Reinforcement Learning for Crypto Portfolio Management

Owen Chaffard
David Siang-Li Jheng
Megang Nkamga Junile Staures

Ladislaus von Bortkiewicz Professor of Statistics
Humboldt-Universität zu Berlin
CardoAI
Bucharest University of Economic Studies

20250604 Cha Jhe Jun RL.kay

RL for Portfolio Optimisation : High Claimed Returns ?

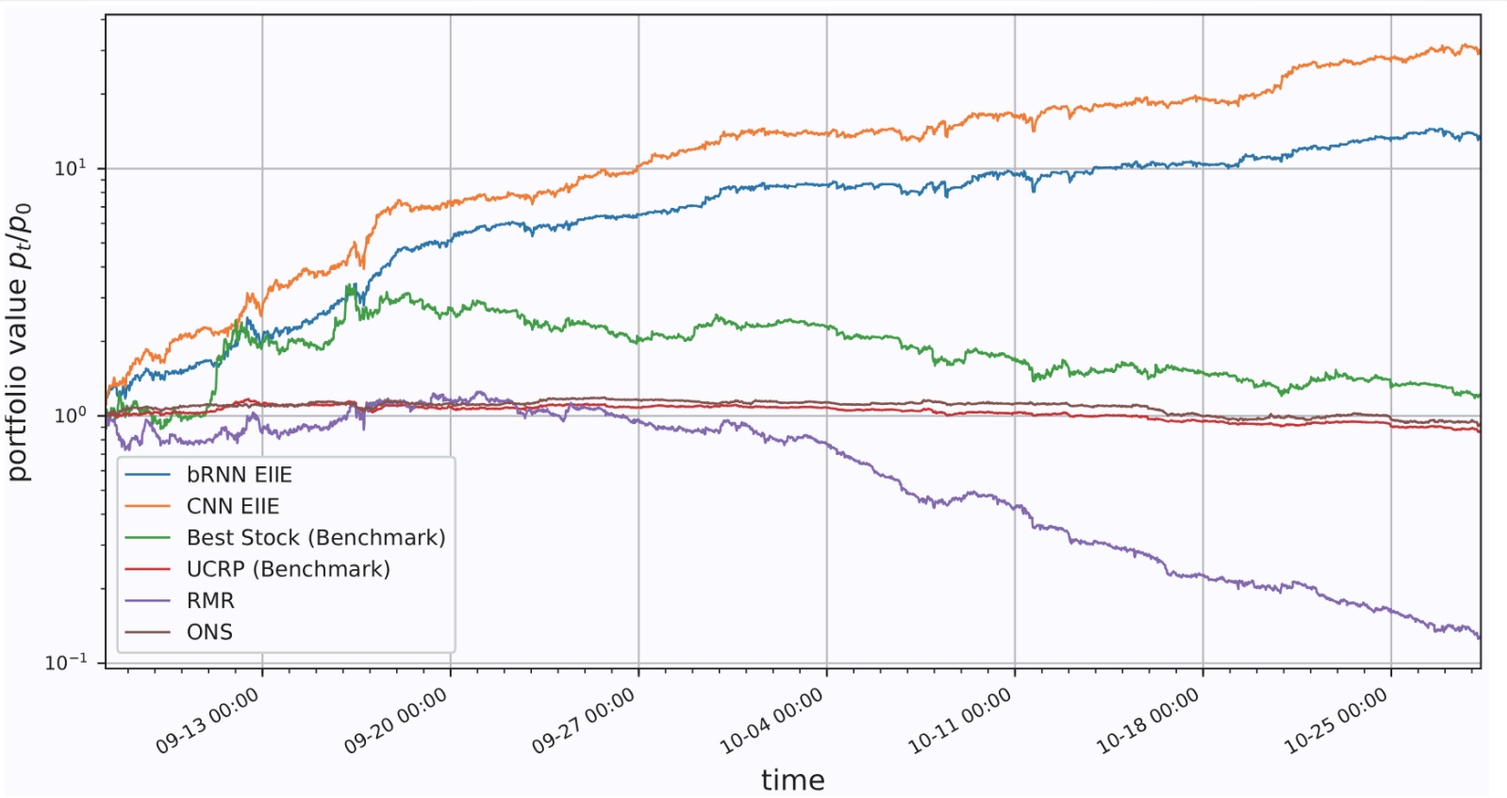
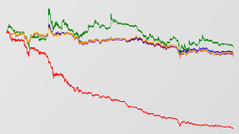


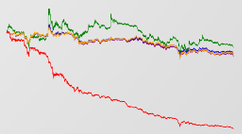
Figure 5: Back-Test 1: 2016-09-07-4:00 to 2016-10-28-8:00 (UTC). Accumulated portfolio values (APV, p_t/p_0) over the interval of Back-Test 1 for the CNN and basic RNN EIIEs, the Best Stock, the UCRP, RMR, and the ONS are plotted in log-10 scale here. The two EIIEs are leading throughout the entire time-span, growing consistently only with a few drawdown incidents.

Source: A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem, Jian *et al.*



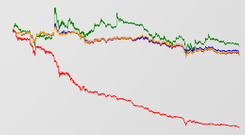
Cryptocurrency Market : Overview

Data Source: S&P Cryptocurrency MegaCap Index - May 09, 2025

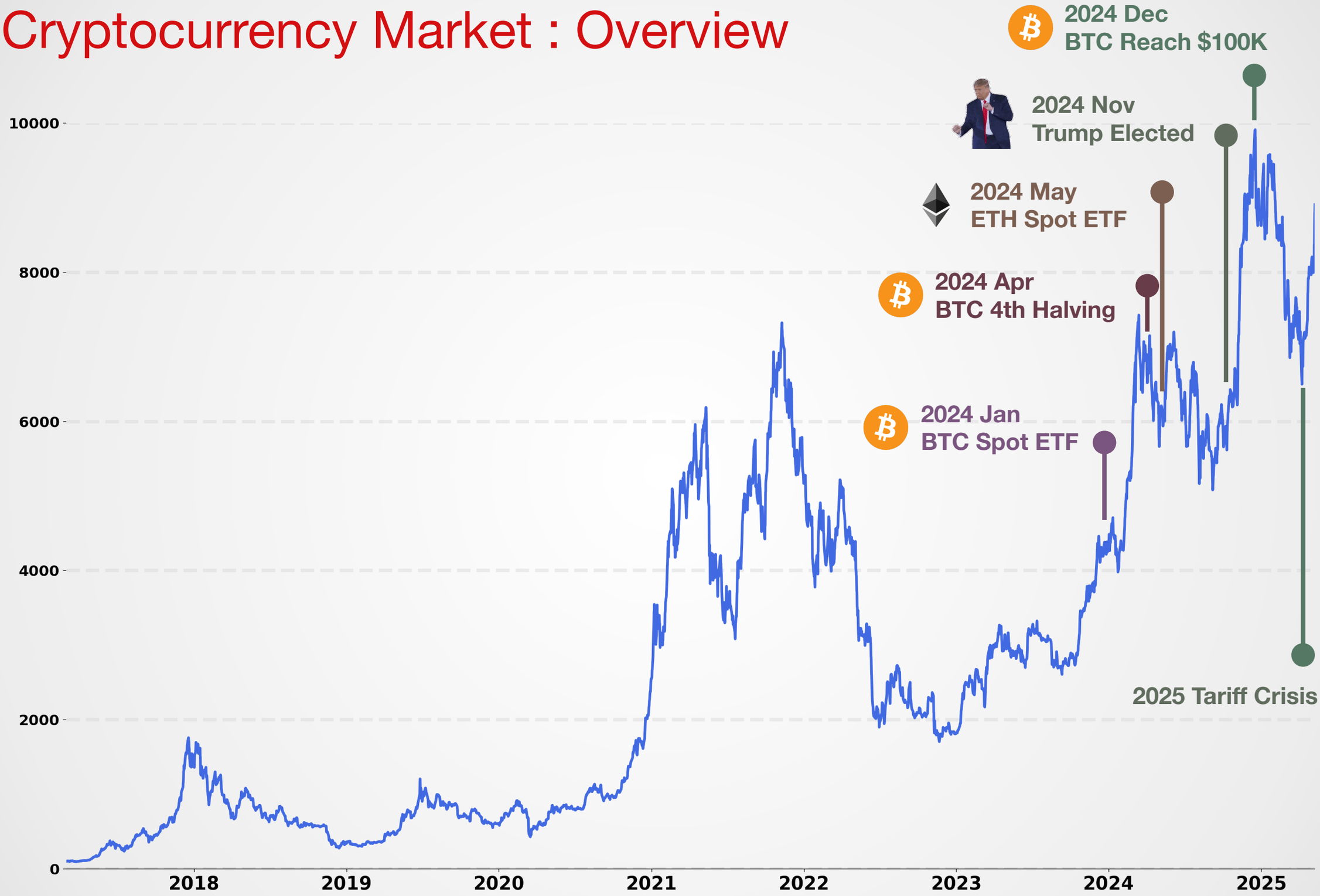


Cryptocurrency Market : Overview

Data Source: S&P Cryptocurrency MegaCap Index - May 09, 2025



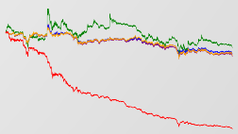
Cryptocurrency Market : Overview



Data Source: S&P Cryptocurrency MegaCap Index - May 09, 2025

Outline

1. Motivation ✓
2. Building a cryptocurrency portfolio
3. Exploratory Data Analysis
4. Methodology - MDP framework
5. Methodology - Model design
6. Evaluation using synthetic data
7. Experimental results and Discussion

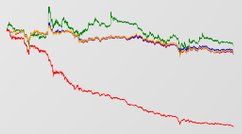


Collecting Cryptocurrency Market Data

- Collect from top market cap coins
  - Collect high-frequency data (30 minutes) from 8 currencies
  - Cover all types of crypto assets
  - Liquidity assumptions (no slippage, no market impact)

- Binance API: Kline data (HLC)
  - Pairs against Bitcoin (cash asset)

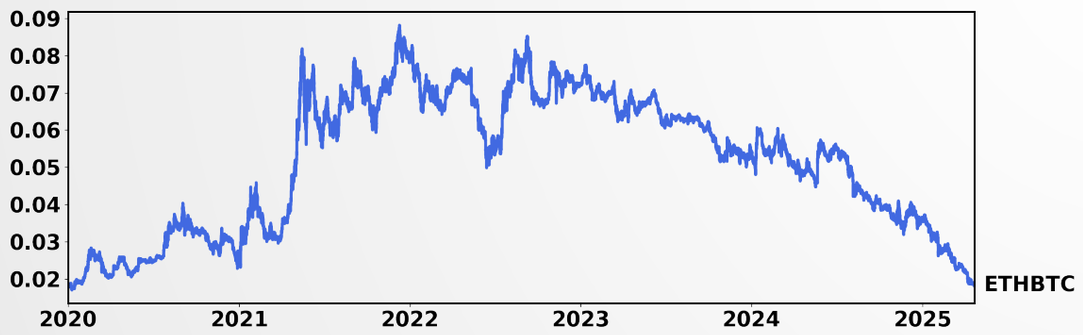
- Time period
  - Collect data from Jan. 2020 to Apr. 2025 (92,941 data points)
  - Use data from Jun. 2024 to Apr. 2025 to train and test RL (15,000 data points)

