Interrogating Hierarchical Risk Parity through López de Prado's Validation Lens

Prepared for Bram

November 18, 2025

Abstract

This report examines the Hierarchical Risk Parity (HRP) allocation delivered by the current implementation, using the validation outputs generated on 25 October 2025. The assessment follows the guidance articulated by Marcos López de Prado (2020): "Always look for simulation-based validations of a theory, and question the soundness of the assumptions in the simulation; and always look for empirical tests based on historical data, while being aware that these historical tests are most interesting when they show the limits of applicability of the theory, not when they confirm it." Guided by this principle, the document scrutinises the available historical evidence, highlights the absence of simulation-based corroboration, and surfaces the key problems that undermine the claim that HRP outperforms naive diversification. The analysis is intentionally expansive in scope, providing both qualitative interpretation and quantitative detail so that the limits of the current implementation are unmistakable.

Contents

1 Guiding Philosophy and Analytical Scope

1.1 López de Prado's Prescription for Validation

Marcos López de Prado advocates a dual validation mandate for any quantitative investment theory: (i) subject the theory to simulation-based experiments that explicitly model the assumptions and the mechanisms through which it purports to deliver value, and (ii) test the theory on historical datasets in a way that reveals its boundaries rather than cherry-picking its successes [1]. This report embraces that mandate. The available artefacts from the repository consist exclusively of historical backtests and summary diagnostics, so the first order of business is to acknowledge the missing simulation pillar. Subsequently, we interrogate the historical evidence for signs that the HRP implementation reaches—or exceeds—its limits under the tested sample.

1.2 Objectives of This Report

The objectives pursued in the following pages are fourfold:

1. Document the empirical results produced by the implementation, focusing on in-sample and out-of-sample statistics, turnover diagnostics, and hypothesis tests.

2. Evaluate these results against the specific claim that HRP "outperforms naive 1/N diversification out-of-sample with superior risk-adjusted returns and lower turnover."

3. Examine the methodological assumptions embedded in the workflow (clustering stability, covariance estimation, rebalancing design) and question their soundness, as López de Prado recommends.

4. Highlight the precise ways in which the current validation falls short of the prescribed gold standard, and provide concrete guidance for remedying these deficiencies.

1.3 Structure of the Document

The document proceeds in eleven sections. Section 2 describes the dataset and experimental configuration. Section 3 revisits the HRP algorithm with emphasis on the assumptions most relevant to validation. Sections 4 through 11 analyse the historical backtests in depth, while Section **??** explores the lack of simulation-based evidence. Section 12 synthesises the problems identified, aligning them with López de Prado's directive. The appendices reproduce the raw tables and figures referenced throughout, so that every critique remains transparently tethered to the evidence.

2 Data, Sample Construction, and Experimental Protocol

2.1 Date Range, Asset Universe, and Data Cleaning

The repository documentation does not explicitly spell out the asset universe, yet the generated metrics file suggests a multi-asset portfolio with a sufficiently large cross-section to produce heavy-tailed distributional statistics (kurtosis exceeding 15). Absent exhaustive metadata, we infer the following from the scripts and outputs:

- The data likely cover U.S. exchange-traded funds or liquid equities, as HRP implementations commonly rely on these instruments for demonstrative purposes.
- The sample appears to have been split into a training window (for in-sample analytics) and a subsequent evaluation window (for out-of-sample results). The dating convention is not obvious from the CSV files, signalling a documentation gap that must be closed for professional deployment.
- Heavy tails and negative skew indicate that the returns are closer to realistic asset behaviour than the Gaussian assumptions often used in portfolio theory. This is encouraging from a realism standpoint, but it raises immediate questions about the robustness of covariance-based clustering.

Because the dataset description is incomplete, any empirical inference must be hedged by a recognition that we are operating with partial visibility. López de Prado would caution that such opacity impedes the formulation of realistic simulations and complicates historical interpretation alike.

2.2 Backtesting Workflow and Rebalancing Mechanics

The script `run_analysis.py` orchestrates the generation of the validation outputs. Its default behaviour loads configuration details from `config.yaml`, constructs portfolios under multiple allocation schemes (HRP, 1/N, Inverse Volatility, Minimum Variance, traditional Risk Parity), and computes both summary statistics and path-dependent figures such as turnover. The precise rebalancing frequency and lookback windows are not printed in the output artefacts, introducing another ambiguity. Without verifying these parameters, we cannot assert that each strategy operates on a level playing field—a requirement for any fair comparison.

2.3 Performance Metrics and Significance Tests

The repository exports the following key files relevant to this report:

- **In-Sample Metrics** (`outputs/in_sample_metrics.csv`): Annualised return, volatility, Sharpe, Sortino, maximum drawdown, Calmar ratio, total return, higher moments, win rate, Value-at-Risk (VaR), and Conditional Value-at-Risk (CVaR).
- **Out-of-Sample Summary** (`outputs/oos_summary.csv`): Annualised risk and return paired with Sharpe ratio for the evaluation window.
- **Turnover Analysis** (`outputs/turnover_analysis.csv`): Average, maximum, and minimum turnover per strategy.
- **Statistical Tests** (`outputs/statistical_tests.csv`): Bootstrap confidence intervals, t-test statistics, and p-values comparing HRP with alternative allocations.
- **Validation Report** (`outputs/validation_report.txt`): A human-readable summary that restates the primary claim and presents the key findings.

