Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Asset-Universe Design and Action-Space Granularity in Cointegration-Based Deep Reinforcement Learning for Cryptocurrency Statistical Arbitrage

Veliota Drakopoulou

vdrakopoulou@yahoo.com

Higher Colleges of Technology

Research Article

Keywords: Cryptocurrency statistical arbitrage, Deep reinforcement learning, Cointegration, Asset-universe design, Action-space granularity, Pairs trading, Robustness analysis

Posted Date: July 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9993426/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Asset-Universe Design and Action-Space Granularity in Cointegration-Based Deep Reinforcement Learning for Cryptocurrency Statistical Arbitrage

Veliota Drakopoulou
Higher Colleges of Technology, United Arab Emirates
Embry-Riddle Aeronautical University, United States
ORCID: 0000-0002-1670-8033
Email: vdrakopoulou@yahoo.com
vdrakopoulou@gmail.com
vdrakopoulou@hct.ac.ae
drakopov@erau.edu

Abstract

This study examines whether cointegration-based deep reinforcement learning (DRL) cryptocurrency statistical arbitrage is shaped by two design choices that are often treated as secondary: asset-universe construction and action-space granularity. Using 30-minute USDT-denominated Binance spot data, the study evaluates four cryptocurrency universes: a Baseline 14-Asset Universe, a Broad 20-Asset Universe, a Payment-Coin 6-Asset Universe, and a Layer-1 Platform 9-Asset Universe. Cointegration is used to construct synthetic arbitrage portfolios, which are then traded using Deep Q-Network (DQN), Proximal Policy Optimization (PPO), and Advantage Actor–Critic (A2C) agents. Each agent is tested under binary and granular action-space specifications and compared with a non-DRL COIN benchmark. The results show that asset-universe design materially affects out-of-sample performance. The strongest seed-42 returns occur in the Payment-Coin 6 and Baseline 14 universes, where DQN with binary actions returns 32.26% and 30.24%, respectively. In contrast, the Broad 20 universe is led by the COIN benchmark, while the Layer-1 universe remains negative across all strategies. Granular actions do not universally improve raw returns, but they improve risk control in several weaker settings by allowing partial or flat exposure. Robustness tests across seeds 42, 101, and 202 further show that some high-performing results are sensitive to stochastic training variation. Overall, the findings indicate that DRL-based cryptocurrency statistical arbitrage depends not only on algorithm choice, but also on the structure of the tradable asset universe, the agent's exposure choices, and the stability of training outcomes.

Keywords

Cryptocurrency statistical arbitrage; Deep reinforcement learning; Cointegration; Asset-universe design; Action-space granularity; Pairs trading; Robustness analysis

---

1. Introduction

Cryptocurrency markets have become an important testing ground for statistical-arbitrage and machine-learning trading strategies because they combine continuous trading, high volatility, strong cross-asset co-movement, and frequent short-horizon price dislocations. These features create opportunities for relative-value strategies, but they also make trading-system design difficult. In particular, a strategy that performs well in one group of assets may fail in another if the underlying price relationships are unstable, noisy, or weakly mean-reverting.

Cointegration-based statistical arbitrage offers a structured way to trade relative mispricing. Instead of forecasting the direction of a single cryptocurrency, the strategy constructs a spread or synthetic portfolio from assets that share a stable long-run relationship. Deep reinforcement learning (DRL) can then be used as a dynamic trading layer that decides when and how to trade deviations from that relationship. However, the performance of such a framework depends not only on the DRL algorithm, but also on the environment in which the agent learns. In a cointegration-based setting, that environment is shaped by the asset universe used to construct the arbitrage relationship and by the action space that determines the agent's exposure choices.

This paper argues that two design choices are especially important but often underexamined: asset-universe construction and action-space granularity. Asset-universe construction determines which cryptocurrencies enter the cointegration search and therefore affects the stability and tradability of the synthetic arbitrage portfolios. A broader universe may create more potential relationships, but it may also introduce weaker or noisier combinations. A narrower and more economically coherent universe may produce cleaner relationships, but with fewer trading opportunities. Similarly, action-space design determines whether the agent must always take full long-short exposure or whether it can reduce, scale, or neutralize exposure when signals are weak or costly to trade.

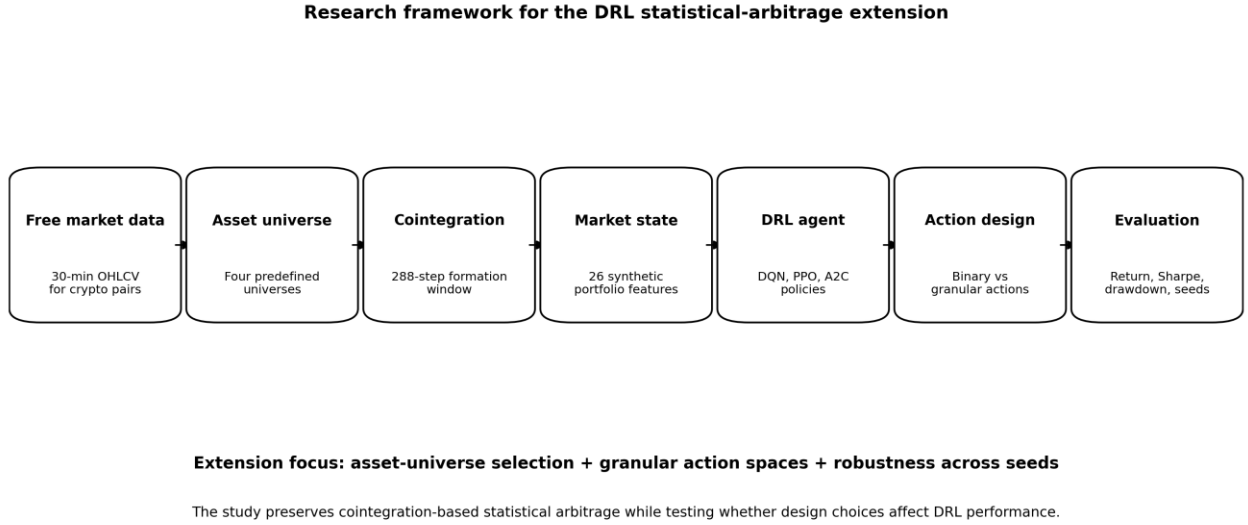
The objective of this study is to examine whether cointegration-based DRL cryptocurrency statistical arbitrage is materially affected by these two design choices. The empirical framework uses 30-minute USDT-denominated Binance spot data and evaluates four predefined cryptocurrency universes: a Baseline 14-Asset Universe, a Broad 20-Asset Universe, a Payment-Coin 6-Asset Universe, and a Layer-1 Platform 9-Asset Universe. The training period runs from 2023-01-01 to 2024-12-10, followed by a short separation window, and the out-of-sample testing period runs from 2024-12-18 to 2025-05-17. The framework constructs cointegration-based synthetic portfolios and evaluates DQN, PPO, and A2C agents against a non-DRL COIN benchmark under both binary and granular action-space specifications.

The paper makes three contributions. First, it treats asset-universe selection as a core empirical design variable rather than as a fixed preprocessing step. This allows the analysis to test whether broader asset coverage or more economically coherent groupings produce more tradable cointegration relationships. Second, it compares binary and granular action spaces within the same cointegration-based DRL framework. The binary specification restricts the agent to full directional exposure, while the granular specification allows full, half, or flat exposure. Third, it reports seed-level robustness for selected key models, reducing reliance on a single stochastic DRL training run.

Figure 1 summarizes the research framework used in this study. The framework begins with public 30-minute cryptocurrency OHLCV data, defines four asset universes, constructs cointegration-based synthetic portfolios, generates market-state features, trains DRL agents, compares binary and granular action-space designs, and evaluates performance using return, risk-adjusted performance, drawdown, turnover, and seed-level robustness.

Figure 1

Research Framework for Cointegration-Based DRL Statistical Arbitrage



Note. *The figure summarizes the empirical workflow used in the study: public 30-minute cryptocurrency OHLCV data, predefined asset-universe selection, cointegration-based portfolio construction, market-state generation, DRL policy learning, action-space design, and out-of-sample evaluation. DRL = deep reinforcement learning; OHLCV = open, high, low, close, and volume; DQN = Deep Q-Network; PPO = Proximal Policy Optimization; A2C = Advantage Actor–Critic.*

The empirical results show that asset-universe design materially affects performance. The strongest seed-42 raw return occurs in the Payment-Coin 6-Asset Universe, where DQN with binary actions returns 32.26%. The Baseline 14-Asset Universe also performs strongly, with DQN binary returning 30.24%. By contrast, the Broad 20-Asset Universe is led by the COIN benchmark rather than by any DRL model, and the Layer-1 Platform 9-Asset Universe remains negative even for its best-performing strategy. These findings indicate that expanding the asset set does not automatically improve arbitrage performance; the coherence and stability of the selected universe matter more than size alone.