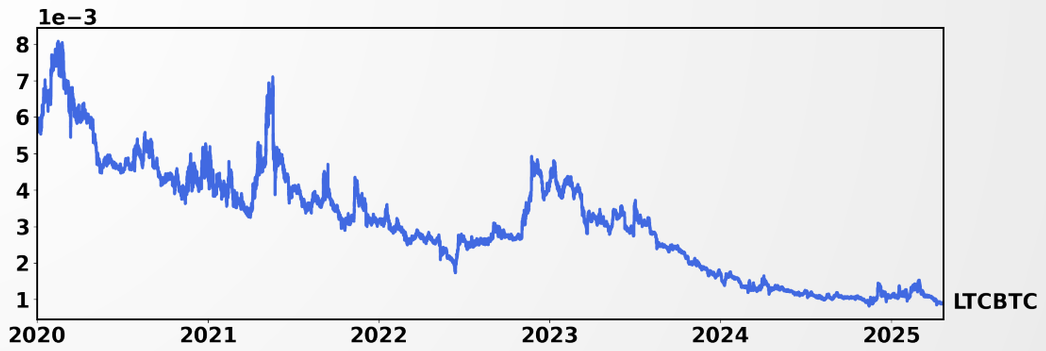


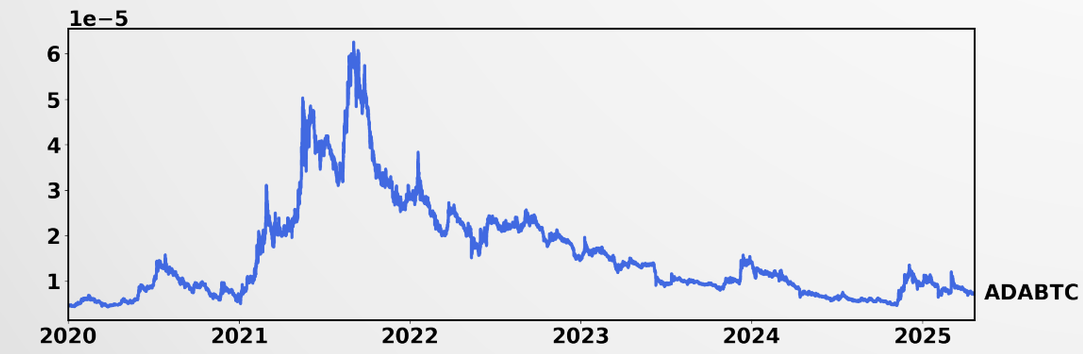
Crypto Portfolio: Technology Coins

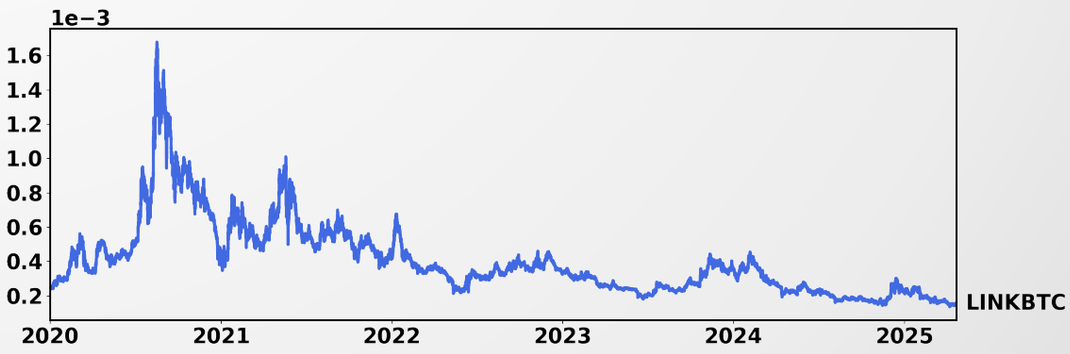
	Ticker	Since	MCap	Supply	Main Technology	Mineable	Max. Supply
	ETH	2013	300B	120M	Smart contracts/ERC-20 (DeFi)	Yes	Infinite
	LTC	2011	7B	76M	Lightweight POW	Yes	84M
	ADA	2017	27B	35B	Native PoS	Yes	45B
	LINK	2017	10B	660M	Oracles	No	1B

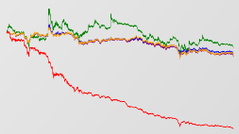
Data updated: May 15, 2025







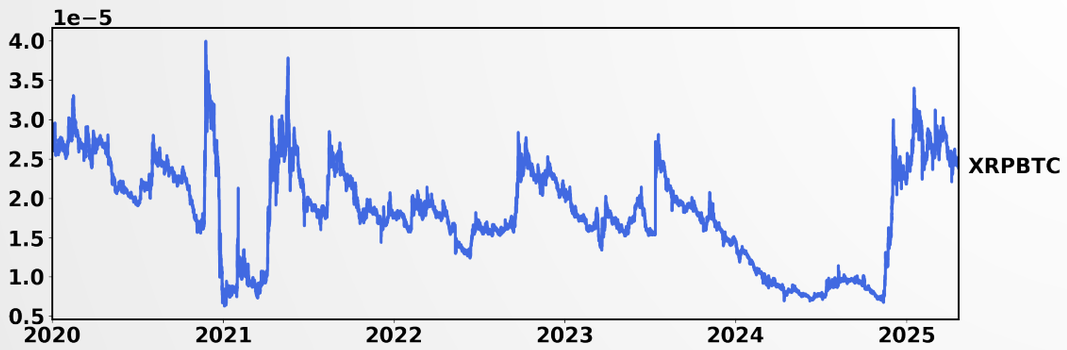


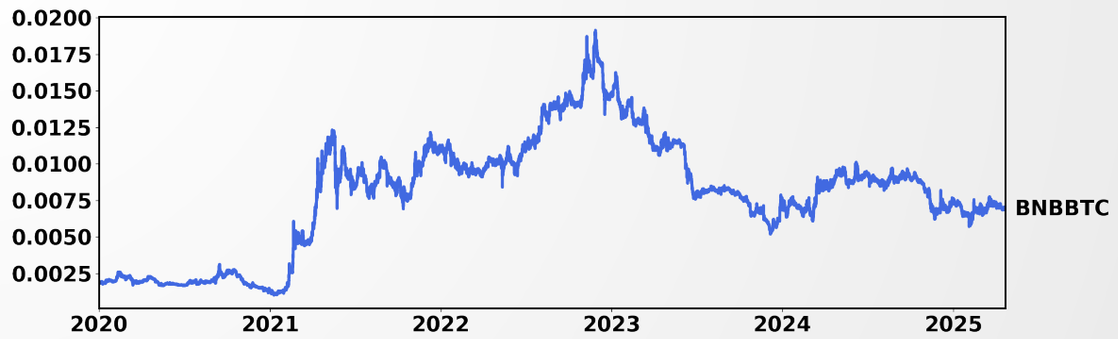


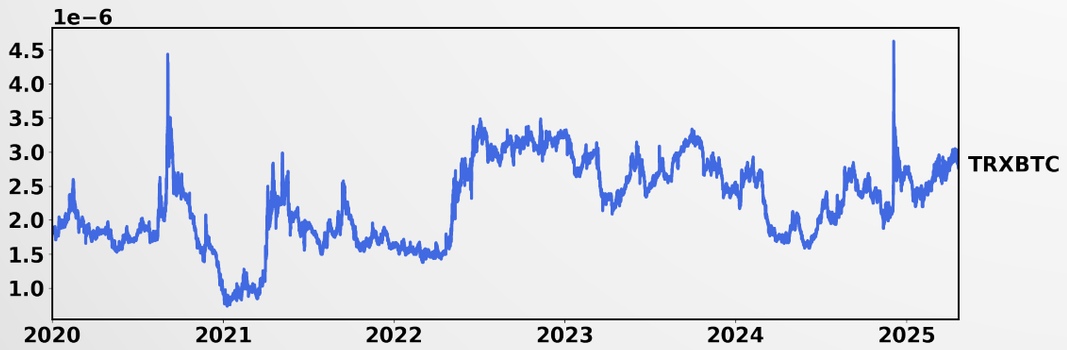
Crypto Portfolio: Influence Coins

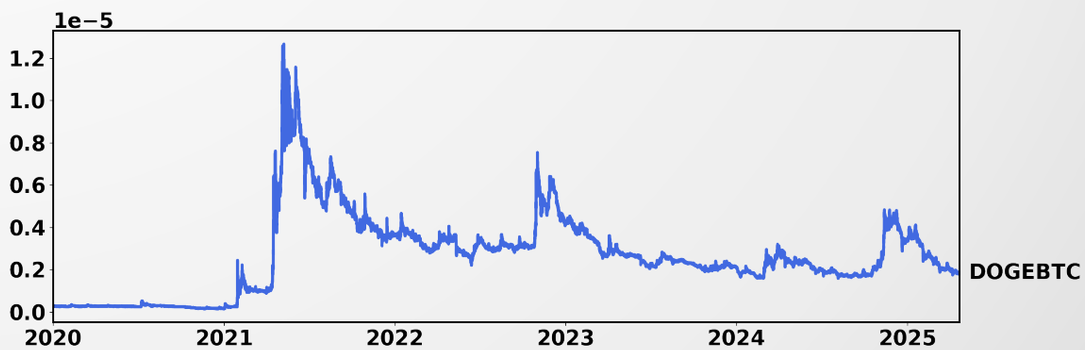
	Ticker	Since	MCap	Supply	Associated company	Primary target/audience
	XRP	2012	150B	150B	RippleNet	Banks
	BNB	2011	7B	76M	Binance	Consumers
	TRX	2017	27B	35B	Tron Foundation	Content creators
	DOGE	2017	10B	660M	Oracles	Social media

