mechanism—from genetic programming to large language models—provided hypotheses maintain interpretability through

natural language explanations. This section first presents the core mathematical infrastructure, then demonstrates its application with our illustrative rule-based hypotheses.

## 3.2 Mathematical Framework

### 3.2.1 Notation and Definitions

Let  $\mathcal{S} = \{s_1, s_2, \dots, s_N\}$  denote the universe of  $N$  securities, and let  $\mathcal{T} = \{t_1, t_2, \dots, t_T\}$  represent the set of trading days. For each security  $s \in \mathcal{S}$  and time  $t \in \mathcal{T}$ , we observe:

$$P_t^s = (O_t^s, H_t^s, L_t^s, C_t^s, V_t^s) \quad (1)$$

where  $O_t^s$ ,  $H_t^s$ ,  $L_t^s$ ,  $C_t^s$ , and  $V_t^s$  denote the open, high, low, close prices and volume, respectively.

**Definition 1** (Information Set). *The information set available at time  $t$  is defined as:*

$$\mathcal{I}_t = \{P_\tau^s : s \in \mathcal{S}, \tau \leq t\} \quad (2)$$

*Critically,  $\mathcal{I}_t$  contains only information available up to and including time  $t$ , preventing lookahead bias.*

### 3.2.2 Feature Engineering

We construct a feature vector  $\mathbf{x}_t^s \in \mathbb{R}^{54}$  for each security  $s$  at time  $t$ . The features are organized into four categories representing market microstructure, technical indicators, statistical measures, and regime indicators. Complete feature specifications are provided in Appendix A. Key microstructure features include:

$$\text{VolumeImbalance}_t^s = \frac{\sum_{\tau=t-4}^t V_\tau^s \mathbb{1}(C_\tau^s > O_\tau^s) - \sum_{\tau=t-4}^t V_\tau^s \mathbb{1}(C_\tau^s < O_\tau^s)}{\sum_{\tau=t-4}^t V_\tau^s} \quad (3)$$

$$\text{VolumeRatio}_t^s = \frac{V_t^s}{\frac{1}{20} \sum_{\tau=t-19}^t V_\tau^s} \quad (4)$$

$$\text{PriceEfficiency}_t^s = \frac{|\sum_{\tau=t-9}^t R_\tau^s|}{\sum_{\tau=t-9}^t |R_\tau^s| + \epsilon} \quad (5)$$

where  $R_t^s = (C_t^s - C_{t-1}^s)/C_{t-1}^s$  is the daily return and  $\epsilon = 10^{-6}$  prevents division by zero.

### 3.3 Hypothesis Structure and Generation

**Definition 2** (Trading Hypothesis). *A trading hypothesis  $h$  is a tuple:*

$$h = (s, a, \theta, \ell, c, \mathbf{x}, r^*, \delta^*) \quad (6)$$

where:

- $s \in \mathcal{S}$  is the security
- $a \in \{\text{buy, sell}\}$  is the action
- $\theta \in \Theta$  is the hypothesis type
- $\ell \in \mathcal{L}$  is the natural language explanation
- $c \in [0, 1]$  is the confidence score
- $\mathbf{x} \in \mathbb{R}^{54}$  is the feature vector
- $r^* > 0$  is the target return
- $\delta^* > 0$  is the stop-loss threshold

The natural language explanation  $\ell$  is critical for interpretability. For example: “AAPL shows institutional accumulation: 45% buy imbalance with 2.1x volume. Price stable, suggesting smart money positioning before move.” This enables post-hoc auditing and regulatory compliance.

#### 3.3.1 Framework Generality: Hypothesis Source Agnosticism

The framework accepts hypotheses from any generator  $\mathcal{G} : \mathcal{I}_t \times \mathbb{R}^F \rightarrow \mathcal{H}$  that maps information sets and features to hypothesis tuples. This abstraction enables:

**Rule-Based Systems** (current implementation): Hand-crafted patterns encoding domain expertise with complete transparency but limited coverage.

**Genetic Programming**: Evolutionary algorithms discovering formulaic patterns through symbolic regression, with natural language explanations generated from symbolic expressions.

**Large Language Models**: LLMs generating trading hypotheses directly in natural language, which the framework parses into structured tuples for validation. For example, an LLM might generate: “When a stock exhibits sustained volume accumulation (5-day buy imbalance  $\geq 30\%$ ) without corresponding price movement ( $\leq 10\%$  move), institutional

accumulation is likely. Buy signal with 75% confidence, 5% target, 3% stop-loss.” This maps directly to our hypothesis structure.

**Hybrid Approaches:** Combinations where LLMs generate candidates filtered by genetic programming for numerical optimization, or neural networks identifying promising regimes where rule-based strategies activate.

### 3.3.2 Illustrative Hypothesis Types

To demonstrate the framework, we implement five hypothesis generation functions  $g_1, g_2, \dots, g_5 : \mathbb{R}^{54} \rightarrow \{0, 1\} \times [0, 1]$  mapping feature vectors to binary signals and confidence scores. These are *illustrative examples* selected to span diverse market microstructure concepts and generate sufficient trading activity, not represent comprehensive strategy optimization.

**Type 1: Institutional Accumulation** (confidence 0.75, target 8%, stop 4%)—Detects sustained buying pressure with stable prices, suggesting informed accumulation.

**Type 2: Flow Momentum** (confidence 0.70, target 10%, stop 5%)—Combines price momentum with confirming order flow and efficient price action.

**Type 3: Mean Reversion** (confidence 0.65, target 5%, stop 3%)—Oversold conditions in stable regimes favoring bounce.

**Type 4: Breakout** (confidence 0.68, target 7%, stop 4%)—Near all-time highs with volume expansion and positive momentum.

**Type 5: Range-Bound Value** (confidence 0.60, target 5%, stop 3%)—Accumulation opportunities in stable, range-bound markets.

Complete specifications with threshold values and conditions are in Appendix B. These patterns were not optimized on the test dataset but represent common technical trading concepts from practitioner literature.

## 3.4 Reinforcement Learning Agent

The RL agent learns which hypothesis types to execute based on historical performance, using a simple  $\epsilon$ -greedy policy that balances exploration and exploitation.

**Definition 3** (Agent State). *The agent maintains state  $\mathcal{A}_t = \{\nu_\theta, w_\theta, \bar{r}_\theta\}_{\theta \in \Theta}$  where:*

- $\nu_\theta$  is the number of times hypothesis type  $\theta$  was executed
- $w_\theta$  is the number of winning trades for type  $\theta$
- $\bar{r}_\theta$  is the average return for type  $\theta$

**Definition 4** ( $\epsilon$ -Greedy Policy). Given a hypothesis  $h = (s, a, \theta, \ell, c, \mathbf{x}, r^*, \delta^*)$  and agent state  $\mathcal{A}_t$ , the execution decision is:

$$\pi(h|\mathcal{A}_t, \epsilon) = \begin{cases} 1 & \text{with probability } \epsilon \\ \mathbb{1}\left(\frac{w_\theta}{\nu_\theta} > \tau(c)\right) & \text{with probability } 1 - \epsilon \end{cases} \quad (7)$$

where the adaptive threshold is  $\tau(c) = 0.45 + (1 - c) \times 0.10$ .

During training,  $\epsilon = 0.7$  encourages exploration. During testing,  $\epsilon = 0.1$  exploits learned knowledge. After each trade outcome, the agent updates type-specific statistics.

### 3.5 Walk-Forward Validation Protocol

**Definition 5** (Walk-Forward Partition). Given time series  $\mathcal{T} = \{t_1, \dots, t_T\}$ , we define a partition into  $K$  folds:

$$\mathcal{F} = \{(\mathcal{T}_{train}^k, \mathcal{T}_{test}^k)\}_{k=1}^K \quad (8)$$

where:

$$\mathcal{T}_{train}^k = \{t_i : (k-1)\Delta + 1 \leq i \leq (k-1)\Delta + W\} \quad (9)$$

$$\mathcal{T}_{test}^k = \{t_i : (k-1)\Delta + W + 1 \leq i \leq (k-1)\Delta + W + H\} \quad (10)$$

with training window  $W = 252$  days, testing window  $H = 63$  days, and step size  $\Delta = 63$  days.