The action-space results are also conditional. Binary actions produce the strongest raw returns in the most favorable DQN settings, suggesting that full exposure can be useful when the arbitrage signal is strong. However, granular actions improve risk control in several weaker environments by allowing the agent to scale down or exit positions. The robustness analysis further qualifies the main results: Payment-Coin DQN binary remains positive on average across seeds but has large return dispersion, while Baseline DQN binary is not stable across robustness seeds. These findings show that DRL arbitrage performance should not be judged from a single successful training run.

Overall, this paper shifts the evaluation question from whether one DRL algorithm is universally superior to when a DRL trading layer can add value to a cointegration-based cryptocurrency arbitrage framework. The results suggest that performance emerges from the interaction between cointegration structure, asset-universe design, action-space specification, and stochastic training stability. This has practical implications for researchers and trading-system developers: algorithm choice matters, but the design of the trading environment may matter just as much.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 describes the data and asset-universe design. Section 4 presents the methodology and experimental framework. Section 5 reports and discusses the empirical results. Section 6 concludes with implications, limitations, and directions for future research.

2. Literature Review

Statistical arbitrage exploits temporary deviations among related assets rather than forecasting the direction of the overall market. In pairs-trading and spread-trading settings, profitability depends on identifying stable relative-price relationships and designing trading rules that can exploit short-term deviations from those relationships. Foundational work on cointegration and pairs trading shows that non-stationary assets can form tradable stationary combinations when their long-run relationship is stable (Engle & Granger, 1987; Johansen, 1988; Gatev et al., 2006). Recent research extends this view by treating statistical arbitrage as a multi-stage pipeline involving portfolio construction, signal extraction, and trading-policy formation. Guijarro-Ordonez, Pelger, and Zanotti (2025), for example, develop a deep-learning statistical-arbitrage framework that constructs arbitrage portfolios, extracts time-series signals, and forms trading policies to maximize risk-adjusted returns. This pipeline perspective is relevant because the performance of an arbitrage strategy depends not only on the trading rule, but also on how the tradable relationship is constructed.

Cointegration remains especially relevant for cryptocurrency statistical arbitrage because crypto assets often share broad market movements while also displaying heterogeneous idiosyncratic behavior. A cointegration-based approach provides a structured way to construct synthetic spreads or portfolios from assets that may be individually non-stationary but stable in combination. Vergara and Kristjanpoller (2024) apply this logic to cryptocurrency markets by constructing cointegration-based synthetic arbitrage portfolios and training DQN, PPO, and A2C agents to trade those portfolios over fixed short-horizon scenarios. Their framework is directly relevant to the present study because it combines classical econometric portfolio construction with DRL-based trading decisions.

However, the asset universe used to estimate cointegration relationships is rarely treated as a central experimental variable. This is a limitation because the selected universe determines the set of possible cointegration vectors, the stability of the resulting spread, and the learning environment faced by the DRL agent. A broader asset universe may increase the number of potential arbitrage relationships, but it may also introduce weaker, noisier, or less stable combinations. Conversely, a smaller and more economically coherent universe may produce cleaner relationships but fewer trading opportunities. This suggests that asset-universe construction should be evaluated as part of the trading-system design rather than treated as a neutral preprocessing choice.

Deep reinforcement learning has become prominent in financial trading because it models trading as a sequential decision problem in which an agent observes a market state, selects an action, receives a reward, and updates its policy over time. Zhang, Zohren, and Roberts (2020) show that action specification and reward design can materially affect DRL trading behavior by comparing discrete and continuous action spaces and incorporating volatility-scaled rewards. Wu et al. (2020) emphasize the role of state representation in adaptive trading strategies, while Liu et al. (2022) introduce FinRL as an open-source framework for DRL trading environments, agents, transaction costs, and backtesting. Yang, Liu, Zhong, and Walid (2020) further demonstrate the use of ensemble DRL methods in stock trading by combining multiple actor–critic approaches.

Cryptocurrency-specific DRL studies also support the use of adaptive sequential trading policies. Liu et al. (2021) propose a PPO-based high-frequency Bitcoin trading framework, Chen and Huang (2021) incorporate price and sentiment information in a multimodal DRL trading system, and Lin et al. (2022) develop a multi-agent DRL framework for portfolio risk shifting. These studies show that DRL performance depends on the interaction between market-state design, reward specification, and policy structure. Yet much of this literature focuses on directional trading, portfolio allocation, or single-market decision-making rather than cointegration-based statistical arbitrage. In a cointegration-based setting, the agent does not simply decide whether to buy or sell one asset; it trades the relative movement between constructed synthetic portfolios. This makes the design of the spread, the asset universe, and the action space particularly important.

Action-space design is one of the most important environment choices in DRL trading. A binary action space is simple and interpretable, but it forces the agent into full-position decisions. More flexible action spaces allow partial exposure, position scaling, or flat positions, which may improve risk control when signals are weak or transaction costs make frequent switching costly. Prior work supports testing action-space design directly. Zhang et al. (2020) compare different trading-action specifications, while Yang and Malik (2024) emphasize observation design, reward shaping, and action-space construction in reinforcement-learning cryptocurrency pair trading. These studies suggest that the action space should not be treated as a technical detail; it determines the set of feasible trading behaviors available to the agent.

Robustness is also a key concern in financial DRL. Financial markets are noisy, non-stationary, and vulnerable to backtesting overfitting. Liu et al. (2022) identify low signal-to-noise ratios, survivorship bias, and backtesting overfitting as major challenges in financial reinforcement-learning benchmarks. Hambly, Xu, and Yang (2023) similarly emphasize that reinforcement learning in finance is difficult to evaluate reliably because market environments change over time. In addition, Henderson et al. (2018) shows that DRL results can vary substantially across random seeds, making single-run performance insufficient for strong claims about model superiority. These concerns motivate the inclusion of seed-level robustness checks in empirical DRL trading studies.

Taken together, the literature establishes that statistical arbitrage is best understood as a pipeline linking portfolio construction, signal design, trading-policy learning, and evaluation. It also shows that DRL can be applied to financial and cryptocurrency trading, but that its performance is sensitive to environment design, action specification, transaction costs, and stochastic training variation. The remaining gap is that existing research has paid limited attention to how asset-

universe construction and action-space granularity jointly affect cointegration-based DRL cryptocurrency arbitrage performance.

The present study addresses this gap by evaluating four predefined cryptocurrency universes and comparing binary versus granular action spaces within the same cointegration-based DRL framework. By including DQN, PPO, A2C, a COIN benchmark, and seed-level robustness checks, the study shifts the evaluation question from identifying the best DRL algorithm in a fixed setting to understanding the conditions under which a DRL trading layer can add value to a cointegration-based cryptocurrency arbitrage system.

3. Data and Asset-Universe Design

This study uses 30-minute cryptocurrency spot-market data to examine whether asset-universe selection affects the performance of DRL-based statistical arbitrage. The data consist of USDT-denominated Binance spot pairs, which keeps the price panel internally consistent by avoiding mixed quote currencies. Each observation contains open, high, low, close, and volume information. These OHLCV fields provide the raw inputs used to construct portfolio prices, return series, spread variables, technical indicators, volatility measures, and liquidity-related proxies.

Binance Public Data is used because it provides downloadable historical cryptocurrency market data in daily and monthly files. The Binance kline/candlestick data structure supports standard candlestick fields and intraday intervals, including 30-minute bars, making it suitable for the intraday scenario-based design used in this study (Binance Public Data, n.d.; Binance Developers, n.d.).

The empirical sample covers a recent cryptocurrency market period. The training sample runs from 2023-01-01 to 2024-12-10, followed by a seven-calendar-day separation window from 2024-12-11 to 2024-12-17. The testing sample then runs from 2024-12-18 to 2025-05-17. This separation window is included to create a short buffer between model training and out-of-sample evaluation, reducing immediate temporal carry-over from the final training observations into the first testing scenarios. The sample design evaluates the framework on a recent out-of-sample period rather than replicating an earlier market calendar. Because cryptocurrency markets change rapidly, the results should be interpreted as evidence from the selected validation period rather than as a claim that the same performance would hold across all market regimes.

Four asset universes are constructed to examine whether universe design affects statistical-arbitrage performance. The Baseline 14-Asset Universe serves as the main reference universe. The Broad 20-Asset Universe expands the asset set by adding additional liquid cryptocurrencies. The Payment-Coin 6-Asset Universe focuses on Bitcoin-like and payment-oriented assets. The Layer-1 Platform 9-Asset Universe focuses on blockchain platform and infrastructure assets. These universes are designed to compare broader asset coverage with more economically coherent asset groupings, allowing the study to test whether cointegration-based trading performance depends on both universe size and asset similarity.

Table 1 reports the composition and purpose of each predefined asset universe. Figure 2 provides a visual summary of the same asset-universe design.

