LINC

LUND UNIVERSITY FINANCE SOCIETY EST 1991

TRADING & QUANTITATIVE RESEARCH REPORT

Reinforcement Learning for Bond Allocation in Global Pension Fund Portfolios

Balancing Duration and Credit Risk for Enhanced Risk-Adjusted Returns

In collaboration with:

AP3

Analysts: Edvin Gunnarsson, Victor Mikkelsen, Carolina Oker-Blom
Supervisor: Joakim Blomqvist

Introduction & Theory

Introduction
This project aims to optimize portfolio strategy for bond allocation by balancing duration and credit risk for enhanced risk-adjusted returns. Today, the Third Swedish National Pension Fund (AP3) has a portfolio strategy which is largely done manually. The operational challenges this causes create a demand for a solution that can work in tandem with an investor to provide a broad asset allocation strategy with the help of reinforcement learning.

Background and Theory
Reinforcement Learning (RL) provides a framework for portfolio optimization as it addresses the sequential decision-making nature of trading, where each rebalancing decision depends on current market conditions and historical performance [1]. The portfolio allocation problem can be formulated as a Markov Decision Process (MDP) where an agent, typically represented by a policy gradient algorithm such as Proximal Policy Optimization (PPO), learns to allocate capital across multiple assets by interacting with a financial market environment [2, 3]. However, the problem is not a true MDP since it assumes that the future state of the market depends only on the current state and action, and not on the full history of past observations [1]. In financial markets, this assumption does not strictly hold as asset returns may show long term dependencies, regime shifts, and latent macroeconomic factors. To account for this fact the state includes a fixed window of historical returns and indicators. This is a standard technique in applied RL that recovers an approximately Markov state by providing the agent with sufficient context to make informed decisions without requiring the full history of past observations [4].

Foundations. The market is represented by a *state space*. The state space represents the information available to the agent at each decision time and includes historical returns for all assets over a specified window, corresponding returns for relevant economic factors and market indicators, plus normalized portfolio state information including current portfolio value relative to initial balance and available cash balance [2]. The *action space* is continuous, representing portfolio weights that must sum to 1, allowing the agent to allocate capital across any set of available assets. It has the constraint that the action must satisfy both budget and non negativity conditions [3]. At each time step t, the agent selects a weight vector $\mathbf{w}_t = [w_1, w_2, \ldots, w_n]$ subject to the constraint $\sum_{i=1}^{n} w_i = 1$ and $w_i \geq 0$ for all i, where n is the number of available assets. The portfolio return at time t is computed as the dot product of the weight vector and the asset returns:

$$r_{p,t} = \mathbf{w}_t^T \mathbf{r}_t = \sum_{i=1}^{n} w_{i,t} \cdot r_{i,t} \tag{1}$$

where $\mathbf{r}_t = [r_{1,t}, r_{2,t}, \ldots, r_{n,t}]$ represents the returns of each asset at time t.

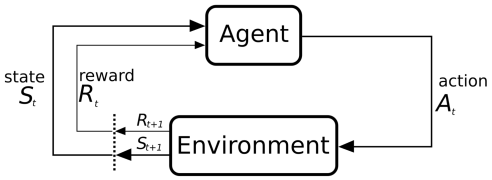

Figure 1: The agent–environment interaction feedback loop in reinforcement learning [1].

Transaction costs, C_t, are explicitly modeled as a percentage of portfolio value proportional to the absolute change in position weights [2, 3]:

$$C_t = c \cdot V_t \cdot \sum_{i=1}^{n} |w_{i,t} - w_{i,t-1}| \tag{2}$$

where c is the transaction cost rate and V_t is the portfolio value at time t. The portfolio value evolves according to:

$$V_{t+1} = V_t \cdot (1 + r_{p,t}) - C_t \tag{3}$$

The transaction cost is important from an RL perspective to discourage the model from excessive rebalancing of the portfolio. The agent is incentivized to trade only when the expected performance improvement outweighs the incurred costs. This also aligns the learned strategy with realistic trading behaviour observed in practice. The reward function combines immediate portfolio returns with a risk-adjusted component based on the Sharpe ratio of recent returns. The Sharpe ratio is a widely used risk adjusted performance measure defined as the ratio between excess return and volatility. This encourages the agent to not only maximize returns but maintain a favorable risk-adjusted-return [4, 3]. The Sharpe ratio, R_t, is given by

$$R_t = 100 \cdot r_{p,t} + \alpha \cdot \frac{\bar{r}_{recent}}{\sigma_{recent} + \epsilon}, \tag{4}$$

where $\bar{r}_{recent}$ and σ_{recent} are the mean and standard deviation, respectively, of returns over a recent time window; α is a scaling factor that determines the relative importance of risk-adjusted returns, and ϵ is a small constant that prevents division by zero.

Equation 4 illustrates the general structure of a risk-adjusted reward signal; the exact formulation used in this project, which replaces the raw return with

excess return over the risk-free rate and annualizes the volatility scaling, is given in Equation 14 in the Reward Function subsection.

Portfolio strategy. The theoretical foundation draws from Modern Portfolio Theory. A more diverse portfolio minimizes risk because security returns are often correlated. Portfolios containing highly correlated securities are very sensitive to market fluctuations, so diversification increases the chances of more stable returns [5]. The risk of bond investments are primarily driven by interest rate fluctuations and credit spread movement. Credit spread risk arises from the possibility that the yield premium demanded by investors for holding non-government debt widens, reducing bond prices independently of changes in the risk-free rate. High yield bonds carry substantially greater spread sensitivity than investment grade bonds, causing them to behave more like equities during stress regimes, a distinction that is critical for portfolio construction when rate shocks and credit shocks coincide [6]. Managing these risks help avoid losses and protect investor interests. To quantify an asset's sensitivity to interest rate changes, practitioners commonly use duration and convexity. Duration measures the linear price sensitivity of a bond to fluctuations in the risk-free rate, while convexity accounts for the non-linear curvature of this price-yield relationship [7, 8].

The RL approach is an extension of Modern Portfolio Theory by including dynamic rebalancing strategies that adapt to changing market conditions, rather than relying on static mean-variance optimization [3]. To ensure stable learning dynamics, asset returns are typically normalized by their historical mean and standard deviation [2]. For each asset i, the normalized return $\tilde{r}_{i,t}$ is computed as:

$$\tilde{r}_{i,t} = \frac{r_{i,t} - \mu_i}{\sigma_i + \epsilon} \tag{5}$$

where μ_i and σ_i are the historical mean and standard deviation of asset i's returns computed over the training dataset. The dataset is typically split into training and validation sets, so that the agent learns patterns from historical data and is evaluated on out-of-sample performance [2]. To interpret which features drive the learned policy, permutation importance is used as a model-agnostic explainability method. For each feature, the feature values are randomly shuffled across the test period and the resulting drop in Sharpe ratio is recorded. A large drop indicates that the feature was important to the policy; a negligible drop indicates redundancy. This approach requires no assumptions about the model architecture and directly measures the contribution of each input to out-of-sample performance [1].

Principal Component Analysis. To mitigate the risk of overfitting in high-dimensional state spaces, Principal Component Analysis (PCA) is applied as a linear dimensionality reduction technique. PCA transforms a set of correlated variables into a smaller set of linearly uncorrelated features known as principal components. Mathematically, this is achieved through an orthogonal transformation that projects the data along the directions of maximum variance, which correspond to the eigenvectors of the data's covariance matrix. The first component (PC_1) is oriented to capture the highest possible variance, while each subsequent component captures the remaining variance under a strict orthogonality constraint (see Appendix II-B). In fixed-income analysis, this decomposition simplifies correlated macroeconomic factors and asset returns into a few stable components representing underlying market regimes [9].

In fixed-income markets, the first three principal components of the yield curve are classically associated with *level*, *slope*, and *curvature* shifts [10], providing an economically interpretable basis for dimensionality reduction.

Proximal Policy Optimization. PPO is an actor-critic algorithm that maintains two networks: a *policy network* (actor) that maps states to actions, and a *value network* (critic) that estimates the expected cumulative reward from a given state. The advantage $A_t = R_t - V(s_t)$ measures how much better or worse an action performed relative to the critic's expectation $V(s_t)$. A positive advantage prompts the policy to increase the probability of that action; a negative advantage prompts a decrease.

Importantly, training stability is ensured through the objective L^{CLIP}, which penalises large deviations from the previous policy.

$$\text{clip}(r_t(\theta),\, 1-\varepsilon,\, 1+\varepsilon) \tag{6}$$

where $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ is the probability ratio between the new and old policy and ε is the clip range. The clipping prevents the ratio from moving outside $[1-\varepsilon, 1+\varepsilon]$, bounding the policy update and improving stability in noisy environments such as financial markets [1].

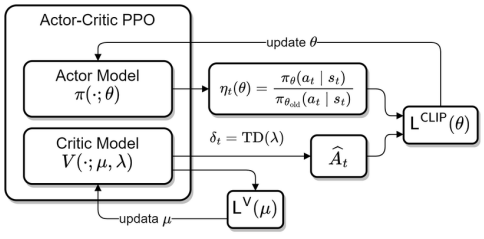

Figure 2: The actor-critic model. The actor maps the current state to portfolio weights, while the critic estimates future rewards to guide training [11].

The PPO algorithm is particularly well-suited for this continuous control problem as it accommodates the continuous action space required for portfolio weight allocation [3]. The agent's objective is to maximize the expected cumulative discounted reward:

$$J(\pi) = E_\pi \left[\sum_{t=0}^{T} \gamma^t R_t \right] \tag{7}$$

where $\gamma \in [0, 1]$ is the discount factor and T is the episode length. This objective translates to finding a policy $\pi(\mathbf{a}_t|\mathbf{s}_t)$ that balances immediate returns, risk-adjusted performance metrics, and transaction cost efficiency over the entire trading window, making it suitable for automated portfolio optimization that can adapt to changing market sentiment and economic conditions [4, 3].

The theoretical framework relies on several simplifying assumptions. The agents trades do not affect market prices, it has zero market impact. Sufficient market liquidity is assumed at all times, allowing the agents actions to be carried out without delay. These assumptions are standard in theoretical portfolio optimization and make the learning problem manageable while retaining the essential parts of market conditions.