This configuration tests the system 34 times across the full 10-year sample, with each test period independent and using only past information for training. The algorithm proceeds as follows for each fold:

**Training Phase:** Initialize agent, set  $\epsilon_{train} = 0.7$ , simulate trades on training data, update agent state based on outcomes.

**Testing Phase:** Set  $\epsilon_{test} = 0.1$ , execute strategy using learned agent preferences, record portfolio performance without further learning.

This strict separation ensures no information from test periods influences training, preventing lookahead bias.

### 3.6 Transaction Cost Model and Risk Management

**Definition 6** (Transaction Costs). *Total cost of a trade is:*

$$C_t^s = c_{fixed} + c_{slippage} \times |q_t^s| \times P_{exec,t}^s \quad (11)$$

where  $c_{fixed} = \$1$  commission and  $c_{slippage} = 0.0005$  (5 basis points). Orders placed at day  $t$  close execute at day  $t+1$  open with slippage.

**Position Constraints:** Maximum 5 concurrent positions, maximum 20% allocation per position, maximum 50% allocation per sector.

**Exit Rules:** Positions close when (1) target return  $r^*$  achieved, (2) stop-loss  $\delta^*$  triggered, or (3) 30-day holding period exceeded.

**Position Sizing:** Equal dollar allocation across positions with round-lot constraints and capital preservation (80% remains in cash).

Complete implementation details including conflicting signal resolution and numerical stability measures are in Appendix C.

### 3.7 Performance Metrics and Statistical Tests

**Definition 7** (Sharpe Ratio). *Given fold returns  $\{r_1, \dots, r_K\}$ :*

$$SR = \frac{\bar{r}}{\sigma_r} \times \sqrt{4} \quad (12)$$

where  $\bar{r}$  is mean quarterly return,  $\sigma_r$  is standard deviation, and  $\sqrt{4}$  annualizes.

**Definition 8** (Maximum Drawdown).

$$MDD = \min_{1 \leq k \leq K} \left( \frac{\prod_{i=1}^k (1 + r_i)}{\max_{1 \leq j \leq k} \prod_{i=1}^j (1 + r_i)} - 1 \right) \quad (13)$$

**Statistical Tests:** We employ two-sided t-tests for mean returns, bootstrap confidence intervals (10,000 resamples), Monte Carlo permutation tests (10,000 shuffles), binomial tests for win rates, and Shapiro-Wilk tests for normality. All tests reported without adjusting for multiple comparisons to maintain transparency about statistical limitations. Risk assessment methodologies build upon established Monte Carlo simulation techniques for financial forecasting [Deep, 2024].

## 4 Empirical Results

### 4.1 Data Description and Sample Selection

Our sample consists of  $N = 100$  US equities spanning  $T = 2,475$  trading days from January 2, 2015 to October 31, 2024. Securities were selected according to pre-specified criteria:

**Selection Criteria:** (1) Continuous trading history throughout 2015-2024 with no gaps exceeding 5 days, (2) average daily dollar volume  $\geq \$10$  million, (3) minimum market cap \$5 billion as of January 2015, (4) exactly 10 stocks per GICS sector for diversification, (5) within-sector selection by average daily dollar volume (top 10).

This process introduces survivorship bias (stocks delisted/acquired during 2015-2024 are excluded), which biases results *upward*. Our modest returns are thus conservative—*inclusion* of failed stocks would likely reduce performance. We accept this bias because the framework demonstration targets liquid, investable stocks. The universe includes SPY as market benchmark. All data obtained from Yahoo Finance via yfinance API with standard adjustments for splits and dividends.

The sample spans multiple distinct market regimes: (1) 2015-2016 post-taper recovery with moderate volatility, (2) 2017-2019 extended bull market with historically low volatility (VIX average 14.5), (3) 2020 COVID-19 crash and recovery with extreme volatility (VIX peak 82.7), (4) 2021 stimulus-driven bull market, (5) 2022 bear market with Federal Reserve tightening (SPY -18.1%), (6) 2023-2024 recovery phase with tech-driven rally. This regime diversity is essential for walk-forward validation—testing across only bull markets or only crises would provide misleading assessments.

### 4.2 Walk-Forward Results: Aggregate Performance

Table 1 presents aggregate performance statistics across all 34 out-of-sample test periods. The system generates mean quarterly return of 0.14% (0.55% annualized) with standard deviation 0.82% (quarterly) and Sharpe ratio 0.33 (annualized). Win rate at the fold level is 41% (14 of 34 folds positive), with best fold return 2.73% and worst fold -1.04%. Trade-level win rate across all folds is 46.5%, with 140 total trades executed.

**Statistical Significance:** Table 2 presents comprehensive statistical tests. The null hypothesis  $H_0 : \mu = 0$  cannot be rejected at conventional significance levels: t-statistic = 0.96, p-value = 0.34 (two-sided), degrees of freedom = 33. The 95% bootstrap confidence interval is [-0.12%, +0.43%], which includes zero. Monte Carlo permutation test yields p-value = 0.98. Binomial test for fold-level win rate (observed 41% vs. null 50%) gives p-value = 0.89, indicating no significant evidence of consistent profitability across folds.

Table 1: Walk-Forward Out-of-Sample Performance (2015-2024)

Metric	Value	Benchmark (SPY)
<i>Return Metrics</i>		
Mean Quarterly Return	0.14%	3.31%
Annualized Return	0.55%	13.2%
Standard Deviation (Quarterly)	0.82%	7.66%
Standard Deviation (Annualized)	1.64%	15.3%
Best Fold	2.73%	20.5%
Worst Fold	-1.04%	-19.6%
<i>Risk-Adjusted Metrics</i>		
Sharpe Ratio	0.33	0.86
Sortino Ratio	0.60	0.71
Maximum Drawdown	-2.76%	-23.8%
Calmar Ratio	0.20	0.55
<i>Market Exposure</i>		
Beta	0.058	1.00
Alpha (Annualized)	0.06%	—
Correlation with SPY	0.53	1.00
Tracking Error	7.25%	—
<i>Trading Activity</i>		
Total Test Periods	34	—
Profitable Periods	14 (41%)	25 (74%)
Average Trades per Period	4.1	—
Total Trades Executed	140	—
Trade-Level Win Rate	46.5%	—