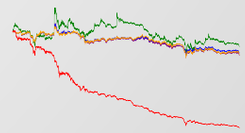
Data updated: May 15, 2025



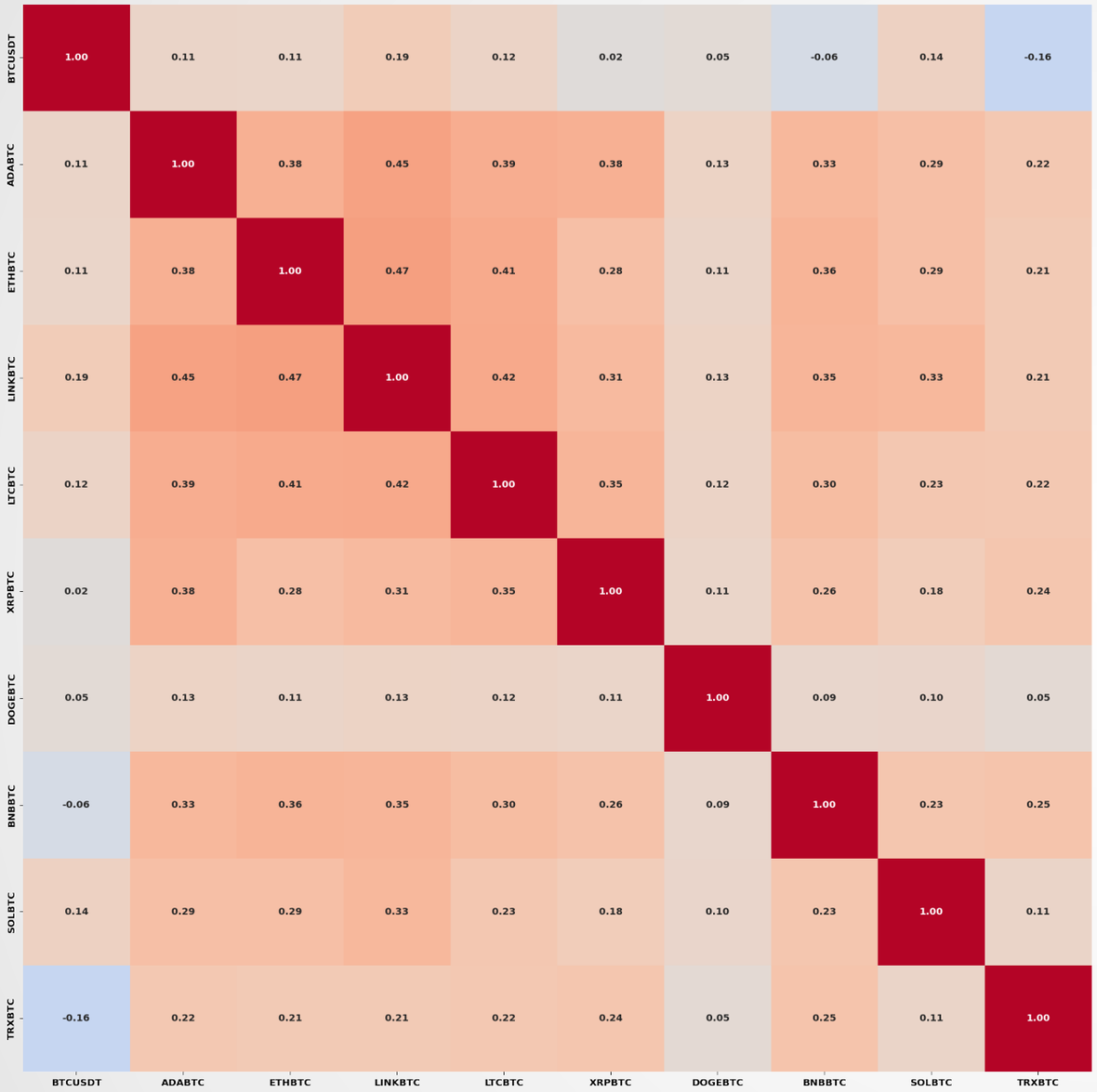


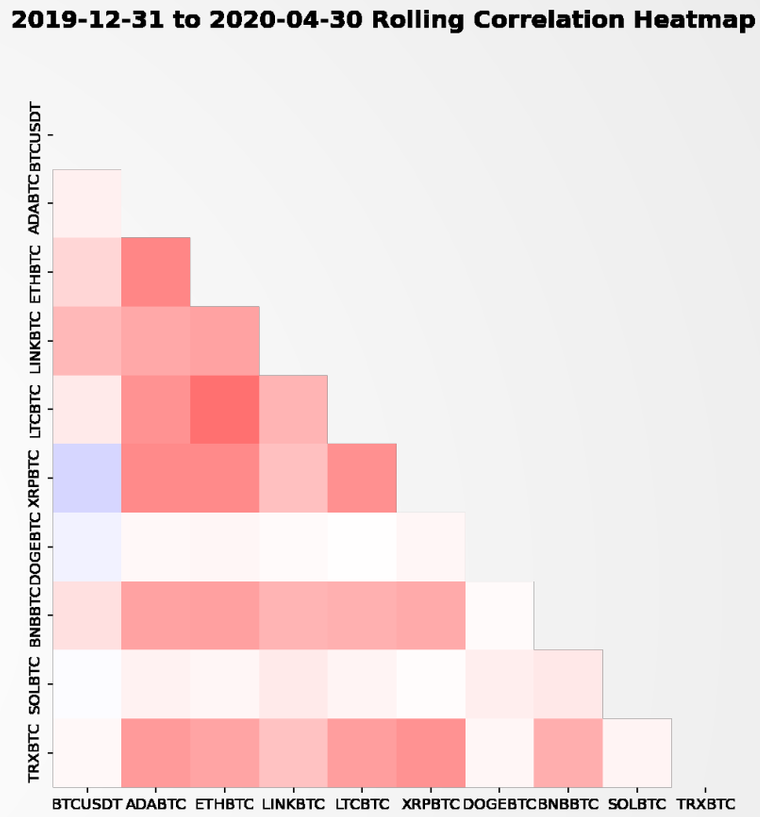


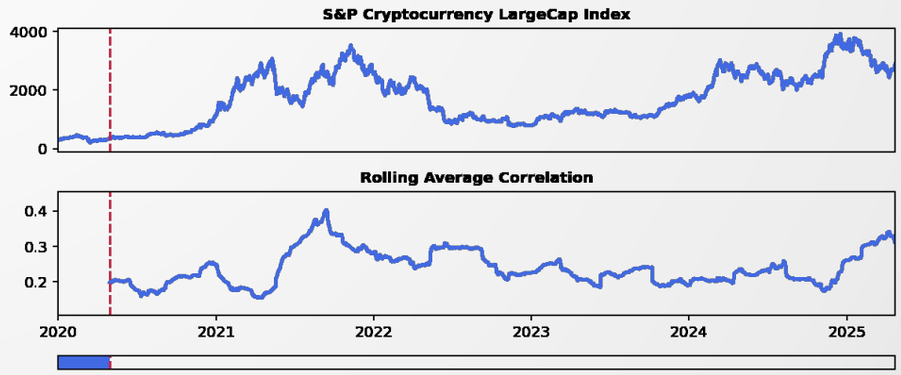




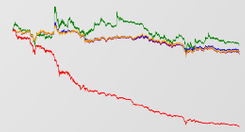
Exploratory Data Analysis





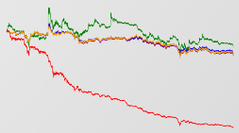


- **Positive correlation across all pairs**



Exploratory Data Analysis

Symbol	Mean	Std	Skew	Kurt	Min	Max	JB	MDD	ACF(1)	Hurst
BTCUSDT	0.00003	0.0050	-0.7448	61.0406	-0.1819	0.1363	14437201.46***	-77.20%	-0.0192	0.5496
ADABTC	0.00000	0.0061	0.4743	42.4179	-0.1535	0.1481	6971121.79***	-92.47%	-0.0411	0.5494
ETHBTC	0.00000	0.0034	0.1610	26.7305	-0.0919	0.0714	2767345.84***	-79.34%	-0.0036	0.5408
LINKBTC	-0.00001	0.0064	0.0130	28.4720	-0.1586	0.1204	3139222.49***	-91.88%	-0.0340	0.5275
LTCBTC	-0.00002	0.0051	0.2741	73.8510	-0.2016	0.1836	21121412.35***	-89.82%	-0.0695	0.4990
XRPBTC	-0.00000	0.0066	0.8806	95.7926	-0.2036	0.2367	35546567.78***	-84.22%	-0.0214	0.5229
DOGEBTC	0.00002	**0.0158**	0.4253	25.7146	-0.3365	0.4238	2563420.26***	-87.46%	-0.3386	0.5259
BNBBTC	0.00001	0.0046	0.1968	40.0950	-0.1083	0.1119	6225992.27***	-72.84%	-0.0234	0.5687
TRXBTC	0.00000	0.0065	1.1543	43.0981	-0.1125	0.2069	7213516.40***	-83.33%	-0.1689	0.5332



Methodology : MDP Formulation

- State Space :

$$\mathcal{S}_t = \left[\text{HLCV}_{t-p}^{(i)}, \text{HLCV}_{t-p+1}^{(i)}, \cdots, \text{HLCV}_{t}^{(i)}\right] \cup \mathbf{w}_t, \quad \forall i. \quad \mathbf{w}_t \in \triangle^{N+1}$$

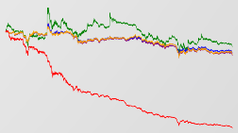
- Action Space :

$$\mathbf{w}_t \in \triangle^{N+1}, \quad \sum_{i=1}^{N+1} w_{i,t} = 1, \quad w_{i,t} \geq 0, \quad \forall i$$

- Immediate Reward:

$$\mathscr{R}_t = \log\left(\mathbf{w}_t^\top \mathbf{y}_t\right) \text{ with } \mathbf{w}_t \in \triangle^{N+1}$$

  - $y_{i,t}$: i-th asset return log return of portfolio



Methodology : Value Function and No Exploration

- (*Action doesn't affect environment)* closed-form discounted value function ($\gamma = 0$):

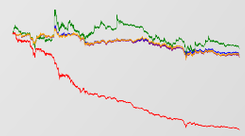
$$V_\pi(S_0) = \sum_{t=0}^{\infty} \gamma^t R_{t+1} = R_1$$

- Closed form enables batch-learning :

$$\frac{1}{t_b} R_{[t,t+t_b]}(\pi_\theta | \dots) = \frac{1}{t_b} \sum_{\tau=t}^{t+t_b} \ln\left(\mathbf{w}_\tau^\top \mathbf{y}_\tau\right)$$ avg reward over batch

- Policy update rule (gradient ascent) :

$$\theta \to \theta + \lambda \nabla_\theta \left[\frac{1}{t_b} R_{[t,t+t_b]}(\pi_\theta | S_t, \dots, S_{t+t_b}) \right]$$



Methodology : Accounting for Transaction Fee

- Transaction fee : actual return is discounted by a factor
  - Depends on amount of asset reallocated
- Define $\mathbf{w}'_t$ asset allocation after price movement at time t

$$\mathbf{w}'_t = \frac{\mathbf{w}_{t-1} \odot \mathbf{y}_t}{\mathbf{w}_{t-1} \cdot \mathbf{y}_t}$$

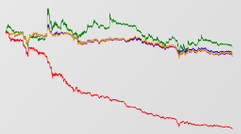
- Portfolio rebalancing is thus $\left| \mathbf{w}_t - \mathbf{w}'_t \right|$ and with fixed rate for selling/buying

  $\eta = 0.025\,\%$ portfolio value is discounted by

$$\mu_t = 1 - \eta \left\| \mathbf{w}_t - \mathbf{w}'_t \right\|_1$$

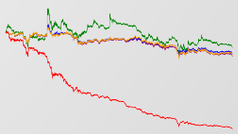
- Thus return becomes :

$$R_t = \log\left(\mu_t \mathbf{w}_\mathbf{t}^\top \mathbf{y}_\mathbf{t}\right) = \log\left(\left(1 - \eta \left\| \mathbf{w}_t - \mathbf{w}'_t \right\|_1\right) \mathbf{w}_\mathbf{t}^\top \mathbf{y}_\mathbf{t}\right)$$



Methodology : Model Selection

- Proposed framework : deterministic policy
  - Based on NN
- Naive choice : CNN to model short term dependencies
  - Fast and simple
- Reversible instance normalisation :
  - Stabilize NN outputs
  - Differentiation kill TS memory
  - Global norm is non causal

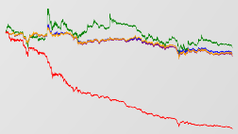


Methodology : Model Selection

- Problems with convergence
  - Add dropout layers to stabilise
  - Initialise weights with higher variance (favour exploration)

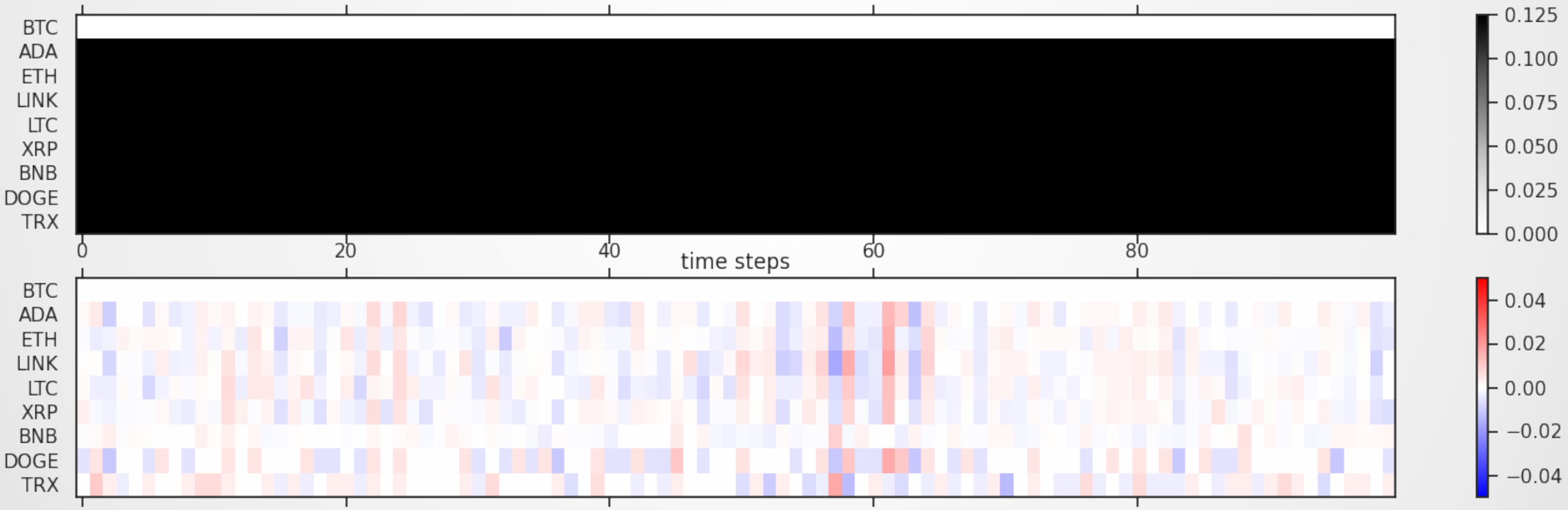
- Determine cash weight ?
  - Mode outputs a score for each asset except cash

