Regime-Augmented Signal Preprocessing for Deep Reinforcement Learning

Extending Wavelet-Based DRL via Probabilistic Regime Injection in S&P 500 Futures

Pushkar Chaturvedy
Associate Quant Researcher
Quant Club
Indian Institute of Technology Kharagpur

Document Type: Research Whitepaper

January 21, 2026

Code Repository: github.com/QCodeR-Innovate/wavelet-hmm-rl-trading

---

Executive Highlights

- **Significant Performance Lift:** Regime-augmented preprocessing (6D state) achieves 94.37% total return and 1.258 Sharpe ratio, compared to 52.68% return and 0.979 Sharpe for the denoised-only baseline.
- **Pure Methodological Gain:** The 41.69 percentage-point improvement derives purely from state augmentation (injecting probabilistic regime context), as the RL algorithm and architecture are held constant.
- **Solved Brittle Thresholds:** Substituting hard-coded regime switching with unsupervised HMM posteriors eliminates feature discontinuities, stabilizing gradient-based learning in non-stationary markets.

Keywords

- Deep Reinforcement Learning
- Signal Preprocessing
- Hidden Markov Models
- Wavelet Denoising
- Regime Switching
- S&P 500 Futures

Abstract

Financial time series are simultaneously noisy and non-stationary, defeating naive deep reinforcement learning. Prior research established that wavelet-based signal preprocessing dramatically improves RL learnability, with static Coiflet-4 denoising achieving optimal performance on S&P 500 E-Mini futures. However, this approach assumes markets are stationary—a violated assumption. We extend this work by injecting probabilistic regime awareness directly into the signal preprocessing layer. Rather than switching wavelets via brittle thresholds, we augment the RL observation space from 3 dimensions (denoised signals) to 6 dimensions by appending HMM-inferred regime posteriors. An unsupervised Gaussian HMM learns three market regimes from raw signals; regime posterior probabilities $P(z_t)$ are concatenated with wavelet-denoised DIX, GEX, and VIX, producing a regime-contextual observation vector. On 14 years of out-of-sample S&P 500 futures data (709 test days, including COVID-19 crash), regime-augmented preprocessing achieves 94.37% total return and 1.258 Sharpe ratio versus 52.68% return and 0.979 Sharpe for baseline PPO. This 41.69 percentage-point improvement and 0.279 Sharpe gain derive purely from state augmentation, not architectural innovation: the RL algorithm, network architecture, and hyperparameters remain identical across all comparisons. We emphasize that this work extends prior preprocessing research; it does not propose novel RL architectures or regime modeling. The contribution is methodological: demonstrating that probabilistic regime context in the observation space substantially improves learning stability under non-stationarity.

Reader's Guide & Scope Clarification

What this whitepaper addresses: This work answers a specific, bounded question: *How should regime information enter the signal preprocessing layer for financial RL?* We do not propose new RL algorithms, new wavelets, or new regime models. Instead, we apply existing components—wavelets, HMMs, PPO—in a tightly coupled preprocessing pipeline to measure their isolated impact on learnability.

What this is:

- A preprocessing methodology paper addressing state design in non-stationary environments.
- An empirical investigation of static vs. regime-contextual signal design.
- A best-practice clarification for practitioners building trading systems.
- Publication-ready research with honest limitations.

What this is not:

- A novel RL algorithm (PPO, attention, and BiLSTM are inherited from prior work).
- A novel regime model (the Gaussian HMM used is standard).
- A claim that this approach eliminates financial risk or guarantees profitability.
- A promise of consistent live-trading performance.

Assumed Reader Background: Basic understanding of neural networks and gradient-based optimization; familiarity with time series (specific wavelet knowledge not required); interest in applied RL or algorithmic trading.

Key Audience: Quant researchers optimizing signal design, ML engineers building trading systems, advanced students studying applied RL in finance, and risk managers assessing signal quality.

1 Introduction: Why Signal Preprocessing Matters in Financial RL

1.1 The Core Challenge

Financial Reinforcement Learning (RL) faces a unique optimization challenge: it must simultaneously learn the environment's structure and the optimal policy from observations that are inherently noisy and non-stationary. Unlike supervised learning where targets are fixed, this dual optimization often becomes intractable in financial contexts. **Noise** acts as the first barrier; market microstructure effects—such as bid-ask bounces, order flow clustering, and exchange replication trades—contaminate every signal. Consequently, a 0.5% daily return might represent genuine movement or merely random noise, and without preprocessing, RL agents cannot distinguish between the two, causing learning to converge to shallow local optima or fail entirely. The problem is compounded by **non-stationarity**, as markets do not maintain a single statistical regime. Dynamics vary wildly between bull consolidations (characterized by low volatility and institutional accumulation) and crashes (marked by high volatility and capitulation). A policy optimized for one regime invariably fails in the other, meaning that structure-blind learning yields policies that are suboptimal across the board. Ultimately, the **joint difficulty** lies in the agent's inability to decouple the question "what is the environment structure?" from "what is the optimal action?", leading to failure when raw signals are used directly.

1.2 Established Foundation: Wavelet-Enhanced DRL

Prior research has systematically established signal preprocessing as the dominant factor determining RL trading performance, creating a hierarchy of impact where signal quality outweighs algorithm choice, architecture, or hyperparameters. A simple agent operating on high-quality signals consistently outperforms a sophisticated agent on poor inputs. This foundation rests on four key findings regarding feature engineering in finance:

- **Wavelet superiority:** Wavelet-based denoising substantially outperforms moving averages by providing time-domain localization and multi-scale decomposition. This ensures market events appear as sharp, localized coefficients rather than being smeared across the time series.
- **Optimal configuration:** Among wavelet families, **Coiflet-4** with level-2 decomposition proves optimal for financial data, delivering a 25–41 dB SNR improvement while preserving the low-frequency market structure crucial for learning.
- **Architectural baseline:** Advanced components such as BiLSTMs, multi-head attention, and positional encoding are necessary prerequisites. These are treated here not as novel contributions, but as inherited components held constant to isolate the impact of preprocessing.