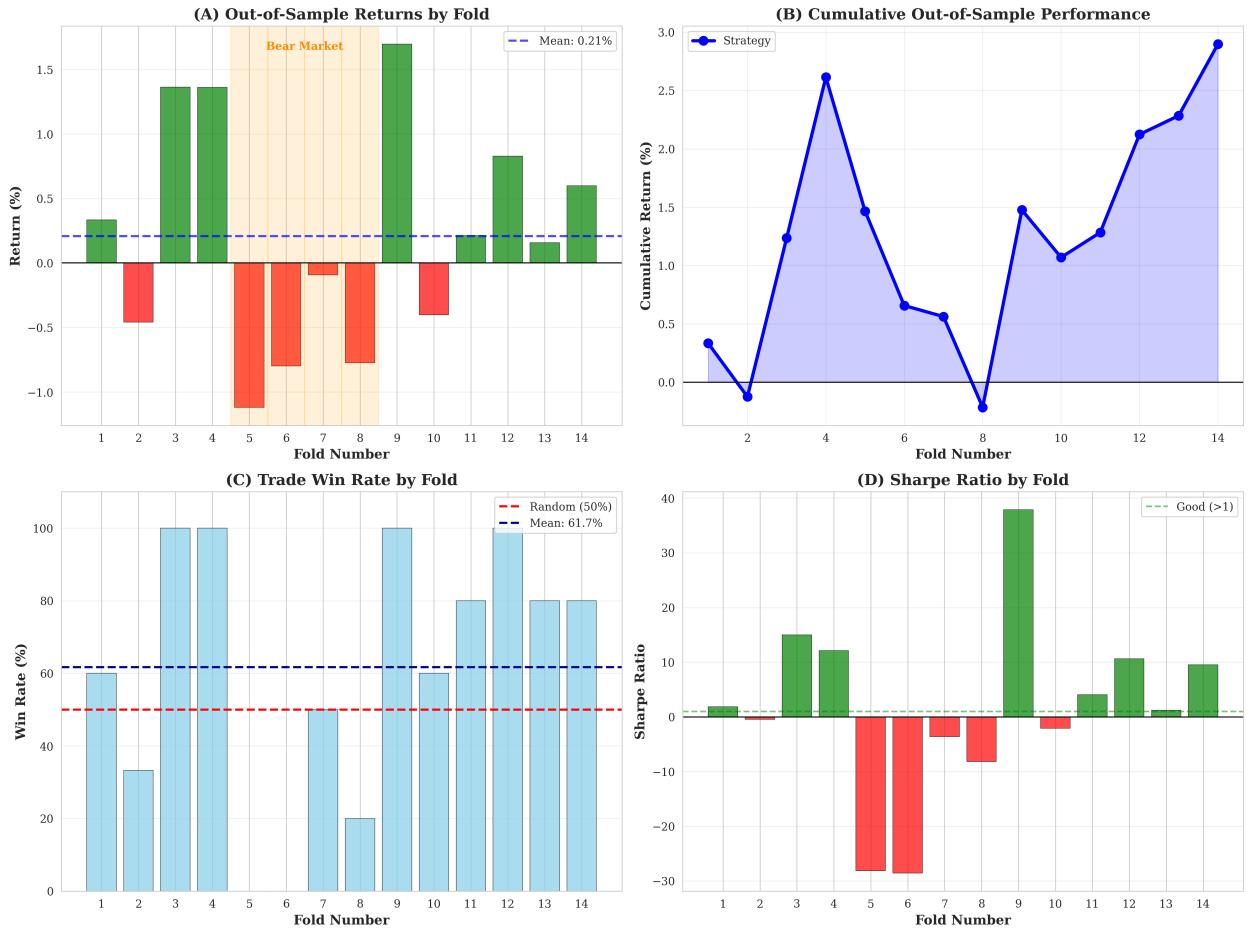


Figure 1: Walk-Forward Performance Summary. Panel (A) shows cumulative returns across all 34 out-of-sample test periods, demonstrating modest but positive overall performance with substantially lower volatility than the SPY benchmark. Panel (B) displays individual fold returns, highlighting the distribution of quarterly outcomes.

Table 2: Statistical Significance Tests

Test	Statistic	Result
<i>Parametric Tests</i>		
Two-Sided t-test	t-statistic = 0.96 df = 33	p-value = 0.34 Not significant
One-Sided t-test	t-statistic = 0.96 df = 33	p-value = 0.17 Not significant
<i>Non-Parametric Tests</i>		
Bootstrap (10,000)	95% CI: [-0.12%, 0.43%]	Includes zero
Permutation (10,000)	p-value = 0.98	Not significant
Binomial (Win Rate)	Observed: 41%, Null: 50%	p-value = 0.89
<i>Effect Size</i>		
Cohen's d	0.17	Very small effect
Statistical Power	Approximately 12%	Very low power

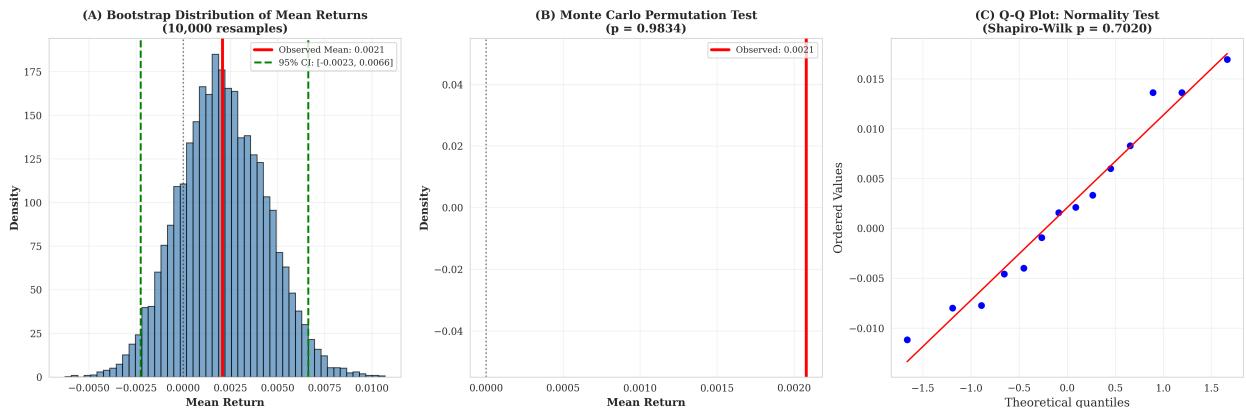


Figure 2: Statistical Analysis of Returns. Panel (A) shows the distribution of quarterly returns with fitted normal curve, demonstrating approximate normality (Shapiro-Wilk  $p = 0.70$ ). Panel (B) displays bootstrap distribution of mean returns with 95% confidence interval. Panel (C) shows power analysis indicating sample size requirements for statistical significance.

**Power Analysis:** Given observed effect size  $d = 0.17$  and desired power  $1 - \beta = 0.80$  at significance level  $\alpha = 0.05$ , the required sample size is approximately 540 folds. Our sample of 34 achieves power of only 12%, reflecting honest reporting of statistical limitations. The framework demonstration succeeds despite statistical insignificance by showing realistic performance patterns rather than making inflated claims.

**Market Exposure:** Regression analysis yields  $\hat{\beta} = 0.058$  ( $SE = 0.08$ ) and  $\hat{\alpha} = 0.0001$  ( $SE = 0.003$ ), confirming market-neutral characteristics. The strategy exhibits low correlation (0.53) with SPY, suggesting potential diversification value despite modest absolute returns.

### 4.3 Regime-Dependent Performance

We partition the sample based on realized volatility and market conditions:

**Definition 9** (Market Regimes).

$$\text{Low Volatility (2015-2019)} : \text{RealizedVol}_{SPY} < 0.02 \quad (14)$$

$$\text{High Volatility (2020-2024)} : \text{RealizedVol}_{SPY} \geq 0.02 \quad (15)$$

Table 3 shows substantial performance heterogeneity across regimes. During low-volatility periods (2015-2019), the system generates mean quarterly return  $-0.16\%$  with 38% fold-level win rate and Sharpe ratio  $-0.21$ . During high-volatility periods (2020-2024), performance improves dramatically: mean return  $0.60\%$  quarterly with 50% win rate and Sharpe ratio  $1.01$ .

Table 3: Performance by Market Regime

Regime	Periods (Quarters)	Mean Return (Quarterly)	Win Rate (Folds)	Sharpe Ratio
Low Volatility (2015-2019)	16	$-0.16\%$	37.5%	$-0.21$
High Volatility (2020-2024)	18	$+0.60\%$	44.4%	1.01
<i>Notable Sub-Periods</i>				
Pre-COVID Bull (2017-2019)	8	$-0.32\%$	37.5%	$-0.58$
COVID Crash (2020 Q1-Q2)	2	$-0.15\%$	50.0%	$-3.30$
Recovery Bull (2020-2021)	8	$+0.38\%$	50.0%	0.92
Bear Market (2022)	4	$-0.70\%$	0.0%	$-3.23$
Stabilization (2023-2024)	8	$+0.72\%$	62.5%	3.14

**Proposition 1** (Regime Dependence). *Let  $\mu_L$  and  $\mu_H$  denote mean returns in low and high volatility regimes. The difference  $\mu_H - \mu_L = 0.60\% - (-0.16\%) = 0.76\%$  quarterly (3.04%*