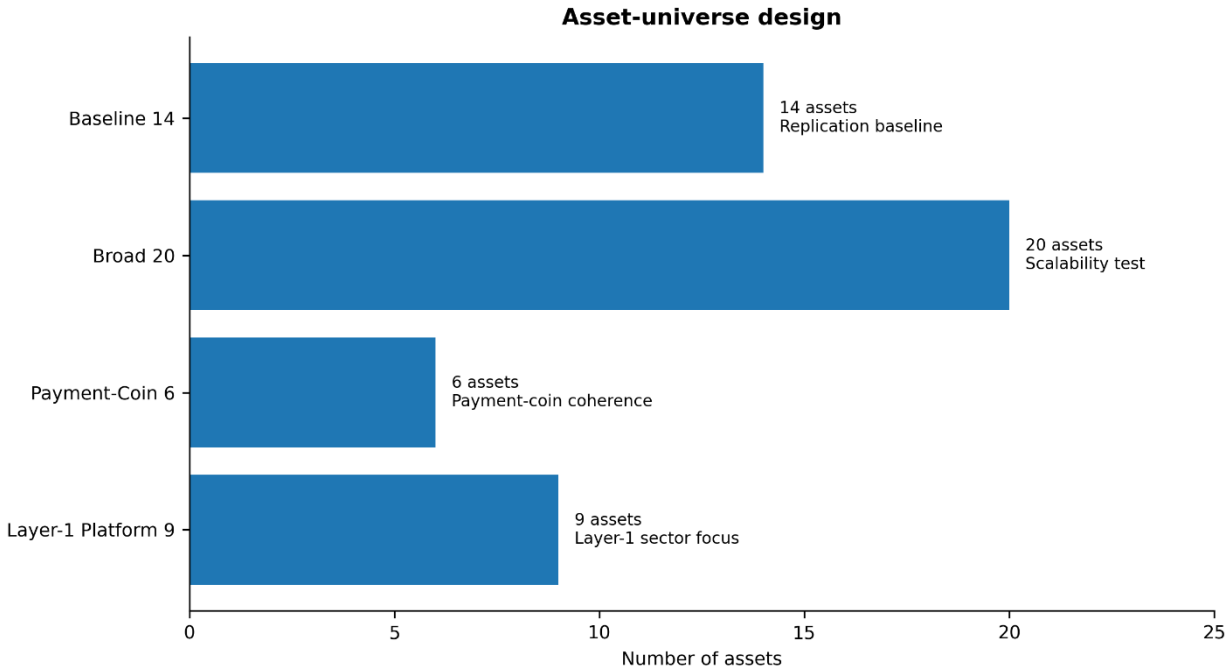
Table 1
Asset universes used in the empirical design.

Universe label	No. of assets	Assets	Purpose
Baseline 14-Asset Universe	14	BTC, ETH, BNB, XRP, ADA, SOL, DOT, BCH, LTC, AVAX, ALGO, AAVE, UNI, CAKE	Replication baseline matching the published 14-asset cryptocurrency universe.
Broad 20-Asset Universe	20	BTC, ETH, BNB, XRP, ADA, SOL, DOT, BCH, LTC, AVAX, ALGO, AAVE, UNI, CAKE, DOGE, LINK, TRX, XLM, ATOM, FIL	Broad-market extension used to test scalability and asset-set noise.
Payment-Coin 6-Asset Universe	6	BTC, BCH, LTC, DOGE, XRP, XLM	Payment-focused subset used to test whether economically related assets improve cointegration stability.
Layer-1 Platform 9-Asset Universe	9	ETH, BNB, ADA, SOL, DOT, AVAX, ALGO, TRX, ATOM	Blockchain-platform sector subset used to test sector-specific arbitrage performance.

Note. Each universe uses USDT-denominated Binance spot pairs at a 30-minute frequency. The Baseline 14-Asset Universe provides the closest replication baseline; the other universes test universe size and economic similarity.

Figure 2

Asset-Universe Design for the Empirical Analysis



Note. Bars show the number of cryptocurrencies in each predefined asset universe. The Baseline 14 universe serves as the replication baseline, the Broad 20 universe tests broader asset coverage and scalability, the Payment-Coin 6 universe tests payment-coin coherence, and the Layer-1 Platform 9 universe tests sector-specific arbitrage performance. All universes use USDT-denominated Binance spot pairs at a 30-minute frequency.

The use of multiple universes allows the study to test whether performance depends on the number and economic similarity of the assets included in the cointegration search. A broader universe may increase the number of potential arbitrage relationships, but it may also introduce noisier or less stable combinations. A smaller and more economically coherent universe may produce cleaner relationships, but it may provide fewer trading opportunities. Therefore, this study treats asset-universe selection as an empirical design choice rather than assuming that either a larger or smaller universe is automatically superior.

The analysis does not use historical Level-2 order-book depth because full-depth historical data are not freely available for the entire sample period. As a result, the study is based on OHLCV data rather than reconstructed historical order books. This improves transparency and reproducibility, but it limits the ability to estimate realized slippage precisely. The transaction-cost treatment is therefore specified separately in the methodology section.

4. Methodology and Experimental Design

This study applies a cointegration-based deep reinforcement learning framework to examine how asset-universe selection and action-space granularity affect cryptocurrency statistical-arbitrage performance. The framework combines classical cointegration portfolio construction with DRL-based trading decisions. Cointegration is used to form two synthetic portfolios, denoted portfolio A and portfolio B, while DRL agents learn whether to trade the relative movement between these portfolios and, in the granular setting, how much exposure to take.

The methodology consists of four stages: cointegration portfolio construction, market-state generation, DRL trading, and out-of-sample evaluation. The experimental design applies this framework across four asset universes and two action-space specifications. This structure allows the study to test whether performance depends on the selected cryptocurrency universe and whether allowing partial or flat positions improves risk-adjusted performance. The design follows the scenario-based logic of cointegration-based statistical arbitrage and DRL trading applications, while extending the comparison across asset-universe and action-space treatments (Engle & Granger, 1987; Johansen, 1988; Sutton & Barto, 2018; Mnih et al., 2015; Schulman et al., 2017; Vergara & Kristjanpoller, 2024).

4.1 Scenario Structure

The empirical framework is organized around fixed trading scenarios constructed from 30-minute observations. Each scenario contains 336 observations in total, divided into a 288-observation formation window and a 48-observation trading horizon. Because each observation represents 30 minutes, the formation window covers six days of historical market data, while the trading horizon covers one calendar day of intraday trading states.

A distinction is made between trading observations and realized reward intervals. The trading horizon contains 48 observations, but rewards are earned over transitions between consecutive observations. Therefore, the 48 trading observations generate 47 realized 30-minute reward intervals. In this study, the term “48-observation trading horizon” refers to the number of out-of-formation market states, while the performance calculations are based on the 47 realized return transitions within that horizon.

This scenario-based structure follows the logic of intraday statistical-arbitrage and reinforcement-learning trading designs, where historical windows are used to estimate trading relationships and later observations are used to evaluate sequential trading decisions (Gatev et al., 2006; Sutton & Barto, 2018; Vergara & Kristjanpoller, 2024).

The formation window is used to estimate the cointegration relationship among the assets in the selected universe and to construct synthetic portfolios A and B. During this stage, portfolio weights are determined from the historical price relationship. Once the formation window ends, these portfolio weights are fixed and carried into the trading horizon. This separation is important

because cointegration-based statistical arbitrage depends on identifying a stable relationship during a formation period and then testing whether deviations from that relationship can be exploited during a later trading period (Engle & Granger, 1987; Johansen, 1988; Gatev et al., 2006).

The trading horizon is used for sequential DRL decision-making. At each 30-minute trading observation, the agent observes the market state, selects an action, receives a reward based on the realized relative return between portfolios A and B, and transitions to the next state. The first trading observation initializes the trading state, while realized rewards begin once the next observation is available. This indexing is necessary because returns cannot be calculated from a single observation; they require a transition from one observation to the next. This formulation is consistent with a finite-horizon Markov decision process, where the agent learns a policy by linking states, actions, rewards, and transitions over time (Sutton & Barto, 2018).

This scenario structure separates portfolio construction from trading evaluation. The first part of each scenario identifies the statistical-arbitrage relationship, while the second part tests whether the agent can exploit that relationship out of formation. The same 288-observation formation window and 48-observation trading horizon are used across all asset universes, DRL agents, and action-space specifications. This keeps the experimental setup consistent and allows performance differences to be attributed to universe selection, model choice, or action-space design rather than to changes in the scenario construction process (Montgomery, 2017; Vergara & Kristjanpoller, 2024).

4.2 Cointegration Portfolio Construction

For each scenario, the first 288 30-minute observations are used as the formation window. Log prices are used to estimate cointegration relationships among the assets in the selected universe. Cointegration identifies combinations of non-stationary asset prices that share a stable long-run relationship, which supports their use in statistical-arbitrage strategies (Engle & Granger, 1987; Johansen, 1988; Gatev et al., 2006).

The estimated cointegration vector is converted into two synthetic portfolios, portfolio A and portfolio B. Positive vector weights form one side of the trade, while negative vector weights form the opposite side. The arbitrage opportunity is therefore defined by the relative movement between portfolios A and B rather than by the movement of a single asset. Once the portfolios are constructed, their weights remain fixed during the trading horizon. The DRL agent does not choose the individual cryptocurrency weights; it only decides the direction and exposure of the synthetic arbitrage position.