Data & Method

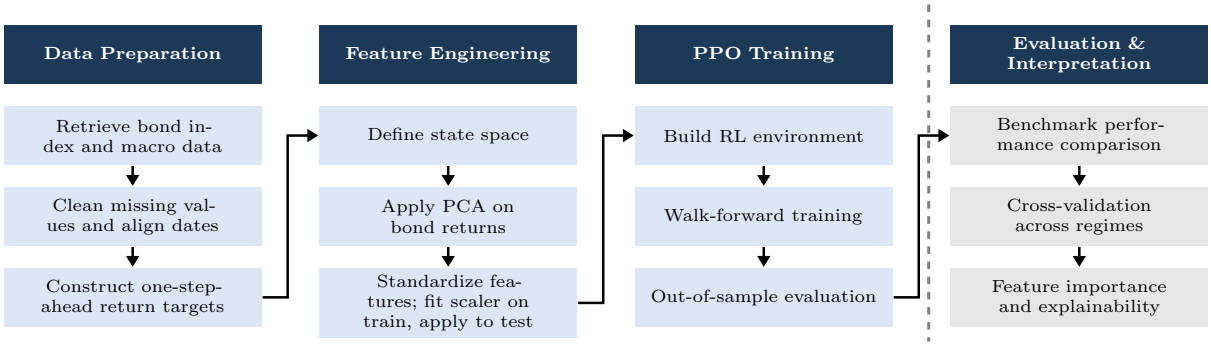

Figure 3: Flowchart of the PPO-based portfolio allocation methodology.

Data

For this project, the data used have all been gathered from Bloomberg ensuring high quality and reliable financial time series data. The dataset consisted of daily Total Return Index data for government and corporate bond indices. The asset pool was categorized into four segments:

1. US Government bonds (USD)
2. Pan-European Government bonds (USD Hedged)
3. Corporate credit (US Investment Grade, US High Yield, Pan-European High Yield Hedged)
4. Cash / risk-free proxy (US 1–3M Treasury Bills, BIL_US)

All data points were collected daily over the period from $[2005-11-21]$ to $[2025-11-19]$. In addition to the tradable asset series, the feature set included daily macro/market indicators (oil, equity indices, FX, policy rates, volatility indices) and monthly macro indicators (inflation, money supply, industrial production, fiscal indicators, housing-related activity) (details of the used macroeconomic features are given in Appendix I-B). The dataset also includes bond risk-sensitivity inputs in the form of duration and convexity, loaded from dedicated Bloomberg folders and matched to each asset ticker.

Method

The method used for this project was structured into five main parts. The first included data preprocessing, the second, feature engineering. The third part detailed the design of the RL environment, structured as a MDP which consisted of the state space, action space, and reward function. The fourth part covered model training and hyper parameter tuning and the fifth and final part the evaluation framework used.

1. Data Processing

Preprocessing. The raw financial time series data gathered from Bloomberg required synchronization and transformation to construct a valid reinforcement learning environment. The preprocessing pipeline, implemented in Python, was executed in three primary stages:

Firstly, the raw asset data, consisting of the daily Total Return Index P_t for each bond, was used to calculate daily simple returns according to $R_t = \frac{P_t}{P_{t-1}} - 1$. The dataset was then merged into a master data frame sorted by date, while handling duplicated timestamps and removing invalid values to ensure continuity.

Secondly, the data frame was combined with macro and market factors (daily and monthly) using backward merge logic, so each trading day only receives information available at or before that date. Duration and convexity were similarly mapped to each tradable bond ticker and merged with strict backward alignment to avoid look-ahead bias.

Thirdly, supervised labels were constructed as one-step-ahead returns, i.e. $X_t \rightarrow R_{t+1}$, by shifting each asset return column forward by one day. The sample was then split chronologically into training and testing periods (train up to 2019-12-31, test from 2020-01-01 onward). To ensure stable convergence of the PPO algorithm, input features were standardized using Z-score normalization based on training-set statistics only, and the same scaler was applied to the test set.

2. Feature Engineering

Interest Rate Risk Metrics: Duration and Convexity. Unlike human portfolio managers, the RL agent has no concept of a bond's maturity or its mathematical relationship to interest rate changes. To address this, duration and convexity were added as time varying features.

Duration Tilt. To encourage interest-rate-aware allocation, a duration tilt signal was constructed at each time step as the duration-weighted deviation of the current portfolio from a neutral baseline. It signals changes in the Federal Funds Target Rate (FDTR), CPI inflation, and the MOVE bond volatility index, normalized as a rolling z-score over a 252-day window. At each time step, assets were ranked by their current duration (percentile rank), and portfolio weights were rescaled according to:

$$\tilde{w}_i = w_i \cdot (1 - \lambda \cdot s_t \cdot 2(r_i - 0.5)) \tag{8}$$

where w_i is the model's raw weight, $r_i \in [0, 1]$ is the duration percentile rank of asset i, $s_t \in [0, 1]$ is the normalized stress level, and $\lambda = 0.12$ is the tilt strength. When stress is elevated, the signal reduces weights on high-duration assets and increases weights on low-duration assets. This signal was included as an additional feature in the state vector, providing the agent with a direct measure of its aggregate interest rate exposure. Weights are renormalized to sum to one after the adjustment.

Principal Component Analysis. Firstly, Principal Component Analysis (PCA) was performed on the bond universe returns. The selection of the first two principal components was based on their collective ability to account for the vast majority of the portfolio's variance. As demonstrated in Table 1, PC_1 and PC_2 explain 71.22% of the total variance during static analysis. In fixed-income analysis, these components are traditionally associated with the level and slope of the yield curve. By excluding these higher order components, we effectively perform dimensionality reduction that prevents the RL agent from overfitting.

Table 1: PCA variance explained by component (PC1–PC2 retained; see Appendix II-C for full decomposition).

Comp.	Variance (%)	Cum. (%)
PC1	45.2956	45.2956
PC2	25.9252	71.2208

The loading comparison in Figure 4 supports excluding higher-order PCs: after PC2, components are materially noisier and less stable across windows. Keeping only PC1 and PC2 improved robustness while preserving most of the information content.

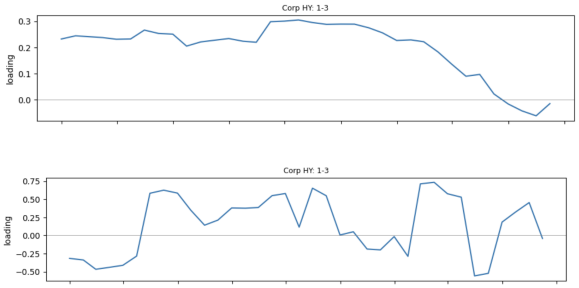

Figure 4: Example PCA loading stability comparison used for component selection. PC1 (top) is stable and interpretable, while PC4 (bottom) and continuing up to PC4 are visibly choppy/high-variance. Therefore, only PC1 & PC2 are retained in the model.

To prevent look-ahead bias, the PCA model was fitted exclusively on training data (prior to 2020-01-01). The resulting transformation parameters were then applied to both the training and the out-of-sample test data. For each macroeconomic feature, a rolling correlation against both of the PCs was calculated using a 104-week lookback window. The window length was chosen to capture long-term regime shifts while filtering out short-term market noise. Two model variants were evaluated: one using PCA-transformed bond returns as state features, and one retaining raw bond returns directly, to isolate the contribution of dimensionality reduction to out-of-sample performance.

3. Reinforcement Learning Framework

Environment Setup. The trading simulation was constructed using the Gymnasium library by implementing a custom environment that follows the standard Gymnasium interface. The environment represented a daily rebalancing bond allocator, where each time step corresponded to one trading day. At each step, the agent observed the

current state and outputted portfolio weights that were applied to next-period returns. Transaction costs were included in the environment dynamics, and BIL_US was used as risk-free proxy in reward computation.

State Space. At each trading day t, the state vector $\mathbf{s}_t$ contained standardized macro/market factors together with bond-sensitivity inputs (duration and convexity), for example:

$$\mathbf{s}_t = [\mathbf{x}_t^{macro}, \mathbf{x}_t^{market}, \mathbf{x}_t^{dur}, \mathbf{x}_t^{conv}]. \tag{9}$$

Furthermore, the environment included previous portfolio weights $\mathbf{w}_{t-1}$ as part of the observation to make the process closer to Markovian and to allow the policy to internalize trading frictions as

$$\tilde{\mathbf{s}}_t = [\mathbf{s}_t, \mathbf{w}_{t-1}]. \tag{10}$$

Action Space. At each trading day t, the agent chose a vector of portfolio *allocation scores*,

$$a_t \in [0,1]^N,$$

where N is the number of assets (columns) in the return matrix. Importantly, the asset universe included a proxy for the risk-free rate (BIL US) as the N:th asset, allowing the agent to reduce duration and credit exposure during stressed market regimes. In the implementation, the action space was defined as a continuous Box space:

$$\mathcal{A} = [0,1]^N.$$

These raw actions were *not* guaranteed to sum to one, so they were converted into valid portfolio weights by normalizing:

$$w_t = \frac{a_t}{\sum_{i=1}^{N} a_{t,i} + \varepsilon}. \tag{11}$$

Here, $\varepsilon = 10^{-8}$ was a small constant used only for numerical stability. It prevented division by zero in the rare case where the agent outputted $a_t = 0$ for all assets. After normalization, the weights approximately satisfied

$$w_{t,i} \geq 0 \quad \text{and} \quad \sum_{i=1}^{N} w_{t,i} \approx 1,$$

meaning the portfolio was long-only and (almost) fully invested. The resulting vector w_t represented target portfolio weights for the next period, and the environment computed turnover relative to w_{t-1} for cost deduction. If $N = 1$ (i.e., the return matrix contains only one asset column), then the action space becomes one-dimensional and the normalization makes the weight essentially $w_t \approx 1$ every time. In that case, the agent cannot meaningfully reallocate risk across assets because there is only one asset to hold.

To further limit excessive turnover, portfolio weights were exponentially smoothed toward the previous allocation at each step using a convex combination $w_t^{\text{smooth}} = \alpha \cdot w_t + (1-\alpha) \cdot w_{t-1}$, where α controls how quickly the portfolio adapts to new signals. During the warmup period $\alpha = 0.6$ was used, giving equal weight to both the new and previous allocation, allowing faster initial exploration. Once the portfolio had stabilized, α was increased to 0.9, retaining 90% of the new signal while anchoring 10% to the prior allocation to suppress noise-driven turnover.