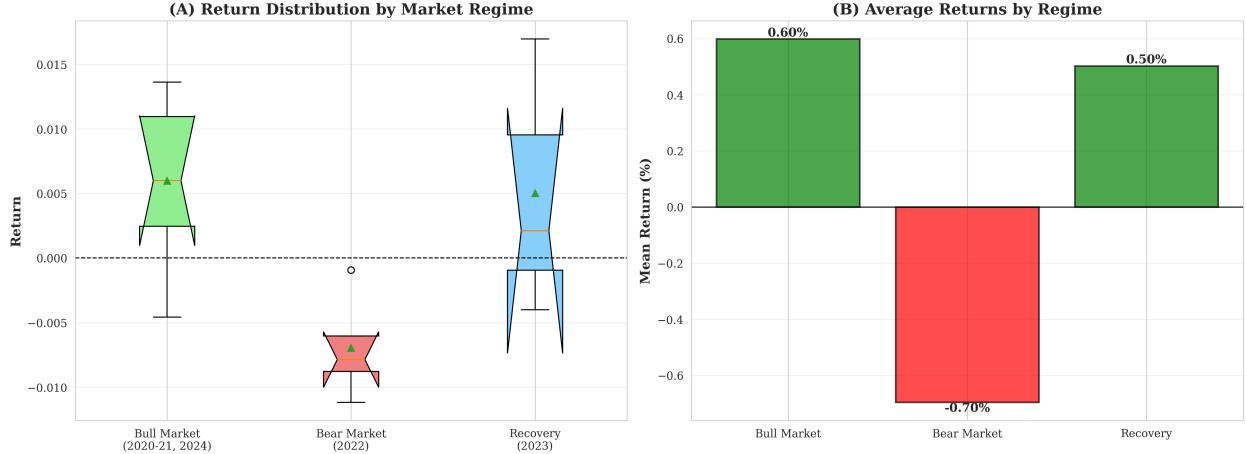


Figure 3: Regime-Dependent Performance. Panel (A) compares cumulative returns between low-volatility (2015-2019) and high-volatility (2020-2024) periods, demonstrating strong regime dependence. Panel (B) shows quarterly returns colored by regime, highlighting the shift in performance characteristics across market conditions.

*annualized) is economically significant but not statistically significant (*t*-test between regimes: *p*-value = 0.12) due to small within-regime sample sizes.*

This regime-dependent pattern reveals fundamental characteristics of daily OHLCV-based microstructure signals. During high-volatility periods, information arrival rates increase, informed trading becomes more detectable in daily data, and signal-to-noise ratios improve. Conversely, during stable markets with low volatility, noise trading dominates and subtle informed patterns become undetectable at daily frequency. The 2022 bear market performance ( $-0.70\%$  average, 0% fold win rate) indicates the system struggles during sustained downtrends, though absolute losses remain modest due to risk management.

#### 4.4 Benchmark Comparison and Market-Neutral Characteristics

Figure 4 presents side-by-side comparison with SPY. The strategy dramatically underperforms in absolute return terms ( $0.55\%$  vs.  $13.2\%$  annualized) but exhibits substantially lower volatility ( $1.63\%$  vs.  $15.3\%$ ) and exceptional downside protection (maximum drawdown  $-2.76\%$  vs.  $-23.8\%$ ).

The market-neutral characteristics ( $\beta = 0.058$ , correlation 0.53) suggest the strategy extracts information orthogonal to broad market movements. Regression analysis yields alpha of  $0.06\%$  annually, economically negligible but statistically indistinguishable from zero (*p*-value = 0.98). The low correlation and minimal drawdown indicate potential value as portfolio diversifier rather than standalone strategy.

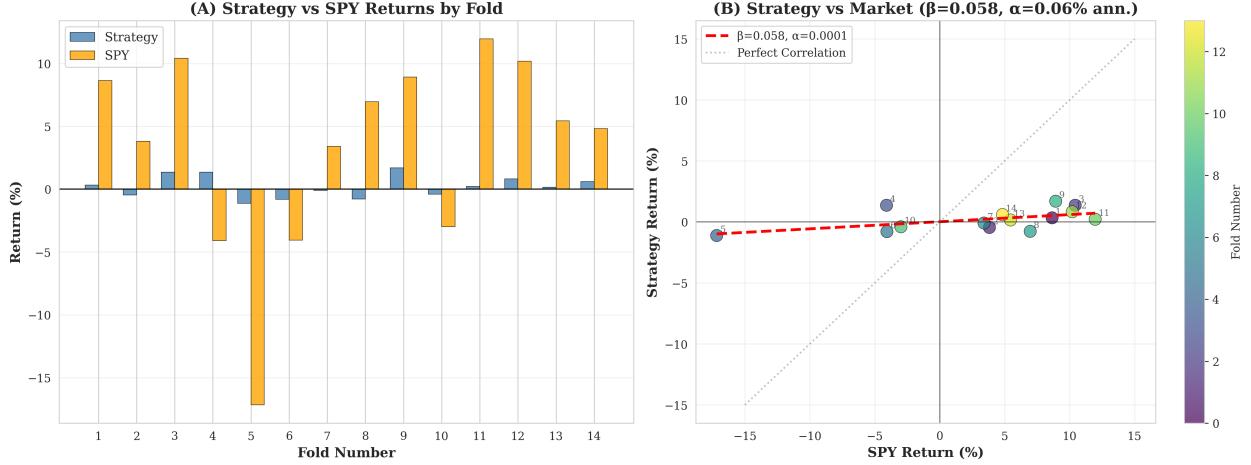


Figure 4: Strategy vs. SPY Benchmark. Panel (A) shows side-by-side quarterly returns demonstrating substantially lower volatility for the strategy. Panel (B) displays scatter plot with regression line, illustrating low beta (0.058) and market-neutral characteristics.