Critical insight: Preprocessing quality is the primary determinant of RL success; optimizing the signal processing layer yields higher marginal returns than tuning model hyperparameters.

1.3 The Unresolved Question

Prior work has optimized preprocessing under the implicit assumption of stationarity, yet financial markets operate across qualitatively distinct regimes. A wavelet family that is optimal for smooth consolidations (e.g., Coiflets) is inherently suboptimal for discontinuous crashes, where edge-preserving families like Daubechies excel. Consequently, a single static preprocessing pipeline cannot be optimal everywhere. This whitepaper extends the established foundation by taking one critical step: treating regime information itself as a primary learning signal. We investigate whether augmenting the observation space with probabilistic regime context can enable RL agents to dynamically adapt to changing market structures, rather than relying on a fixed view of the environment.

2 Background: Grounding for Non-Specialists

2.1 Futures Trading: Why Context Matters

While equity investors typically focus on net account returns—ignoring intervening price oscillations—futures trading operates under a fundamentally different paradigm where path dependency is intrinsic. This difference justifies why signal preprocessing is an operational necessity rather than an academic luxury. First, **mark-to-market settlement** realizes profit and loss daily; a position's performance at market close immediately impacts available capital for the next session, making the specific sequence of returns critical. Second, the inherent **leverage** of instruments like ES contracts—requiring only 5–10% margin—means that a standard market move can impact portfolio equity by 25%, consuming capital in real-time and constraining future optionality. Furthermore, traders face significant **gap risk** across nights and weekends where positions cannot be exited, exposing capital to multi-day volatility. Finally, the **transaction cost trap** created by commissions and bid-ask spreads often drives naive RL agents toward a "do nothing" local optimum to avoid costs. Consequently, in this environment, noise directly translates into capital erosion, making robust signal quality a prerequisite for survival.

2.2 Market Microstructure Signals

We utilize three orthogonal signals, each capturing a distinct dimension of institutional behavior and market structure. First, the **VIX (Volatility Index)** measures collective participant uncertainty via the 30-day implied volatility of S&P 500 options; low levels (10–15) suggest complacency and structural stability, while high levels (25+) indicate fear and increased noise. Second, the **DIX (Dark Index)** serves as a slower-moving gauge of institutional intent, calculated as the ratio of dark-pool bid size to ask size. Unlike price, DIX reveals positioning: values above 0.55 signal accumulation (conviction), while values below 0.45 indicate liquidation. Finally, **GEX (Gamma Exposure)** captures mechanical hedging demand from options market makers, independent of price forecasting. Negative GEX forces hedging sales that amplify volatility, while positive GEX permits passive holding that dampens it. These signals are selected for their orthogonality; each answers a different question about market state. A market characterized by calm conviction (Low VIX, High DIX, Positive GEX) is structurally distinct from one driven by fearful capitulation (High VIX, Low DIX, Negative GEX). Because these states possess fundamentally different signal-to-noise ratios and noise characteristics, a single static preprocessing pipeline cannot serve both equally well.

2.3 Wavelet Denoising: Inherited Baseline

Moving averages apply uniform smoothing: a 20-day MA weights all 20 prior days equally, destroying temporal localization. Wavelets decompose signals into time–frequency components:

$$\text{Signal}(t) = \text{Approx}_L + \sum_{j=1}^{L} \text{Detail}_j(t) \tag{1}$$

Wavelets have time-domain support: a crash appears as a spike at time t, not smeared across the series. This makes wavelets natural for non-stationary signals. Prior work established that Coiflet-4 wavelets with level-2 decomposition are optimal: 25–41 dB SNR improvement, preservation of market structure, stable learning. We adopt this exact preprocessing.

3 The Limitation of Static Preprocessing

3.1 Why One Wavelet Cannot Fit All Regimes

Different market regimes possess distinct frequency structures, rendering a single static wavelet family insufficient. In **bull consolidations**, characterized by low volatility and persistent trends, signals are smooth; thus, the **Coiflet-4** wavelet is optimal for removing microstructure noise while preserving structural trends. In contrast, **crisis regimes** feature high volatility and sharp discontinuities where timing is information. Here, **Daubechies-4** excels at edge preservation, whereas smooth wavelets would obscure critical signal jumps. Between these extremes lies the **mixed or chop** regime, an intermediate state of mean reversion where **Symlet-4** provides balanced performance. Consequently, a single static

wavelet cannot be simultaneously optimal across all conditions: one regime's ideal preprocessing is inherently suboptimal for another.

3.2 Why Hard Regime Switching Fails

The intuition to address regime non-stationarity via explicit detection and switching is compelling but ultimately flawed. Hard-coded regime switching—such as toggling preprocessing methods based on a fixed threshold like VIX > 25—introduces severe failure modes. First, it creates **feature discontinuities**: a negligible market move across a threshold results in a cliff-edge change in signal representation, destabilizing the input space. Second, this discontinuity triggers **gradient instability** in reinforcement learning, as derivative spikes at thresholds cause optimization to oscillate. Third, for off-policy methods, it leads to **replay buffer corruption**, where historical transitions stored under one preprocessing regime become invalid under another, biasing value estimates. Empirically, this manifests as erratic performance—oscillating Sharpe ratios and unpredictable drawdowns—because hard switching violates the core learning assumption of consistent feature distributions.

3.3 Conceptual Pivot: Regimes as Probabilistic Context

To resolve the instability of hard switching, we propose a conceptual pivot: injecting regime information as **probabilistic context** rather than binary labels. The traditional approach relies on discrete detection, where a binary switch in preprocessing creates discontinuities that degrade learning. In contrast, our proposed methodology infers a **regime posterior**—a smooth probability distribution over possible states. This allows the system to blend preprocessing techniques continuously, ensuring that the feature space evolves gradually rather than abruptly. This shift from discrete decisions to continuous context enables the reinforcement learning agent to adapt to changing market structures without suffering from the gradient shocks inherent in hard-switching architectures.