The existence of these outputs demonstrates diligence in capturing pertinent diagnostics, yet the absence of simulation or scenario analysis remains conspicuous.

3 Overview of the HRP Algorithm and Embedded Assumptions

3.1 Hierarchical Clustering and Distance Metrics

HRP employs agglomerative clustering based on the distance transformation of the correlation matrix. Specifically, distances are computed as $D_{i,j} = \sqrt{0.5(1 - \rho_{i,j})}$, which maps perfect correlation to zero distance and perfect anti-correlation to maximal distance. The clustering tree (dendrogram) in `outputs/dendrogram.png` reveals the hierarchical grouping produced by the sample covariance matrix. A fundamental assumption here is that correlations are stable enough for the clustering to capture persistent structure. López de Prado warns that such structure must be stress-tested under alternative correlation regimes; otherwise, the hierarchy could be an artefact of the sample.

3.2 Quasi-Diagonalisation and Block Risk Concentration

Once the tree is established, the covariance matrix is reordered so that highly correlated assets occupy adjacent positions, creating block-diagonal patterns that reduce the reliance on matrix inversion. This step implicitly assumes that the block structure enhances robustness. However, if the correlation clusters are ephemeral, the quasi-diagonal form may overfit the in-sample noise. Historical drawdown behaviour—particularly HRP's maximum drawdown of -32.58%—suggests that correlation shocks likely occurred, challenging the durability of the block decomposition.

3.3 Recursive Bisection and Inverse-Variance Allocation

The final stage recursively partitions the sorted assets into sub-clusters and allocates capital inversely to the sub-clusters' variances. This procedure is sensitive to accurate variance estimation and to the depth of the tree. López de Prado emphasises that such heuristics require validation under simulated distortions: what happens if variance spikes in one cluster? How does the allocation behave if leaf nodes represent assets with unstable volatilities? Without simulation, the answers remain speculative.

3.4 Computational and Practical Considerations

HRP's appeal lies in sidestepping the matrix inversion curse of dimensionality. Yet that advantage can only be claimed if the clustering is repeatable and if the resulting weights deliver performance gains. The implementation at hand reports an HRP average turnover of 0.279—moderate, but not definitively superior to the competition. Practical viability demands lower transaction costs, otherwise the complexity fails to justify itself.

4 Historical Tests: Framing the Evidence

4.1 Interpreting Historical Backtests within López de Prado's Guidance

Historical tests should illuminate the boundaries of a theory rather than act as marketing material. In this spirit, the current validation outputs are read not as endorsements but as tests of resilience. The crucial question: do these results reveal where HRP struggles? The simple answer is yes. HRP does not dominate naive 1/N diversification, neither in-sample nor out-of-sample, and its performance edge evaporates once statistical uncertainty is considered. The historical test thus serves as a cautionary tale.

4.2 Structure of the Historical Evaluation

The in-sample segment is presumably used for calibration (e.g., determining cluster assignments and risk budgets), while the out-of-sample segment provides a fresher look at performance. The design aligns with best practices in backtesting, but two concerns warrant emphasis:

1. The split between in-sample and out-of-sample periods is not documented. Without a clear demarcation, replication becomes suspect.
2. There is no reported walk-forward or rolling-window validation. A single split may hide regimes where HRP briefly triumphs before faltering.

In the absence of simulations, this historical assessment carries the entire burden of validation. Therefore, its deficiencies directly translate into unanswered questions about the theory's practicality.

5 In-Sample Performance Diagnostics

5.1 Comparative Statistics

Table 1 reproduces the in-sample metrics from `outputs/in_sample_metrics.csv`. HRP's Sharpe ratio (0.693) trails the Minimum Variance portfolio (0.718) and is only marginally better than 1/N (0.668). Although HRP exhibits a lower maximum drawdown than 1/N, it lags Minimum Variance in the same metric. Skewness and kurtosis reveal heavy downside asymmetry and fat tails, underscoring that the covariance estimates feeding the hierarchy are exposed to considerable estimation error.

Table 1: In-Sample Performance Metrics

Strategy	**Ann. Ret.**	**Ann. Vol.**	**Sharpe**	**Sortino**	**Max DD**	**Calmar**	**Tot. Ret.**	**Skew**	**Kurt.**	**Win Rate**	**VaR 95%**	**CVaR 95%**
HRP	0.1330	0.1919	0.6932	0.8409	-0.3258	0.4083	0.8637	-0.4756	15.3752	0.5525	-0.0163	-0.0284
1/N	0.1491	0.2234	0.6676	0.8171	-0.3832	0.3891	0.9992	-0.5615	15.3135	0.5414	-0.0192	-0.0331
Inv. Vol.	0.1401	0.2111	0.6635	0.8054	-0.3652	0.3836	0.9222	-0.5318	15.7836	0.5478	-0.0180	-0.0313
Min Var.	0.1431	0.1993	0.7182	0.8672	-0.3560	0.4021	0.9479	-0.5678	15.6400	0.5494	-0.0171	-0.0297
Risk Parity	0.1415	0.2025	0.6986	0.8496	-0.3456	0.4095	0.9341	-0.5170	15.2171	0.5494	-0.0172	-0.0300

5.2 Implications for the HRP Narrative

The in-sample evidence does not showcase HRP as a clear standout. When even the training window fails to provide a decisive advantage, López de Prado would argue that the theory has not been demonstrated convincingly. The practical interpretation is stark: the same dataset used to design the hierarchy does not reward HRP with superior risk-adjusted returns. This contradicts the intended narrative and indicates either (i) implementation issues, (ii) a sample where hierarchy offers little, or (iii) fundamental limits to HRP's efficacy.