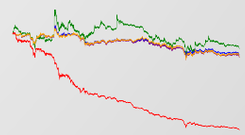
- *Cash bias before softmax*
- *Dense layer*



Methodology : Model Selection

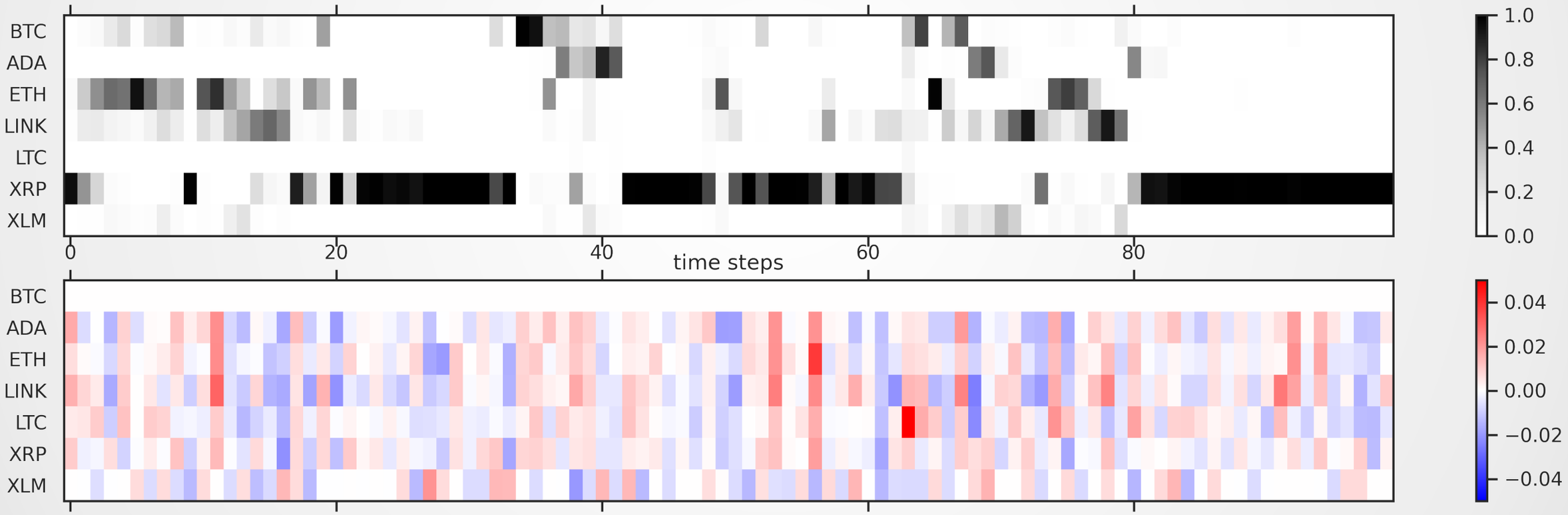
- *Cash bias before softmax*
  - All in bias or UCRP





Methodology : Model Selection

- *Dense layer*
  - Model is biased towards one coin (historical best performer)

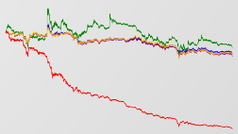


Methodology : Final Model Design

- Univariate model
  - Independent voting scores
  - Cannot hold cash

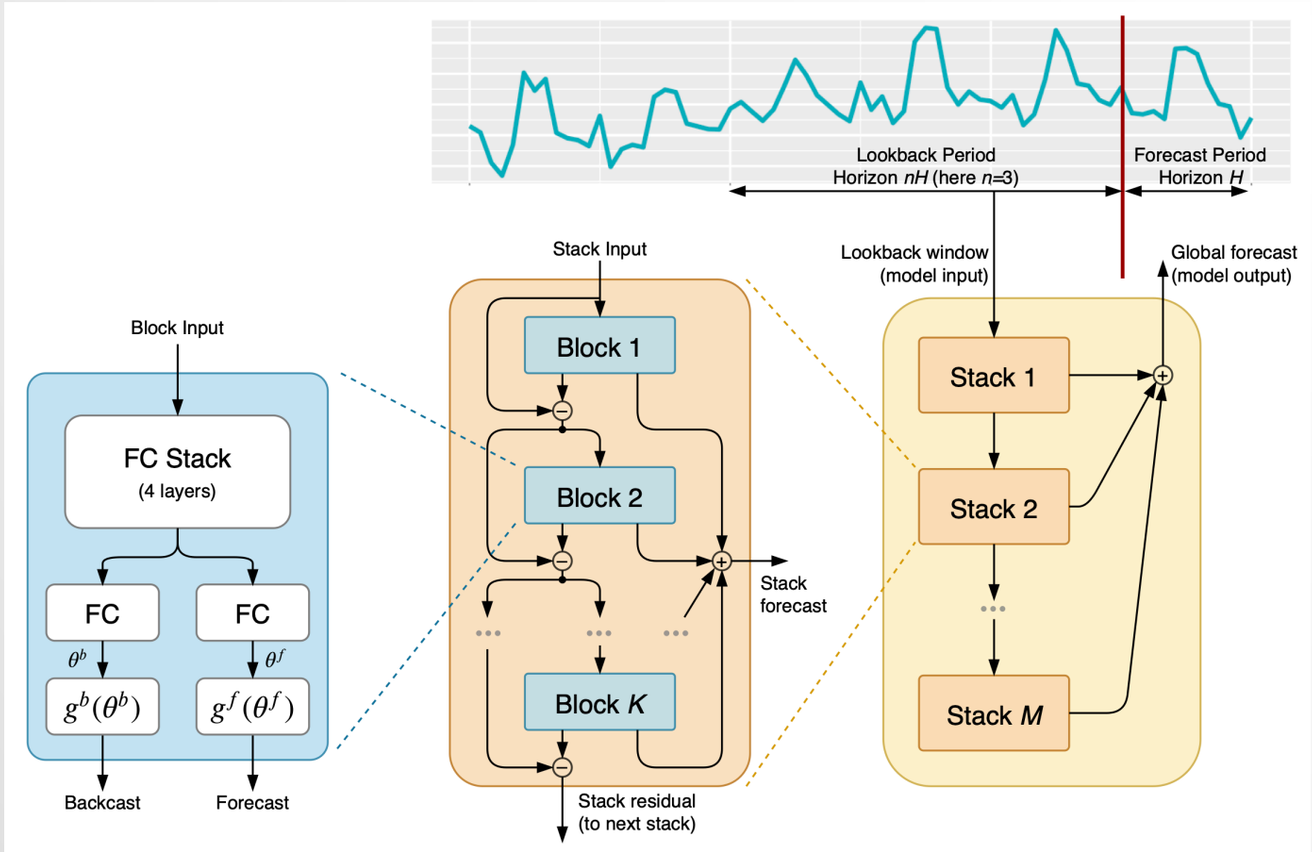
- SOTA forecasting model
  - NBEATS : capture complex seasonalities
  - Forecast is used to set voting score

- Weight sharing
  - Augment training set (time + variates)
  - No historical bias



Methodology : Final Model Design

- Each block j receives a backcast (residual) $\hat{b}^{(j-1)}$ and outputs :
  - A backcast $\hat{b}^{(j)}$ and a forecast $\hat{f}^{(j)}$
  - $\hat{b}^{(j)}, \hat{f}^{(j)} = \text{Block}^{(j)}\left(b^{(j)}\right), b^{(0)} = x_t$



- Trend Block :

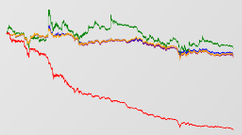
$$\hat{f}^{(j)} = \sum_{i=0}^{p} \theta_i^j \mathbf{t}^i$$

low order polynomial

- Seasonal Block :

$$\hat{f}^{(j)} = \sum_{i=0}^{h/2-1} \theta_i^j \cos(2\pi i t) + \theta_{i+h/2}^j \sin(2\pi i t)$$

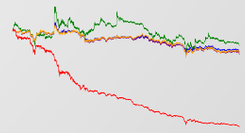
Periodic function



Evaluation Using Synthetic Data

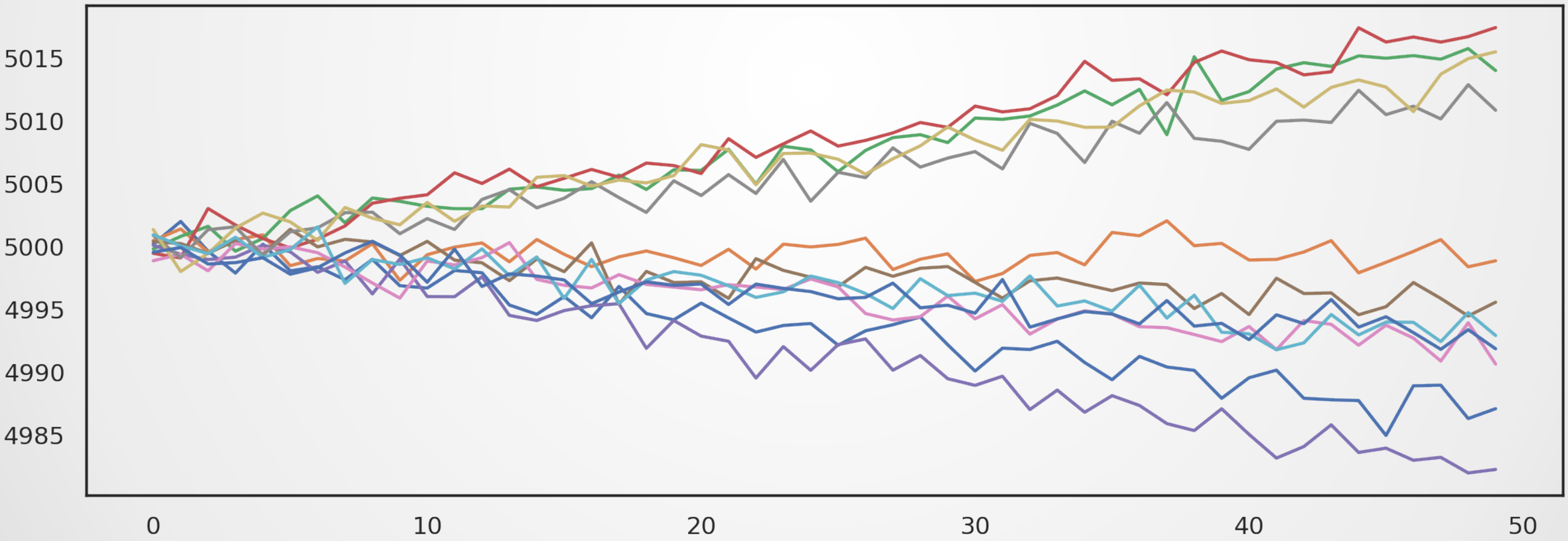
- Use a generated synthetic dataset to test the model
  - Signal + gaussian noise

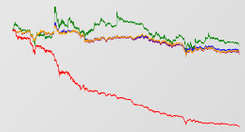
- Check learned policy
  - Performance metrics (Sharpe ratio, final portfolio value)
  - Allocation weight distribution