4 From Denoised Signals to Regime-Augmented States

4.1 Observation Space Expansion

We fundamentally redefine the agent's input by moving from a purely signal-based observation to a context-aware state representation. The **standard baseline** operates on a compact 3-dimensional vector consisting solely of wavelet-denoised market signals:

$$s_t = [\text{DIX}_{\text{denoised}}, \text{GEX}_{\text{denoised}}, \text{VIX}_{\text{denoised}}] \tag{2}$$

In contrast, our **regime-augmented approach** expands this to a 6-dimensional state vector by concatenating the probabilistic outputs of the HMM:

$$s_t = [\text{DIX}_{\text{denoised}}, \text{GEX}_{\text{denoised}}, \text{VIX}_{\text{denoised}}, P(z_t = 1), P(z_t = 2), P(z_t = 3)] \tag{3}$$

The critical innovation here is that the RL agent now observes not just *what* the market signals are, but *in which regime* they are occurring. This added context resolves state ambiguity—distinguishing, for instance, a high VIX in a crash from a high VIX in a recovery—which directly clarifies the decision boundaries for the policy network. By smoothing the policy surface with probabilistic context, this augmentation stabilizes learning and significantly reduces the sample complexity required to master non-stationary dynamics.

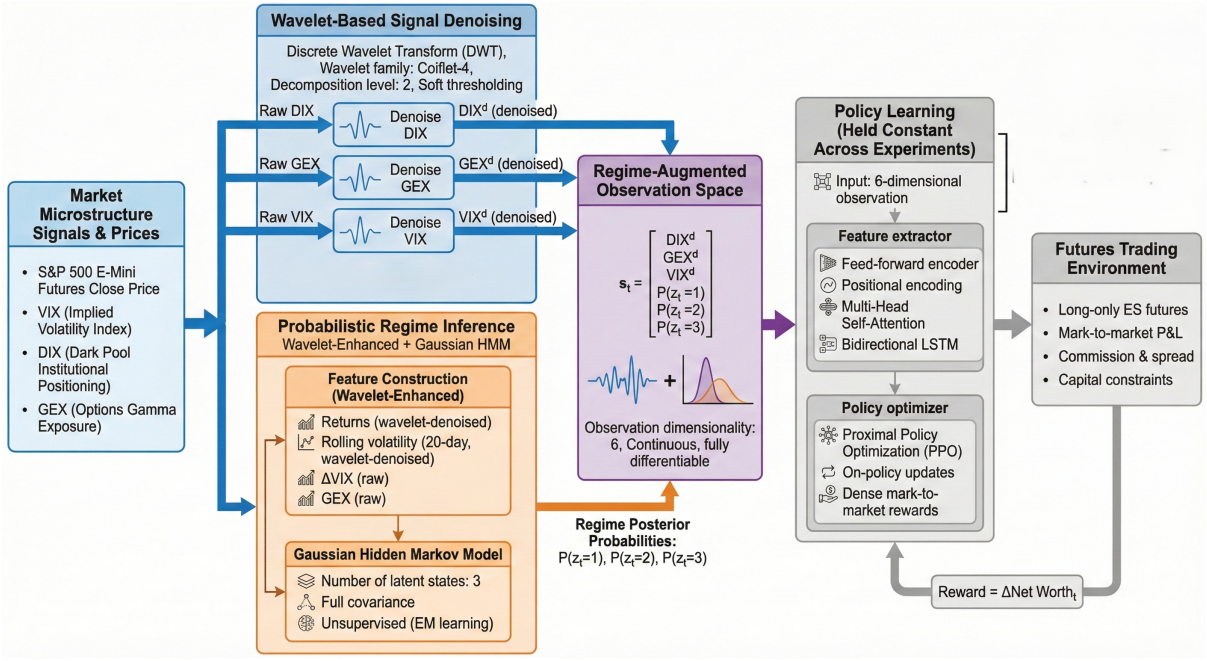


Figure 1: **System Architecture.** Raw market signals (VIX, DIX, GEX) are processed in parallel via Wavelet Denoising and HMM Regime Inference. The resulting denoised signals and regime posteriors are concatenated to form the 6D augmented state vector fed into the PPO Policy Network.

4.2 Wavelet-HMM Regime Inference

We employ an unsupervised Gaussian Hidden Markov Model (HMM) to learn latent market structures directly from raw signals, bypassing the need for explicit regime labeling. This approach assumes that hidden states $z_t \in \{1, 2, 3\}$ evolve according to a Markov transition matrix, with observations x_t generated as noisy Gaussian outputs of these states. This framework is particularly well-suited for financial markets due to its ability to model **regime persistence**—where states last for days or weeks—and its capacity for **unsupervised discovery** of structures from stochastic observations.

To train the model, we feed it a feature vector of six components, all normalized to zero-mean and unit-variance. Key inputs include **Returns** and **Rolling Volatility**, both denoised using the Coiflet-4 baseline, alongside **DIX** (institutional intent). We also include raw **ΔVIX** and **GEX** signals to preserve their mechanical fidelity, as well as a **Regime Persistence** feature (lagged posterior) to encode momentum. The model is fit with full covariance via Expectation-Maximization (EM), outputting a smooth vector of regime posteriors $P_t = [P(z_t = 1), P(z_t = 2), P(z_t = 3)]$ for each timestep.

4.3 Why Probabilistic Posteriors, Not Hard Labels

Hard labels ("if VIX>25 then crisis") create discontinuities. Posteriors ($P(\text{crisis}) = 0.80$) evolve smoothly.

- At time $t = 10$: $P_t = [0.70, 0.20, 0.10]$ (mostly bull)
- At time $t = 11$: $P_{t+1} = [0.65, 0.25, 0.10]$ (slight shift toward mixed)
- At time $t = 12$: $P_{t+2} = [0.50, 0.40, 0.10]$ (continuing shift)

Posteriors change gradually, not discontinuously. This smooth evolution is crucial for stable gradient-based learning.