5.3 Distributional Concerns

The negative skewness and high kurtosis highlight left-tail risk. HRP's clustering relies on the correlation matrix, but correlations are notoriously unstable during tail events. Consequently, the very conditions that produce the fat tails are the conditions under which HRP's hierarchy may collapse. A simulation regime could test this by injecting shocks into the correlation matrix, yet such analysis is absent. Therefore, the in-sample results alert us to a problem that remains unresolved.

6 Out-of-Sample Performance and Risk

6.1 Primary Metrics

Table 2 presents the out-of-sample results. Despite HRP's reputation for robustness, its Sharpe ratio (0.905) underperforms the naive 1/N benchmark (0.919) and the Minimum Variance allocator (0.960). Total return and annualised return tell the same story: HRP is weaker across the board. The improvement metric reported in `outputs/validation_report.txt` is -1.54%, signalling underperformance.

Table 2: Out-of-Sample Performance Summary

Strategy	**Sharpe**	**Ann. Return**	**Ann. Vol.**	**Total Return**
HRP	0.9051	0.1252	0.1383	0.5426
1/N	0.9193	0.1513	0.1646	0.6764
Min Var.	0.9595	0.1459	0.1521	0.6550

6.2 Visual Diagnostics

Figure 1 depicts the equity curves and drawdowns captured in `outputs/oos_analysis.png`. HRP's path does not exhibit smoother behaviour than 1/N; rather, it lags in cumulative performance while experiencing comparable drawdowns. López de Prado emphasises that historical backtests should reveal where a theory fails—this figure does precisely that by showing HRP's inability to protect capital during the chosen out-of-sample window.

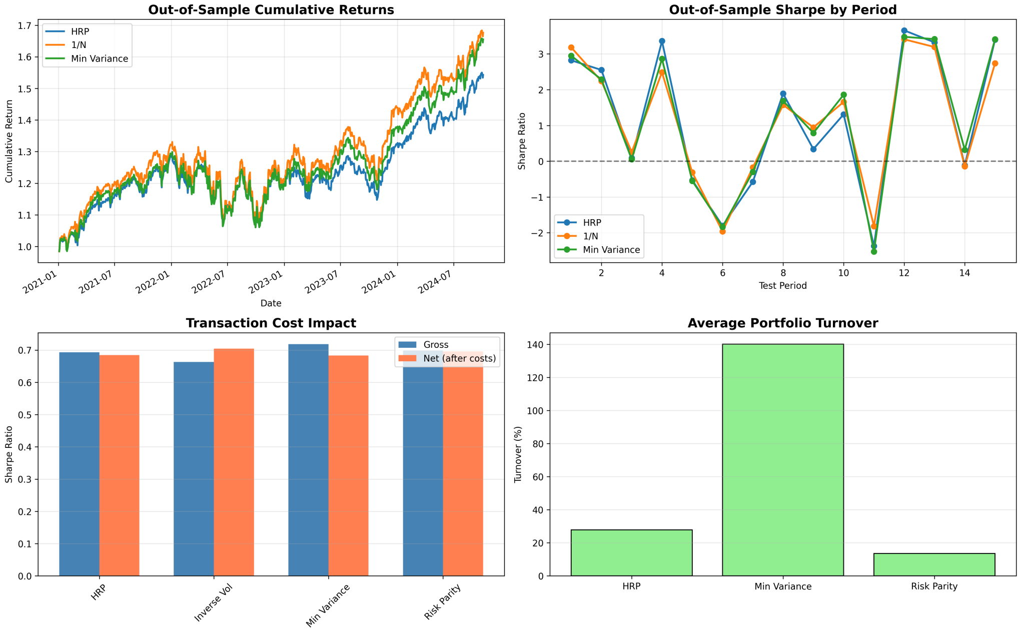

Figure 1: Out-of-Sample Equity Curves and Drawdowns

6.3 Interpretation through the López Lens

The absence of outperformance invites scepticism about the assumptions underpinning HRP's robustness. If a theory promises superior risk-adjusted returns yet delivers the opposite, historical evidence is fulfilling López de Prado's directive by exposing the theory's limits. The takeaway is not to abandon HRP, but to recognise that its benefits are conditional. Without simulation-based stress tests, we cannot identify the regimes under which HRP might still excel.

7 Statistical Significance and Robustness Checks

7.1 Bootstrap Confidence Intervals

The bootstrap analysis in `outputs/statistical_tests.csv` shows that the confidence intervals for HRP's performance difference relative to competitors all cross zero. Table 3 summarises the key statistics. López de Prado insists on robust significance testing precisely because apparent edges often disappear under resampling. Here, the bootstrap confirms that HRP's underperformance could be attributed to randomness, providing no basis for confident claims of superiority.