For universes with 12 or fewer assets, Johansen rank testing is applied. For larger universes, the implemented Johansen routine does not provide standard critical values. To avoid an inconsistent treatment across scenarios, the study applies a pre-specified fixed-rank construction rule based on the leading cointegration vector. This rule is used as an implementation constraint that preserves consistency and reproducibility across asset universes. It should not be interpreted as a claim that the leading vector is always theoretically optimal, nor as evidence that higher-dimensional systems necessarily contain only one economically meaningful cointegration relationship (Johansen, 1988; Lutkepohl, 2005; Statsmodels Developers, 2025).

This treatment ensures that larger universes remain comparable with smaller universes in the empirical design. The fixed-rank rule allows the study to evaluate broad asset sets without changing the basic portfolio-construction logic or introducing discretionary rank choices across scenarios.

4.3 Market-State Generation

After portfolios A and B are constructed, market-state variables are generated for each 30-minute trading observation. These variables summarize the information available to the agent when selecting an action. The state representation includes portfolio prices, spread variables, technical indicators, volatility measures, return signs, candlestick features, and market turbulence.

The technical and statistical features include MACD, RSI, stochastic RSI, Kalman-filter moving averages, EWMA volatility, bias, candlestick shadows, spread deviation, and turbulence. These variables capture trend, momentum, volatility, relative mispricing, and abnormal market conditions. Such features are consistent with technical-analysis, volatility-modeling, and statistical-arbitrage applications in financial trading (Murphy, 1999; Chande & Kroll, 1994; Kalman, 1960; J.P. Morgan/Reuters, 1996; Nison, 2001; Gatev et al., 2006; Kritzman & Li, 2010).

The state representation is held constant across all asset universes, DRL agents, and action-space designs. Holding the information set constant ensures that performance differences are driven by the experimental treatments rather than by changes in model inputs.

4.4 DRL Trading Environment and Action Spaces

The trading problem is formulated as a finite-horizon Markov decision process. At each 30-minute trading observation, the agent observes the current market state S_t, selects an action a_t, receives a reward, and transitions to the next state. The portfolio weights inside portfolios A and B are fixed by the cointegration construction step. The agent therefore decides only which side of the arbitrage relation to trade and, in the granular setting, how much exposure to take. This setup is consistent with reinforcement-learning trading environments in which actions are learned from observed market states and rewards (Sutton & Barto, 2018; Vergara & Kristjanpoller, 2024).

The timing convention is defined as follows. At trading observation t, the agent observes S_t and selects a_t. The selected action determines the position to be held over the next return interval. When rewards are indexed at the end of the interval, the reward at t is therefore generated by the action selected at the previous observation, a_{t-1}. This convention explains why the reward function in Section 4.5 uses a_{t-1} rather than a_t.

The baseline specification uses a binary action space:

$$a_t \in \{-1,1\}$$

$$a_t = \begin{cases} 1, & \text{long A / short B,} \\ -1, & \text{short A / long B.} \end{cases}$$

This binary setting forces the agent to maintain one of the two directional arbitrage positions. It follows the original framework in which the agent does not choose individual asset weights but only selects the direction of the synthetic arbitrage trade.

This study also evaluates a granular action space:

$$a_t \in \{-1, -0.5, 0, 0.5, 1\}$$

$$a_t = \begin{cases} 1, & \text{full long A / short B,} \\ 0.5, & \text{half long A / short B,} \\ 0, & \text{no exposure,} \\ -0.5, & \text{half short A / long B,} \\ -1, & \text{full short A / long B.} \end{cases}$$

The granular specification allows the agent to take full exposure, half exposure, or no exposure. This extension tests whether partial and flat positions improve performance or risk control, especially when arbitrage signals are weak or when transaction costs reduce the benefit of frequent switching. The study does not assume that granular actions will outperform binary actions; the comparison is evaluated empirically.

4.5 Reward Function and Transaction Costs

The reward function is based on the realized relative return between portfolios A and B. Let r_t^A and r_t^B denote the log returns of portfolios A and B over trading interval t. Because returns are earned over intervals rather than at isolated observations, r_t^A and r_t^B are calculated between two consecutive trading observations. With 48 trading observations, this produces 47 realized return intervals.

The gross reward is calculated using the action selected at the previous trading observation, a_{t-1}, because that position determines the return earned during interval t:

$$R_t^{\text{gross}} = a_{t-1}\left(e^{r_t^A} - e^{r_t^B}\right)$$

This reward structure links profitability to the direction of the synthetic arbitrage position. When $a_{t-1} = 1$, the agent is long portfolio A and short portfolio B, so the reward is positive when portfolio A outperforms portfolio B. When $a_{t-1} = -1$, the agent is short portfolio A and long portfolio B, so the reward is positive when portfolio B outperforms portfolio A. In the granular action space, the same equation scales exposure according to the selected action. For example, $a_{t-1} = 0.5$ represents half long-A/short-B exposure, while $a_{t-1} = 0$ represents no exposure.

To account for trading frictions, the net reward subtracts a transaction-cost penalty whenever the agent changes its exposure. The net reward at each decision point combines the return earned from the position held over the just-realized interval with the cost of updating exposure for the next interval. The net reward is defined as:

$$R_t^{\text{net}} = R_t^{\text{gross}} - c|\Delta a_t|$$

where c is the transaction-cost coefficient and $|\Delta a_t|$ measures the absolute change in exposure between consecutive actions? The change in exposure is:

$$\Delta a_t = a_t - a_{t-1}$$

This specification penalizes frequent position switching and encourages the agent to change exposure only when the expected improvement in the arbitrage position is large enough to justify

the cost. In the binary action space, switching directly from -1 to 1, or from 1 to -1, produces a larger exposure change than maintaining the same position. In the granular action space, the agent can adjust exposure more gradually through half or flat positions, which may reduce unnecessary turnover during weak or uncertain signals.

This timing convention separates three elements of the trading process: the previous action determines the return earned over the current interval, the current action determines the exposure to be held in the next interval, and the transaction-cost term penalizes the change from the previous exposure to the newly selected exposure.

4.6 DRL Agents, Benchmark, and Training Protocol

Three DRL agents are evaluated: Deep Q-Network, Proximal Policy Optimization, and Advantage Actor-Critic. DQN represents a value-based method that learns a state-action value function. PPO and A2C represent policy-gradient or actor-critic approaches that learn trading policies from interaction with the environment (Mnih et al., 2015; Mnih et al., 2016; Schulman et al., 2017). These algorithm families are also used in recent DRL applications to cryptocurrency statistical arbitrage (Vergara & Kristjanpoller, 2024).

The primary benchmark is the COIN strategy. COIN follows the cointegration-based arbitrage signal without learning a dynamic DRL policy. It represents the classical statistical-arbitrage component of the framework and is included to test whether the DRL decision layer improves performance relative to a non-DRL trading rule.

For each asset universe and action-space specification, trading scenarios are generated from the training period. Scenarios are sampled during training so that agents learn across different historical market conditions. The main model comparison is based on seed 42. Selected key models are then re-estimated using seeds 42, 101, and 202 to evaluate whether the main findings are sensitive to stochastic training variation. This robustness step is included because DRL performance can vary with random initialization, scenario sampling, and exploration behavior (Henderson et al., 2018).

The separation between the training sample and testing sample ensures that agents are evaluated on a later out-of-sample period that is not used for policy learning. The short buffer between the two samples further reduces the risk that the first testing scenarios are mechanically adjacent to the final training observations.

4.7 Evaluation Metrics and Reporting Structure

All strategies are evaluated out of sample using profitability, risk, drawdown, downside-risk, risk-adjusted-performance, and turnover measures. The evaluation metrics include total return, annualized mean return, annualized volatility, Sharpe ratio, Calmar ratio, maximum drawdown, 5% Value at Risk, 5% Expected Shortfall, and turnover. These metrics are selected to capture both return generation and risk exposure. Turnover is included because granular actions may improve performance not only by increasing returns, but also by reducing unnecessary position changes.

The results are reported in three layers. First, the seed-42 results compare all DRL agents and the COIN benchmark across the four asset universes. Second, the binary and granular action spaces are compared directly to evaluate whether the action-space extension improves performance or

risk control. Third, selected key models are tested across seeds 42, 101, and 202 to assess robustness to stochastic DRL training variation.

This reporting structure separates the main model comparison, action-space analysis, and seed-level robustness checks. It also keeps the experimental design consistent across all models, so that observed differences can be interpreted as the result of asset-universe selection, action-space specification, or DRL model choice.

5. Empirical Results and Discussion

This section reports the out-of-sample performance of the cointegration-based DRL statistical-arbitrage framework across four asset universes, three DRL agents, two action-space specifications, and the COIN benchmark. The analysis focuses on three empirical questions: whether performance varies across asset universes, whether granular actions improve trading outcomes, and whether the strongest results remain stable across random seeds. Performance is evaluated using total return, Sharpe ratio, Calmar ratio, maximum drawdown, annualized volatility, and turnover, rather than total return alone.

5.1 Cross-Universe and Model Performance

The results show that asset-universe design materially affects out-of-sample performance. As defined in Table 1, the four universes differ by both size and economic coherence. The Baseline 14-Asset Universe provides the closest replication baseline, the Broad 20-Asset Universe tests whether broader asset coverage improves opportunity discovery, and the Payment-Coin 6-Asset and Layer-1 Platform 9-Asset universes test whether more economically focused groups produce more tradable cointegration relationships.