Evaluation Using Synthetic Data : Noisy Trends

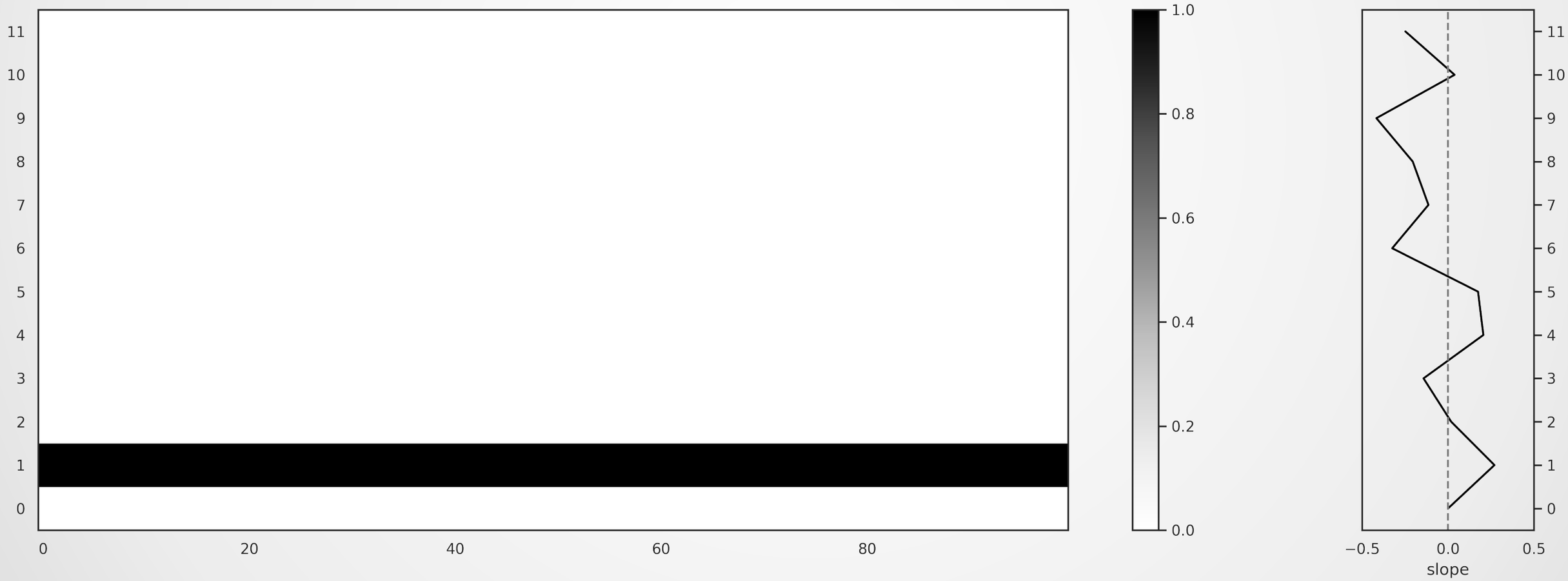
- Collection of assets with randomly selected slopes $\alpha \in [-0.5, 0.5]$
  - Gaussian noise $\sigma = 1$

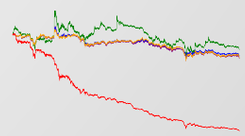




Evaluation Using Synthetic Data : Noisy Trends

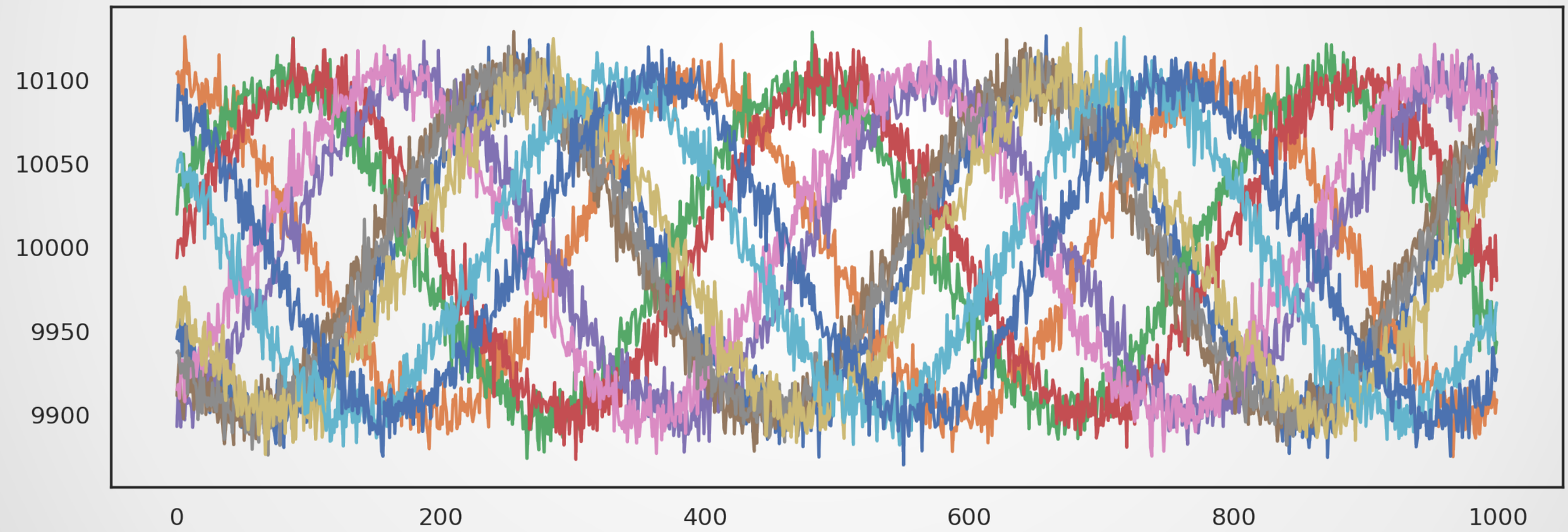
- Collection of assets with randomly selected slopes $\alpha \in [-0.5, 0.5]$
  - Model learns to hold only asset w/ highest slope

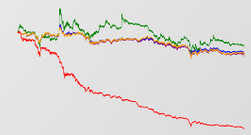




Evaluation Using Synthetic Data : Noisy Periodic Data

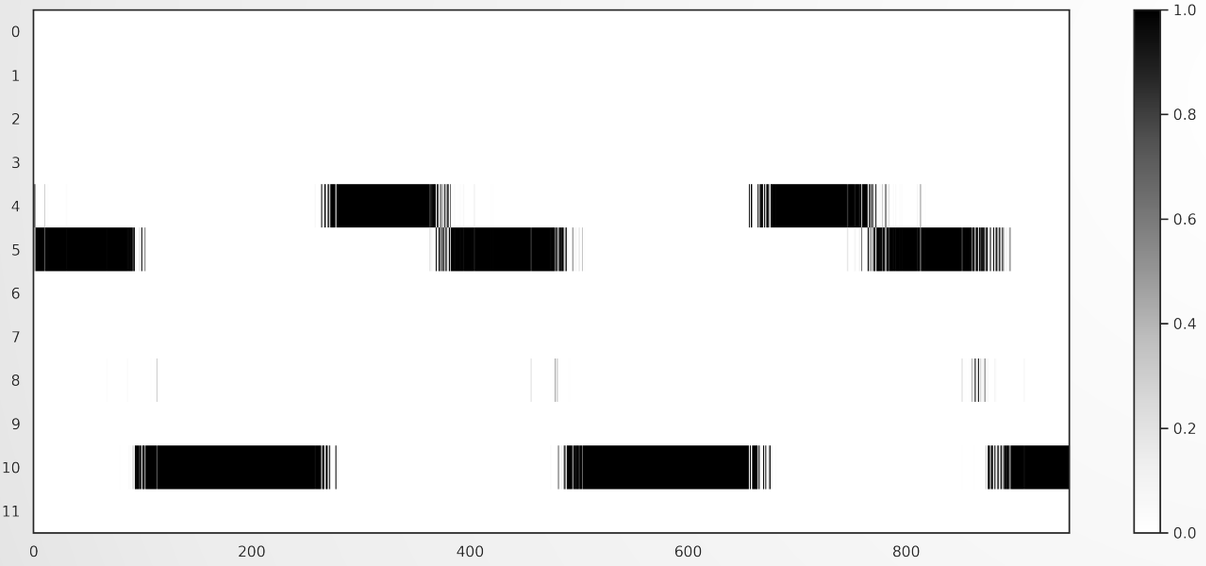
- Collection of periodic signals with randomly selected phase
  - Gaussian noise $\sigma = 1$

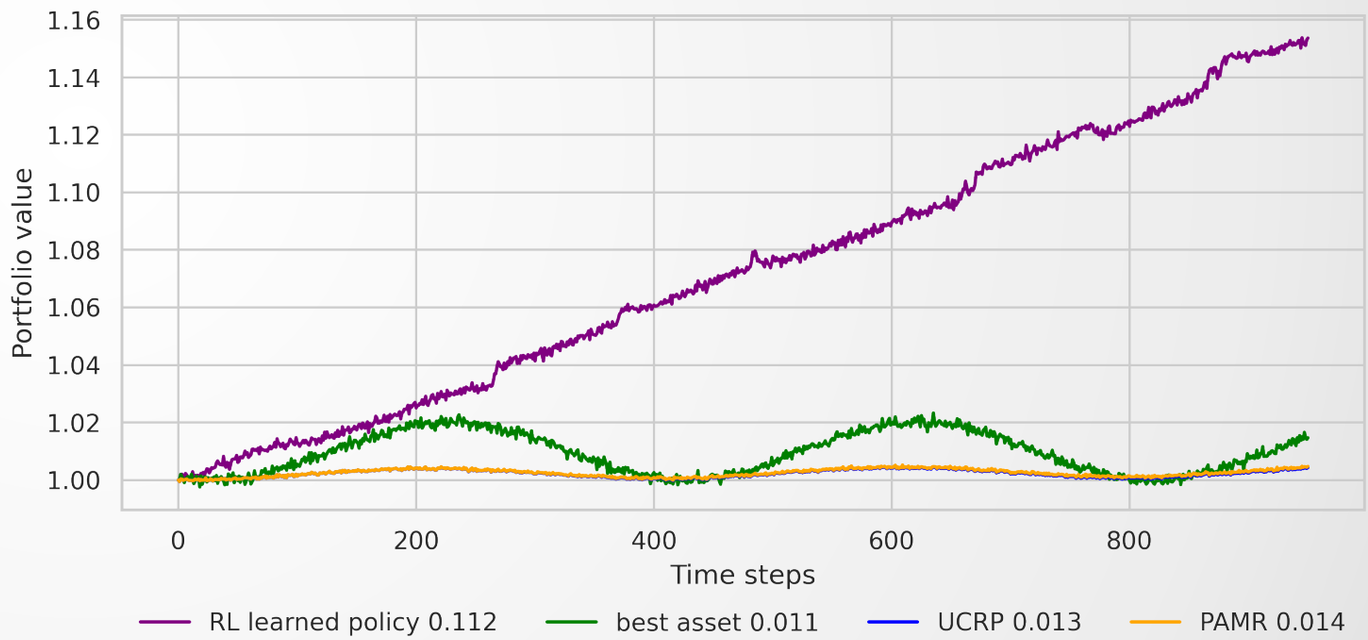


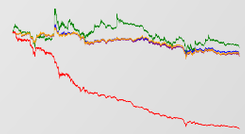


Evaluation Using Synthetic Data : Noisy Periodic Data

- Periodic reallocation to hold assets on uptrend
  - Linearly growing portfolio value