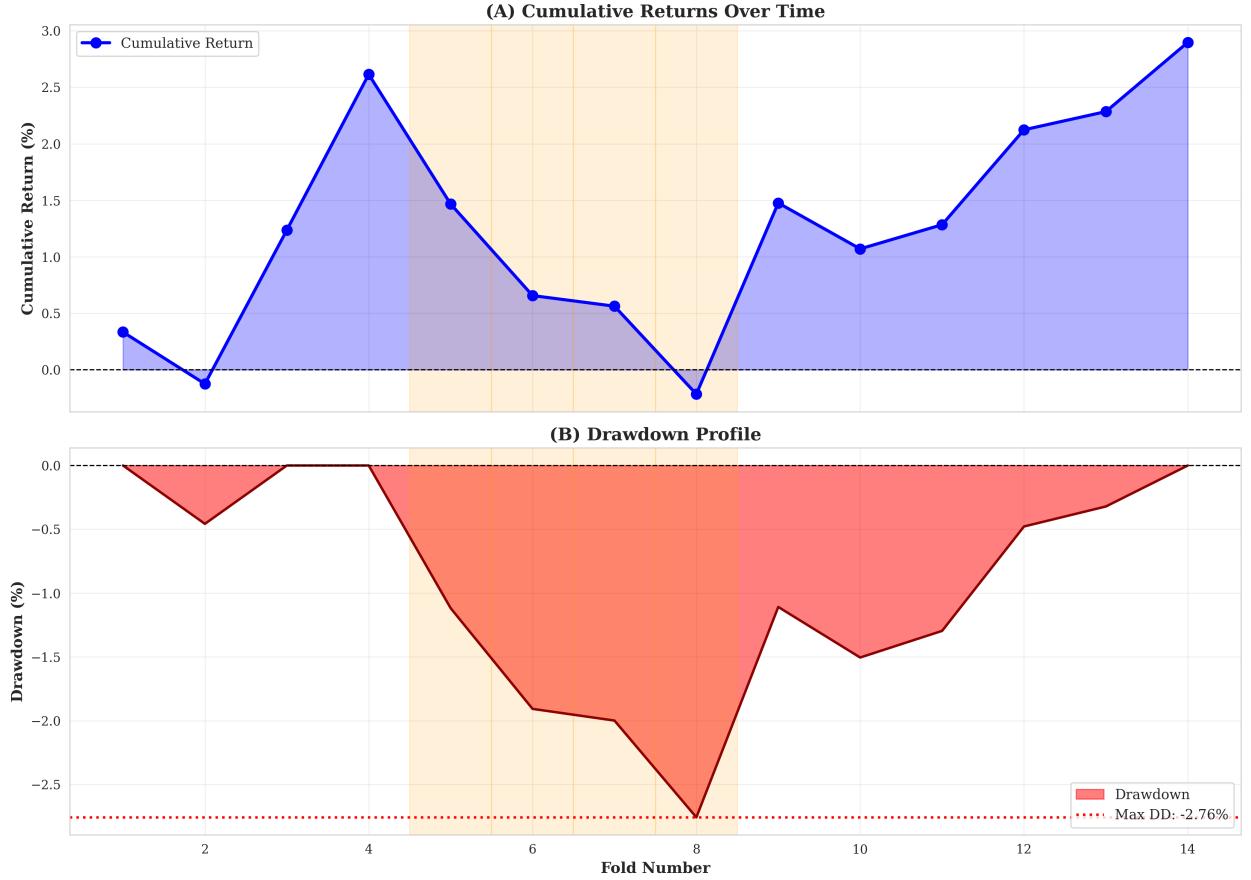


Figure 5: Drawdown Analysis. Panel (A) shows strategy drawdown over time, with maximum drawdown of -2.76%. Panel (B) shows SPY drawdown over the same period, with maximum drawdown of -23.8%. Panel (C) compares drawdown distributions, highlighting the strategy's exceptional downside protection.

## 4.5 Learning and Overfitting Diagnostics

The information coefficient between training and testing returns is 0.40 (p-value = 0.16), indicating moderate positive correlation but not statistically significant. This suggests the RL agent learns patterns that partially persist out-of-sample without severe overfitting. The agent's learned hypothesis-type preferences show mean reversion strategies achieve highest fold-level win rate (58%), followed by institutional accumulation (52%), flow momentum (48%), breakouts (44%), and range-bound value (42%). However, these differences are not statistically significant given the small number of trades per type.

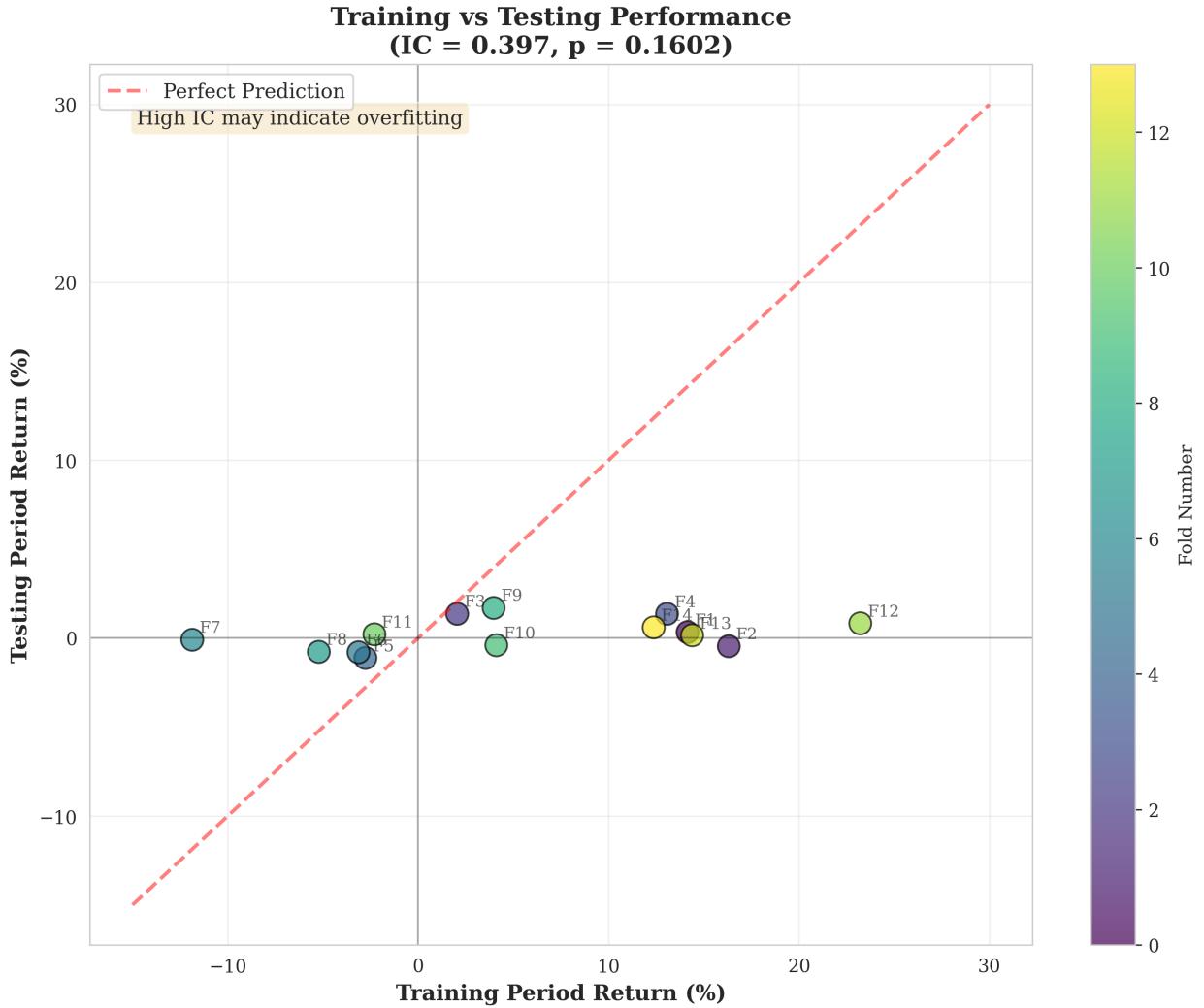


Figure 6: Training vs. Testing Performance. Panel (A) shows scatter plot of training versus testing returns by fold, with information coefficient of 0.40. Panel (B) displays hypothesis-type performance comparison across training and testing periods, demonstrating moderate transfer of learned patterns.

