Portfolio Construction Is Not Alpha: A Hierarchical Long–Short Construction Case Study

[TODO: author names and affiliation]

Working paper — first draft, September 2, 2026

Abstract

We run a leakage-hardened field experiment that isolates a question the anomalies literature almost never tests: holding a real proprietary alpha fixed, how much can portfolio construction alone add at production scale, net of costs? Our answer is a rigorous negative result — hierarchical construction's estimation-error benefits are real but cannot be converted into returns, because construction only moves a transfer coefficient bounded above by one. The vehicle, RHPort, builds a large cash-neutral long–short equity portfolio as a hierarchy of smaller problems: the tradable universe is partitioned into small, deliberately decorrelated subsets; a cash-neutral portfolio is optimized within each subset against a local shrunk covariance; and the resulting local portfolios are treated as a basis recombined under final stock-level risk, gross, name-cap, neutrality, and turnover constraints. The hypothesis — that covariance estimation is more reliable in small cross-sections and that the hierarchy diversifies estimation error — is well grounded in the literature on estimation risk and hierarchical risk parity. Under a point-in-time research protocol with sealed as-of inputs, pre-registered staged experiments, and a single-consumption holdout, we find that the hierarchy improves construction *mechanics* but not returns: local conditioning improves, yet the time-varying basis induces mechanical turnover, richer bases add noise rank rather than information rank, and no recombination, regularization, or transition scheme recovers an edge over a flat stock-space optimizer. The implication for readers of the construction literature is that most claimed gains from clever construction — hierarchical risk parity included — are risk engineering benchmarked against weak baselines, not alpha. We document the failure modes, the protocol that made the negative result credible, and the boundary conditions under which hierarchical construction remains worth revisiting.

Keywords: portfolio optimization, estimation risk, hierarchical risk parity, transaction costs, transfer coefficient, backtest overfitting.

1 Introduction

Mean–variance portfolio construction is notoriously fragile. Sample moments enter the optimizer as if they were known, and the optimizer rewards exactly the directions in which they are most wrong (Michaud, 1989; Best and Grauer, 1991; Chopra and Ziemba, 1993). The difficulty grows with the cross-section: for a universe of several thousand stocks, the sample covariance matrix is singular or nearly so at any realistic lookback, and even shrunk estimators (Ledoit and Wolf, 2004, 2017) leave the optimizer free to bet on poorly estimated low-variance directions. A natural response, with a substantial recent literature, is to impose *hierarchy*: cluster the universe, solve

small problems within clusters where estimation is easier, and combine the cluster-level solutions (López de Prado, 2016; Raffinot, 2017; López de Prado, 2019).

This paper reports what happened when we pushed that idea considerably further than the risk-parity setting in which it is usually studied, and evaluated it the way a production trading system would be evaluated: against a fixed proprietary cross-sectional return forecast, net of turnover penalties, under point-in-time data discipline, and against a competitive flat (non-hierarchical) stock-space optimizer rather than against $1/N$.

The construction, which we call RHPort, proceeds in three layers. First, a *partition engine* divides the rolling tradable universe into small subsets — by deliberate decorrelation, by correlation clustering, or by stratified randomization — with stable subset identities maintained through quarterly refreshes and daily membership changes. Second, a *local layer* optimizes a cash-neutral portfolio within each subset, maximizing the (global) signal against a local Ledoit–Wolf covariance under a local volatility cap and gross constraint. Third, a *meta layer* treats the normalized local portfolios as a basis and solves for non-negative combination coefficients under final stock-level constraints: a portfolio volatility cap using a blend of projected, prequential-process, and structural factor covariances; per-name caps; exact cash neutrality; a gross cap; and L_1/L_2 turnover penalties. The design also admits signed-coefficient challengers that separate positive alpha allocation from internal hedging, cross-subset feedback sweeps, and a final stock-space transition layer that repairs yesterday's realized portfolio into today's constraint set before moving toward today's target.

The a priori case for the architecture had three parts. Covariance estimation should be more reliable in cross-sections of 25–200 names than in one cross-section of 3,000. Many small, partially independent optimizations should diversify estimation error, in the spirit of bagging. And the construction should scale to more names and more frequent bets without one large dense stock-covariance solve.