Reward Function. The reward function was designed to balance profitability, risk-adjusted performance, and trading efficiency. Firstly, portfolio net return was computed after transaction costs and then converted to excess return by subtracting the risk-free proxy (BIL total return index), denoted $r_{f,t}$:

$$r_t^{excess} = r_{p,t}^{net} - r_{f,t}. \tag{12}$$

Intuitively, this meant the agent was only rewarded for performance *above* what it could have earned by staying in the risk-free asset.

Next, risk was accounted for by scaling excess return by recent volatility. The environment kept a rolling history of recent returns (in the implementation, a 60-trading-day lookback) and computed the standard deviation:

$$\sigma_t = \text{std}\big(r_{p,t-L+1}^{net}, \ldots, r_{p,t}^{net}\big) + \epsilon, \tag{13}$$

where $L = 60$ and ϵ was a small constant (e.g., 10^{-6}) added to avoid division by zero.

This made the reward higher when the agent achieved the same excess return with lower fluctuations, and lower when the returns were unstable. Finally, the risk-adjusted signal was annualized and defined as a Sharpe-like reward, where $days = 252$:

$$R_t = \frac{r_t^{excess}}{\sigma_t}\sqrt{days}. \tag{14}$$

The Sharpe ratio was selected as the primary reward function because it provided a well-established, computationally efficient baseline for evaluating risk-adjusted performance. An additional cash penalty of 0.02 was subtracted from the reward proportional to the weight allocated to the risk-free proxy (BIL US), discouraging the agent from holding excessive cash and incentivizing active allocation across the bond universe.

Table 2: Environment and execution parameters.

Parameter	Value
Lookback window	60 steps
Steps per year	252
Transaction fee	1 bp
Cash penalty	0.02
Smoothing (steady)	0.9 / 0.1
Smoothing (warmup)	0.6 / 0.4
Duration tilt strength	0.12

4. Model Training

Training Procedure & Hyperparameter Selection. To prevent overfitting to recent market trends, the PPO model was trained using a Walk-Forward methodology with Regime Randomization. Instead of sequential training, the 3-year historical windows were reshuffled each epoch. This exposed the model to diverse macroeconomic conditions, forcing it to generalize its policy rather than memorize a single timeline.

In each training run, the model was exposed to rolling 3-year windows (`WF_TRAIN_YEARS=3`) stepped by 1 year (`WF_STEP_YEARS=1`), repeated for 2 randomized epochs (`WF_EPOCHS=2`). The total training budget of 350,000 timesteps was distributed across fold-updates using

$$\texttt{per\_fold} = \max\left(1024, \left\lfloor \frac{\texttt{total\_timesteps}}{\#\texttt{windows} \times \texttt{WF\_EPOCHS}} \right\rfloor\right)$$

and fold order was reshuffled each epoch to reduce path dependence and overfitting to a single historical sequence. The discount factor (gamma) $\gamma = 0.99$ was chosen to reflect long-horizon reward accumulation, appropriate for a daily rebalancing strategy where decisions compound over multi-year periods. The entropy coefficient $\varepsilon_{\text{ent}} = 0.01$ was set to maintain sufficient exploration without preventing policy convergence as too high a value prevents the agent from committing to allocations, while too low risks premature convergence to a suboptimal policy. All remaining PPO parameters were left at Stable-Baselines3 defaults, which are well-established for continuous control tasks.

350,000 timesteps were selected empirically to balance learning and overfitting. While the mean episode reward continues to grow stochastically beyond 350,000 steps, the value function reaches an explained variance of ≈0.99 and both train and value losses stabilize well within this range, indicating that the value function has converged and further training yields diminishing returns. More precisely, `value_loss` converged to near zero while `ep_rew_mean` continued rising, indicating the value function had fully learned the state space but the policy was continuing to fit to training-specific return sequences. This follows the principle that policy updates without corresponding value function improvement represent memorization rather than generalization. A comparison against a 512,000 timestep baseline confirms negligible improvement across all performance metrics beyond this point (see Appendix II-A).

Table 3: PPO training hyperparameters (See full setup including Stable-Baselines3 defaults in Appendix II-C).

Hyperparameter	Value
Training timesteps	350,000
Walk-forward window	3 years
Walk-forward step	1 year
Walk-forward epochs	2
gamma	0.99
ent_coef	0.01

Choosing PPO as Algorithm. While the theoretical foundation of PPO is outlined in the background section, its practical application was deliberately chosen for this project due to its structural advantages in handling financial time-series data. Portfolio allocation requires continuous action spaces to assign weights to various bond assets. While other algorithms also support continuous spaces, PPO was selected primarily for its superior training stability. PPO is highly suited for noisy markets, as its clipped objective function restricts excessively large policy updates during sudden market changes.

Trading Frequency Variants. In addition to the primary daily rebalancing setup, weekly and monthly rebalancing variants were evaluated to assess the sensitivity of performance to trading frequency. To maintain comparable learning budgets, total timesteps were scaled proportionally: 350,000 for daily, 75,000 for weekly, and 17,000 for monthly, reflecting the ratios of 252/52 and 252/12 trading days per year respectively. A fixed random seed was used throughout training and evaluation to ensure reproducibility of all reported results.

5. Evaluation Framework

Time Series Cross-Validation. The walk-forward procedure described in Section 4 governs the *training* of the model across historical windows. To evaluate *out-of-sample* performance, a separate 5-fold time series cross-validation was applied, splitting the data chronologically into strictly separated training and future test periods to prevent data leakage and look-ahead bias [12].

The time-series cross-validation setup used five chronological folds with a fixed initial train start (2011-10-12) and expanding train end dates. Test windows moved forward without overlap leakage,

ending at 2025-11-18 in Fold 5 (See Appendix II-F). This provided a consistent out-of-sample sequence across changing market periods. Crucially, to prevent look-ahead bias during feature scaling, the standard scaler was fitted exclusively on the training data. The subsequent test data was then transformed using only these historical parameters.

Feature Importance. To interpret which features drive the learned policy, permutation importance was used as a model-agnostic explainability method. Each feature was independently shuffled across the test period and the resulting Sharpe drop recorded. Results were averaged over 30 independent permutations to reduce variance. A feature group ablation was also performed by zeroing out entire groups of related inputs to isolate the contribution of duration/convexity versus macro signals.

Transaction Cost Sensitivity. To assess the robustness of model performance to execution costs, the PCA model was repriced post-hoc at fee levels ranging from 1 bp to 50 bp per trade. Portfolio returns were recomputed at each fee level using the original allocation path, and the resulting Sharpe ratio was recorded. This isolates the impact of transaction costs from the learning process, since the policy itself was trained at a fixed 1 bp cost.

Performance Metrics. To evaluate out-of-sample performance against the benchmark, the following metrics were mainly utilized (assuming 252 trading days per year):

- Annualized Return: The geometric average yearly return of the portfolio,

$$R = \left(\prod_{t=1}^{N} (1 + R_t) \right)^{\frac{252}{N}} - 1,$$

  (where R_t was the daily return and N was the total number of trading days).
- Sharpe Ratio: The annualized risk-adjusted return, calculated using the mean daily return (μ) and standard deviation (σ),

$$\text{Sharpe} = \frac{\mu}{\sigma} \sqrt{252}.$$

- Max Draw Down: Measured as the largest peak-to-trough drop in equity,

$$MDD = \min_t \left(\frac{E_t - P_t}{P_t} \right),$$

  (where E_t was the equity value and P_t was the highest historical peak up to time t).
- Benchmark: The Bloomberg Global Aggregate Total Return Index (USD Hedged, ticker: LEGATRUH), sourced from Bloomberg, representing a broad investment-grade global bond market index used as the performance reference throughout the evaluation.

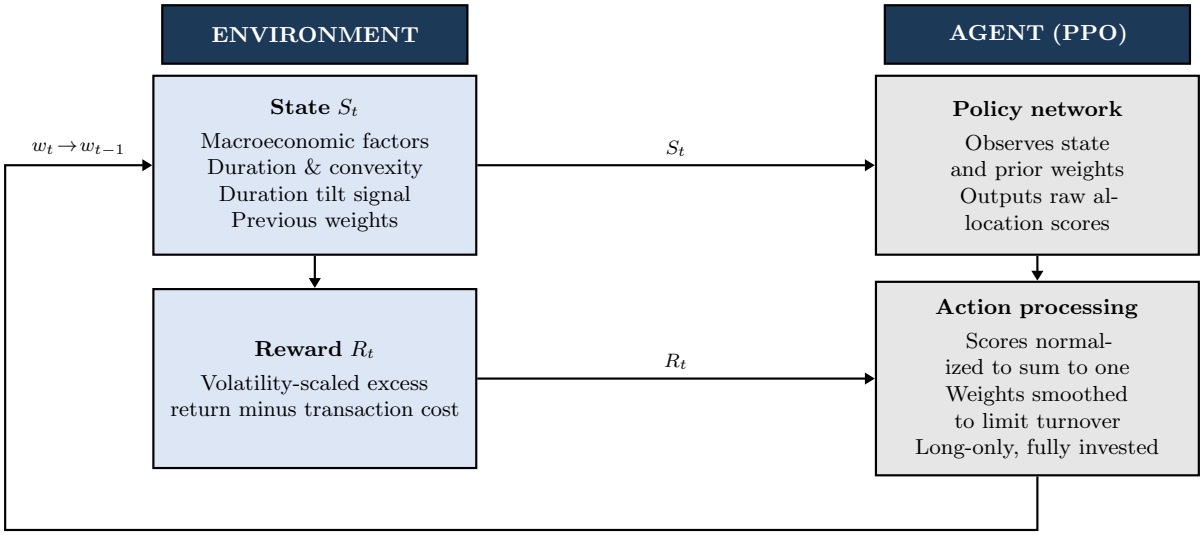

Figure 5: Reinforcement learning framework for bond portfolio allocation.

Results

1. Performance

In this section, the results of two distinct PPO configurations, one utilizing the full set of 19 macroeconomic features ("PPO no PCA") and one employing the PCA-based correlation framework ("PPO PCA"), are evaluated against the LEGATRUH benchmark. Both models incorporate bond-specific risk metrics, duration and convexity, to navigate varying interest rate environments.

As summarized in Table 4, both PPO agents significantly outperform the benchmark across nearly all risk-adjusted metrics. The PCA configuration achieved the highest annualized return of 6.8% and a Sharpe ratio of 1.25, compared to the benchmark's 1.39% return and 0.38 Sharpe ratio. This outperformance suggests that the compressed state space created by the PCA framework may have acted as a noise filter, removing false correlations within the high-dimensional macro data and preventing the agent from overfitting to underlying market noise.