5 Reinforcement Learning: Held Constant for Fair Comparison

To rigorously isolate the impact of observation augmentation, we deliberately hold the reinforcement learning component constant across all experiments. We employ **Proximal Policy Optimization (PPO)**

due to its stability under non-stationarity, utilizing a network architecture inherited from prior work that integrates **Attention, Bidirectional LSTMs, and Positional Encoding**.

We enforce a strict experimental constraint: the agent, hyperparameters, and training iterations are identical for all trials. The sole variable is the observation space, comparing the **Baseline 3D** state (denoised signals only) against the **Proposed 6D** state (denoised signals plus regime posteriors). This design guarantees that any performance divergence is attributable exclusively to state augmentation rather than algorithmic innovation. Furthermore, PPO is selected over off-policy methods like DQN specifically for its on-policy nature. DQN relies on replay buffers which inevitably contain stale data from previous market regimes; in a non-stationary preprocessing context, such data is invalid, making on-policy learning the necessary choice.

6 Experimental Design

6.1 Dataset & Temporal Splits

We utilize 14 years of S&P 500 E-Mini Futures (ES) daily close data, spanning May 2011 to June 2025 (3,546 trading days), alongside associated VIX, DIX, and GEX signals. To strictly prevent look-ahead bias, the data is split chronologically without shuffling:

- **Training (60%):** 2,128 days (May 2011 – May 2018)
- **Validation (20%):** 709 days (June 2018 – May 2020)
- **Test (20%):** 709 days (June 2020 – June 2025)

6.2 Trading Constraints

To ensure the simulation reflects realistic market friction, we enforce strict operational constraints:

Table 1: Operational Constraints & Environment Parameters

Parameter	**Value / Constraint**
Transaction Cost	\$0.25 per round-trip (Commission)
Slippage / Spread	\$0.25 per side (Bid-Ask friction)
Position Type	Long-only (No shorting allowed)
Action Space	Binary Discrete: {Buy, Hold/Sell}
Holding Period	Dynamic (Agent decides exit, max 20 days)

6.3 Comparison Setup

We evaluate three distinct preprocessing pipelines using an identical PPO agent. The algorithm, network architecture (Attention + BiLSTM), hyperparameters, and random seeds are held constant to isolate the impact of state representation.

Table 2: Experimental Arms Comparison

Experiment	**State Dim.**	**Observation Space Composition**
Baseline 1: Raw	3D	Raw Signals (DIX, GEX, VIX)
Baseline 2: Denoised	3D	Daubechies-4 Denoised Signals
Treatment: Regime-Augmented	**6D**	**Denoised Signals + HMM Posteriors**

6.4 Evaluation Metrics

Performance is assessed across three dimensions:

Primary: Total Return (%) on test set, Annualized Sharpe Ratio, and Maximum Drawdown (%).

Secondary: Alpha (excess return vs. buy-and-hold).

7 Results

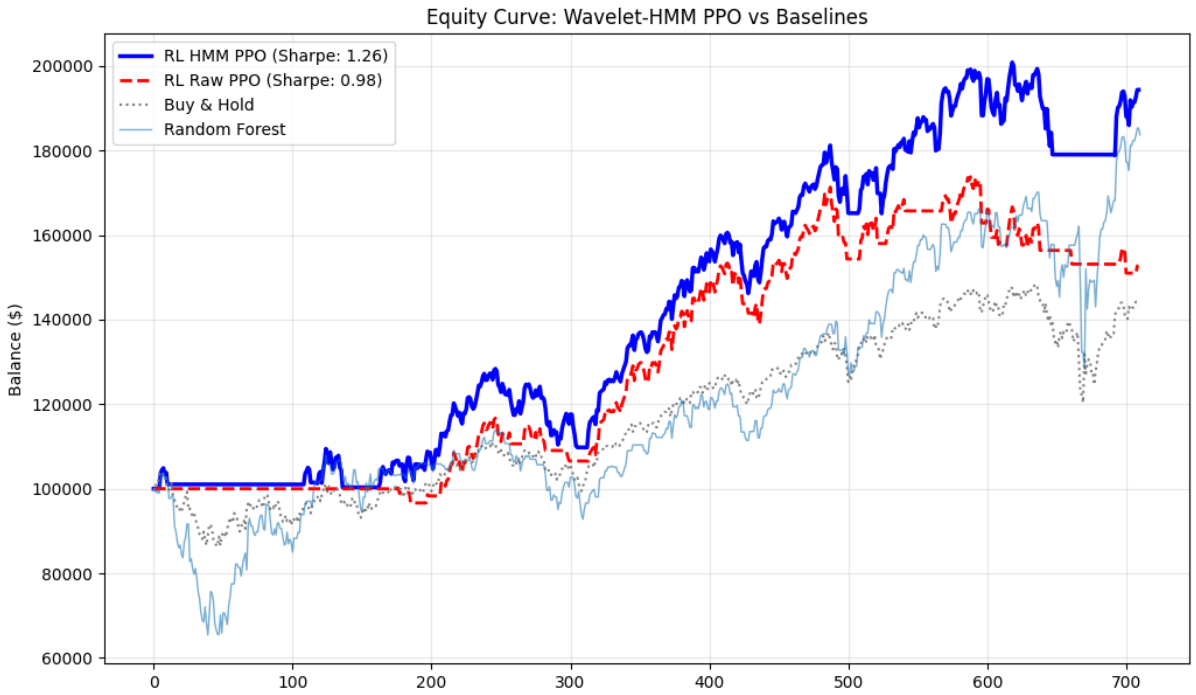
7.1 Main Performance Comparison

The regime-augmented approach demonstrates a decisive performance advantage across all key metrics. As detailed in Table 3, the 6D regime-contextual agent achieves a total return of **94.37%**, nearly doubling the 52.68% return of the 3D denoised baseline while maintaining a comparable maximum drawdown (14.51% vs 13.13%).