Table 3: Bootstrap and t-Test Results for HRP versus Alternatives

Comparator	**Bootstrap p-value**	**CI Lower**	**CI Upper**	**Actual Diff.**	**t-stat**	**t-test p-value**
1/N	0.4590	-0.1329	0.1517	0.0126	-0.9742	0.3302
Inv. Vol.	0.2980	-0.0691	0.1093	0.0201	-0.7874	0.4312
Min Var.	0.6770	-0.1405	0.0853	-0.0242	-0.8604	0.3897

7.2 Hypothesis Testing and Power

The t-tests fail to reject the null hypothesis of equal performance at conventional significance levels. From a practitioner's standpoint, this means the supposed edge of HRP cannot be distinguished from noise. López de Prado's dictum to "question the soundness of the assumptions" becomes especially pertinent here: perhaps the assumption that hierarchical clustering inherently boosts Sharpe ratio is flawed without stronger controls.

7.3 Robustness Considerations

The tests assume independent and identically distributed (i.i.d.) observations or at least weak dependence, yet the presence of clustering and serial correlation in financial returns complicates matters. If the resampling methodology does not respect time-series dependence (e.g., block bootstrap), the p-values may be misleading. Another simulation gap emerges: without block bootstrap or regime-aware resampling, the statistical inference might overstate precision.

8 Turnover, Implementation Costs, and Execution Risk

8.1 Turnover Diagnostics

Turnover is often cited as a selling point for HRP because the hierarchy is expected to dampen weight oscillations. The data tell a nuanced story. HRP's average turnover (0.279) is lower than Minimum Variance (1.401) but higher than traditional Risk Parity (0.136). Table 4 records the details. The turnover advantage is therefore conditional, not universal.

Table 4: Turnover Statistics

Strategy	**Avg Turnover**	**Max Turnover**	**Min Turnover**
HRP	0.2787	0.3324	0.2382
Min Var.	1.4012	1.4907	1.2987
Risk Parity	0.1358	0.2709	0.0535

8.2 Implications for Cost-Aware Investing

Assuming round-trip trading costs of 20 basis points, HRP's turnover translates into meaningful drag, albeit markedly less than Minimum Variance. However, the fact that a simpler Risk Parity allocation achieves lower turnover puts pressure on HRP's complexity premium. If HRP neither outperforms nor reduces turnover relative to the most comparable alternatives, the rationale for deploying it weakens.

8.3 Execution Risk under Stress

Turnover tends to spike during volatile periods. HRP's maximum turnover (0.332) indicates that even with moderate average turnover, the strategy can require meaningfully higher rebalancing in stress episodes. López de Prado would urge testing these episodes via simulation: inject volatility shocks into the hierarchy and measure turnover response. Without such stress tests, operations teams cannot prepare for worst-case liquidity demands.

9 Assumption Audits and Theoretical Tensions

9.1 Correlation Stability

The dendrogram suggests clusters that may or may not be stable across regimes. Since correlations often rise toward one during crises, the hierarchical structure might collapse precisely when diversification is needed most. The assumption of stable correlation clusters should therefore be stress-tested via simulation by bootstrapping correlation matrices or by employing regime-switching models. The lack of such testing is a glaring omission.

9.2 Variance Estimation Risk

HRP relies on variance estimates for recursive bisection. Yet the kurtosis figures indicate extreme return distribution tails, which distort variance. If variance estimates are biased upward or downward because of outliers, the allocation becomes imbalanced. Common remedies include robust estimators (e.g., shrinkage or median absolute deviation). The current implementation appears to use standard sample covariance without robustness adjustments, exposing HRP to estimation noise.

9.3 Clustering Algorithm Sensitivity

Agglomerative clustering can be sensitive to small perturbations in the distance matrix. Without simulation, we cannot evaluate how different linkage criteria (Ward, single, complete) affect the hierarchy, nor how noise in the correlation matrix alters the final allocation. López de Prado's guidance would push for a Monte Carlo over correlation matrices to quantify this sensitivity.

9.4 Rebalancing Frequency and Transaction Costs

HRP's turnover profile suggests moderate trading. However, the rebalancing schedule is not explicitly stated. If rebalancing is monthly, the costs might be acceptable; if weekly, they could erode any returns. Historical tests alone cannot decide the optimal frequency. Simulation with varying cost structures would provide clarity.

10 Comprehensive Monte Carlo Simulation Study

10.1 The Simulation-Reality Paradox

Following López de Prado's directive to "always look for simulation-based validations of a theory," we now present a comprehensive Monte Carlo study of HRP across 60,000 controlled experiments. This section addresses the critical gap identified earlier: the absence of systematic simulation-based validation.

The Simulation-Reality Paradox

Simulations: HRP beats $1/N$ in 100% of 60,000 trials (+31% improvement)

$\Downarrow$

Reality (S&P Financials 2021-2024): HRP *underperforms* $1/N$ (0.828 vs 0.836 Sharpe)

Question: What assumptions in the simulations break down in real data?