Table 4: Backtest performance metrics comparison, PCA, no PCA and Benchmark.

Metric	PCA	No PCA	Bench.
Annual ret.	0.0680	0.0573	0.0139
Cum. ret.	0.4405	0.3623	0.0795
Ann. vol.	0.0537	0.0547	0.0380
Sharpe ratio	1.2520	1.0468	0.3816
Calmar rat.	0.4442	0.3621	0.1023
Stability	0.5065	0.5096	0.0114
MaxDD	-0.1531	-0.1583	-0.1357
Omega ratio	1.2592	1.2067	1.0654
Sortino ratio	1.8477	1.5000	0.6125
Skew	0.4655	0.1202	0.1872
Kurtosis	5.9219	6.8168	1.5286
Tail ratio	1.0559	1.0546	0.9990
Daily VaR	-0.0052	-0.0054	-0.0039

Table 4 shows that both PPO configurations outperform the benchmark on return and risk-adjusted metrics over the full test sample. Both PPO variants also report higher Sortino and Omega ratios than the benchmark, with different drawdown and volatility profiles.

To stress test robustness in a difficult macro regime, Table 5 isolates the 2022–2024 window where rate shocks dominated bond market behaviour.

Table 5: Regime performance comparison 2022–2024: PPO no PCA vs PPO PCA vs benchmark.

Metric	PCA	No PCA	Bench.
Annual ret.	0.018	0.0168	0.0005
Cum. ret.	0.0533	0.0479	0.0016
Ann. vol.	0.0653	0.0677	0.0467
Sharpe ratio	0.3025	0.2687	0.0347
Calmar rat.	0.1164	0.1014	0.0044
Stability	0.2349	0.5837	0.3496
MaxDD	-0.1527	-0.1579	-0.1205
Omega ratio	1.0525	1.8449	1.0055
Sortino ratio	0.4949	0.4370	0.0623
Skew	1.3585	0.8738	0.2353
Kurtosis	3.3456	4.6634	3.5015
Tail ratio	1.0735	1.0546	1.0821
Daily VaR	-0.0061	-0.0064	-0.0045

In the 2022–2024 subset, the two PPO configuration columns show similar performance levels across most metrics. The "PPO PCA" reports higher annual return and Sharpe than "PPO no PCA", and both are above the benchmark column in the same window. Maximum drawdown and daily value at risk remain negative for all three columns, with similar magnitudes.

2. Returns and Allocations

Below follows results obtained from a qualitative analysis of the PPO configurations allocation paths. Figure 6 reports the full out-of-sample path for both configurations. The "no PCA" model (Figure 6 left) shows broader and more frequent reallocations across sleeves, while the PCA model (Figure 6 right) exhibits longer persistence in a smaller set of dominant allocations before rotating. Both paths include clear regime shifts during 2022–2024, followed by re-risking behavior in the later part of the sample.

The top-row cumulative return plots show both strategies ending above the benchmark path over the displayed horizon. The bottom-row allocation panels show persistent time variation in sleeve weights rather than static allocations. Together, the panels document both return-path and allocation-path outputs for the two PPO specifications. A detailed breakdown of returns is provided in Appendix II-E.

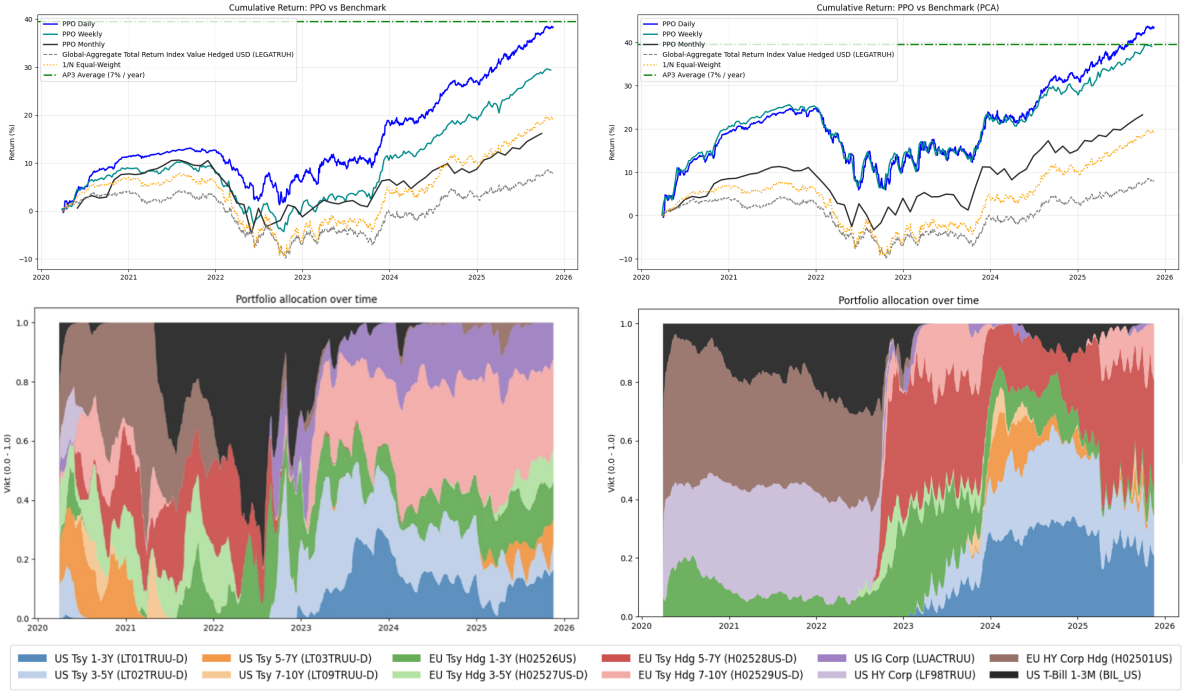

Figure 6: Out-of-sample PPO comparison. Top row: cumulative returns (no PCA left, PCA right). Bottom row: time-varying portfolio allocations (no PCA left, PCA right).

3. Cross-Validation and Robustness

To verify that these results were not merely a product of overfitting to a specific timeframe, a 5-fold Time Series Cross-Validation was performed (see Figure 7). The diagnostic plots (Figure 7) reveal that both models maintain a positive average Sharpe ratio across folds (1.09 "no PCA" vs 1.23 "PCA" vs 1.10 Benchmark), though performance varies significantly depending on the market regime.

A critical observation is found in Fold 4, corresponding to the aggressive rate-hiking cycle of 2021–2023. During this period, both PPO models and the benchmark experienced negative Sharpe ratios. This underscores a fundamental limitation in long-only bond strategies: during periods of rapid, systemic interest rate shocks, diversification across duration and credit sleeves provides limited protection, as correlations often converge toward unity.

Feature ranking differs materially between the two periods. In the full sample, bond-specific metrics (Convexity US IG Corp, MOVE Index) and housing indicators (Case-Shiller) dominate at modest magnitudes, with importance spread across many features. Under the 2022–2024 stress regime, the MOVE Index alone accounts for 22.3% of the baseline Sharpe, with importance highly concentrated in the top three features. To further isolate the contribution of each feature group, Table 6 reports the Sharpe drop when each group is zeroed out across both model variants. The dashes mean the mutually exclusive state representations: the no-PCA model uses raw macro features directly, while the PCA model replaces them with rolling correlations against the principal components

Table 6: Feature group ablation: Sharpe ratio change (%) when each group is zeroed out (set to training mean in standardised space).

Feature Group	No-PCA (%)	PCA (%)
Duration / Convexity	−31.7	−9.3
Macro / Stress	−13.5	—
Corr Overall PCs	—	−9.7

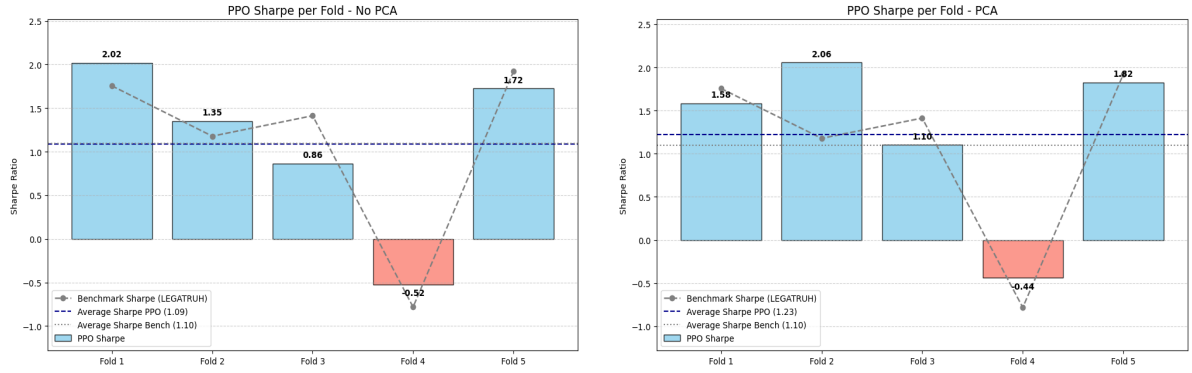

Figure 7: Cross-validation. Left: no-PCA model. Right: PCA model. The testing phase starts in 2014, utilizing five folds with a standard duration of approximately two years each. (Specific date ranges for training and testing sets are detailed in Appendix II-F).

Transaction costs. The PCA model is repriced post-hoc at fee levels from 1 bp to 50 bp per trade. As shown in Figure 8, the Sharpe ratio degrades linearly from 0.97 at 1 bp to 0.47 at 50 bp, with no threshold at which returns collapse. This reflects mean daily turnover of 2.3%, confirming the strategy is robust to realistic spread widening.