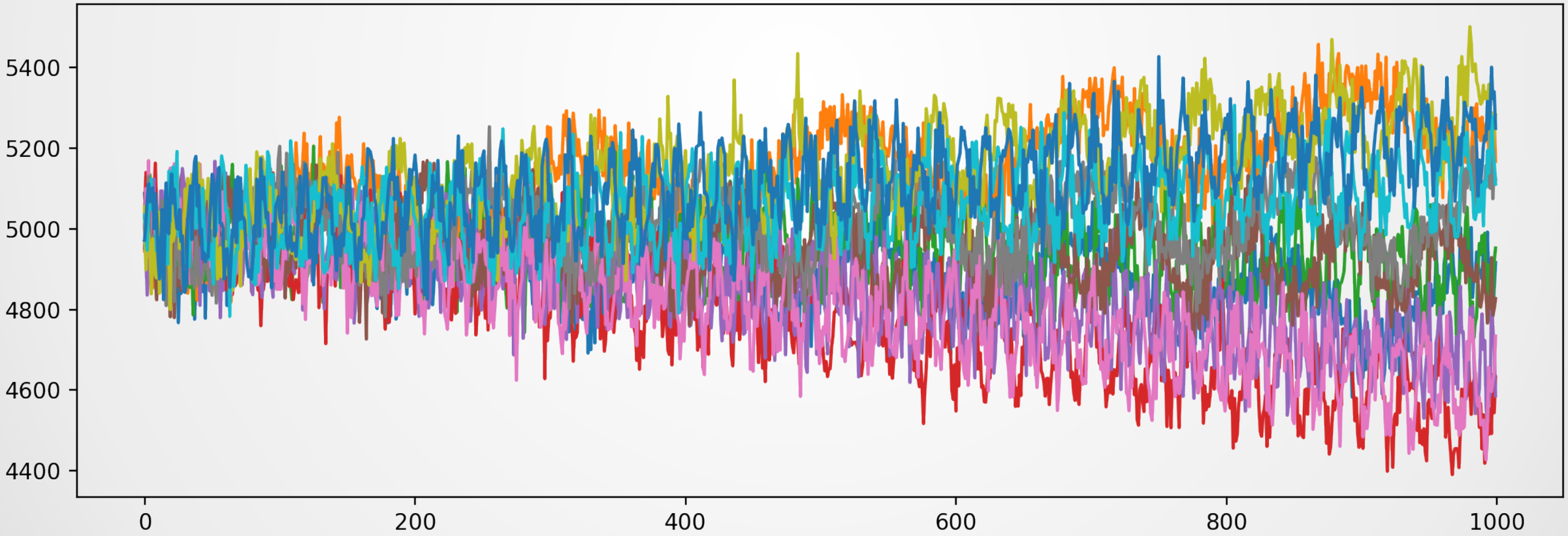






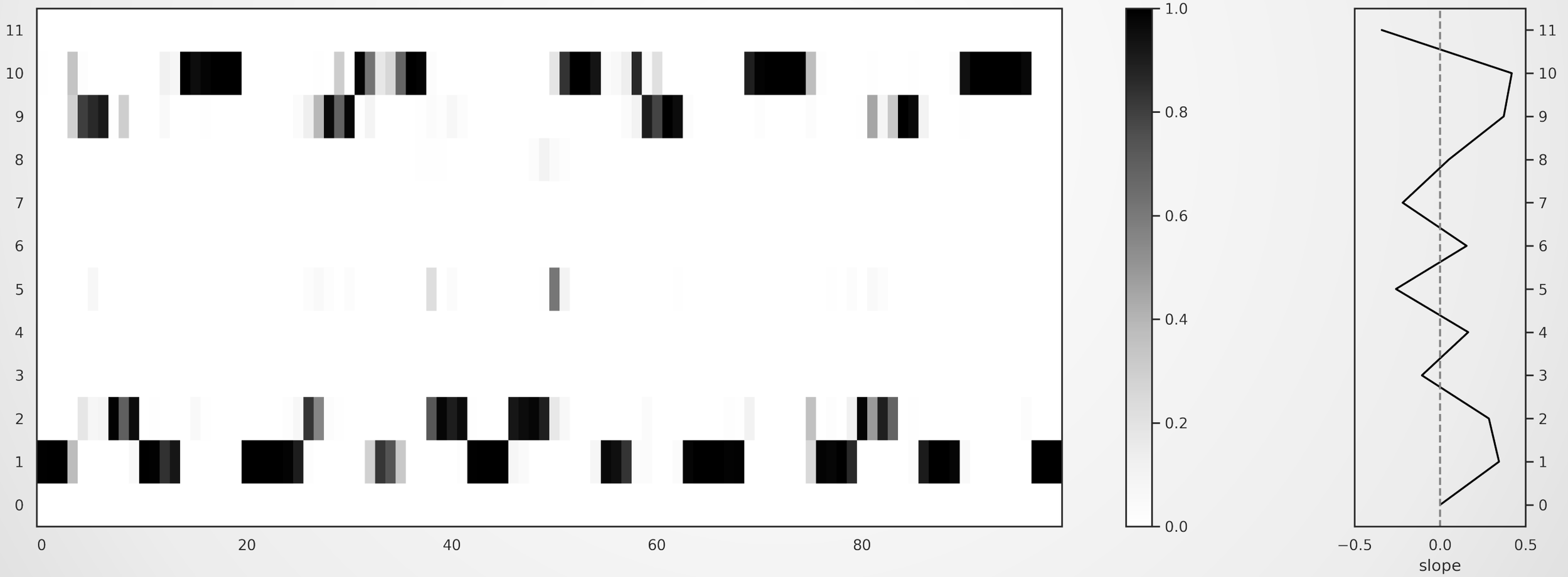
Evaluation Using Synthetic Data : Trend Seasonal Data

- Randomly selected slope and frequency
  - Gaussian noise $\sigma = 1$



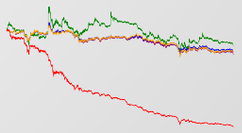
Evaluation Using Synthetic Data : Trend Seasonal Data

- Periodic reallocation to hold assets on uptrend
- Allocation between assets with highest slopes



Evaluation Using Synthetic Data

- Model convergence is shown
- Outperforms typical portfolio policies on synthetic data
- Works with crypto ?
  - Non stationary volatility
  - Highly correlated assets
  - Complex seasonality



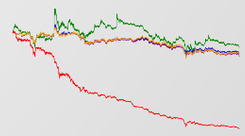
Experimental Results : Benchmark Policies

- Choose policies to benchmark model performance
  - Online Portfolio Selection: A Survey Li *et al.*

- **Basic Policies :**

- *Best Stock* (hindsight) : $\mathrm{fPV} = \max_i \left(\bigodot_{t=1}^{n} x_{i,t} \right)$

- *CRP* : $\mathbf{w}_t = \left(\frac{1}{m}, \ldots, \frac{1}{m} \right).$ $\mathrm{fPV} = \frac{1}{m} \prod_{t=1}^{n} \left(\sum_{i=1}^{m} x_{i,t} \right)$



Experimental Results : Benchmark Policies

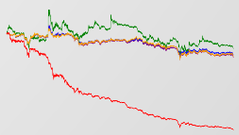
- *PAMR*: either passively maintains previous portfolio or aggressively reallocates based on threshold on returns
  - Update Rule : $\mathbf{w}_{t+1} = \mathbf{w}_t - \tau_t \left(\mathbf{x}_t - \bar{x}_t \mathbf{1} \right)$

$$\text{with } \tau_t = \max \left(0, \frac{\mathbf{b}_t \cdot \mathbf{x}_t - \varepsilon}{\| \mathbf{x}_t - \bar{x}_t \mathbf{1} \|_2} \right)$$

- *OLMAR*: Predict next returns with SMA and reallocates:

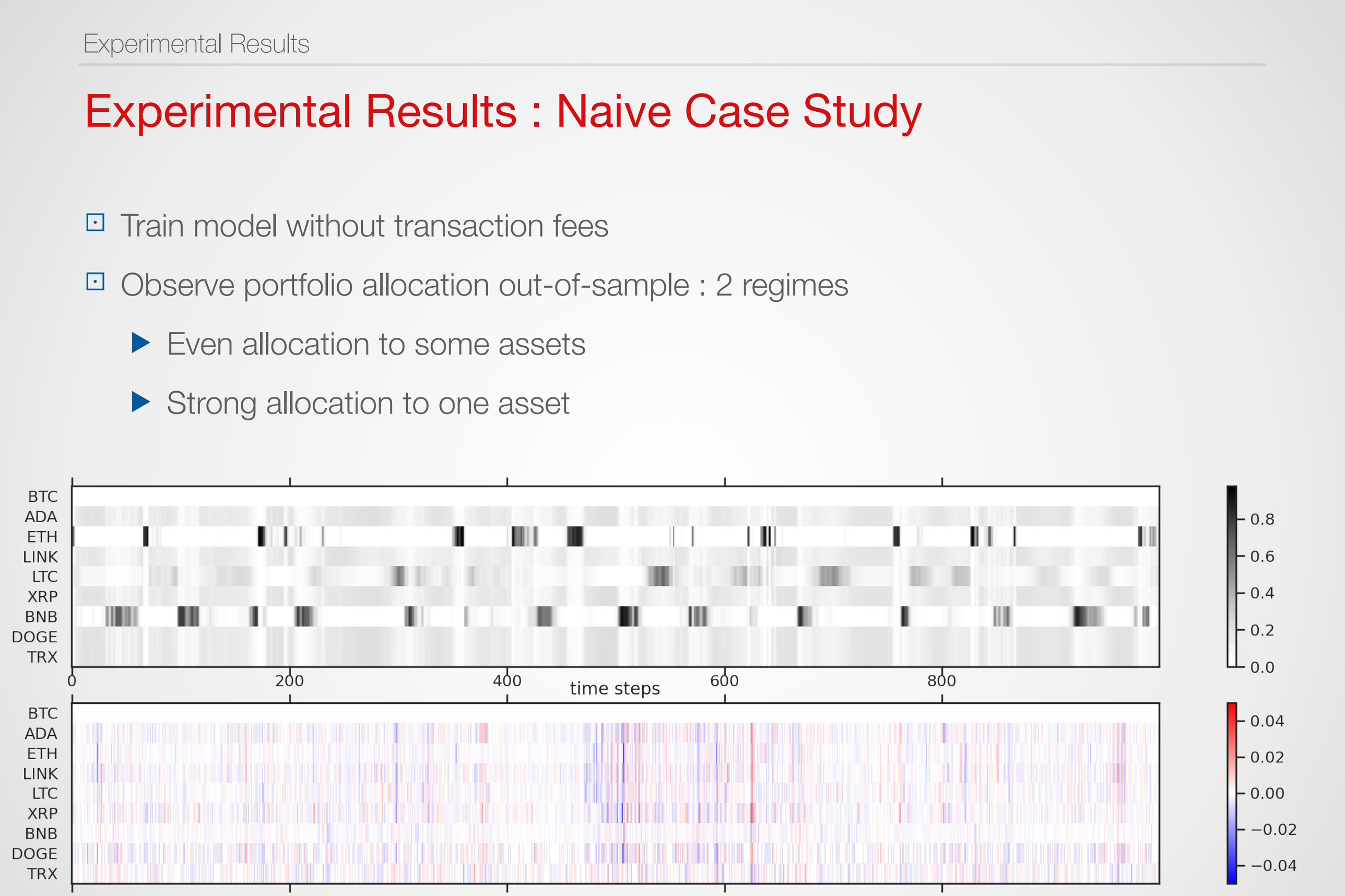
$$\mathbf{x}_{t+1} = \frac{\text{SMA}_t(T)}{\mathbf{p}_t} = \frac{\frac{1}{T} \sum_{i=t-T+1}^{t} \mathbf{p}_i}{\mathbf{p}_t}$$

$$\mathbf{w}_{t+1} = \arg\min_{w} \frac{1}{2} \| \mathbf{w} - \mathbf{w}_t \|_2 \quad \text{s.t.} \quad \mathbf{w} \cdot \mathbf{x}_{t+1} \geq \varepsilon$$



Experimental Results : Naive Case Study

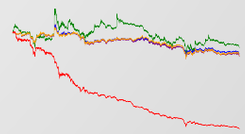
- Train model without transaction fees
- Observe portfolio allocation out-of-sample : 2 regimes
  - Even allocation to some assets
  - Strong allocation to one asset