Our result is negative, and we believe usefully so. Across partition schemes, recombination modes, oversampled bases, risk blends, coefficient regularizations, alpha/hedge separations, and transition-layer penalties — explored under a pre-registered staged design rather than a Cartesian grid — the hierarchy improved several construction mechanics but did not produce a stronger or more stable return profile than the flat benchmark after turnover costs. Five recurring failure modes, documented in Section 6, account for the outcome. The most instructive is that the basis itself moves through time: yesterday's stock portfolio, though still reasonable in stock space, is frequently outside today's positive sleeve cone, generating mechanical turnover that no warm start or coefficient regularization removes, and that post-hoc smoothing can suppress only by destroying the trades that carry the signal — precisely the trade-off that dynamic trading theory predicts for repair-after-the-fact schemes (Gârleanu and Pedersen, 2013).

We frame the economic interpretation with the fundamental law of active management (Grinold, 1989; Clarke et al., 2002). Writing the information ratio as $\mathrm{IR} \approx \mathrm{IC} \cdot \sqrt{\mathrm{BR}} \cdot \mathrm{TC}$, every quantity RHPort manipulates — partitions, bases, recombination, risk blends, penalties — acts on the transfer coefficient TC, which is bounded above by one. Construction can stop a portfolio from wasting a forecast; it cannot strengthen the forecast. When the underlying signal is weak or unstable, the ceiling is low and the elaborate machinery competes for basis points against its own turnover. The case study is, to our knowledge, one of the few published attempts to measure that ceiling directly by holding the forecast fixed and varying only construction, at realistic scale and cost.

A negative result is only as credible as the process that produced it, so the research protocol is

a co-equal contribution. Construction code received sealed as-of snapshots rather than complete panels, with every read recorded in an availability audit; peekahead and cross-sectional sentinel tests were mandatory at module and pipeline level; experiments were staged with immutable selections rather than run as one grid; candidates were frozen — code, configuration, data vintage, cost assumptions, and metrics — before a single-consumption holdout could be opened; and the decision to stop was made *without* consuming the holdout, on the multiple-testing grounds that further tuning against the same development history would fit the process rather than test a hypothesis (Bailey and López de Prado, 2014; Harvey et al., 2016; Arnott et al., 2019).

The paper makes three contributions. First, a documented negative result: a hierarchical long–short construction, considerably richer than hierarchical risk parity, does not beat a competent flat optimizer when the return forecast is held fixed and turnover is costed. Second, a mechanism-level account of *why*, connecting each observed failure mode to known results in estimation risk, random-matrix theory, and dynamic trading. Third, a transferable protocol for producing trustworthy negative results in proprietary quantitative research, where the temptation to keep tuning is strongest.

Section 2 situates the work. Section 3 defines the construction. Section 4 describes the research protocol. Section 5 describes the staged experimental design and benchmarks. Section 6 reports results. Section 7 interprets them and states reopening conditions. Section 8 concludes.

2 Related work

Estimation risk in portfolio choice. That plug-in mean–variance optimization amplifies estimation error is one of the oldest results in the field (Jobson and Korkie, 1980; Michaud, 1989). Sensitivity analyses established that errors in means dominate errors in covariances at moderate risk aversion (Best and Grauer, 1991; Chopra and Ziemba, 1993), while Merton (1980) showed that expected returns are close to unestimable from realized history — both reasons to suspect, from the outset, that covariance-side machinery has limited room to add value. Repair strategies include shrinkage of the covariance (Ledoit and Wolf, 2003, 2004, 2017), random-matrix cleaning of the correlation spectrum (Laloux et al., 1999; Bun et al., 2017), factor structure (Ross, 1976; Fama and French, 1993), and priors over expected returns (Black and Litterman, 1992). The sobering benchmark of DeMiguel et al. (2009) — that $1/N$ is hard to beat out of sample — frames the entire strand. RHPort differs from this literature in its benchmark: we compare against a flat optimizer already equipped with shrinkage and a hierarchical factor covariance, not against naive diversification, so any hierarchy premium must be incremental to state-of-practice risk modeling.

Hierarchical and clustered construction. Hierarchical risk parity (López de Prado, 2016) and its successors — hierarchical clustering-based allocation (Raffinot, 2017) and nested clustered optimization (López de Prado, 2019) — argue that quasi-diagonalizing the correlation matrix by clustering, then allocating recursively, sidesteps the instability of inverting a full covariance. The correlation-network literature (Mantegna, 1999; Tumminello et al., 2010) supplies the clustering machinery. Evidence on out-of-sample gains is mixed and concentrated in *risk* metrics rather than returns. RHPort generalizes the idea from long-only risk allocation to signal-driven cash-neutral long–short construction, with explicit local optimizations rather than recursive bisection, and with the basis recombined under production constraints. Our neg-

ative result is therefore not a refutation of HRP; it is evidence about what happens when the hierarchical premise is asked to produce *alpha efficiency* rather than risk efficiency.