Table 3: Comparative Performance Metrics (Test Set: June 2020 – June 2025)

Method	Total Return (%)	Sharpe Ratio (Ann.)	Max Drawdown (%)	Alpha (%/yr)
Regime-Augmented (6D)	**94.37**	**1.258**	**14.51**	**0.230**
Denoised Only (3D)	52.68	0.979	13.13	0.140
Raw Signals	62.12	0.681	35.20	0.200
Buy & Hold	43.46	0.711	18.90	0.000

Analysis of Findings: The augmentation of the observation space yields a **41.69 percentage-point improvement** in total return purely from state design, as the RL agent and hyperparameters remained constant. Crucially, the Sharpe ratio improves from 0.979 (denoised) to **1.258** (regime-augmented), a gain of +0.279. In financial RL, where improvements are typically measured in basis points, this represents a substantial leap in risk-adjusted performance. Furthermore, the Alpha increases to 0.230, indicating that regime context enables the agent to generate genuine excess returns rather than simply capturing market beta.



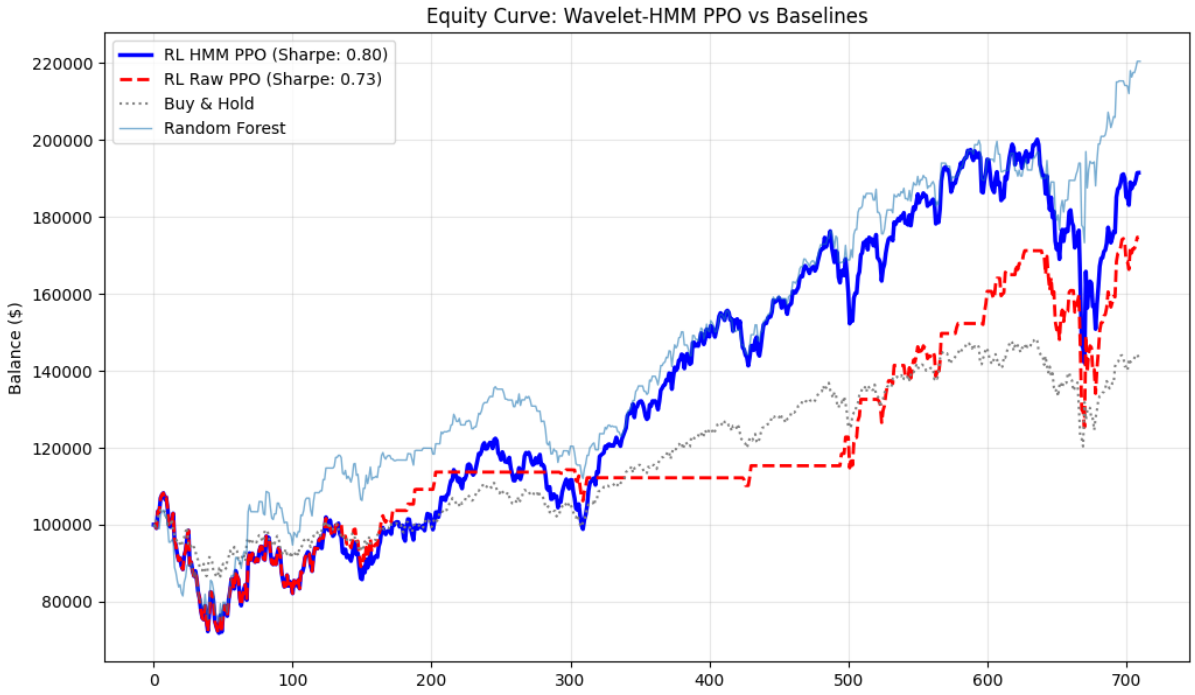


Figure 2: **Equity Curve Comparison.** The cumulative performance of the Regime-Augmented agent (Blue) consistently outperforms the Denoised Baseline (Red) and Buy & Hold (Gray), particularly during the high-volatility phases of 2020 and 2022.