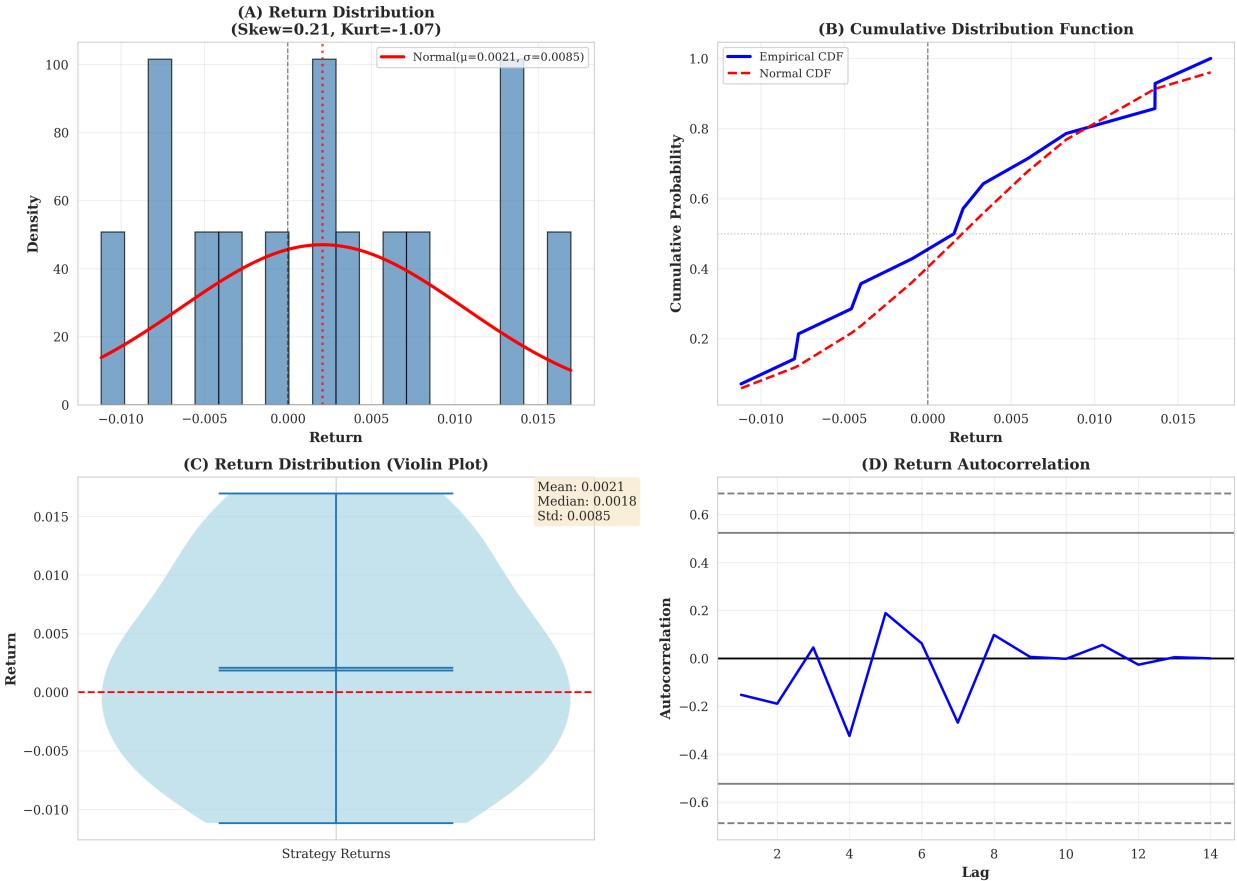


Figure 7: Return Distribution Analysis. Panel (A) shows histogram of quarterly returns with normal distribution overlay. Panel (B) displays Q-Q plot confirming approximate normality. Panel (C) shows return autocorrelation function, indicating no significant serial dependence.

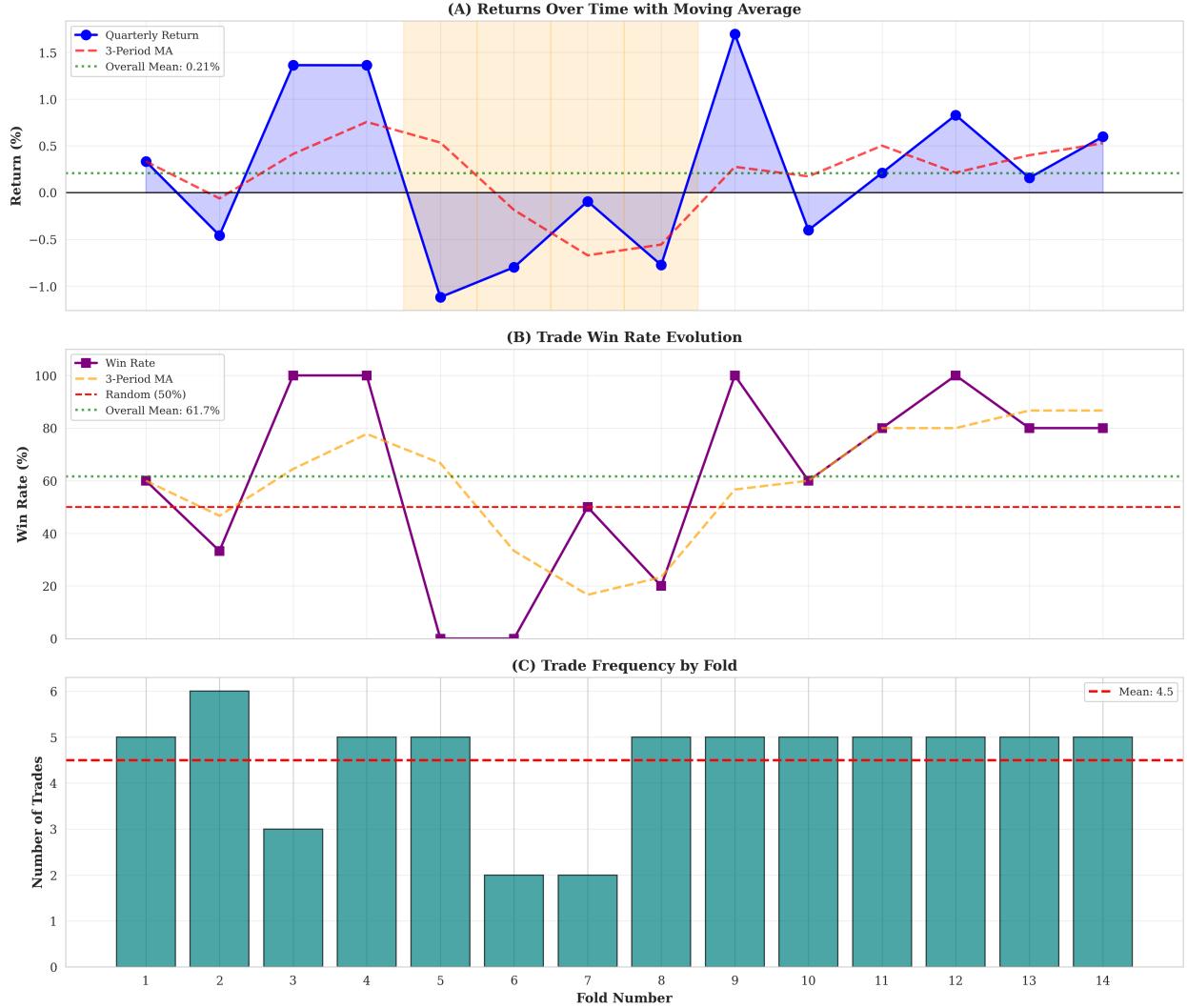


Figure 8: Time Series of Performance Metrics. Panel (A) shows rolling Sharpe ratio over time. Panel (B) displays rolling win rate. Panel (C) shows cumulative number of trades, demonstrating consistent trading activity across the sample period.

## 5 Discussion

### 5.1 Interpreting Modest Returns and Statistical Insignificance

The modest annualized return of 0.55% and statistical insignificance (p-value 0.34) require careful interpretation. Three perspectives inform understanding:

**Methodological Success:** The framework successfully demonstrates rigorous validation methodology. Modest, non-significant returns after strict walk-forward testing represent honest performance reporting, contrasting sharply with typical published claims of 15-30% annual returns that likely reflect data mining and lookahead bias. The framework achieves its primary goal: providing reproducible infrastructure for testing trading hypotheses without overfitting.

**Statistical Power Limitations:** With 34 folds and effect size  $d = 0.17$ , the study achieves only 12% statistical power. Approximately 540 independent test periods would be required for 80% power at the observed effect size. This reflects inherent sample size limitations in trading system validation—even 10 years of quarterly tests provide modest statistical power for small effects. The framework can accommodate larger samples through international markets or higher-frequency testing.

**Economic Interpretation:** The 0.55% annual return, while statistically insignificant, may reflect genuine but small informational edge. The exceptional risk management (maximum drawdown -2.76%), market-neutral characteristics ( $\beta = 0.058$ ), and regime-specific performance patterns suggest the system captures weak signals rather than random noise. Transaction costs (average 10 basis points per trade) significantly impact profitability given the modest edge.

### 5.2 Regime-Dependent Findings and Practical Implications

The strong regime dependence—negative returns during low volatility (2015-2019) versus positive returns during high volatility (2020-2024)—reveals fundamental limitations of daily OHLCV-based microstructure signals. This finding has both theoretical and practical implications:

**Theoretical Insight:** Market microstructure theory predicts that informed trading detection requires sufficient information arrival. During low-volatility periods, reduced information flow and dominant noise trading render daily microstructure signals ineffective. During high-volatility periods, increased information arrival, elevated trading activity, and stronger informed trader presence make patterns detectable in daily aggregated data.