Table 2
Main model results by asset universe using the seed-42 run.

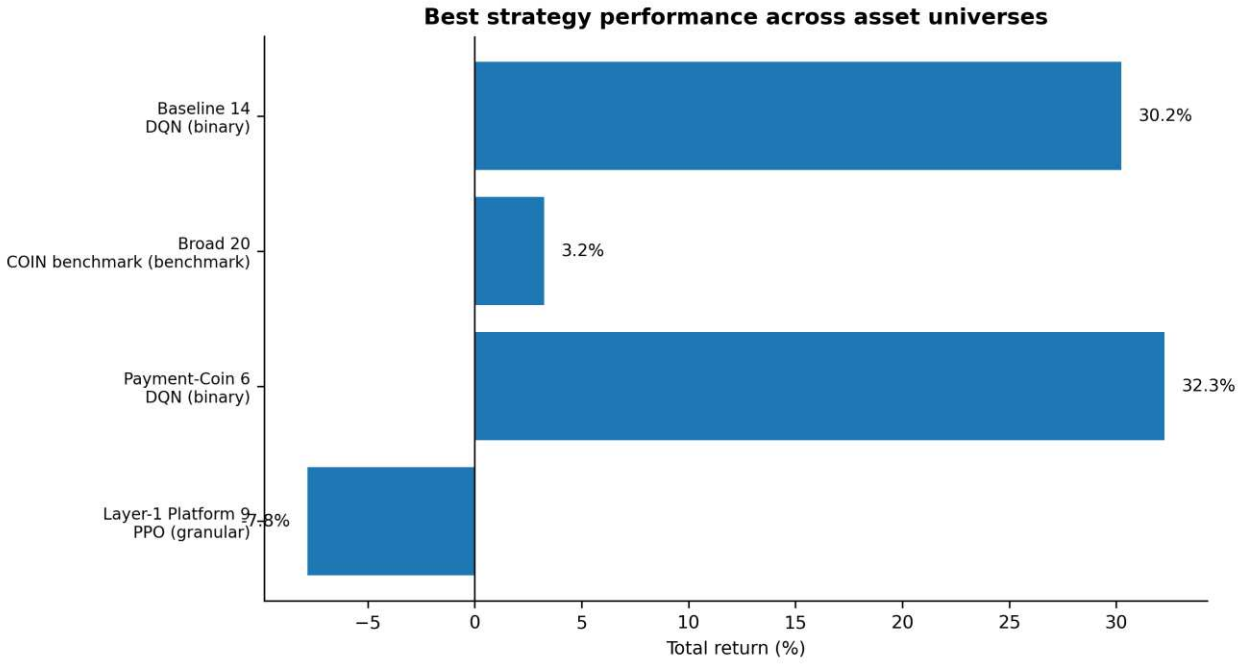
Universe	Strategy	Action	Return (%)	Sharpe	Calmar	Max DD (%)	Vol. (%)	Turnover
Baseline 14	DQN	binary	30.24	2.68	7.36	-9.56	26.26	148.0
Baseline 14	DQN	granular	8.69	1.16	1.90	-12.16	19.81	413.0
Baseline 14	PPO	binary	-30.97	-3.73	-2.20	-41.30	24.37	1242.0
Baseline 14	PPO	granular	-3.06	-0.32	-0.38	-15.97	18.98	721.5
Baseline 14	A2C	binary	-38.30	-4.32	-3.00	-39.49	27.42	1228.0
Baseline 14	A2C	granular	-28.50	-5.53	-2.84	-29.54	15.16	903.0
Baseline 14	COIN benchmark	benchmark	0.97	0.22	0.36	-16.89	26.80	576.0
Broad 20	DQN	binary	-12.97	-1.01	-1.18	-25.91	30.29	148.0
Broad 20	DQN	granular	-5.82	-0.92	-1.06	-13.31	15.20	72.0
Broad 20	PPO	binary	-45.00	-5.57	-3.22	-45.93	26.50	1468.0
Broad 20	PPO	granular	-9.48	-1.28	-1.57	-15.01	18.42	872.5
Broad 20	A2C	binary	-21.10	-2.09	-1.75	-32.32	27.03	1264.0
Broad 20	A2C	granular	-25.54	-3.56	-2.75	-26.37	20.39	1023.0
Broad 20	COIN benchmark	benchmark	3.25	0.45	0.76	-14.73	25.24	548.0
Payment-Coin 6	DQN	binary	32.26	1.72	3.89	-21.05	47.49	206.0
Payment-Coin 6	DQN	granular	24.59	1.80	3.77	-16.26	34.03	401.5
Payment-Coin 6	PPO	binary	-39.91	-2.57	-2.84	-41.66	45.90	1622.0

Payment-Coin 6	PPO	granular	-5.83	-0.42	-0.72	-16.14	27.24	1630.5
Payment-Coin 6	A2C	binary	-28.12	-1.65	-2.12	-34.83	44.78	1282.0
Payment-Coin 6	A2C	granular	-13.01	-1.09	-1.50	-20.84	28.69	936.5
Payment-Coin 6	COIN benchmark	benchmark	-12.26	-0.51	-0.63	-36.70	44.92	484.0
Layer-1 Platform 9	DQN	binary	-23.29	-1.67	-2.07	-29.18	36.10	212.0
Layer-1 Platform 9	DQN	granular	-8.66	-1.15	-1.64	-12.91	18.52	95.5
Layer-1 Platform 9	PPO	binary	-13.44	-0.93	-1.06	-29.16	33.40	1860.0
Layer-1 Platform 9	PPO	granular	-7.83	-0.51	-0.51	-31.29	31.06	1304.5
Layer-1 Platform 9	A2C	binary	-46.76	-4.30	-3.10	-49.43	35.55	1330.0
Layer-1 Platform 9	A2C	granular	-18.57	-1.58	-1.57	-30.24	30.14	1389.0
Layer-1 Platform 9	COIN benchmark	benchmark	-21.67	-2.37	-1.90	-30.94	24.77	478.0

Note. The main table reports the seed-42 run for each universe and strategy. COIN is the cointegration benchmark. Max DD = maximum drawdown; Vol. = annualized volatility.

Figure 3

Best strategy performance across asset universes.



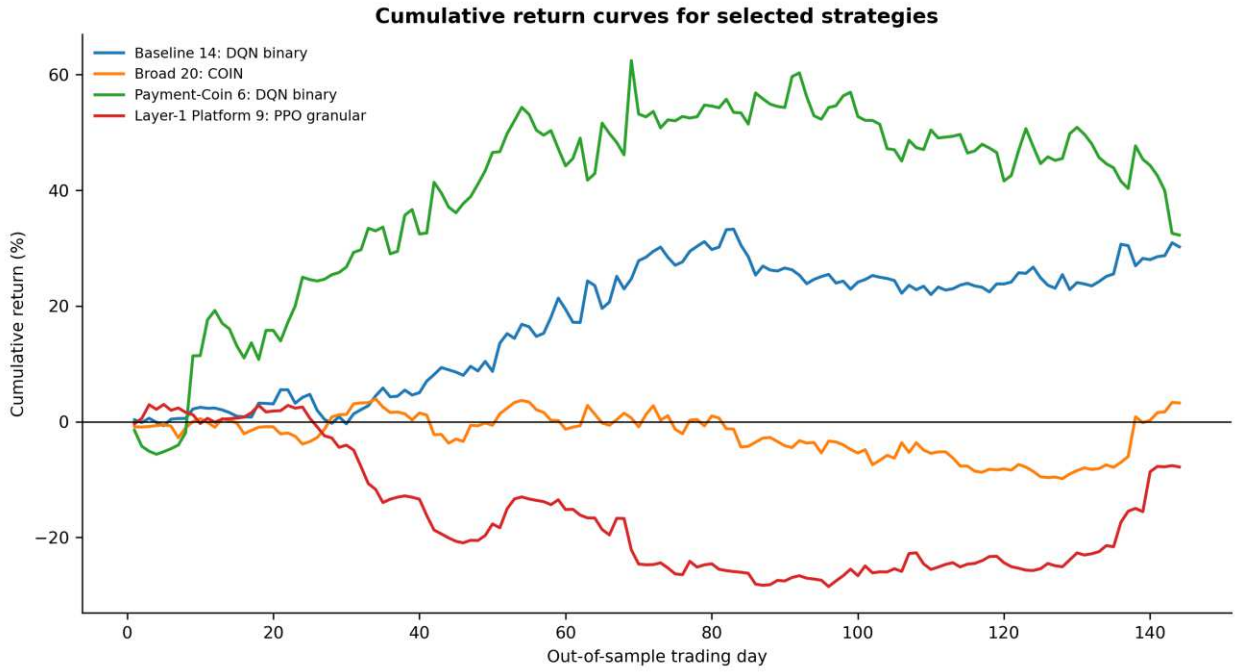
Note. Bars report the seed-42 out-of-sample total return for the best-performing strategy in each universe.