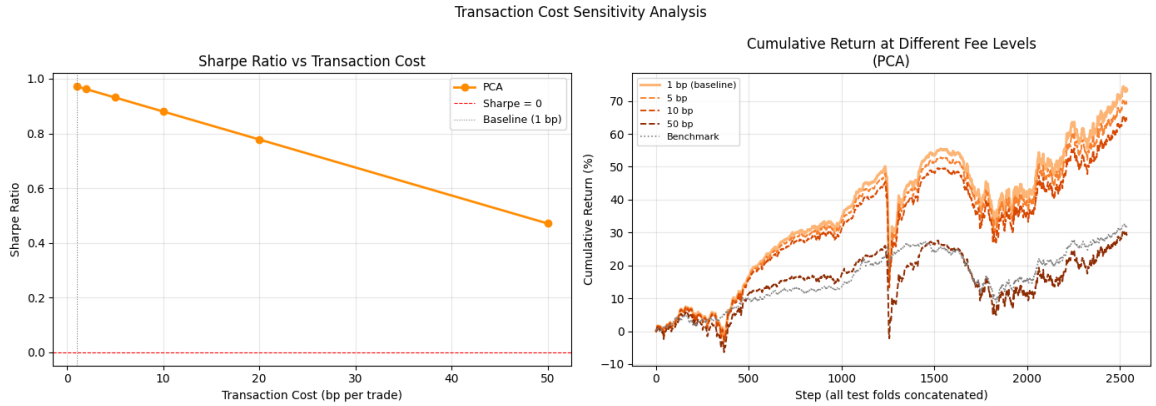

Figure 8: Sharpe ratio as a function of transaction cost per trade (PCA model). The strategy remains profitable up to 50 bp per trade.

Analysis & Conclusion

Analysis

Model Performance and Benchmark Comparison. The final "no PCA" model achieved an annualized return of 5.7%, and "PCA" achieved 6.8%, which are on par with AP3's annual average return of 7%. This represents a significant evolution from the initial iterations, which relied solely on a broad set of macro indicators. The no-PCA variant reached 5.7% annualized at a Sharpe of 1.05. Both configurations materially outperform the LEGATRUH benchmark (1.4% annualized, Sharpe 0.38) across all reported risk-adjusted metrics including Sortino ratio, Omega ratio, and Calmar ratio (Table 4).

While the early models struggled to generate actionable signals from noisy global data, the inclusion of asset specific features specifically duration and convexity, provided the necessary structure. The performance gap between PCA and no-PCA is consistent with the hypothesis that dimensionality reduction acts as a noise filter. By compressing 19 correlated macro features into two stable principal components capturing the level and slope of the yield curve, the PCA model reduces the risk of the policy overfitting spurious correlations present in the training sample.

While the current model retains only PC1 and PC2, the next two components jointly account for a meaningful share of remaining variance (Table 10). Including them would give the agent more information to train on, such as yield curve curvature, and may improve the Sharpe in calmer periods. But as Figure 4 shows, the loadings on higher-order PCs are unstable and shift depending

on the window. In practice that means the agent would be learning from features whose meaning changes over time, which is essentially adding noise back in. This is most dangerous during regime transitions like 2022, where the correlation structure breaks down and unreliable inputs are more likely to confuse the policy than help it. Keeping only two components sacrifices some explanatory power but gives the agent a cleaner, more stable view of the market. This is a tradeoff we think favors robustness over fit.

Allocation Dynamics and Regime Adaptation. As demonstrated in the allocation plots (Figure 6), the PPO models exhibit defensive behaviors that align with classic 'Flight to Quality' strategies. A defining characteristic of the model without PCA is its proactive rotation into the risk-free proxy during periods of extreme market stress. This is most evident during the 2022 rate-hiking cycle, where the agent shifted nearly 100% of the portfolio to avoid the systemic drawdown in long duration bonds. Post 2023, as macro regimes stabilized, the agent successfully reallocated capital across sovereign and credit sleeves, capturing the "risk on" momentum.

The PCA model shows a different pattern. During 2020–2022, it concentrates heavily in high-yield credit rather than rotating into cash. With only two principal components in the state, assets appear more similar to the agent, and the policy converges toward whichever sleeve offered the best risk-adjusted return during training which was HY in the low-rate environment. From 2022 onward, as short-end yields rose sharply, the model reduced HY exposure and shifted toward shorter-duration sovereign bonds. That this tilt persisted out-of-sample and contributed positively to Sharpe suggests it reflects the rate environment rather than overfitting to training data. That said, the specific allocation paths shown here are sensitive to the training seed, and different runs can produce materially different weight trajectories while showing similar aggregate performance (see Limitations).

Feature Importance and Regime Interpretation. Feature importance analysis (Tables 11 and 12) reveals a clear regime shift in the features driving policy decisions. Over the full test sample, Convexity of US Investment Grade Corporates (7.6%) and the MOVE Index (4.8%) account for the largest Sharpe drops, with importance spread across many features at modest magnitudes, suggesting the policy draws on a broad set of signals in normal conditions. In the 2022–2024 stress window, the MOVE Index dominates at 22.3%, followed by the Case-Shiller index (12.8%) and Convexity US IG Corp (10.3%), with importance far more concentrated in the top features. This concentration reflects the degree to which the rate-hiking cycle created a single dominant macro regime signal, and is consistent with the observed defensive allocation behavior during that period.

It is however worth noting that permutation importance is one of several approaches to measuring feature contributions, and alternative methods such as SHAP values or drop-column importance can produce different rankings depending on how they handle feature interactions and correlations [14].

The feature group ablation (Table 6) further supports the central role of bond-specific risk metrics: removing duration and convexity causes a 31.7% Sharpe decline in the no-PCA model and a 9.3% decline in the PCA model. The smaller impact in the PCA variant reflects the partial redundancy between duration features and the PCA correlation components, which already reflect yield curve sensitivity indirectly.

Performance Robustness and Cross-validation. The results for the cross validation (see Table 6) show that the PPO model outperforms the benchmark in terms of average annualized return. It shows the PCA model maintaining a positive average Sharpe of 1.23 across folds versus 1.09 for no-PCA and 1.10 for the benchmark. However, performance is highly regime-dependent. The high standard deviation (approximately 1.2 for both model configurations) in performance across folds highlights the challenge of RL in non stationary markets. Fold 4 (2021–2023), coinciding with the most aggressive Federal Reserve tightening cycle in four decades, produced negative Sharpe ratios for both PPO models and the benchmark. This may be attributed to a structural limitation of long-only bond strategies, namely that when yields rise systemically across all maturities, cross-asset correlations converge and duration diversification provides limited protection.

During the aggressive rate-hiking cycle of 2022–2024 (fold 4) both PPO models and benchmark suffered (Sharpe $-0,44$ (PCA) and $-0,52$ (no PCA)). It is important to note that the PCA and no PCA had a similar Sharpe according to Figure 7. The similar drawdown demonstrates that while the agent learns defensive actions, and PCA helps eliminate noise, certain systematic shocks are too large for many traditional portfolio strategies.

The explainability section in Appendix II-D support the regime interpretation in the main results. The permutation importance results (Tables 11 and 12) indicates that policy decisions are mainly driven by allocation weights together with duration/convexity and macro-policy signals, which is consistent with rate-risk-sensitive reallocation be-

havior. The feature-importance comparison and correlation matrices (Appendix II-B) show that the 2022–2024 structure differs materially from the full sample, with stronger defensive weight concentration and shifted dependency patterns, supporting the observed performance dispersion across regimes.

In the most recent period (2023–2025, fold 5), the PPO models significantly outperformed the benchmark (Sharpe 1.82 (PCA) and 1.72 (no PCA). This suggests that the model is particularly good at identifying once a new macro regime has stabilized and also what risks to take during such a regime.

The 2022–2024 asset–factor matrix provides direct evidence for the defensive cash allocation behavior. In this matrix, the risk-free sleeve (BIL_US) shows consistently low mean rolling correlation magnitudes versus macro/market factors relative to credit sleeves (especially HY), which show materially higher factor coupling. This is consistent with the allocation paths, where BIL weight increases during the stress phase (2021–2023) and declines as the regime normalizes after 2023 and risk sleeves are reintroduced. Since the heatmap reports correlation magnitude, this supports differences in sensitivity strength rather than directional sign.

Transaction Cost Robustness. The strategy's low mean daily turnover of 2.3% makes it resilient to transaction cost uncertainty. Post-hoc repricing at fee levels from 1 bp to 50 bp per trade shows a gradual linear Sharpe decline from 0.97 to 0.47, with no threshold at which performance collapses (Figure 8). At 10–20 bp — a realistic range for stressed bond market spreads — the Sharpe remains above 0.78, meaning the performance doesn't depend on unrealistically low trading costs.

Risk Mitigation and the Duration Tilt. Despite the improvements gained from duration based features, the model initially struggled with the significant drawdowns of the post-Covid era (2022). To address this, we implemented a heuristic signal termed a "duration tilt." This adjusted the agent's allocations based on market stress levels, specifically reducing exposure to high duration assets during periods of rising inflation and interest rate volatility.

Both PPO configurations experienced a maximum drawdown of approximately 15%, somewhat larger than the benchmark's 13.6%. The duration tilt reduced exposure to long-duration bonds during stress but could not fully offset the systemic 2022 sell-off, where yields rose sharply across all maturities simultaneously. We believe that incorporating this classical risk management principle during training nonetheless helped coerce the agent into a more robust understanding of market dynamics. Given that high duration bonds are more sensitive to interest rate hikes, this tilt could act as a defensive shield, perhaps if the tilt sensitivity was increased. From a pension fund perspective, the higher maximum drawdown warrants careful consideration. That said, both drawdowns hit during the same 2022 rate-hiking episode so the agent isn't creating new risk, it's taking slightly more of the same systematic risk. The Calmar ratio (0.44 PCA vs 0.10 benchmark), however, shows the extra drawdown is well compensated.

Portfolio Stability and Noise Reduction. A recurring challenge in the development phase was the high turnover and erratic nature of the agent's allocations. To stabilize the portfolio, we implemented three distinct strategies:

State Space Feedback: Including lagged portfolio weights in the state vector made the policy path-dependent, reducing unmotivated rebalancing.

Trading Constraints: We imposed a restriction on the maximum allowable change in portfolio weights per time step. Exponential smoothing of weights (0.9/0.1 in steady state) directly penalized high turnover. This mimics real world constraints where a rational asset manager avoids liquidating entire positions daily. The turnover profile in Figure 12 confirms this works in practice, average daily weight changes spike during the 2022 stress period as the agent rotates defensively, then fall back to baseline once the regime stabilizes. While this could theoretically limit the agent's responsiveness during unanticipated events, the model's performance remained adequate, likely aided by the high frequency of daily rebalancing.

Time step frequency: We experimented with daily, weekly, and monthly time steps. While monthly data proved too lagging to react to major market shifts, weekly data was more promising and offered a smoother allocation profile, the daily frequency ultimately yielded the highest returns. The choice between daily and weekly steps remains a trade-off between reactivity and practical execution costs in a real world setting.