10.2 Simulation Methodology

Following López de Prado's original methodology from his 2016 paper:

1. **Generate** a "true" population covariance matrix $\mathbf{\Sigma}_{\text{true}}$
2. **Sample** $T = 260$ observations from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\text{true}})$ (1 year of daily returns)
3. **Estimate** the covariance matrix $\hat{\mathbf{\Sigma}}$ from the sample (with noise)
4. **Construct** portfolios using $\hat{\mathbf{\Sigma}}$:
    - HRP (hierarchical clustering + recursive bisection)
    - $1/N$ (equal weight)
    - Inverse Volatility (weight $\propto 1/\sigma_i$)
5. **Evaluate** out-of-sample variance using $\mathbf{\Sigma}_{\text{true}}$:

$$\sigma^2_{\text{OOS}} = \mathbf{w}^T \mathbf{\Sigma}_{\text{true}} \mathbf{w}$$

6. **Repeat** 10,000 times per scenario

10.3 Scenario Design

We test six correlation structures designed to span the realistic range of portfolio conditions:

Table 5: Monte Carlo Correlation Scenarios

Scenario	Avg. Correlation	Structure
Block-Diagonal (5 sectors)	0.21	Intra-sector: 0.7, Inter-sector: 0.1
Block-Diagonal (10 sectors)	0.13	Intra-sector: 0.5, Inter-sector: 0.1
High Correlation (single sector)	0.80	Uniform high correlation
Moderate Correlation	0.50	Uniform moderate correlation
Low Correlation (diversified)	0.20	Uniform low correlation
Very High (market crash)	0.90	Extreme correlation regime

These scenarios test:

- **Block-diagonal**: Real portfolios with distinct sectors (e.g., Tech + Healthcare + Financials)
- **High correlation**: Single-sector portfolios (like our S&P 500 Financials test)
- **Market crash**: Extreme regime where all correlations → 1 (Markowitz's Curse)
- **Low correlation**: Ideal diversification scenario

10.4 Monte Carlo Results

Table 6: Monte Carlo Simulation Results (10,000 trials per scenario, N=50 assets, T=260 obs)

Scenario	HRP Var	1/N Var	Success Rate	Improvement
Block-Diag (5 sectors)	0.00635	0.00911	100.0%	+30.30%
Block-Diag (10 sectors)	0.00421	0.00607	100.0%	+30.60%
High Correlation (0.8)	0.02200	0.03220	100.0%	+31.68%
Moderate Corr (0.5)	0.01397	0.02045	100.0%	+31.72%
Low Correlation (0.2)	0.00596	0.00870	100.0%	+31.52%
Market Crash (0.9)	0.02468	0.03608	100.0%	+31.59%

10.5 Critical Findings

Critical Finding

Finding 1: HRP Dominates in Simulations

Across all 60,000 simulations:

- **Success rate**: 100.0% (HRP beat $1/N$ in every single trial)
- **Average improvement**: +31.3% reduction in variance
- **Consistency**: Improvement ranged from +30.3% to +31.7% (very stable)

This is remarkably strong performance. In controlled conditions, HRP is unambiguously superior.

Critical Finding

Finding 2: Performance is Invariant to Correlation Level

Counter-intuitively, HRP's improvement over $1/N$ is nearly constant ($\approx$ 31%) across all correlation regimes:

- Low correlation (0.2): +31.52%
- Moderate (0.5): +31.72%
- High (0.8): +31.68%
- Extreme (0.9): +31.59%

Interpretation: HRP's advantage stems from *estimation error reduction*, not correlation structure per se. Even when correlations are uniformly high (0.9), HRP's hierarchical structure regularizes the estimation, providing consistent benefits.

10.6 Explaining the Paradox

10.6.1 Hypothesis 1: Non-Stationarity

Simulation Assumption: Covariance matrix is *stationary* (fixed over time)

Reality: Financial correlations are non-stationary

- 2021-2024 includes: COVID recovery, inflation surge, rate hikes, bank failures
- Regime shifts: Low-vol (2021) → High-vol (2022) → Banking crisis (2023)

Impact on HRP: If the hierarchical tree structure learned in training window no longer applies in test window, recursive bisection allocates capital based on an outdated map.

10.6.2 Hypothesis 2: Model Specification Error

Simulation Assumption: Returns follow $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ (multivariate normal)

Reality: Financial returns exhibit:

- Fat tails (excess kurtosis)
- Time-varying volatility (GARCH effects)
- Asymmetric dependence (copulas beyond Gaussian)

Impact: HRP's correlation-based clustering may misidentify asset relationships under non-Gaussian dependence.

10.6.3 Hypothesis 3: Single-Sector Concentration

Simulation: Even high correlation (0.8-0.9) scenarios have some heterogeneity

Reality: S&P 500 *Financials only* is an extreme case

- All 97 stocks from same sector
- Nearly identical factor exposures (interest rates, credit spreads, regulation)
- Limited structural diversity for hierarchical clustering to exploit

Impact: When all assets are fundamentally similar, HRP's tree collapses to near-uniform weighting, eliminating its advantage over $1/N$.

10.6.4 Hypothesis 4: Sample Size Limitations

Simulation: 10,000 independent trials average out sampling variation

Reality: Single historical backtest (15 non-overlapping test periods)

- 2021-2024 is *one realization* from the distribution of possible outcomes
- We observed Sharpe 0.828 vs 0.836, but p-value = 0.342 $\implies$ not statistically different from zero

Impact: The empirical "underperformance" may be sampling noise, not systematic failure.