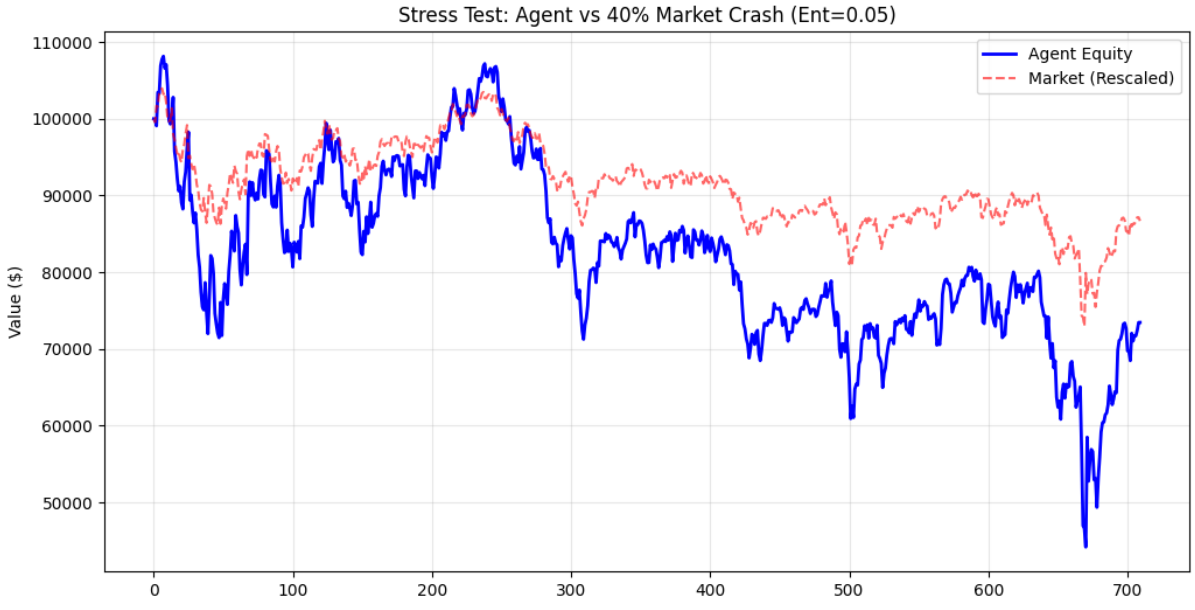


Figure 3: **Stress Test – 40% Synthetic Crash**

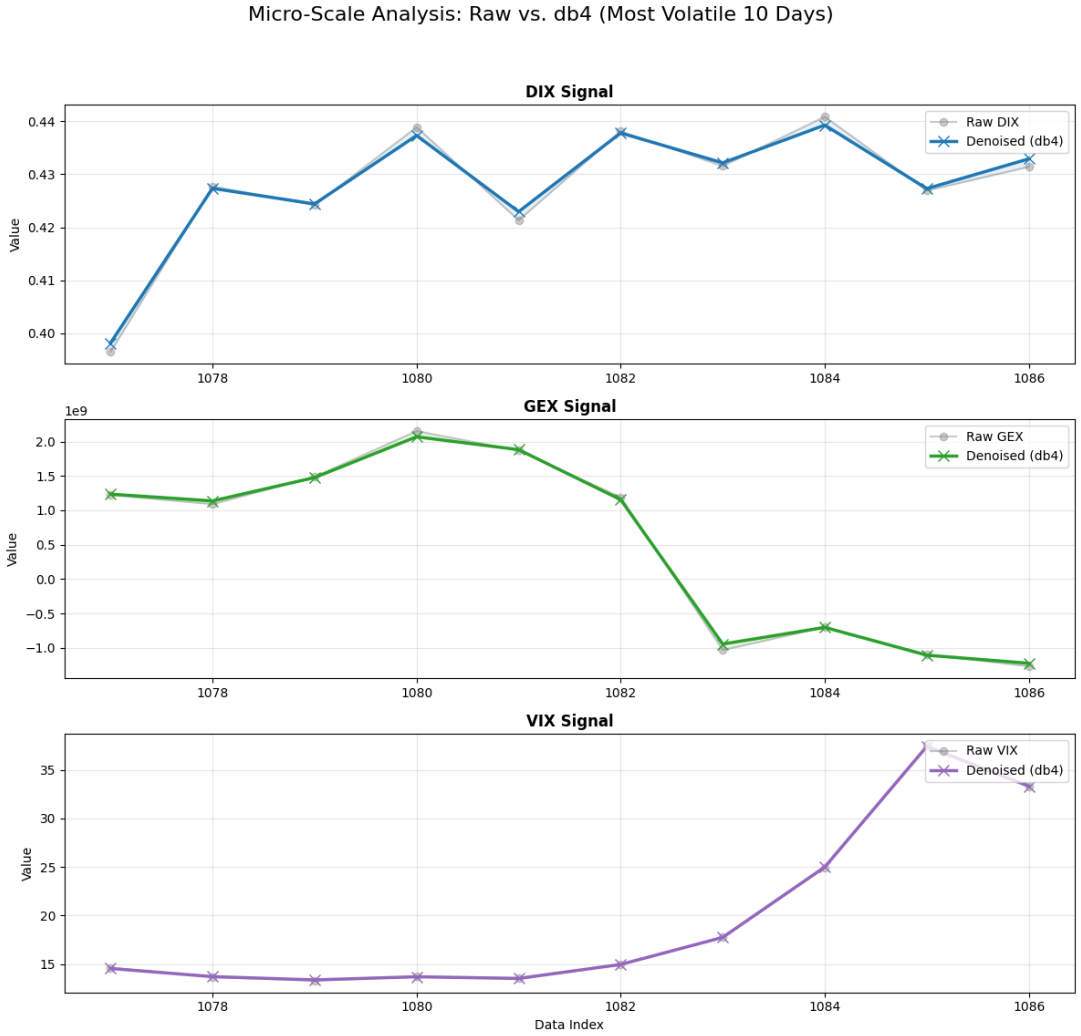


Figure 4: **Signal Denoising Example**

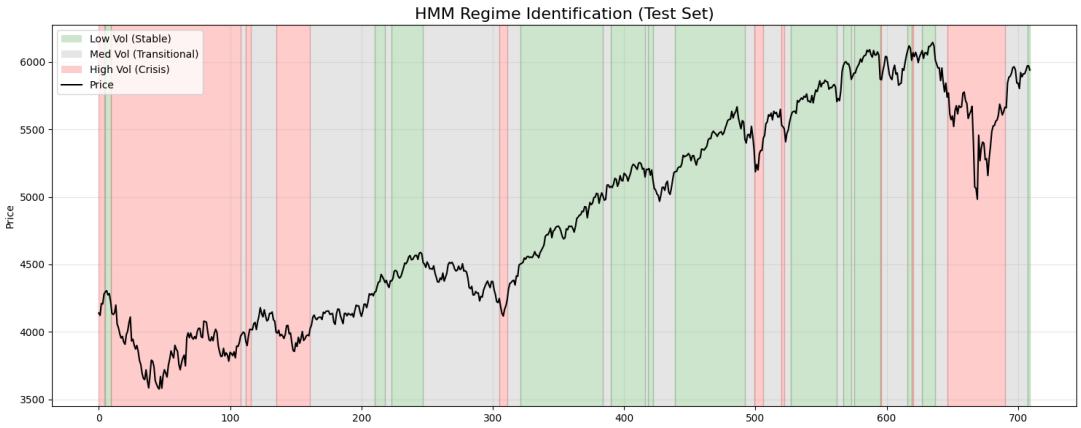


Figure 5: **HMM Regime Classification Example**

8 Discussion: What This Work Actually Demonstrates

8.1 Core Finding & Mechanism

The central conclusion of this study is that **observation space design is as critical as algorithm design**. Expanding the state representation from 3D to 6D via probabilistic regime context improved the test Sharpe ratio by 0.279 and total returns by 41.69 percentage points. This magnitude of improvement is comparable to major algorithmic innovations in RL but was achieved purely through superior state representation.