The sensitivity of allocation behavior to training budget also deserves attention. As noted in Appendix II-E, reducing timesteps below 350,000 produces visibly more diversified weight paths, while the current budget leads to concentrated allocations in a small number of sleeves. This raises the question of whether the concentrated positions observed in the main results reflect a genuinely learned regime signal or partial overfitting to the training sample. The training diagnostics (Appendix II-A) suggest the latter is a real concern: beyond roughly 300,000 steps, the value function has

fully converged while the policy continues updating, consistent with the agent memorizing training-specific return sequences rather than learning generalizable patterns. A more diversified allocation profile at lower timesteps may actually represent a more generalizable policy, even if it scores slightly lower on in-sample reward.

Limitations and Methodological Integrity. It is important to acknowledge that the model's results are not yet conclusive. The high number of arbitrary parameters and the lack of live market testing mean that our assumptions, while economically grounded, remain unproven in a high stakes environment.

Furthermore, there is a risk of overtraining. Although we utilized isolated test data, the iterative nature of our development where decisions were informed by test performance introduces risk for overfitting. To verify the model's true generalization capabilities, further validation on entirely new, out of sample data would be required to confirm that the observed returns are robust and not merely a result of hyper parameter tuning against the test set. It should also be noted that results vary between training runs even with identical hyperparameters, as PPO is sensitive to random initialization and the stochastic ordering of walk-forward windows. The results reported here are based on a single fixed seed, and different seeds may produce materially different allocation paths and performance metrics.

The high cross-fold variance (standard deviation $\approx$1.2 in Sharpe across folds) indicates the policy has not achieved stable generalization across all regimes. The training set (2005–2019) covers primarily a low-rate, quantitative-easing environment, which means the 2022 hiking shock represents an out-of-distribution event for which the policy had limited prior exposure. Finally, the long-only constraint prevents the agent from directly hedging rate risk through short positions, a limitation that is most binding precisely when systemic shocks cause cross-asset correlations to converge.

Furthermore, the PCA decomposition was fitted exclusively on pre-2020 training data and held static across the test period. Given that the 2022–2024 correlation structure differs materially from the pre-COVID sample, the static components may not fully capture yield curve dynamics during the hiking cycle.

Lastly, monthly macroeconomic features were aligned by reference date rather than release date, which may introduce a minor look-ahead bias since indicators such as CPI and industrial production are typically published weeks after the reference period.

Conclusion

This project investigated the application of Deep Reinforcement Learning, specifically PPO, to optimize bond allocation within a global pension fund portfolio. By modeling the trading environment as a Markov Decision Process, we successfully developed an agent capable of dynamically balancing duration and credit risk across sovereign and corporate bonds, and a risk-free cash proxy.

The results demonstrate that incorporating asset specific risk metrics, such as duration and convexity, provides the agent with the essential information required to navigate complex fixed income environments. The PCA framework successfully compressed the macroeconomic state space into structural market regimes (specifically capturing the Level and Slope of the yield curve). Rather than losing key data points, this dimensionality reduction functioned as a noise filter that limited overfitting and directly contributed to the model's improved risk-adjusted returns across the entire sample period. Furthermore, the inclusion of lagged portfolio weights and transaction costs successfully penalized excessive turnover, promoting path-dependency and leading to distinct, persistent regime-allocation shifts during periods of market stress.

The 5-fold Time Series Cross-Validation provided a robust assessment of the model, revealing an average outperformance over the benchmark in terms of both annualized return and Sharpe ratio. However, the high variance across folds highlights the inherent challenge of RL in non-stationary financial markets. While the agent excelled in stabilizing and capturing alpha in post-crisis recovery phases (2023–2025), it faced significant challenges during the aggressive rate-hiking cycle of 2022. This period underscored a fundamental limitation, rate-hikes of such magnitude are nearly impossible to fully mitigate, even with more advanced defensive methods.

The project was constrained by certain parameters. One such limitation was the lack of access to more advanced data. We only had publicly available datasets. In a scenario involving access to specialized industry data, the agent's predictive capabilities could have been further improved.

Ultimately all of these results are based on historical data, to fully assess the model, a live trading test would have to be carried out. Only then could one find out whether the hypothesis made in the earlier parts of this text hold true. It is very much possible that real market dynamics such as slippage and limited liquidity restricts the model to such a degree that no meaningful use can be made of it.

To improve the generality and robustness of the model in future applications, the optimization process could try to expose the agent to a wider variety of extreme market conditions. Future work should explore Sliding Window PCA and Rolling/Expanding training windows, because by continuously updating the model, the agent would be able to adapt to new regimes, such as the 2022 rate hikes, much faster. Additionally, the state space could try and be expanded through including an even larger set of fixed income metrics beyond duration and convexity. This would make the model slower but perhaps more comprehensive by providing the agent with an even more nuanced understanding of the market. Furthermore, algorithm design and reward function engineering present significant room for improvement. Experiments with alternative reward functions, such as Value at Risk (VaR), Expected Shortfall (ES), or explicit Maximum Drawdown penalties, to better manage tail control would be valuable. It would also be of value to dynamically optimize which reward function is applied depending on the identified market regime.

In an even broader scope research should extend beyond the PPO algorithm. Exploring other reinforcement learning architectures or building an ensemble framework that can seamlessly switch between different model types based on the prevailing macroeconomic regime could substantially improve adaptability in practice, and create a more resilient portfolio allocation strategy. A significant structural constraint in the current study is the exclusion of derivative instruments and short selling capabilities. By operating within a long only framework restricted to physical assets (ETFs), the agent is inherently vulnerable to periods of high cross-asset correlation. This limitation was particularly evident during the 2022 inflationary shock, where the simultaneous rise in yields across all maturities caused a systemic sell off in fixed income markets. In such a regime, a long only portfolio lacks the necessary tools to hedge against rising interest rates. Integrating interest rate futures would allow the model to manage duration more dynamically through short positions, providing an important mechanism for downside protection when traditional diversification fails [13]. In addition to improving execution efficiency, futures provide unique predictive value. By discounting future market expectations directly into their price, they offer a 'lead' signal on interest rate shifts that is often missing from traditional spot market data.

References

[1] Sutton, R. S. and Barto, A. G., *Reinforcement Learning: An Introduction*, MIT Press, 2nd edition, 2018.

[2] Jiang, Z., Xu, D., and Liang, J., *A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem*, arXiv preprint arXiv:1706.10059, 2017.

[3] Li, Y., Zheng, Y., and Zheng, Y., *Dynamic Portfolio Optimization with Deep Reinforcement Learning*, IEEE Intelligent Systems, 34(4), 2019.

[4] Moody, J. and Saffell, M., *Reinforcement Learning for Trading*, Advances in Neural Information Processing Systems (NeurIPS), 1999.

[5] Markowitz, *Portfolio Selection Efficient Diversification of Investments*, Yale University Press, 2008.

[6] Liu, S. *Risk in Bond Investment and Risk Management Methods*, Transactions on Economics Business and Management Research, 2024.

[7] Fabozzi, F. J., *Fixed Income Analysis*, CFA Institute Investment Series, John Wiley & Sons, 2nd edition, 2007.

[8] Hull, J. C., *Options, Futures, and Other Derivatives*, Pearson, 11th edition, 2022.

[9] Jolliffe, I. T., *Principal Component Analysis*, Springer Series in Statistics, Springer-Verlag New York, 2nd edition, 2002.

[10] Litterman, R. and Scheinkman, J., *Common Factors Affecting Bond Returns*, Journal of Fixed Income, 1(1), 1991.

[11] Kaviani, S., Ryu, B., Ahmed, E., Kim, D., Kim, J., Spiker, C., & Harnden, B. (2023). *DeepMPR: Enhancing opportunistic routing in wireless networks through multi-agent deep reinforcement learning.* arXiv preprint arXiv:2306.09637. `https://arxiv.org/abs/2306.09637`

[12] Bergmeir, C., & Benítez, J. M. (2012). *On the use of cross-validation for time series predictor evaluation.* Information Sciences, 191, 192–213. `https://doi.org/10.1016/j.ins.2011.12.028`

[13] CME Group, *Using Bloomberg Credit Futures in Credit Portfolio Risk Management*, 2024. `https://www.cmegroup.com/articles/2024/using-bloomberg-credit-futures-in-credit-portfolio-risk-management.html`

[14] Molnar, C., *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd edition, 2022. `https://christophm.github.io/interpretable-ml-book/`

Appendix I - Bloomberg Data

Appendix I-A: Tickers for Asset Universe

This appendix details the specific Bloomberg tickers used to construct the asset pool for the reinforcement learning environment. The assets are divided into U.S. and Pan-European government and corporate bond indices (including USD-hedged variants), as well as a U.S. Treasury Bill cash proxy.

Table 7: Tradable Asset Universe. *Note:* The listed Bloomberg tickers serve as the base identifiers for each asset. For every ticker, the daily Total Return index value, Duration, and Convexity were extracted

Ticker	Asset Name	Region/Currency	Bucket
LT01TRUU-D	US Treasury 1–3Y TR	US / USD	Sovereign
LT02TRUU-D	US Treasury 3–5Y TR	US / USD	Sovereign
LT03TRUU-D	US Treasury 5–7Y TR	US / USD	Sovereign
LT09TRUU-D	US Treasury 7–10Y TR	US / USD	Sovereign
H02526US	Pan-European Treasury 1–3Y TR Hedged	EU / USD-Hedged	Sovereign
H02527US-D	Pan-European Treasury 3–5Y TR Hedged	EU / USD-Hedged	Sovereign
H02528US-D	Pan-European Treasury 5–7Y TR Hedged	EU / USD-Hedged	Sovereign
H02529US-D	Pan-European Treasury 7–10Y TR Hedged	EU / USD-Hedged	Sovereign
LUACTRUU	US Investment Grade Corporates TR	US / USD	Credit (IG)
LF98TRUU	US High Yield Corporates TR	US / USD	Credit (HY)
H02501US	Pan-European High Yield TR Hedged	EU / USD-Hedged	Credit (HY)
BIL_US	US 1–3M Treasury Bills	US / USD	Cash / Risk-free Proxy

Appendix I-B: Macroeconomic Features Details

This part features the specific macroeconomic feature series used to construct the state space for the reinforcement learning environment. The variables are categorized into diverse themes, including real economic activity, monetary conditions, and market volatility across U.S. and Eurozone regions, providing the agent with a broad view of market conditions.

Table 8: Macro Factor Series Used in the RL State Space.