Experimental Results : Naive Case Study

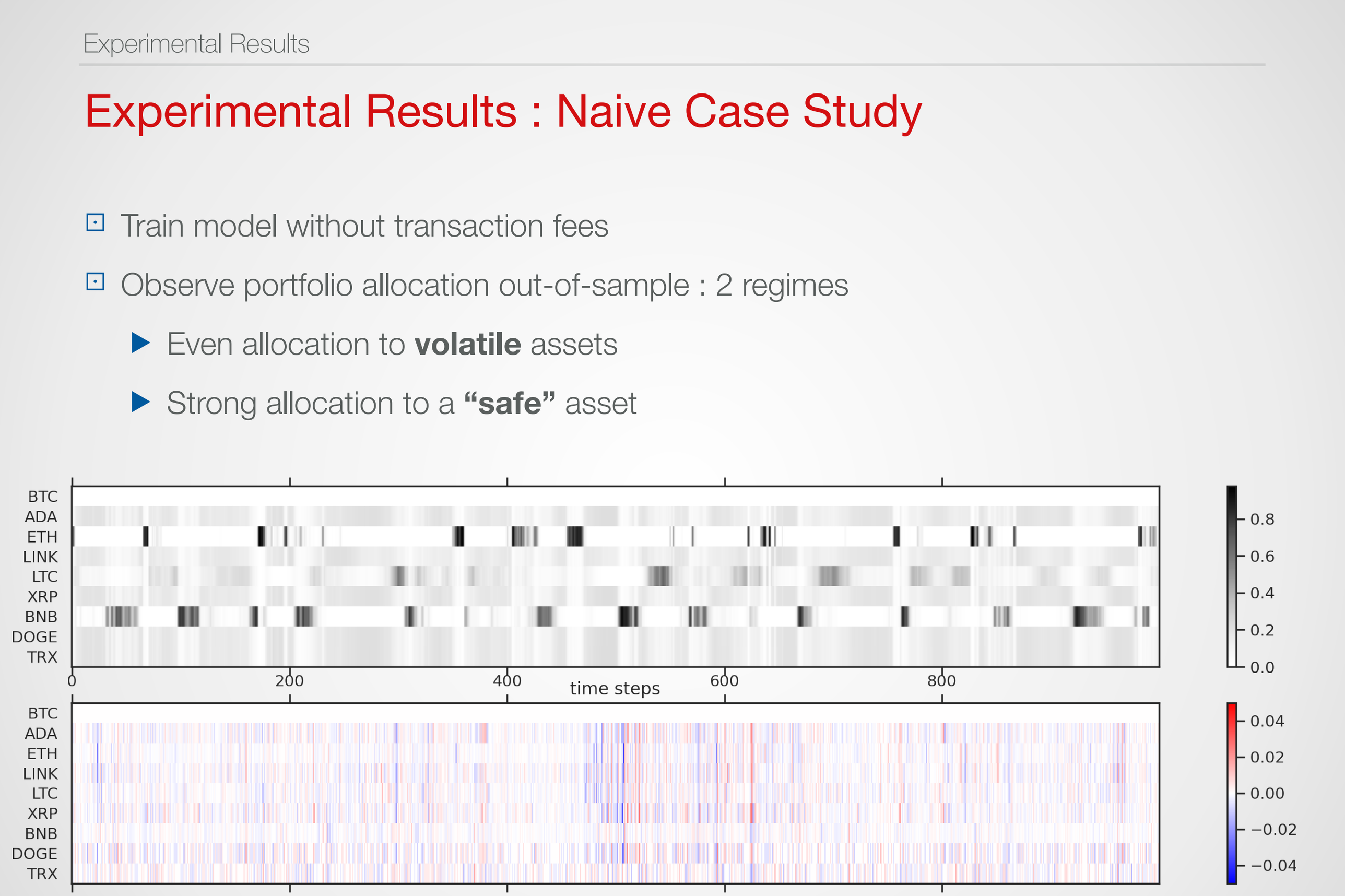
- Callback to EDA :

Symbol	Mean	Std	Skew	Kurt	Min	Max	JB	MDD	ACF(1)	Hurst
BTCUSDT	0.00003	0.0050	-0.7448	61.0406	-0.1819	0.1363	14437201.46***	-77.20%	-0.0192	0.5496
ADABTC	0.00000	0.0061	0.4743	42.4179	-0.1535	0.1481	6971121.79***	-92.47%	-0.0411	0.5494
ETHBTC	0.00000	0.0034	0.1610	26.7305	-0.0919	0.0714	2767345.84***	-79.34%	-0.0036	0.5408
LINKBTC	-0.00001	0.0064	0.0130	28.4720	-0.1586	0.1204	3139222.49***	-91.88%	-0.0340	0.5275
LTCBTC	-0.00002	0.0051	0.2741	73.8510	-0.2016	0.1836	21121412.35***	-89.82%	-0.0695	0.4990
XRPBTC	-0.00000	0.0066	0.8806	95.7926	-0.2036	0.2367	35546567.78***	-84.22%	-0.0214	0.5229
DOGEBTC	0.00002	0.0158	0.4253	25.7146	-0.3365	0.4238	2563420.26***	-87.46%	-0.3386	0.5259
BNBBTC	0.00001	0.0046	0.1968	40.0950	-0.1083	0.1119	6225992.27***	-72.84%	-0.0234	0.5687
TRXBTC	0.00000	0.0065	1.1543	43.0981	-0.1125	0.2069	7213516.40***	-83.33%	-0.1689	0.5332



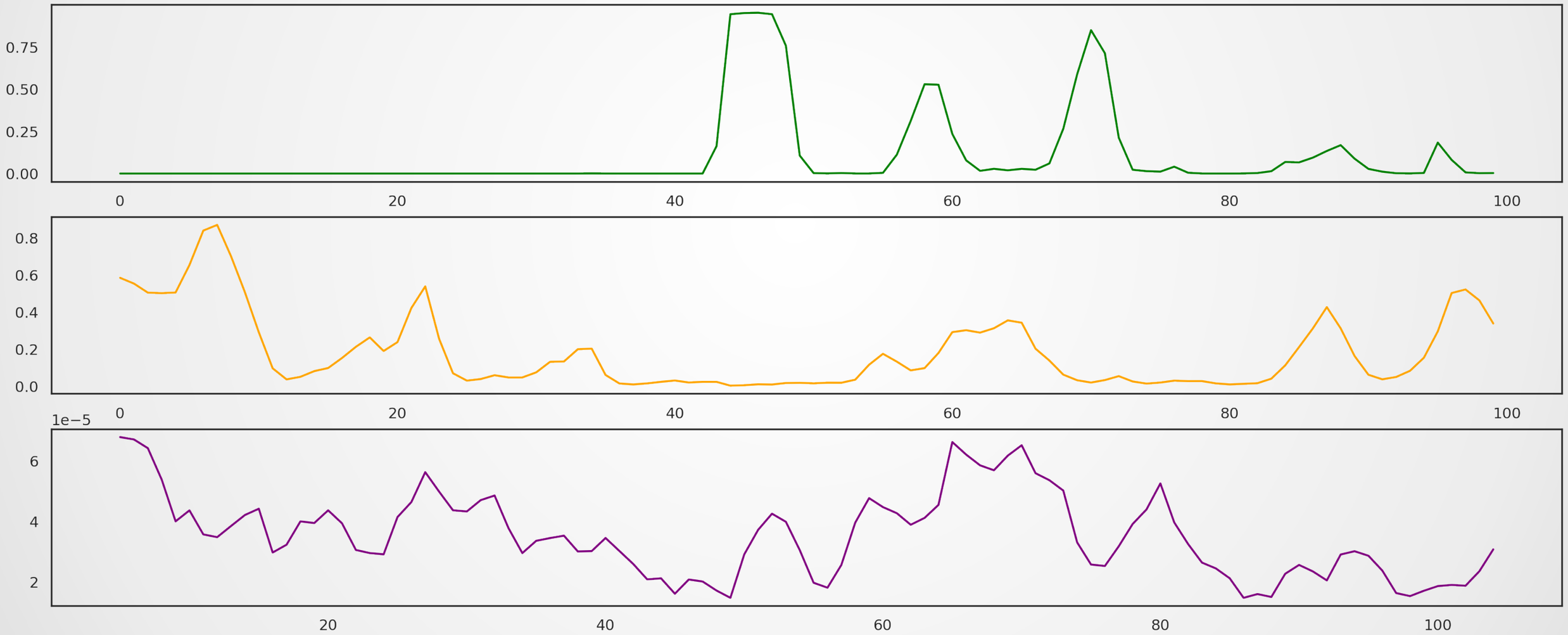
Experimental Results : Naive Case Study

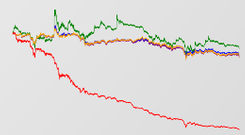
- Train model without transaction fees
- Observe portfolio allocation out-of-sample : 2 regimes
  - Even allocation to **volatile** assets
  - Strong allocation to a **“safe”** asset



Experimental Results : Naive Case Study

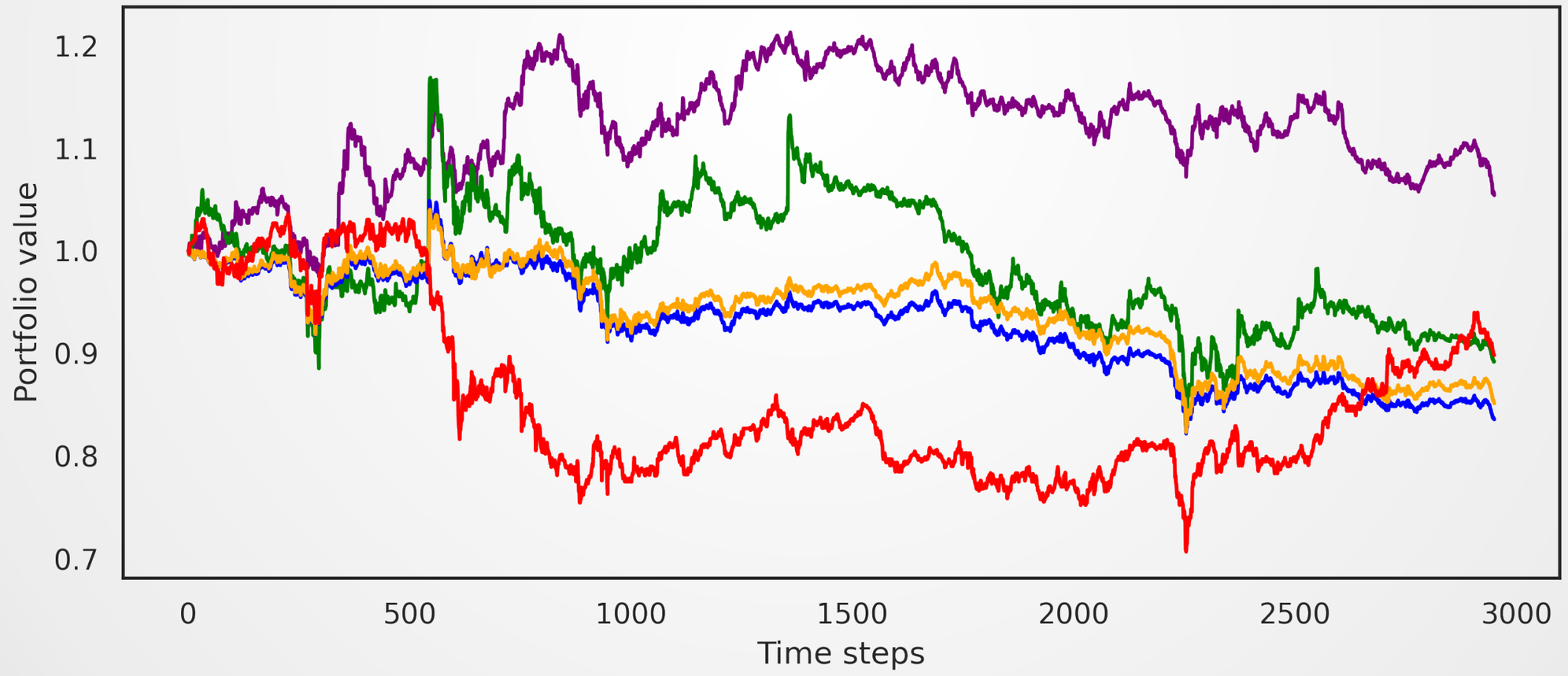
- Correlation between total pf volatility and allocation to safe assets
  - ETH and LTC examples