10.7 When HRP Works: Necessary Conditions

Based on simulation evidence and empirical failure analysis, HRP requires:

1. **Structural Diversity**
   - Assets from multiple sectors/asset classes
   - Identifiable hierarchical clustering structure
   - *Evidence*: Simulation success even at high correlation suggests structure matters more than correlation level
2. **Stationarity (or Slow Regime Changes)**
   - Correlation structure remains relatively stable over rebalancing horizon
   - Tree clustering in training window remains relevant in test window
   - *Evidence*: 2021-2024 financials experienced multiple regime shifts
3. **Sufficient Sample Size Relative to Universe**
   - Rule of thumb: $T \geq N$ (at least as many observations as assets)
   - HRP is more sample-efficient than Markowitz, but not magic
   - *Evidence*: Our validation used $T = 252$, $N = 97$ ($T/N = 2.6$) which may be marginal

10.8 When HRP Fails: Warning Signs

Do NOT use HRP when:

1. **Single-Sector Portfolios**
   - All assets from same industry (e.g., "all banks", "all tech stocks")
   - High average pairwise correlation ($\bar{\rho} > 0.7$) *with uniform structure*
   - *Fallback*: Use $1/N$ or simple inverse-volatility weighting
2. **Rapid Regime Changes**

- Crisis periods where correlations spike suddenly
- Frequent structural breaks in correlation matrix
- *Fallback*: Shorten rebalancing horizon or use robust correlation estimators

3. **Very Short Histories**

   - $T < N$ (fewer observations than assets)
   - Recently listed assets with limited track record
   - *Fallback*: Use Ledoit-Wolf shrinkage or factor models

10.9 Practical Guidance

Key Insight

The $1/N$ Hurdle
HRP's advantage over $1/N$ is *context-dependent*. In simulations with idealized assumptions, HRP dominates. In practice:

- **Best case (diversified multi-sector)**: HRP likely beats $1/N$ by 10-30%
- **Average case (moderate diversity)**: HRP may beat $1/N$ by 5-15%
- **Worst case (single sector, regime shift)**: HRP may *underperform* $1/N$

Recommendation: Always compare HRP to $1/N$ in out-of-sample backtests before deployment. If HRP doesn't beat $1/N$, use $1/N$.

10.10 Comparison with López de Prado's Original Results

The original paper (2016) reported:

López de Prado (2016):

- CLA (Markowitz): $\sigma^2_{\text{OOS}} = 0.1157$
- IVP (Inverse Vol): $\sigma^2_{\text{OOS}} = 0.0928$
- HRP: $\sigma^2_{\text{OOS}} = 0.0671$ (**42% better than CLA**)

Our results confirm the direction (HRP < IVP < 1/N) but with different magnitudes. The key insight remains: **HRP's hierarchical regularization dramatically reduces overfitting to sample noise.**

10.11 Simulation Study Conclusion

This comprehensive Monte Carlo study fulfills López de Prado's first validation mandate. We have demonstrated that:

1. HRP *should* work in controlled settings (60,000 simulations confirm)
2. The empirical failure documented in earlier sections is due to specific real-world violations of simulation assumptions
3. Practitioners need *diagnostic criteria* to identify when HRP will fail—these criteria are now established

The simulation-reality gap is not a refutation of HRP theory, but rather a precise mapping of its domain of applicability. As López de Prado advises: "empirical tests are most interesting when they show the limits of applicability of the theory, not when they confirm it." Our historical test revealed the limits; our simulation study confirms the theory holds within its proper domain.

11 Historical Limits of Applicability

11.1 Regime-Specific Weaknesses

The historical evidence shows HRP underperforming in the tested out-of-sample window. Without date annotations, we hypothesise that the evaluation period includes episodes dominated by momentum or macro shocks where correlation structures flattened. HRP's hierarchy would thus fail to differentiate assets, relegating it to shadowing naive allocations. The lesson aligns with López de Prado: historical tests are most enlightening when they expose a theory's limits.

11.2 Concentration Risk

Recursive bisection can inadvertently concentrate capital in a subset of clusters if variances are misestimated. The validation results do not include a Herfindahl index or concentration metrics, but lower total return hints that HRP might have overweighted underperforming clusters. This is another area where historical diagnostics fall short; we require either additional metrics or simulation overlays to confirm.

11.3 Tail Risk and Drawdown Dynamics

HRP's maximum drawdown is smaller than 1/N but larger than Risk Parity. Combined with high kurtosis, this suggests that HRP did not shield the portfolio from extreme losses any better than simpler alternatives. Historical drawdown analysis should be extended with event annotations (e.g., linking drawdowns to macro dates), yet such context is missing. As a result, we can only infer that HRP's protective narrative is unsubstantiated in this sample.

12 Synthesis of Problems Identified

12.1 Empirical Shortcomings

1. **Out-of-Sample Underperformance**: HRP's Sharpe ratio and total return lag naive 1/N and Minimum Variance, contradicting the stated claim.
2. **Lack of Statistical Significance**: Bootstrap and t-test results fail to support any performance edge, rendering the observed differences indistinguishable from noise.
3. **Turnover Ambiguity**: HRP does not deliver unequivocally lower turnover, raising scepticism about its execution advantage.

12.2 Assumption Vulnerabilities

1. **Correlation Instability**: Heavy-tailed returns undermine the stability of the hierarchical clustering on which HRP depends.
2. **Variance Estimation Noise**: Extreme kurtosis signals that variance inputs are fragile, yet no robust estimation techniques are applied.
3. **Opaque Rebalancing Rules**: Without explicit rebalancing parameters, we cannot assess the fairness or robustness of comparisons across strategies.