Series	Frequency	Theme	Description
C01	Daily	Commodities	Crude oil front-month benchmark (oil futures proxy)
OIL	Daily	Commodities	Oil price benchmark (CL1 front month)
DAX	Daily	Equity Market Index	German large-cap equity index
EURODEPO	Daily	Monetary Policy (EU)	ECB deposit/policy rate
EURUSD	Daily	FX	EUR/USD foreign exchange rate
FDTR	Daily	Monetary Policy (US)	US Fed policy/target rate
MOVE	Daily	Rates Volatility	US Treasury bond volatility index
SP500	Daily	Equity Market Index	US S&P 500 equity index
STOXX600	Daily	Equity Market Index	Pan-European STOXX 600 equity index
VIX	Daily	Equity Volatility	US equity implied volatility index
dVIX	Daily	Volatility Change	VIX first difference (daily change)
ECOREAN	Monthly	Money / Credit (EU)	Euro area money supply
EUIPEMU	Monthly	Real Activity (EU)	Euro area industrial production
FDDSSD	Monthly	Fiscal Conditions	Government budget deficit series
IP_CHNG	Monthly	Real Activity (US)	US industrial production change
M2_YOY	Monthly	Money Supply (US)	US M2 money supply year-over-year growth
PCE	Monthly	Inflation (US)	US Personal Consumption Expenditures inflation series
PCE_CYOY	Monthly	Inflation (US)	US PCE inflation year-over-year
SPCS20M	Monthly	Housing / Real Economy	US Case-Shiller 20-city home price index
CPI	Monthly	Inflation (US)	Consumer Price Index inflation

Appendix II - Training & Model Diagnostics

The following subsections cover different diagnostics used to determine performance and training- and model specifications.

Appendix II-A: Training Diagnostics

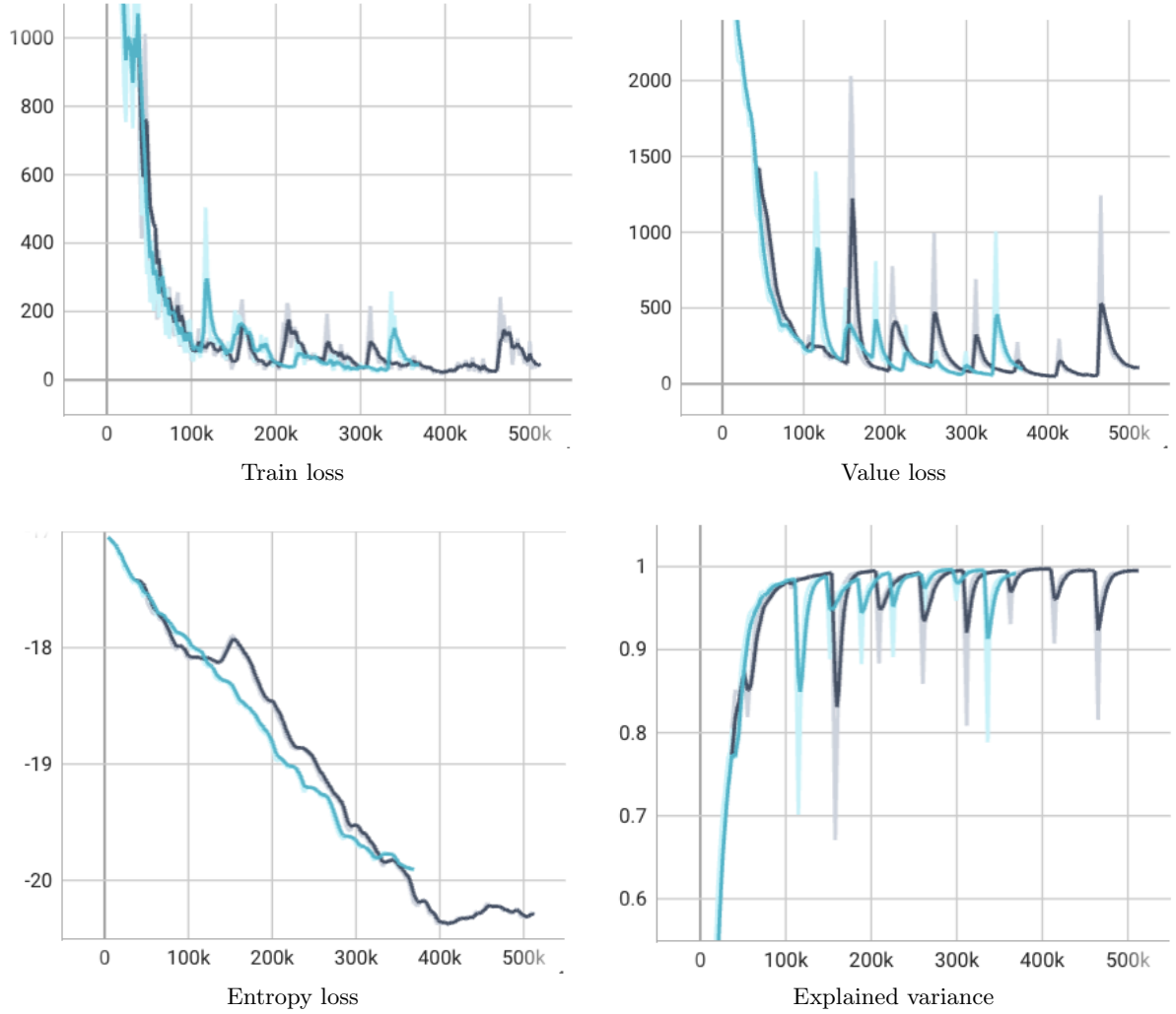

Figure 9: Training convergence diagnostics comparing ≈350k steps vs ≈512k time steps (TensorBoard).

Appendix II-B: Correlations

This appendix presents the feature correlation structure across two perspectives.

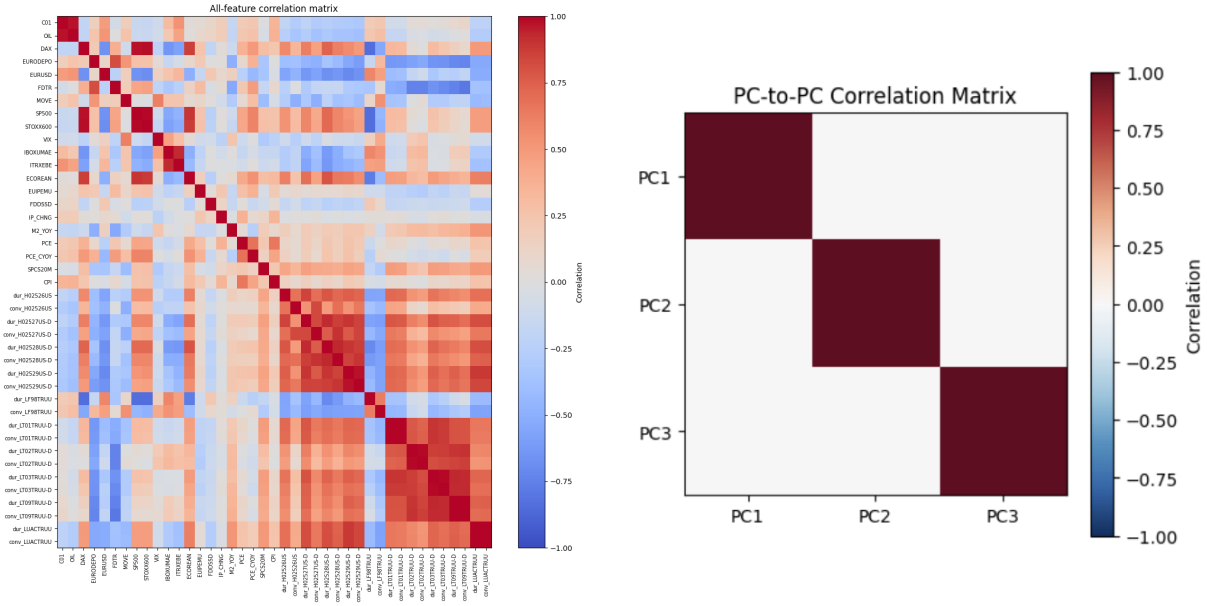

Figure 10: Correlation (global). Left: all-feature correlation matrix (full sample). Right: PC-to-PC correlation matrix (orthogonal to each-other)

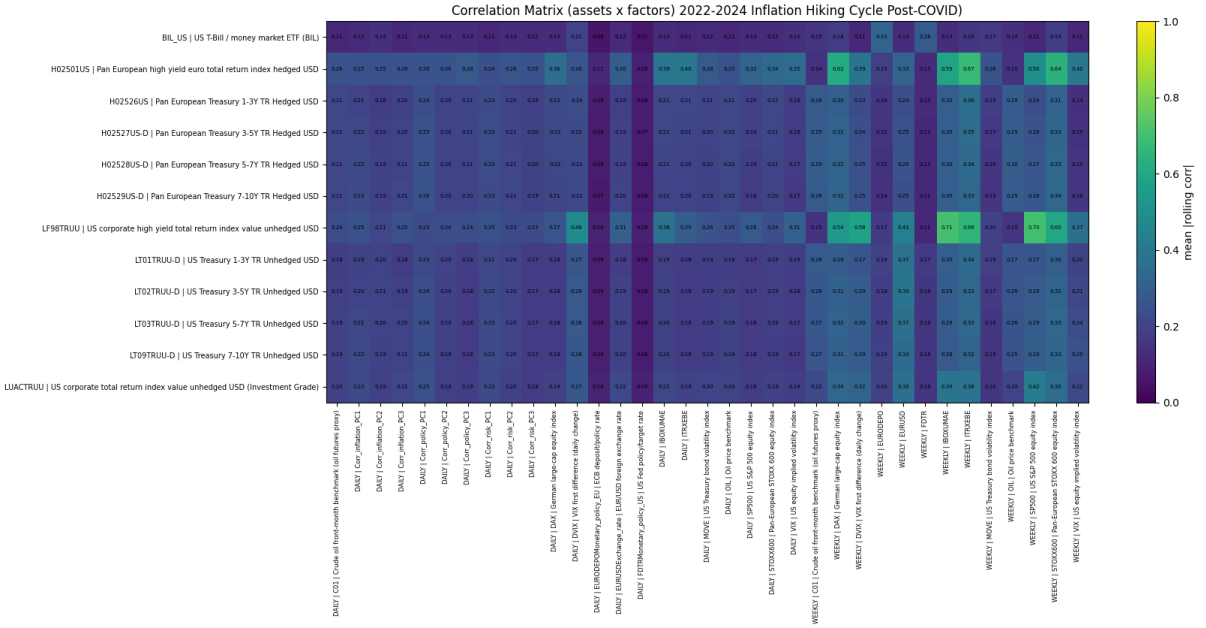

Figure 11: Correlation matrix in the 2022–2024 regime.