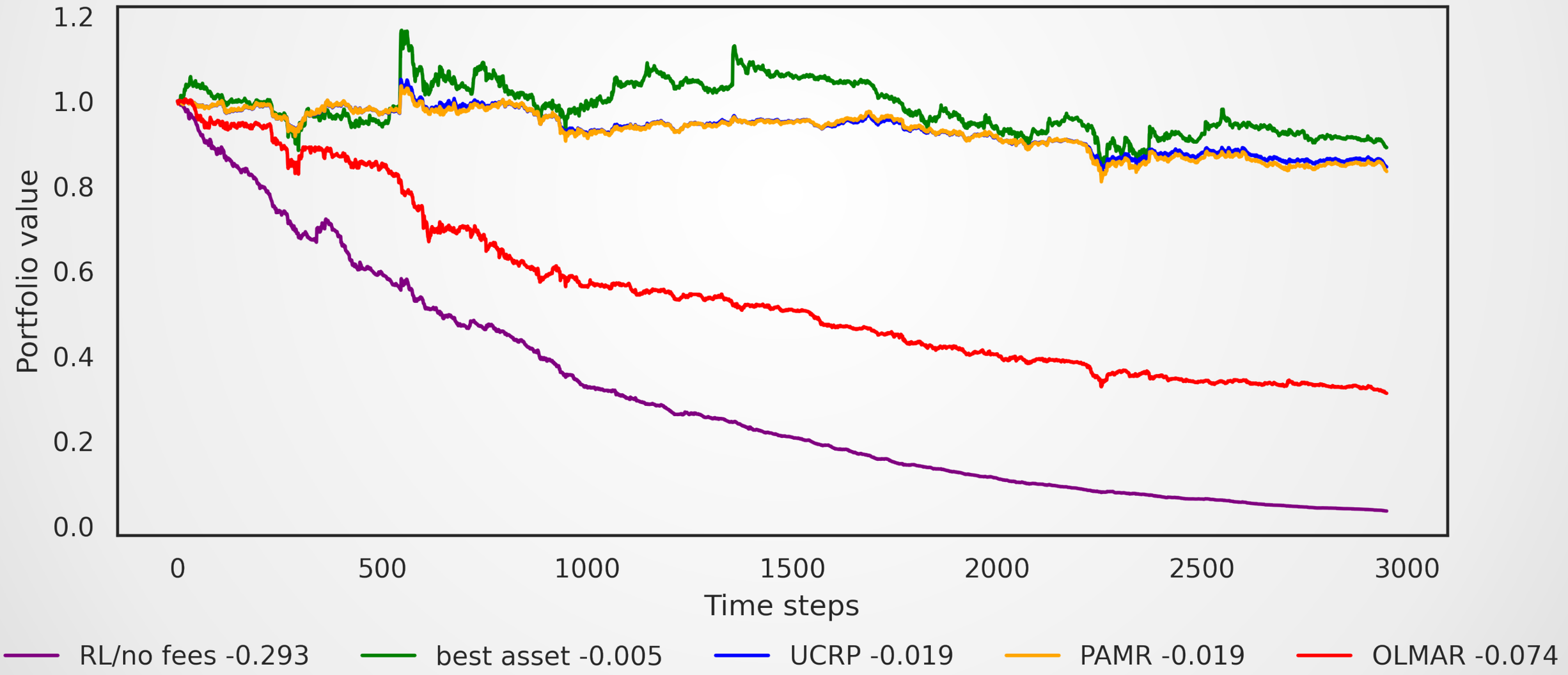
Experimental Results : Naive Case Study

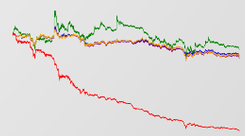
- Benchmark on recent data (w/o fees)
- Observe portfolio value out-of-sample
  - Training: June 2024 - Jan 2025
  - Test: Feb 2025 - May 2025



Experimental Results : Reality Check

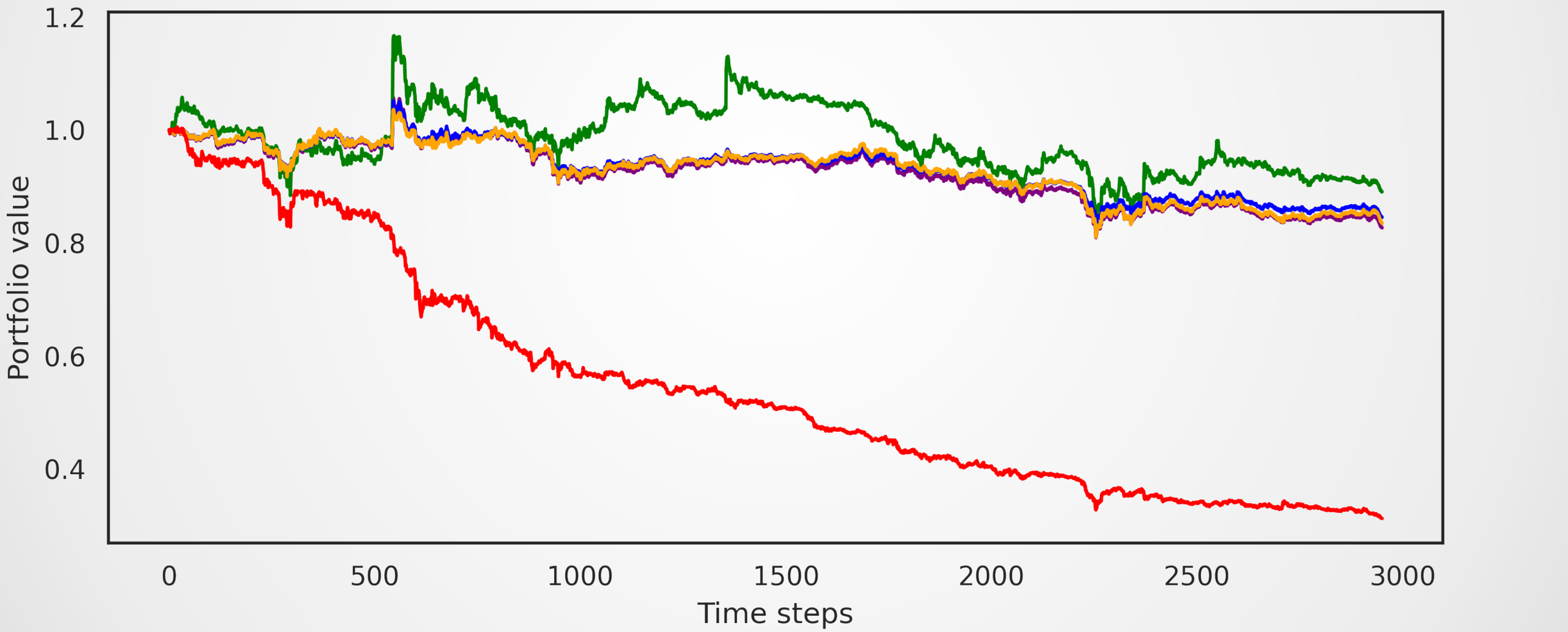
- Benchmark on recent data (with fee 0.001)
- RL policy loses most of pf value (too frequent rebalancing)

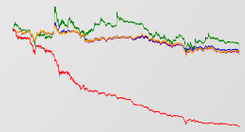




Experimental Results : Adding Transaction Fees

- Transaction fees : policy converges to holding all assets
- The RL algorithm doesn't find better than CRP

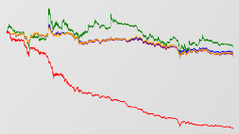




Experimental Results : Discussion

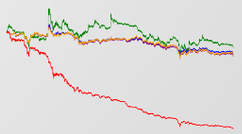
- Time-sensitive results
  - Bull market => all policies win (Convergence problems)
  - Correlation in CC market

- RL policy loses most of pf value (too frequent rebalancing)

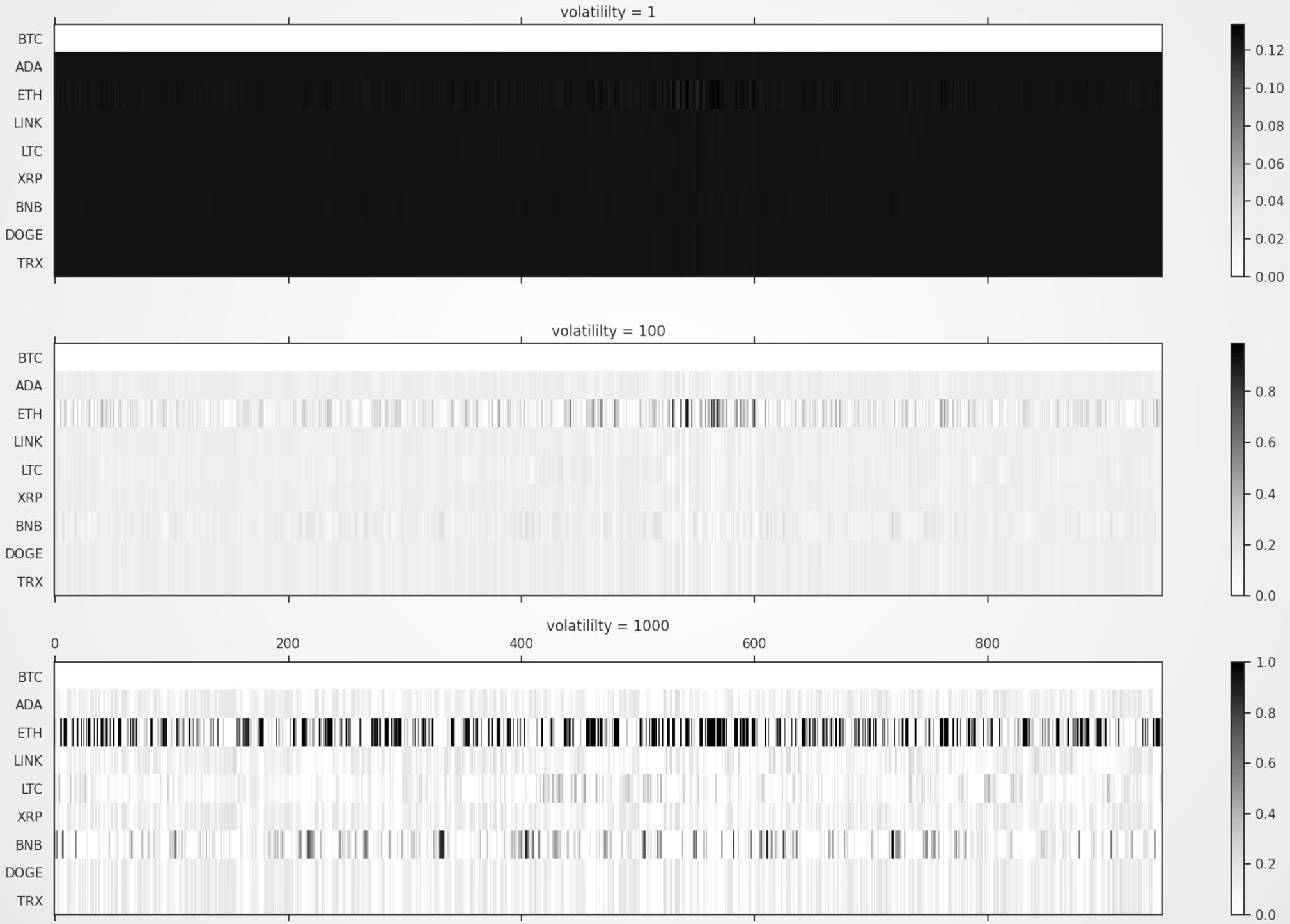


Discussion : TS forecasting results

- Idea : train NBEATS to forecast next price return, use as policy
  - Use simplex projection to obtain weights
- We observe low volatility in predictions (returns close to 1)
  - Tune volatility to study patterns ?



Discussion : TS forecasting results

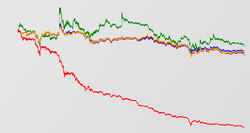


Discussion : TS forecasting results

- NBEATS forecasts the least volatile coins with the highest std
  - Explains allocation to “safe” assets
- Problem with sparseness of data (returns = 0)

	Ticker	Sparseness (%)	Mean Volume ($)	Median Volume ($)
	ETH	5.7	2120381	1169886
	LTC	11.2	115926	64006
	ADA	15.9	151004	39424
	LINK	5.9	81621	31646
	XRP	10.9	717887	218843
	BNB	2.6	402777	231924
	TRX	40	87666	34205
	DOGE	36	451477	73114

Trading Dataset (2024-06-20 - 2025-04-31)



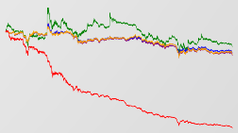
Discussion : Perspectives

- Longer Timeframes ?
  - Stability of statistical results in time

- Correlation between crypto assets
  - Possibility of hedging ?

- Pairs against BTC
  - Good proxy for crypto risk free rate?

- Volatility in bursts
  - Forecast possible ?



Reinforcement Learning for crypto portfolio management

Owen Chaffard
Siang-Li Jheng
Megang Nkamga Junile Staures

Ladislaus von Bortkiewicz Professor of Statistics
Humboldt-Universität zu Berlin
CardoAI
Bucharest University of Economic Studies

20250312 Cha Kor Hae Cardo Loan Forecasting.key