Table 2 and Figure 3 show that the strongest raw return occurs in the Payment-Coin 6-Asset Universe, where DQN with binary actions returns 32.26%. The Baseline 14-Asset Universe also performs strongly, with DQN binary returning 30.24%, a Sharpe ratio of 2.68, a Calmar ratio of 7.36, and a maximum drawdown of -9.56%. In contrast, the Broad 20-Asset Universe is led by the COIN benchmark, which returns 3.25%, while all DRL models are negative. The Layer-1 Platform 9-Asset Universe is the weakest environment, with no strategy producing a positive return and the COIN benchmark losing 21.67%.

Figure 7 complements Table 2 by showing the cumulative out-of-sample return paths for the best seed-42 strategy in each asset universe. This allows the comparison to reflect not only final performance, but also the stability and timing of gains and losses during the testing period.

Figure 7

Cumulative Return Curves for the Best Seed-42 Strategy in Each Asset Universe



Note. Curves report cumulative out-of-sample returns for the best-performing seed-42 strategy in each asset universe. Returns are shown after transaction costs over the out-of-sample trading period. COIN refers to the non-DRL cointegration benchmark.

These results indicate that expanding the asset set does not automatically improve statistical-arbitrage performance. The Broad 20-Asset Universe may increase the number of potential combinations, but it may also introduce weaker or noisier cointegration relationships that are harder for DRL agents to exploit. By contrast, the Payment-Coin 6-Asset Universe suggests that a smaller and more economically coherent asset set can provide a more favorable learning environment. However, this interpretation is sample-specific and should not be read as evidence that payment-oriented assets will dominate across all market regimes.

The full model comparison also shows that no single strategy dominates across all universes. DQN is the strongest agent in the Baseline 14 and Payment-Coin 6 universes, but it underperforms in the Broad 20 and Layer-1 Platform 9 universes. PPO and A2C generally perform poorly, although PPO granular is the least negative strategy in the Layer-1 universe. This supports a conditional interpretation: the DRL trading layer can add value in selected settings, but its performance depends on the interaction between universe design, model choice, and the structure of the underlying cointegration signal.

5.2 Effect of Action-Space Granularity

Table 3 and Figure 4 compare binary and granular action spaces for each DRL agent and asset universe. Delta values are calculated as granular minus binary. A positive delta therefore means that the granular specification produces a higher value for the metric; whether this represents an improvement depends on the metric being evaluated. Higher return, Sharpe ratio, and Calmar ratio are generally favorable, whereas higher turnover, volatility, or a more negative drawdown may not indicate better performance.

Table 3

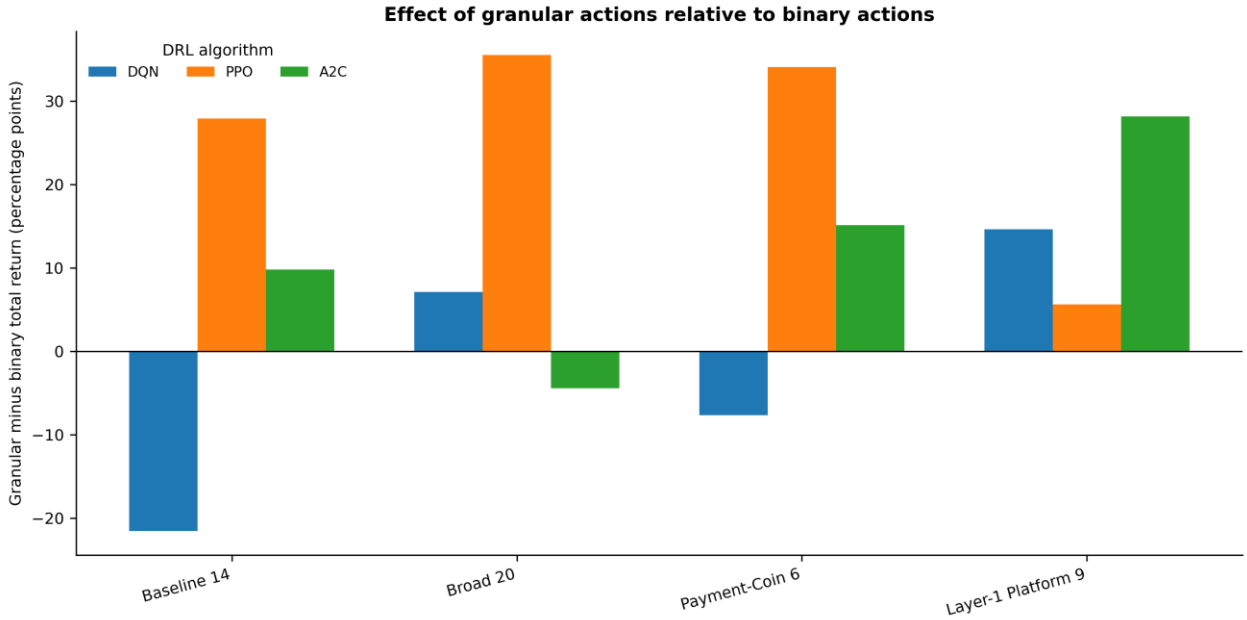
Binary versus granular action-space comparison

Universe	Alg.	Ret. Binary (%)	Ret. Granular (%)	Delta Ret.	Sharpe Binary	Sharpe Granular	Delta Sharpe	DD Binary (%)	DD Granular (%)	Turn. Binary	Turn. Granular
Baseline 14	A2C	-38.30	-28.50	9.79	-4.32	-5.53	-1.21	-39.49	-29.54	1228.0	903.0
Baseline 14	DQN	30.24	8.69	-21.55	2.68	1.16	-1.52	-9.56	-12.16	148.0	413.0
Baseline 14	PPO	-30.97	-3.06	27.91	-3.73	-0.32	3.41	-41.30	-15.97	1242.0	721.5
Broad 20	A2C	-21.10	-25.54	-4.44	-2.09	-3.56	-1.47	-32.32	-26.37	1264.0	1023.0
Broad 20	DQN	-12.97	-5.82	7.15	-1.01	-0.92	0.09	-25.91	-13.31	148.0	72.0
Broad 20	PPO	-45.00	-9.48	35.52	-5.57	-1.28	4.29	-45.93	-15.01	1468.0	872.5
Payment-Coin 6	A2C	-28.12	-13.01	15.11	-1.65	-1.09	0.56	-34.83	-20.84	1282.0	936.5
Payment-Coin 6	DQN	32.26	24.59	-7.67	1.72	1.80	0.08	-21.05	-16.26	206.0	401.5
Payment-Coin 6	PPO	-39.91	-5.83	34.08	-2.57	-0.42	2.15	-41.66	-16.14	1622.0	1630.5
Layer-1 Platform 9	A2C	-46.76	-18.57	28.19	-4.30	-1.58	2.72	-49.43	-30.24	1330.0	1389.0
Layer-1 Platform 9	DQN	-23.29	-8.66	14.63	-1.67	-1.15	0.53	-29.18	-12.91	212.0	95.5
Layer-1 Platform 9	PPO	-13.44	-7.83	5.61	-0.93	-0.51	0.42	-29.16	-31.29	1860.0	1304.5

Note. Delta values are calculated as granular minus binary. Granular actions allow partial and flat positions; binary actions restrict the agent to full long-A/short-B or short-A/long-B exposure.

Figure 4

Effect of Granular Actions Relative to Binary Actions



Note. *Bars show the granular-minus-binary difference in total return for each DRL agent and asset universe. Positive values indicate that the granular action space produced a higher total return than the binary action space, while negative values indicate that the binary action space produced a higher total return. Differences are reported in percentage points. DQN = Deep Q-Network; PPO = Proximal Policy Optimization; A2C = Advantage Actor–Critic.*

The results show (see Figure 4) that granular actions do not universally outperform binary actions. For DQN, binary actions produce higher raw returns in the two strongest universes. In the Baseline 14 universe, DQN binary outperforms DQN granular by 21.55 percentage points. In the Payment-Coin 6 universe, DQN binary outperforms DQN granular by 7.67 percentage points. These results suggest that full directional exposure can be beneficial when the arbitrage signal is sufficiently strong.

However, the return comparison alone does not capture the full effect of granularity. In the Payment-Coin 6 universe, DQN granular produces a slightly higher Sharpe ratio than DQN binary and has a smaller maximum drawdown. Granular actions also improve several weaker PPO and A2C configurations. For example, PPO granular substantially reduces losses relative to PPO binary in the Baseline 14, Broad 20, and Payment-Coin 6 universes. In the Layer-1 Platform 9 universe, all three DRL agents produce better total returns under the granular specification, although the best result remains negative.

The action-space evidence therefore supports a risk-management interpretation. Binary actions can be preferable when the signal is strong and directional, but granular actions become more useful when signals are weaker, noisier, or costly to trade. Granularity should therefore be interpreted as an exposure-control mechanism rather than as a universally superior trading specification.

5.3 Robustness Across Random Seeds

Because DRL training is stochastic, robustness across random seeds is important for distinguishing stable model behavior from seed-specific outcomes. Table 4 reports robustness statistics for selected key models across seeds 42, 101, and 202. These models are selected because they represent the strongest main-run DRL findings, the most relevant granular-action comparison, and the weakest universe case. Table 4 reports robustness statistics for selected key models across seeds 42, 101, and 202, while Figure 5 visualizes the corresponding dispersion in total returns across those robustness runs.