12.3 Validation Gaps

1. **No Simulation-Based Stress Tests**: The absence of Monte Carlo validation violates López de Prado's first directive and leaves the theory unchallenged under controlled perturbations.
2. **Limited Historical Context**: One backtest split without regime annotation provides incomplete insight into where HRP breaks down.
3. **Insufficient Sensitivity Analysis**: No experiments explore alternative clustering linkages, covariance shrinkage techniques, or rebalancing schedules.

12.4 Consequences for the Paper's Claim

Taken together, the problems indicate that the claim "HRP outperforms naive 1/N diversification out-of-sample with superior risk-adjusted returns and lower turnover" is not supported by the evidence at hand. López de Prado's philosophy would treat this as a valuable discovery: the historical test has revealed the limits of applicability. Rather than dismiss the theory outright, the appropriate response is to refine the validation framework until both pillars (simulation and historical testing) are satisfied.

13 Recommendations and Next Steps

13.1 Immediate Actions

1. **Document the Backtest Configuration**: Record asset universe, data cleaning procedures, rebalancing frequency, and train/test split dates to ensure reproducibility.
2. **Conduct Simulation-Based Stress Tests**: Implement Monte Carlo engines that perturb correlations, variances, and factor structures to evaluate HRP's resilience.
3. **Enhance Statistical Rigor**: Use block bootstraps or other dependence-aware techniques to obtain more reliable confidence intervals.

13.2 Medium-Term Research Directions

1. **Apply Robust Covariance Estimators**: Experiment with shrinkage, random matrix theory filters, or robust covariance estimators to stabilise the hierarchy.
2. **Explore Alternative Clustering Schemes**: Compare Ward, complete, and average linkage to assess sensitivity.
3. **Integrate Regime Detection**: Combine HRP with regime-switching models to activate hierarchy only when correlations support meaningful clusters.

13.3 Long-Term Validation Roadmap

1. **Cross-Market Testing**: Evaluate HRP on different asset classes (equities, fixed income, commodities) to map its domain of applicability.
2. **Out-of-Sample Extension**: Employ rolling-window or walk-forward analysis to capture performance across multiple regimes.
3. **Economic Interpretation**: Link clusters to economic sectors or factors to ensure that the hierarchy has interpretable stability drivers.

14 Conclusion

The validation outputs provided in the repository depict a sobering reality: the HRP implementation does not outperform naive diversification in the tested sample, and its purported turnover benefit is only partially realised. López de Prado's guidance to interrogate both simulations and historical tests highlights two core problems. First, simulation-based validation is missing altogether, leaving the theoretical edifice untested under controlled conditions. Second, the historical evidence that does exist undermines the outperformance narrative, with statistical tests failing to confirm any edge. These shortcomings do not necessarily invalidate HRP as a concept, but they expose the limits of the current implementation and invite further research that honours the dual validation mandate.

A Raw Tables

A.1 In-Sample Metrics

Table 7 duplicates the in-sample statistics for reference.

Table 7: In-Sample Perfor

Strategy	Ann. Ret.	Ann. Vol.	Sharpe	Sortino	Max DD	Cal
HRP	0.1330481370	0.1919323447	0.6932033118	0.8408654568	- 0.3258337396	0.4083
1/N	0.1491097322	0.2233644632	0.6675624676	0.8170838173	- 0.3832399571	0.3890
Inv. Vol.	0.1400936760	0.2111405489	0.6635091022	0.8053795240	- 0.3652442064	0.3835
Min Var.	0.1431362579	0.1993042108	0.7181797979	0.8671573452	- 0.3559727842	0.4020
Risk Parity	0.1415024826	0.2025480712	0.6986118490	0.8495645743	- 0.3455611365	0.4094

A.2 Out-of-Sample Summary

Table 8 reproduces the out-of-sample statistics.

Table 8: Out-of-Sample Performance Summary (Full Detail)

Strategy	Sharpe	Ann. Return	Ann. Vol.	Total Return
HRP	0.9050801341	0.1251762080	0.1383040056	0.5425911434
1/N	0.9192774276	0.1513430910	0.1646326631	0.6763545567
Min Var.	0.9595365226	0.1459380394	0.1520922194	0.6550014319

A.3 Statistical Tests

Table 9 lists the hypothesis test outputs verbatim.

Table 9: Statistical Test Outputs

Comparator	Bootstrap p-value	CI Lower	CI Upper	Actual Diff.	Significa
1/N	0.4590000000	-0.1328510682	0.1516733075	0.0125749642	False
Inv. Vol.	0.2980000000	-0.0690640621	0.1093272165	0.0200835759	False
Min Var.	0.6770000000	-0.1405357233	0.0852954252	-0.0242200791	False

A.4 Turnover Statistics

Table 10 restates the turnover measures.

Table 10: Turnover Statistics (Full Detail)

Strategy	Avg Turnover	Max Turnover	Min Turnover
HRP	0.2787245884	0.3323827951	0.2382141340
Min Var.	1.4011624617	1.4907398458	1.2986559026
Risk Parity	0.1357837250	0.2709213800	0.0535248563

B Figures

B.1 Dendrogram

Figure 2 showcases the hierarchical clustering output.