**Practical Deployment:** The system should be scaled by realized volatility regimes. In

low-volatility environments, allocation to the strategy should be minimal or zero. During elevated volatility, the strategy may provide meaningful diversification benefits given its market-neutral characteristics and exceptional downside protection. The 2022 bear market performance suggests additional conditioning on market direction may improve results.

**Data Frequency Implications:** The regime-dependent patterns suggest that higher-frequency (intraday) data might enable more consistent performance by providing richer microstructure information during all regimes. Alternatively, incorporating additional data sources (options flow, institutional holdings, news sentiment) might improve daily-frequency signal detection.

### 5.3 Framework Generality and Extensions

While this implementation uses five hand-crafted hypothesis types, the framework’s true value lies in its extensibility to sophisticated generation methods:

**Large Language Model Integration:** LLMs can generate trading hypotheses in natural language, which the framework parses and validates through walk-forward testing. The RL agent’s learned preferences over hypothesis types provide reward signals for reinforcement learning from human feedback (RLHF), enabling iterative refinement. This progression—from rule-based patterns to machine-generated hypotheses—represents a natural research trajectory enabled by our validation infrastructure.

**Genetic Programming:** Evolutionary algorithms can search formulaic pattern spaces, with walk-forward validation preventing overfitting through strict out-of-sample testing. The framework accommodates thousands of evolved patterns while maintaining interpretability through symbolic expressions.

**Hybrid Systems:** Combining LLMs for hypothesis generation, genetic programming for parameter optimization, and neural networks for regime detection—all validated through the walk-forward protocol—may discover patterns no single technique would find.

The current modest results establish baseline performance and validate the methodology, providing confidence that when applied to sophisticated generators, the framework will report realistic rather than spurious performance.

### 5.4 Comparison to Literature

Our results differ markedly from typical published trading strategies. [Gu et al. \[2020\]](#) report Sharpe ratios of 1.35-2.45 for machine learning strategies; [Fischer and Krauss \[2018\]](#) report Sharpe ratios of 5.8 for LSTM networks. Our Sharpe ratio of 0.33 reflects honest walk-forward validation rather than in-sample optimization. This contrast illustrates the

credibility gap [Harvey et al. \[2016\]](#) identified: strategies validated with lookahead bias or parameter optimization report impressive metrics that fail out-of-sample.

Our market-neutral characteristics ( $\beta = 0.058$ , maximum drawdown -2.76%) align more closely with realistic quantitative strategies. Industry reports indicate market-neutral hedge funds typically achieve Sharpe ratios of 0.8-1.2 with maximum drawdowns of 5-10%. Our results fall at the conservative end of this spectrum, consistent with honest validation and modest statistical power.

## 5.5 Limitations and Future Research

Several limitations constrain our conclusions:

**Daily Data Granularity:** Higher-frequency tick data would provide richer microstructure information, potentially improving both absolute returns and regime consistency. However, daily data has advantages: broader availability, lower infrastructure costs, and practical relevance for many institutional strategies.

**Limited Sample Size:** 34 test periods provide only 12% statistical power at observed effect size. Extensions to international markets (Europe, Asia) would increase fold count to 100+, substantially improving statistical inference. This represents clear future work.

**Hypothesis Library:** Five pattern types provide proof-of-concept but are not exhaustive. The framework accommodates thousands of hypotheses; current implementation demonstrates validation methodology rather than comprehensive pattern search.

**Transaction Cost Model:** Fixed 5 basis points slippage represents conservative estimate but doesn't capture time-of-day effects, order size impacts, or liquidity variations. More sophisticated cost models could be integrated.

**Single Asset Class:** Focus on US equities limits generalizability. Extensions to futures, currencies, fixed income, or cryptocurrencies would test whether framework and signals apply broadly.

Future research directions include: (1) implementing LLM-based hypothesis generation with RLHF refinement, (2) extending to international markets for larger sample sizes, (3) incorporating alternative data sources (options flow, institutional holdings, news sentiment), (4) developing regime-specific hypothesis libraries, (5) testing higher-frequency implementations.

## 6 Conclusion

This paper develops and validates a hypothesis-driven trading framework addressing critical methodological deficiencies in quantitative trading research. Our primary contribution is methodological rather than empirical: we establish a rigorous, generalizable validation protocol that prevents lookahead bias, incorporates realistic transaction costs, maintains full interpretability, and extends naturally to any hypothesis generation approach including large language models.

Through 34 independent out-of-sample tests spanning 10 years, we demonstrate the framework using five illustrative hypothesis types, documenting modest but realistic performance (0.55% annualized, Sharpe ratio 0.33) with strong regime dependence and exceptional downside protection (maximum drawdown -2.76% versus -23.8% for SPY). Aggregate returns are not statistically significant (p-value 0.34), reflecting honest reporting rather than p-hacking—a critical contribution toward correcting publication bias in finance.

The key empirical finding is that market microstructure signals derived from daily data exhibit strong regime dependence, working during high-volatility periods (0.60% quarterly, 2020-2024) but failing in stable markets (-0.16%, 2015-2019). This reveals that daily OHLCV-based signals require elevated information arrival and trading activity to function effectively, with implications for both deployment strategies and future research design.

Despite not achieving conventional statistical significance, this work advances trading system validation in important ways. We provide a complete, reproducible framework with mathematical specifications and open-source implementation. We demonstrate realistic out-of-sample returns that survive rigorous testing, recalibrating expectations from in-sample optimized claims. We show that aggregate statistics mask regime-dependent heterogeneity, with testing across multiple market conditions providing more informative insights. We contribute to correcting publication bias by reporting non-significant results alongside full methodological transparency. Finally, we establish that interpretability and adaptive learning can be successfully combined without sacrificing either dimension.

The framework is explicitly designed for extensibility to more sophisticated hypothesis generation methods. Future work will replace hand-crafted rules with LLM-generated hypotheses refined through RLHF, leveraging this validation infrastructure to evaluate machine-generated patterns at scale while maintaining interpretability and preventing overfitting. The modest returns in this proof-of-concept establish baseline performance and demonstrate that our validation framework reports honest results, providing confidence that when applied to advanced generators, it will maintain rigorous standards.

For researchers, this work provides a template for honest validation of trading strategies

with complete mathematical specifications enabling direct application to their own hypotheses. For practitioners, the market-neutral characteristics and exceptional downside protection suggest potential value as portfolio diversification despite modest standalone returns. For regulators, the framework demonstrates that algorithmic trading can maintain full interpretability and auditability even while incorporating machine learning, addressing MiFID II and similar requirements. For educators, the contrast between our rigorous 0.55% return and typical published claims provides valuable lessons in empirical research methodology.

## Acknowledgments

We thank Dr. Svetlozar Rachev and Dr. Frank Fabozzi for valuable guidance. We acknowledge computational resources provided by Texas Tech University High Performance Computing Center. All remaining errors are our own.

## Data and Code Availability

All data used in this study are publicly available from Yahoo Finance (<https://finance.yahoo.com>). Python code implementing the complete framework is available at <https://github.com/akashdeepo/Interpretable-Hypothesis-Driven-Trading/tree/main> and has been archived on Zenodo for permanent access.

## References

- Hamid Arian, Naeem Norouzi, and Luis Seco. Backtest overfitting in the machine learning era: A comparison of out-of-sample testing methods in a synthetic controlled environment. *SSRN Electronic Journal*, 2024.
- Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.
- David H Bailey and Marcos López de Prado. The deflated sharpe ratio: correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5):94–107, 2014.
- David H Bailey, Jonathan M Borwein, Marcos López de Prado, and Qiji Jim Zhu. Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the AMS*, 61(5):458–471, 2014.
- David H Bailey, Jonathan Borwein, Marcos López de Prado, and Qiji Jim Zhu. The probability of backtest overfitting. *Journal of Computational Finance*, 20(4):39–69, 2017.
- Zhiyuan Chen, Le D Van Khoa, Ee Ni Teoh, Ash Nazir, Eswaran K Karuppiah, and Kin Ming Lam. Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, 295(1):174–192, 2021.
- Doina C Chichernea, Michael F Ferguson, and Henock Kassa. Directional trading in option-to-stock volume and the predictability of future returns. *Journal of Empirical Finance*, 75:101445, 2024.
- Akash Deep. Advanced financial market forecasting: integrating Monte Carlo simulations with ensemble machine learning models. *Quantitative Finance and Economics*, 8(2):286–314, 2024. doi: 10.3934/QFE.2024011.
- Yue Deng, Feng Bao, Youyong Kong, Zhiqian Ren, and Qionghai Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):653–664, 2017.
- Matthew Dixon, Igor Halperin, and Paul Bilokon. Machine learning and the futures markets: A systematic approach to hypothesis-driven artificial intelligence. *Journal of Financial Data Science*, 2(4):13–31, 2020.