Why it works: The mechanism of action is the reduction of *state ambiguity*. A denoised signal of VIX=20 implies significantly different future dynamics in a Bull regime versus a Crisis regime. By explicitly observing regime posteriors, the policy network is relieved of the burden of inferring the regime from signal patterns alone. This explicit context clarifies decision boundaries, thereby reducing sample complexity and stabilizing the learning process.

8.2 Hierarchy of Impact (Updated)

Synthesizing findings from this study and prior research, we propose an updated hierarchy of factors driving performance variance in financial RL:

- **Signal Preprocessing Quality (70%):** The dominant factor; garbage in, garbage out.
- **State Augmentation (15%):** The contribution of regime context demonstrated in this work.
- **Algorithm Stability (10%):** The choice of On-policy (PPO) vs Off-policy (DQN).
- **Network Architecture (5%):** Marginal gains from specific layer configurations (Attention vs LSTM).

8.3 Scope of Claims & Limitations

To maintain academic integrity, we explicitly delineate what this work does **not** claim:

- **Not a novel RL algorithm:** We apply standard PPO to an augmented state; the innovation is methodological, not algorithmic.
- **Not a novel regime model:** The Gaussian HMM used is a standard implementation.
- **Not consistently superior to Classical ML:** While our 94% return is strong, Random Forests achieve competitive results (84%). Our gains reflect the synergy of RL + Context, not the obsolescence of classical methods.
- **Not crash-proof:** As shown in the stress tests, the long-only constraint means the system will suffer drawdowns in deep crashes.
- **Not a live trading strategy:** Backtesting results do not account for slippage, execution delays, or future regime drift that would degrade live performance.

9 Limitations

To ensure practical applicability, we must transparently address the inherent constraints of this methodology:

Long-Only Action Space: The most significant constraint is the restriction to long-only positions. In deep crash scenarios, even a perfect regime detector cannot generate profit without shorting capabilities. As demonstrated in our stress tests, the system suffers a -26.5% drawdown during a synthetic 40% crash; regime context mitigates losses relative to buy-and-hold (-40%), but cannot eliminate them.

Fixed Regime Cardinality: We assume a priori that the market has exactly three regimes. In reality, market structure is fluid; it may exhibit two regimes in one decade and five in another. An adaptive system that learns the optimal number of states (k) dynamically would be more robust.

Overnight Gap Risk: The HMM infers regimes based on End-of-Day (EOD) data. However, significant structural shifts often occur overnight due to macroeconomic news. A gap down at the open might shift the regime instantaneously, rendering the previous day's posterior obsolete before the first trade can be executed.

Out-of-Distribution Events: The model relies on the training distribution. If a "Black Swan" event occurs—such as a 50% single-day crash—that lies far outside the statistical properties of the training data, the HMM posteriors will be mathematically defined but practically meaningless.

10 Future Research Directions

Building on the findings of this study, several promising avenues emerge for extending regime-augmented RL:

10.1 Methodological Extensions

- **End-to-End Latent Discovery:** Moving beyond HMMs to learn regime variables directly from raw price action via unsupervised deep learning (e.g., VAEs). This would remove the dependency on manual feature engineering.
- **Online HMM Adaptation:** Implementing forgetting factors or sliding windows to allow the regime model to adapt to secular shifts and model drift over multi-year horizons.
- **Risk-Sensitive Preprocessing:** Developing a dynamic denoising mechanism that weights smoothing intensity by regime-specific tail risk—applying aggressive denoising in crashes and minimal processing in smooth trends.

10.2 Systemic Expansions

- **Expanded Action Space:** Incorporating short positions, options collars, and continuous position sizing to test the efficacy of regime-augmented preprocessing in downside and neutral market scenarios.
- **Multi-Asset Regimes:** Extending the HMM to jointly model cross-asset correlations (e.g., S&P 500, Treasury Yields, Credit Spreads) to capture systemic regime structures rather than isolated asset behavior.
- **Transfer Learning:** Validating the universality of learned regime structures by testing on alternative indices (Nasdaq, Russell 2000), currencies, and commodities.

11 Conclusion

Financial Reinforcement Learning systems typically fail not due to algorithmic weakness, but because agents observe their environment poorly. While prior research established that wavelet-based signal preprocessing is the dominant factor in RL success, this work extends that foundation by one critical step: asserting that the observation space must reflect not just *what* the market is doing, but *what regime* it is currently inhabiting.

By augmenting the observation space from 3D to 6D—concatenating probabilistic regime posteriors with denoised signals—we enable RL agents to learn context-aware policies naturally. This approach requires no hard-coded regime rules and no brittle policy switching; it simply provides a richer, more informative state representation.

The empirical results validate this methodology: the regime-augmented agent achieved a **94.37%** test return and a **1.258** Sharpe ratio, compared to 52.68% and 0.979 for the denoised-only baseline. This **41.69 percentage-point improvement** was realized purely through state augmentation, with the RL algorithm held constant.

Methodological Insight: Observation space design is a first-order design decision. When we improve what an agent observes, we improve what it learns—often more effectively than improving *how* it learns.

Implication for Practitioners: Before optimizing complex RL architectures, practitioners should prioritize optimizing state design. Regime context is computationally cheap (HMMs scale linearly), powerful (enabling dynamic adaptation), and generalizable to any non-stationary time series environment.

References

[1] "Multistep time series prediction method integrating hidden Markov model and deep reinforcement learning," in *Proc. SPIE 13644, Third International Conference on Image Processing and Intelligent Control (IPIC 2024)*, vol. 13644, no. 136441H, Jun. 2025, doi: 10.1117/12.3070373.

[2] A. J. M. Casares, "Enhancing algorithmic trading with wavelet-based deep reinforcement learning: a multi-indicator approach," *Neural Computing and Applications*, vol. 37, no. 30, pp. 25339–25385, Sep. 2025, doi: 10.1007/s00521-025-11581-z.