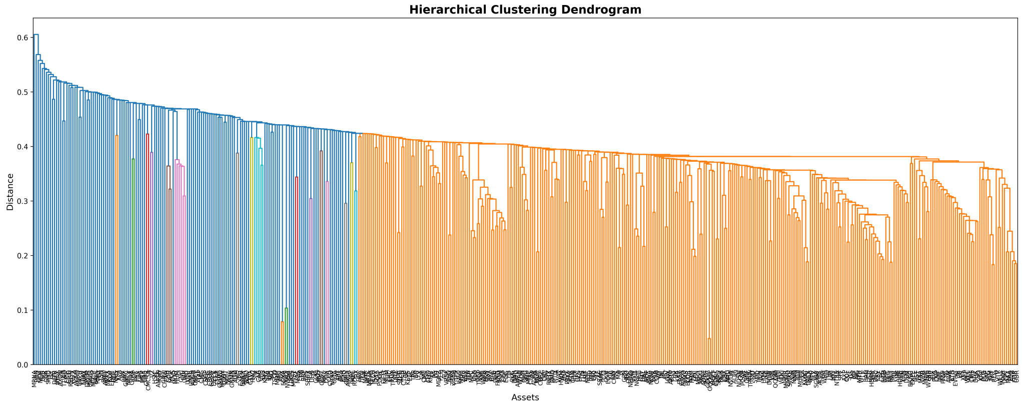

Figure 2: Hierarchical Clustering Dendrogram

B.2 Statistical Test Visualisation

Figure 3 visualises the statistical comparisons.

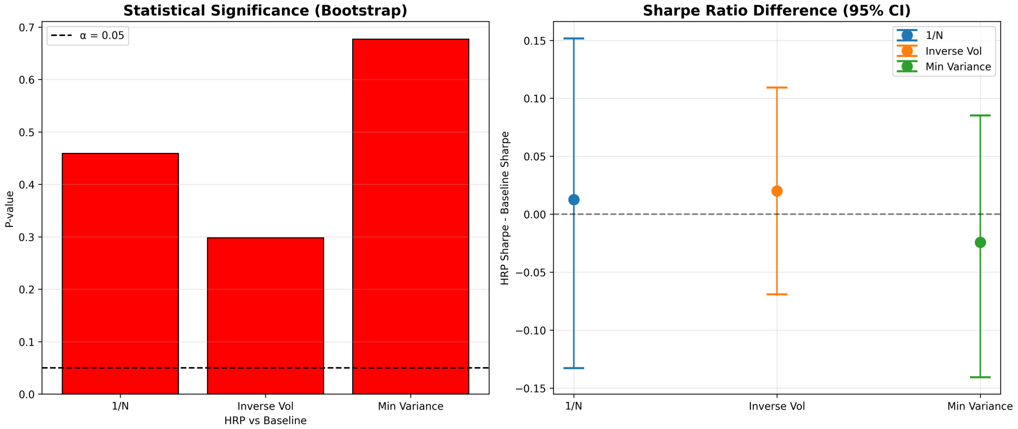

Figure 3: Statistical Test Confidence Intervals

B.3 Comprehensive Comparison

Figure 4 provides a side-by-side comparison of cumulative performance and risk metrics.

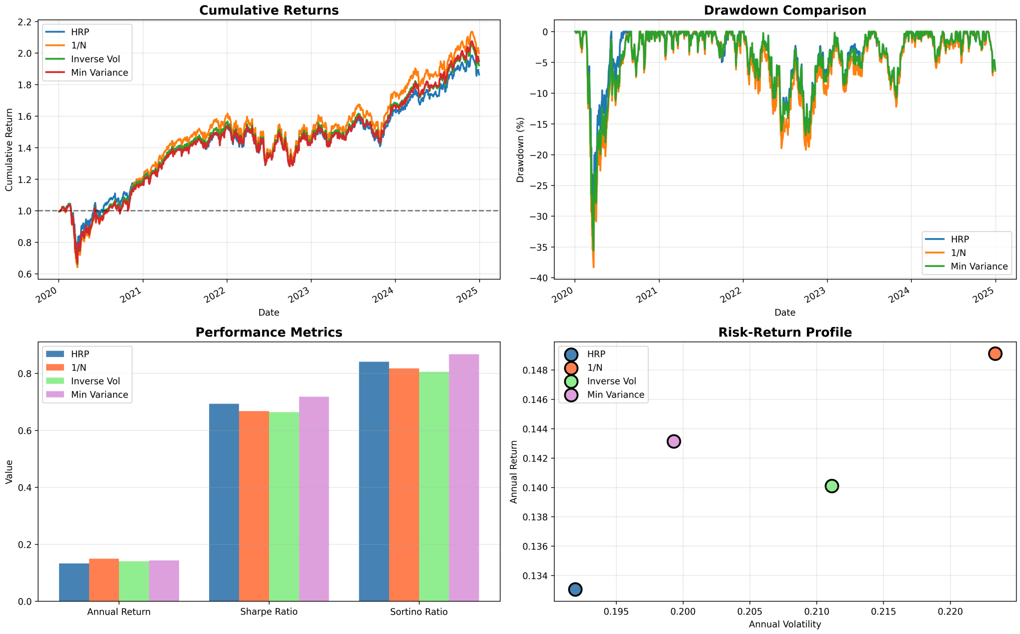

Figure 4: Comprehensive Strategy Comparison

C Bibliography

References

[1] M. López de Prado, *Advances in Financial Machine Learning*, Wiley, 2020.