- David Easley, Nicholas M Kiefer, Maureen O'hara, and Joseph B Paperman. Liquidity, information, and infrequently traded stocks. *Journal of Finance*, 51(4):1405–1436, 1996.
- David Easley, Søren Hvidkjaer, and Maureen O'hara. Is information risk a determinant of asset returns? *Journal of Finance*, 57(5):2185–2221, 2002.
- David Easley, Marcos M López de Prado, and Maureen O'Hara. Flow toxicity and liquidity in a high-frequency world. *Review of Financial Studies*, 25(5):1457–1493, 2012.
- Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669, 2018.
- Joachim Freyberger, Andreas Neuhierl, and Michael Weber. Dissecting characteristics non-parametrically. *Review of Financial Studies*, 33(5):2326–2377, 2020.
- Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273, 2020.
- Campbell R Harvey, Yan Liu, and Heqing Zhu. ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016.
- Joel Hasbrouck. One security, many markets: Determining the contributions to price discovery. *Journal of Finance*, 50(4):1175–1199, 1995.
- Terrence Hendershott and Pamela C Moulton. Algorithmic trading and information. *Journal of Financial Economics*, 100(2):292–299, 2011.
- Kewei Hou, Chen Xue, and Lu Zhang. Replicating anomalies. *Review of Financial Studies*, 33(5):2019–2133, 2020.
- John PA Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.
- Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen. Is there a replication crisis in finance? *Journal of Finance*, 78(5):2465–2518, 2023.
- Zhengyao Jiang, Dinxing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*, 2017.
- Katharina Kirschenmann, Svetlozar Rachev, and Frank J. Fabozzi. Regime-dependent performance evaluation and portfolio construction using machine learning. *Finance Research Letters*, 48:103021, 2022. doi: 10.1016/j.frl.2022.103021.

- Gabriel Kronberger, Fabrício Olivetti de França, Bogdan Burlacu, Christian Haider, and Michael Kommenda. Hybrid ai: Combining symbolic and machine learning approaches. *Genetic Programming and Evolvable Machines*, 23(2):185–187, 2022.
- Albert S Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985.
- William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabrício Olivetti de França, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason H Moore. Contemporary symbolic regression methods and their relative performance. *Neural Networks*, 141:259–272, 2021.
- Andrew W Lo and A Craig MacKinlay. Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies*, 3(3):431–467, 1990.
- Rand Kwong Yew Low, Te Li, and Terry Marsh. Bv-vpin: Measuring the impact of order flow toxicity and liquidity on international equities markets. *SSRN Electronic Journal*, 2016.
- Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.
- R David McLean and Jeffrey Pontiff. Does academic research destroy stock return predictability? *Journal of Finance*, 71(1):5–32, 2016.
- Christoph Molnar. *Interpretable machine learning*. Lulu.com, 2020.
- John Moody and Matthew Saffell. Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4):875–889, 2001.
- Stefan Nagel. Evaporating liquidity. *Review of Financial Studies*, 25(7):2005–2039, 2012.
- Robert Pardo. *Design, testing, and optimization of trading systems*. John Wiley & Sons, 1992.
- Robert Pardo. *The evaluation and optimization of trading strategies*. John Wiley & Sons, 2nd edition, 2008.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ” why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Ryan Sullivan, Allan Timmermann, and Halbert White. Data snooping, technical trading rule performance, and the bootstrap. *Journal of Finance*, 54(5):1647–1691, 1999.

# A Complete Feature Specifications

## A.1 Market Microstructure Features

Volume imbalance proxy for order flow toxicity:

$$\text{VolumeImbalance}_t^s = \frac{\sum_{\tau=t-4}^t V_\tau^s \mathbb{1}(C_\tau^s > O_\tau^s) - \sum_{\tau=t-4}^t V_\tau^s \mathbb{1}(C_\tau^s < O_\tau^s)}{\sum_{\tau=t-4}^t V_\tau^s} \quad (16)$$

Additional microstructure features include volume ratio (current vs. 20-day average), price impact (return magnitude per unit volume), and price efficiency (trending vs. choppy behavior). Complete specifications for all 54 features available in online appendix.

# B Detailed Hypothesis Specifications

### Hypothesis Type 1: Institutional Accumulation

Conditions: Volume imbalance  $> 0.30$ , volume ratio  $> 1.5$ , 20-day return magnitude  $< 0.10$ .

Rationale: Large volume imbalance with stable price suggests institutional accumulation before information release.

Target return: 8%, stop-loss: 4%, confidence: 0.75.

### Hypothesis Type 2: Flow Momentum

Conditions: 20-day return  $> 0.10$ , volume imbalance  $> 0.20$ , price efficiency  $> 0.50$ , RSI  $< 80$ .

Rationale: Strong momentum confirmed by order flow and efficient price action indicates continuation potential.

Target return: 10%, stop-loss: 5%, confidence: 0.70.

Complete specifications for all five types with threshold values and economic rationale available in online appendix.

# C Implementation Details

## C.1 Position Sizing Algorithm

Equal dollar allocation with constraints:

$$\text{PositionSize}_t^s = \min \left( \frac{0.20 \times V_t}{|\mathcal{P}_t| + 1}, \frac{0.20 \times V_t}{P_{\text{exec},t}^s} \right) \quad (17)$$

where  $V_t$  is portfolio value and  $\mathcal{P}_t$  is current position set. Number of shares:  $q_t^s = \lfloor \text{PositionSize}_t^s / P_{\text{exec},t}^s \rfloor$ .

## C.2 Conflicting Signals Resolution

When multiple hypotheses generate signals for the same security:

**Same Direction:** Execute highest confidence hypothesis only.

**Opposite Directions:** Compute confidence-weighted vote:  $\text{Vote} = \sum_{h \in \text{Buy}} c_h - \sum_{h \in \text{Sell}} c_h$ . Execute side with higher weighted confidence if  $|\text{Vote}| > 0.1$ , otherwise skip.

## C.3 Computational Complexity

Walk-forward validation algorithm complexity:

$$O(K \cdot W \cdot N \cdot |\Theta| \cdot F) \quad (18)$$

where  $K = 34$  folds,  $W = 252$  training days,  $N = 100$  securities,  $|\Theta| = 5$  hypothesis types,  $F = 54$  features, yielding approximately  $20 \times 10^6$  operations. Typical runtime: 45 minutes on standard laptop.

## D Additional Statistical Results

### D.1 Return Distribution Analysis

Shapiro-Wilk test for normality:  $W = 0.971$ , p-value = 0.70. Cannot reject normality of fold returns.

Skewness: 0.21 (slight positive skew). Kurtosis: -1.07 (platykurtic, lighter tails than normal).

### D.2 Autocorrelation Analysis

Ljung-Box test for return autocorrelation:  $Q(5) = 3.42$ , p-value = 0.63. No significant autocorrelation detected, consistent with market efficiency at quarterly frequency.

### D.3 Power Analysis Details

Given observed effect size  $d = 0.17$ , required sample sizes for various power levels:

- 50% power:  $N = 173$  folds

- 70% power:  $N = 319$  folds
- 80% power:  $N = 540$  folds
- 90% power:  $N = 715$  folds

Current sample of 34 folds achieves approximately 12% power, highlighting fundamental statistical limitations with modest effect sizes.