Table 4

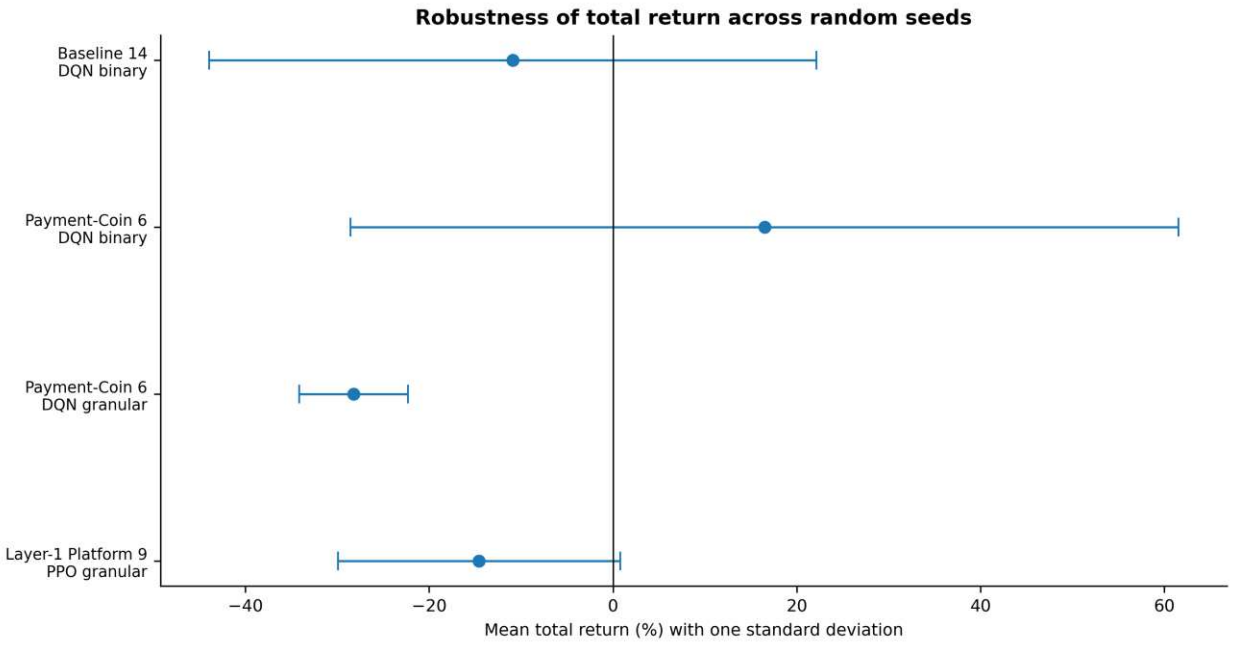
Robustness statistics across random seeds for key models.

Universe	Alg.	Action	n	Seeds	Ret. M (%)	Ret. SD	Ret. Min	Ret. Max	Sharpe M	Sharpe SD	Sharpe Min	Sharpe Max	DD M (%)	DD SD	DD Min	DD Max	Turn. M	Turn. SD
Baseline 14	DQN	binary	3	42;101;202	-10.92	33.05	-30.02	27.24	-1.42	3.35	-3.36	2.45	-24.15	12.63	-31.47	-9.56	164.0	28.0
Payment-Coin 6	DQN	binary	3	42;101;202	16.50	45.05	-35.51	43.04	0.67	2.47	-2.18	2.11	-28.61	15.37	-46.35	-19.61	197.3	35.2
Payment-Coin 6	DQN	granular	3	42;101;202	-28.22	5.94	-33.42	-21.74	-2.70	0.61	-3.39	-2.23	-35.28	12.87	-48.69	-23.02	882.3	173.0
Layer-1 Platform 9	PPO	granular	3	42;101;202	-14.57	15.37	-27.62	2.38	-3.17	3.24	-6.04	0.34	-26.42	3.74	-28.92	-22.12	1154.5	540.6

Note. Robustness statistics are computed across seeds 42, 101, and 202. Ret. = total return; DD = maximum drawdown; M = mean; SD = standard deviation; Turn. = total turnover.

Figure 5

Robustness of Total Return Across Random Seeds



Note. Points show the mean total return across seeds 42, 101, and 202 for selected key models. Horizontal bars show one standard deviation around the mean. The vertical reference line marks zero return. The robustness runs use the robustness training configuration, so values may differ from the main seed-42 results when the training budget differs.

The robustness results qualify the main seed-42 findings. Payment-Coin 6 DQN binary remains positive on average across seeds, with a mean return of 16.50%, but its return dispersion is large.

This makes it the most promising robustness candidate, although not a fully stable result. Baseline 14 DQN binary is less robust: despite its strong seed-42 performance, its mean return across robustness seeds is negative. Payment-Coin 6 DQN granular is also negative on average, while Layer-1 Platform 9 PPO granular remains negative across the robustness analysis.

These findings limit overgeneralization from a single successful run. The seed-42 results identify where the framework can perform well, but the robustness tests show that DRL outcomes remain sensitive to stochastic training variation. The strongest conclusion is therefore not that one DRL model universally dominates, but that cointegration-based DRL can add value in selected settings when asset-universe design, action-space structure, and training stability align.

Overall, the empirical evidence supports three main conclusions. First, asset-universe design materially affects cryptocurrency statistical-arbitrage performance: the Baseline 14 and Payment-Coin 6 universes are more favorable than the Broad 20 and Layer-1 Platform 9 universes. Second, DQN is strongest in favorable universes, but its advantage is not universal. Third, granular actions mainly improve risk control rather than raw return. These findings show that DRL-based cryptocurrency statistical arbitrage is best understood as an interaction between cointegration structure, asset-universe design, action-space specification, and stochastic training stability.

6. Conclusion

This study examined whether cointegration-based deep reinforcement learning cryptocurrency statistical arbitrage is shaped not only by the choice of DRL algorithm, but also by two design decisions that are often treated as secondary: asset-universe construction and action-space granularity. The central contribution is that the study moves the evaluation question from "Which DRL model performs best?" to "Under what asset-universe and action-space conditions can a DRL trading layer add value to a cointegration-based arbitrage framework?" This distinction is important because statistical arbitrage depends on the stability of relationships among assets, while DRL performance depends on the quality of the environment in which the agent learns.

The empirical results show that asset-universe design is a first-order determinant of performance. The strongest seed-42 results occur in the Baseline 14-Asset and Payment-Coin 6-Asset universes, where DQN with the binary action space generates the highest raw returns. In contrast, the Broad 20-Asset Universe is led by the COIN benchmark rather than by any DRL agent, and the Layer-1 Platform 9-Asset Universe remains negative even for the best-performing strategy. These findings show that expanding the number of assets does not automatically improve statistical-arbitrage performance. A larger universe may create more possible combinations, but it can also introduce weaker, noisier, or less stable cointegration relationships. The results therefore suggest that the economic coherence and statistical stability of the universe matter more than universe size alone.

The action-space results provide a second important finding. Granular actions do not universally outperform binary actions. In the strongest DQN cases, the binary action space produces higher total returns, suggesting that full directional exposure can be valuable when the arbitrage signal is sufficiently strong. Granular actions improve risk control in several weaker or more difficult settings. In particular, granular specifications reduce losses or improve drawdown behavior for several PPO and A2C configurations, and they help limit downside exposure in the Layer-1

universe. This means that action-space granularity should not be interpreted simply as a return-enhancing extension. It's more important role is as a risk-management mechanism that allows the agent to reduce or neutralize exposure when the signal is weak, noisy, or costly to trade.

The robustness results qualify the main findings and are central to the interpretation of the study. Payment-Coin DQN binary remains the most promising robustness candidate, with a positive mean return across seeds 42, 101, and 202. However, its high return dispersion shows that performance remains sensitive to stochastic training variation. Baseline DQN binary performs strongly in the main seed-42 run but is not stable across the robustness seeds. Payment-Coin DQN granular and Layer-1 PPO granular are also negative on average in the robustness analysis. These results reinforce a key lesson from DRL research: a successful single run should not be interpreted as definitive evidence of model superiority, especially in noisy financial environments where initialization, sampling, and exploration can materially affect outcomes (Henderson et al., 2018).

The novelty of this study is therefore not simply that it applies DQN, PPO, and A2C to cryptocurrency arbitrage. Its novelty lies in treating the asset universe and the action space as experimental design variables rather than fixed background choices. Prior cointegration and statistical-arbitrage research establishes the importance of stable relationships among assets, while DRL trading research emphasizes learning policies from sequential market states (Engle & Granger, 1987; Johansen, 1988; Gatev et al., 2006; Sutton & Barto, 2018). This study connects those ideas by showing that the learning problem faced by the DRL agent changes materially when the universe becomes broader, more focused, or more sector-specific. It also shows that the design of the action set affects whether the agent behaves aggressively, defensively, or inconsistently across market conditions.