Appendix II-C: Model Configuration

This appendix provides the complete model configuration used in the final training run, including the full PCA variance decomposition across all twelve components and the complete set of PPO hyperparameters including Stable-Baselines3 defaults.

Table 9: PPO training hyperparameters. Asterisked values (*) are Stable-Baselines3 defaults.

Hyperparameter	Value
Training timesteps	350,000
Walk-forward window	3 years
Walk-forward step	1 year
Walk-forward epochs	2
gamma	0.99
ent_coef	0.01
n_steps*	2048
batch_size*	64
n_epochs*	10
gae_lambda*	0.95
clip_range*	0.2
learning_rate*	3e-4

Table 10: PCA variance explained by component (all 12 components).

Comp.	Variance (%)	Cum (%)
PC1	45.2956	45.2956
PC2	25.9252	71.2208
PC3	11.7346	82.9555
PC4	8.2821	91.2376
PC5	3.3138	94.5514
PC6	2.7380	97.2894
PC7	1.8671	99.1565
PC8	0.4871	99.6436
PC9	0.2081	99.8517
PC10	0.0940	99.9456
PC11	0.0403	99.9860
PC12	0.0140	100.0000

Appendix II-D: Explainability

Top 10 features by Sharpe drop, averaged over 30 permutations.

Table 11: Metrics for the full test period (baseline Sharpe = 1.03).

Rank	Feature	% of Base
1	Conv. US IG Corp	7.6%
2	MOVE Index	4.8%
3	Case-Shiller 20-City	4.2%
4	Dur. US IG Corp	2.7%
5	CPI	2.6%
6	iTraxx Europe	2.4%
7	DAX	1.8%
8	EU Ind. Production	1.4%
9	Stress Index	1.3%
10	iBoxx EUR HY	1.3%
11	US Ind. Production	1.1%
12	Crude Oil (C01)	1.1%
13	EUR/USD	0.8%
14	Conv. EU Tsy 5–7Y	0.7%
15	Dur. EU Tsy 7–10Y	0.5%

Table 12: Metrics for 2022–2024 stress regime (baseline Sharpe = 0.65).

Rank	Feature	% of Base
1	MOVE Index	22.3%
2	Case-Shiller 20-City	12.8%
3	Conv. US IG Corp	10.3%
4	EU Ind. Production	6.0%
5	iBoxx EUR HY	5.2%
6	Dur. US Tsy 3–5Y	4.5%
7	iTraxx Europe	4.1%
8	Conv. US Tsy 3–5Y	3.9%
9	ECB Deposit Rate	3.5%
10	VIX	2.6%
11	Dur. EU Tsy 7–10Y	2.3%
12	Conv. EU Tsy 5–7Y	2.3%
13	Stress Index	2.0%
14	Euro M2 Supply	2.0%
15	EUR/USD	1.8%

Appendix II-E: Backtest Performance

This appendix presents a detailed breakdown of out-of-sample backtest performance for the no-PCA model, including annual and monthly returns, maximum drawdown trajectory, average daily portfolio turnover, and the distribution of monthly returns over the full test period.

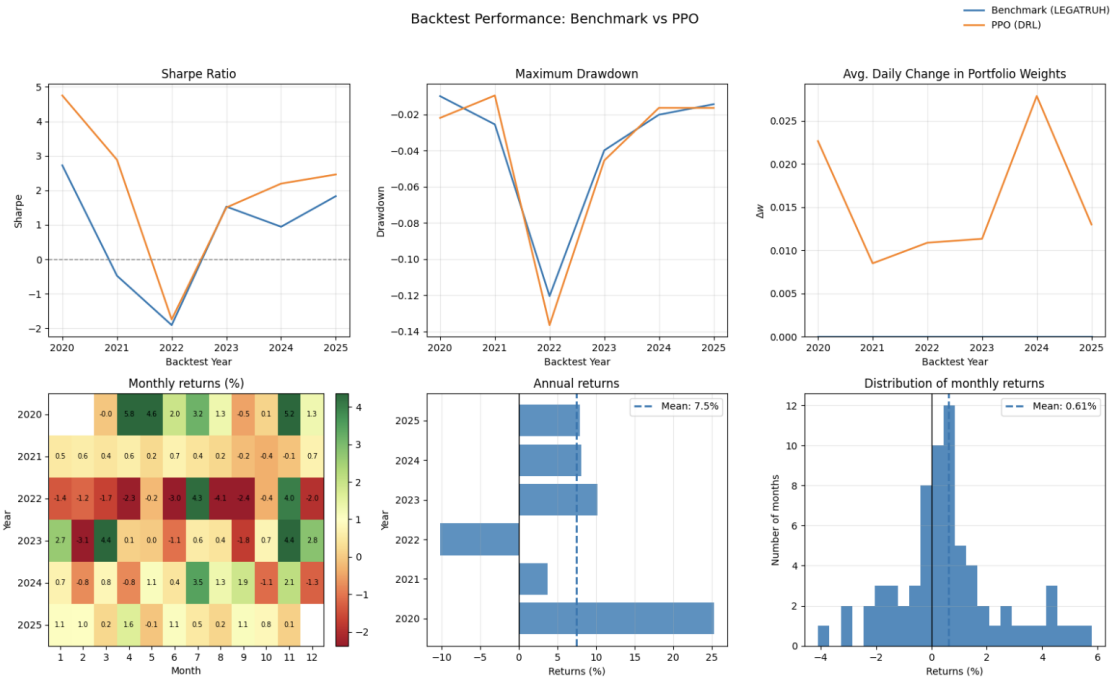

Figure 12: Backtest Results, Monthly Returns, Annual Returns, Monthly Distribution of Return.

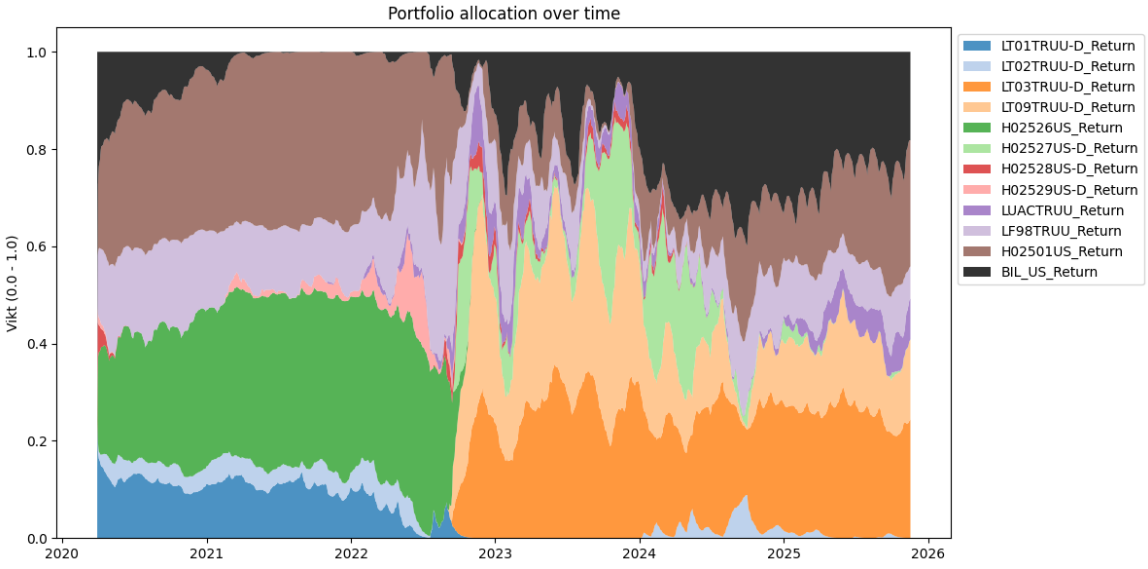

Figure 13: Example of time-varying portfolio allocations (no-PCA model. Reducing timesteps below 350,000 produces visibly more diversified weight paths.

Appendix II-F: Cross-Validation Training and Testing Folds

This section details the walk-forward validation schema. The model was trained and tested across five sequential folds starting in 2014, with each testing period spanning approximately two years.

Table 13: Walk-forward validation periods: training and testing date specifications per fold.

Fold	Train Start	Train End	Test Start	Test End
1	2011-10-12	2014-03-04	2014-03-05	2016-07-08
2	2011-10-12	2016-07-08	2016-07-11	2018-11-08
3	2011-10-12	2018-11-08	2018-11-09	2021-03-19
4	2011-10-12	2021-03-19	2021-03-22	2023-07-27
5	2011-10-12	2023-07-27	2023-07-28	2025-11-18

Disclaimer

Disclaimer
These analyses, documents, and any other information originating from LINC Research & Analysis (henceforth "LINC R&A") are created for information purposes only, for general dissemination, and are not intended to be advisory. The information in the analysis is based on sources, data, and persons which LINC R&A believes to be reliable. LINC R&A can never guarantee the accuracy of the information. The forward-looking information found in this analysis is based on assumptions about the future, and is therefore uncertain by nature, and using information found in the analysis should therefore be done with care. Furthermore, LINC R&A can never guarantee that the projections and forward-looking statements will be fulfilled to any extent. This means that any investment decisions based on information from LINC R&A, any employee or person related to LINC R&A, are to be regarded to be made independently by the investor. These analyses, documents, and any other information derived from LINC R&A are intended to be one of several tools involved in investment decisions regarding all forms of investments regardless of the type of investment involved. Investors are urged to supplement with additional relevant data and information, as well as consult a financial adviser before any investment decision. LINC R&A disclaims all liability for any loss or damage of any kind that may be based on the use of analyses, documents, and any other information derived from LINC R&A.

Conflicts of Interest and Impartiality
To ensure LINC R&A's independence, LINC R&A has established compliance rules for analysts. In addition, all analysts have signed an agreement in which they are required to report any conflicts of interest. These terms have been designed to ensure that COMMISSION DELEGATED REGULATION (EU) 2016/958 of 9 March 2016, supplementing Regulation (EU) No 596/2014 of the European Parliament and of the Council concerning regulatory technical standards for the technical arrangements for objective presentation of investment recommendations or other information recommending or suggesting an investment strategy and for disclosure of particular interests or indications of conflicts of interest.

Other