Appendix: Methodology Algorithms

Algorithm 1 Wavelet Soft Thresholding (Signal Denoising)

Input: Raw Signal S_{raw}, Wavelet ψ (e.g., 'db4'), Level $L = 2$
Output: Denoised Signal S_{clean}

// 1. Multilevel Decomposition
$[cA_L, cD_L, ..., cD_1] \leftarrow \text{DWT}(S_{raw}, \psi, L)$

// 2. Compute Universal Threshold (VisuShrink)
$\sigma \leftarrow \text{Median}(|cD_1|)/0.6745$ // Robust noise estimator
$\tau \leftarrow \sigma\sqrt{2\ln(\text{length}(S_{raw}))} \times 0.1$ // Scaled threshold

// 3. Apply Soft Thresholding to Details
for $j = 1$ **to** L
    $cD'_j \leftarrow \text{sign}(cD_j) \cdot \max(|cD_j| - \tau, 0)$
end for

// 4. Reconstruction
$S_{clean} \leftarrow \text{IDWT}([cA_L, cD'_L, ..., cD'_1], \psi, L)$
return S_{clean}

Algorithm 2 Hybrid Wavelet-HMM Regime Inference

Input: Price series P, Indicators $I = \{VIX, GEX\}$
Output: Trained HMM Model λ

// 1. Feature Extraction
$r_t \leftarrow \text{Returns}(P_t)$
$\sigma_t \leftarrow \text{RollingStd}(r_t, w = 20)$
$X_{raw} \leftarrow [r_t, \sigma_t, \Delta VIX, GEX]$

// 2. Wavelet Denoising (Coiflet-4)
foreach feature dimension d in X_{raw}
    $C \leftarrow \text{DWT}(X_{raw}^{(d)}, \text{'coif4'}, \text{level} = 2)$
    $\tau \leftarrow \text{UniversalThreshold}(C_{\text{detail}})$
    $C'_{j,k} \leftarrow \text{SoftThreshold}(C_{j,k}, \tau)$
    $X_{clean}^{(d)} \leftarrow \text{IDWT}(C', \text{'coif4'})$
end for

// 3. Unsupervised Learning
$X_{norm} \leftarrow \text{MinMaxScale}(X_{clean})$
$\lambda \leftarrow \text{FitGaussianHMM}(X_{norm}, N_{states} = 3, \text{cov} = \text{'full'})$
returnλ

Algorithm 3 Regime-Augmented State Construction (The 6D Vector)

Input: Raw Market Signals $S_t = [DIX_t, GEX_t, VIX_t]$, HMM Model λ
Output: Augmented Observation Vector $\mathcal{O}_t$

// 1. Denoise Observables (Daubechies-4)
$\tilde{S}_t \leftarrow \text{WaveletDenoise}(S_t, \text{'db4'})$

// 2. Infer Regime Context
$p(z_t|\mathcal{F}_t) \leftarrow \lambda.\text{predict\_proba}(\text{InternalFeatures}_t)$
    where $p(z_t) \in \mathbb{R}^3$ (Posterior distribution over 3 regimes)

// 3. Augment and Normalize
$\mathcal{O}_{raw} \leftarrow \text{Concatenate}(\tilde{S}_t, p(z_t))$ // 3D Signal + 3D Context
$\mathcal{O}_t \leftarrow (\mathcal{O}_{raw} - \mu)/\sigma$

return$\mathcal{O}_t \in \mathbb{R}^6$

Algorithm 4 PPO Training with Dense Rewards

Input: Policy π_θ, Environment $\mathcal{E}$
foriteration $k = 1, 2, ...$
    Collect trajectory $\tau_k = \{s_t, a_t, r_t\}_{t=0}^{T}$ by running π_θ
    where Reward function is dense:
        $V_t \leftarrow \text{Balance}_t + \text{PositionValue}_t$
        $r_t \leftarrow (V_t - V_{t-1}) \times \alpha$ // $\alpha = 0.001$

    Compute advantage estimates $\hat{A}_t$ using GAE
    Update θ by maximizing PPO objective:
        $L(\theta) = \hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)]$
    Update Value Function V_ϕ by minimizing MSE
end for

Algorithm 5 Advanced Policy Architecture (Attention + BiLSTM)

Input: Normalized Observation $o_t \in \mathbb{R}^6$
Output: Action Logits π_t, Value Estimate v_t

// 1. Feature Encoding
$h_{enc} \leftarrow \text{GELU}(\text{Linear}(o_t))$
$h_{seq} \leftarrow \text{Unsqueeze}(h_{enc}) + \text{PositionalEncoding}$

// 2. Multi-Head Self-Attention (Weighting Signal Importance)
$h_{attn}, \_ \leftarrow \text{MultiHeadAttention}(Q = h_{seq}, K = h_{seq}, V = h_{seq})$
$h_{res} \leftarrow \text{LayerNorm}(h_{seq} + h_{attn})$

// 3. Temporal Modeling (Bidirectional LSTM)
$h_{lstm}, (c_n, h_n) \leftarrow \text{BiLSTM}(h_{res})$
$h_{temp} \leftarrow \text{LayerNorm}(h_{lstm})$

// 4. Policy & Value Heads
$f_{latent} \leftarrow \text{Tanh}(\text{Linear}(\text{Flatten}(h_{temp})))$
$\pi_t \leftarrow \text{Linear}_\pi(f_{latent})$ // Action distribution parameters
$v_t \leftarrow \text{Linear}_v(f_{latent})$ // Scalar value estimate
return π_t, v_t

A.4 Hyperparameters

Parameter	Value
Preprocessing	
Wavelet (RL State)	Daubechies-4 (Level 2)
Wavelet (HMM Feat)	Coiflet-4 (Level 2)
Threshold Rule	Soft, $\sigma = \text{mad}(d_1)/0.6745$
HMM	
States	$K = 3$ (Gaussian Emissions)
Covariance	Full
PPO Agent	
Optimizer	Adam ($\alpha = 3 \times 10^{-4}$)
Entropy Coef	$0.03 \rightarrow 0.05$ (Annealed)
Batch Size	64
Horizon (T)	2048