The study contributes to the literature in three ways. First, it provides evidence that universe construction should be treated as a core modeling decision in cointegration-based cryptocurrency arbitrage. Second, it extends the DRL trading framework by comparing binary and granular action spaces, showing that partial and flat exposure can improve risk control even when they do not maximize return. Third, it incorporates seed-level robustness analysis, which helps separate promising model behavior from seed-specific outcomes. Together, these contributions provide a more careful evaluation of DRL arbitrage performance than a single-model, single-universe, single-seed comparison.

These findings matter for several groups. For academic researchers, the results show that DRL trading studies should report the asset-universe design, action-space specification, benchmark comparison, and seed sensitivity rather than focusing only on the best-performing algorithm. For quantitative traders and portfolio managers, the findings suggest that asset selection is not merely a preprocessing step; it directly affects whether the arbitrage relation is learnable and tradable. For risk managers, the results highlight the importance of drawdown, turnover, and robustness checks before deploying DRL-based strategies. For developers of trading systems, the study shows that a more flexible action space may be useful not because it always increases return, but because it can reduce harmful exposure in difficult environments.

The practical implication is cautious but important. Cointegration-based DRL can add value in favorable cryptocurrency universes, especially when DQN is applied to asset sets with more stable

arbitrage structure. However, the evidence does not support a universal claim that DRL always improves on a classical cointegration benchmark. The Broad 20 and Layer-1 results show that DRL can underperform when the underlying relationships are noisy or unstable. Therefore, successful deployment requires more than selecting a sophisticated algorithm. It requires disciplined universe design, transaction-cost awareness, downside-risk evaluation, and robustness testing.

The study has several limitations. The analysis is based on one recent out-of-time sample and USDT-denominated Binance spot pairs, so the findings may not generalize to all exchanges, quote currencies, or market regimes. The framework uses OHLCV data rather than full historical Level-2 order-book data, which limits the precision of execution-cost and slippage estimates. The robustness analysis uses selected key models and three seeds, which is useful for detecting instability but does not fully characterize the distribution of possible DRL outcomes.

Future research should therefore focus on broader regime testing, richer execution-cost modeling, and stronger robustness designs. A natural extension would be to test dynamic universe selection, where the asset set changes over time based on liquidity, cointegration stability, volatility, or economic similarity. Another useful extension would be to compare discrete granular actions with continuous position sizing, allowing the agent to learn exposure more flexibly. These extensions would help determine whether the strongest results in this study reflect persistent arbitrage structure or favorable conditions in a specific sample.

Overall, the study shows that DRL-based cryptocurrency statistical arbitrage is best understood as an interaction between cointegration structure, asset-universe design, action-space specification, and training stability. The results support the potential of DRL in selected settings, but they also caution against overgeneralizing from isolated high-return runs. The main conclusion is that the design of the trading environment matters as much as the DRL algorithm itself.

Data Availability and Replication Package

The replication package for this study is archived on Zenodo and includes the code, configuration files, universe definitions, saved result summaries, final tables, figures, validation scripts, and documentation required to reproduce the reported manuscript outputs. The package also includes the universe-label crosswalk linking the manuscript labels to the corresponding data folders and configuration files, which supports reproducibility of the renamed asset-universe design. Raw Binance OHLCV files and trained DRL model checkpoints are not redistributed; instead, the package provides data-acquisition scaffolding and result-generation scripts. The replication package is available at: **INSERT-ZENODO-ID**

Funding Declaration

The authors received no specific funding for this work.

References

Binance. (n.d.). *Binance public data*. Retrieved June 10, 2026, from https://data.binance.vision/

Binance. (n.d.). *binance-public-data*. GitHub. Retrieved June 10, 2026, from https://github.com/binance/binance-public-data

Binance Developers. (n.d.). *Market data endpoints: Kline/candlestick data*. Retrieved June 10, 2026, from https://developers.binance.com/docs/binance-spot-api-docs/rest-api/market-data-endpoints

Chande, T. S., & Kroll, S. (1994). *The new technical trader: Boost your profit by plugging into the latest indicators*. John Wiley & Sons.

Chen, Y.-F., & Huang, S.-H. (2021). Sentiment-influenced trading system based on multimodal deep reinforcement learning. *Applied Soft Computing, 112*, Article 107788. DOI: 10.1016/j.asoc.2021.107788

Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica, 55*(2), 251–276. DOI: 10.2307/1913236

Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies, 19*(3), 797–827. DOI: 10.1093/rfs/hhj020

Guijarro-Ordonez, J., Pelger, M., & Zanotti, G. (2025). Deep learning statistical arbitrage. *Management Science*. Advance online publication. DOI: 10.1287/mnsc.2022.03132

Hambly, B., Xu, R., & Yang, H. (2023). Recent advances in reinforcement learning in finance. *Mathematical Finance, 33*(3), 437–503. DOI: 10.1111/mafi.12382

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1), 3207–3214. DOI: 10.1609/aaai.v32i1.11694

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control, 12*(2–3), 231–254. DOI: 10.1016/0165-1889(88)90041-3

J.P. Morgan/Reuters. (1996). *RiskMetrics—Technical document* (4th ed.). Morgan Guaranty Trust Company.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82*(1), 35–45. DOI: 10.1115/1.3662552

Kritzman, M., & Li, Y. (2010). Skulls, financial turbulence, and risk management. *Financial Analysts Journal, 66*(5), 30–41. DOI: 10.2469/faj.v66.n5.3

Li, Z., Tam, V. W. L., & Yeung, K. L. (2024). Developing a multi-agent and self-adaptive framework with deep reinforcement learning for dynamic portfolio risk management. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (pp. 1174–1182). International Foundation for Autonomous Agents and Multiagent Systems.

Lin, Y.-C., Chen, C.-T., Sang, C.-Y., & Huang, S.-H. (2022). Multiagent-based deep reinforcement learning for risk-shifting portfolio management. *Applied Soft Computing, 123*, Article 108894. DOI: 10.1016/j.asoc.2022.108894

Liu, F., Li, Y., Li, B., Li, J., & Xie, H. (2021). *Bitcoin transaction strategy construction based on deep reinforcement learning*. arXiv. arXiv:2109.14789

Liu, X.-Y., Yang, H., Gao, J., & Wang, C. D. (2021). FinRL: Deep reinforcement learning framework to automate trading in quantitative finance. In *Proceedings of the Second ACM International Conference on AI in Finance* (pp. 1–9). Association for Computing Machinery. DOI: 10.1145/3490354.3494366

Liu, X.-Y., Xia, Z., Rui, J., Gao, J., Yang, H., Zhu, M., Wang, C. D., Wang, Z., & Guo, J. (2022). FinRL-Meta: Market environments and benchmarks for data-driven financial reinforcement learning. *Advances in Neural Information Processing Systems, 35*, 1835–1849. DOI: 10.52202/068431-0134

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer. DOI: 10.1007/978-3-540-27752-1

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of Machine Learning Research: Vol. 48. Proceedings of the 33rd International Conference on Machine Learning* (pp. 1928–1937). PMLR.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. DOI: 10.1038/nature14236

Montgomery, D. C. (2017). *Design and analysis of experiments* (9th ed.). John Wiley & Sons.

Murphy, J. J. (1999). *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. New York Institute of Finance.

Nison, S. (2001). *Japanese candlestick charting techniques: A contemporary guide to the ancient investment techniques of the Far East* (2nd ed.). Prentice Hall Press.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms*. arXiv. arXiv:1707.06347

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference* (pp. 92–96). DOI: 10.25080/majora-92bf1922-011

statsmodels Developers. (2025). *statsmodels.tsa.vector_ar.vecm.coint_johansen*. In *statsmodels 0.14.6 documentation*. Retrieved June 10, 2026, from https://www.statsmodels.org/stable/generated/statsmodels.tsa.vector_ar.vecm.coint_johansen.html

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

Vergara, G., & Kristjanpoller, W. (2024). Deep reinforcement learning applied to statistical arbitrage investment strategy on cryptomarket. *Applied Soft Computing, 153*, Article 111255. DOI: 10.1016/j.asoc.2024.111255

Wu, X., Chen, H., Wang, J., Troiano, L., Loia, V., & Fujita, H. (2020). Adaptive stock trading strategies with deep reinforcement learning methods. *Information Sciences, 538*, 142–158. DOI: 10.1016/j.ins.2020.05.066

Yang, H., Liu, X.-Y., Zhong, S., & Walid, A. (2020). Deep reinforcement learning for automated stock trading: An ensemble strategy. In *Proceedings of the First ACM International Conference on AI in Finance* (Article 31, pp. 1–8). Association for Computing Machinery. DOI: 10.1145/3383455.3422540

Yang, H., & Malik, A. (2024). Reinforcement learning pair trading: A dynamic scaling approach. *Journal of Risk and Financial Management, 17*(12), Article 555. DOI: 10.3390/jrfm17120555

Zhang, Z., Zohren, S., & Roberts, S. (2020). Deep reinforcement learning for trading. *The Journal of Financial Data Science, 2*(2), 25–40. DOI: 10.3905/jfds.2020.1.030

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- apatables.docx
- apafiguresfinalpngsandsourcedata.zip