Active management arithmetic. The fundamental law (Grinold, 1989; Grinold and Kahn, 2000) decomposes the information ratio into skill (IC), breadth, and the transfer coefficient (Clarke et al., 2002) measuring how faithfully construction translates forecasts into holdings. It supplies our organizing claim: partitioning, recombination, and penalties act on TC ≤ 1 only. A construction experiment with the forecast held fixed is, in this language, a measurement of achievable TC across architectures.

Transaction costs and dynamic trading. Gârleanu and Pedersen (2013) show that with persistent forecasts and quadratic costs the optimal policy trades partially toward an *aim* portfolio — turnover control belongs inside the objective, not in a post-processing repair. Multi-period convex formulations operationalize this at scale (Boyd et al., 2017). Cost measurement and capacity are treated by Almgren and Chriss (2001); Almgren et al. (2005); Frazzini et al. (2018), and the survival of anomalies net of costs by Novy-Marx and Velikov (2016). RHPort's final transition layer is deliberately a repair-after-the-fact scheme, and its observed failure — smoothing removes useful trades before it removes mechanical ones — is the Gârleanu and Pedersen (2013) prediction realized in a hierarchical setting where part of the turnover is generated by basis rotation rather than by forecast change.

Backtest overfitting and research process. Multiple testing inflates apparent performance (White, 2000; Harvey et al., 2016; Harvey and Liu, 2015); deflated performance statistics and overfit probability estimates (Bailey and López de Prado, 2014; Bailey et al., 2017) quantify the damage; and protocol proposals (Arnott et al., 2019; López de Prado, 2018) prescribe embargoed validation, pre-registration, and restraint. Post-publication decay (McLean and Pontiff, 2016) warns that even true edges attenuate. This strand justifies two decisions documented below: the staged (non-Cartesian) experiment design with immutable selections, and the choice to stop without consuming the frozen holdout once continued tuning could only fit the development history.

Where this paper sits. In sum: the hypothesis belongs to the estimation-risk and hierarchical construction literature; the evaluation standard comes from the active management and transaction-cost literature; and the epistemics come from the backtest-overfitting literature. The intersection — hierarchical construction evaluated as an alpha-delivery mechanism at production scale, net of costs, under leakage-hardened protocol, with a published negative outcome — is, to our knowledge, unoccupied.

3 The RHPort construction

We describe the construction as evaluated. Throughout, t indexes trading days; $s_t \in \mathbb{R}^{N_t}$ is the cross-sectional return forecast (z-scored, unbiased with respect to cash neutrality) for the N_t names in the point-in-time universe; r_t is the vector of close-to-close returns labeled by bar start; and all statistics used to build weights at t use return rows with timestamp strictly less than t.

3.1 Partition engine

At quarterly refresh dates the engine computes a shrunk full correlation matrix from up to 756 trailing observations and divides the eligible universe into $K = \lceil N/\text{subset\_size} \rceil$ subsets whose sizes differ by at most one. Three constructions are compared with identical downstream solvers:

- **Low-correlation (default).** Symbols are greedily assigned to the subset minimizing

  $$\text{score}(j, S) = \text{mean}_{i \in S}\big(\rho_{ij}^2\big) + \lambda_{\text{repeat}}\ \text{mean}_{i \in S}\,\text{pairCount}(i, j) + \text{strataPenalty}(j, S) + \epsilon\ \text{tieBreak}(j, S),$$

  followed by one or two deterministic pair-swap improvement passes. Squared correlation is penalized because both strongly positive and strongly negative pairs damage local conditioning; a pair-repetition penalty decorrelates repeated (oversampled) partitions; and sector/industry strata are balanced to within one name per subset where feasible.

- **Correlation-clustered.** Positively correlated groups built from correlation distance, rebalanced to fixed capacities — the HRP-style premise, as a controlled ablation.

- **Stratified random.** Balanced and stratified, selected by deterministic seeded hashes — the bagging premise.

Subsets carry stable identities across refreshes (maximum-Jaccard Hungarian matching), and between refreshes membership evolves incrementally: exits are removed, entrants are assigned by the same score using point-in-time correlations, and unaffected incumbents are never reshuffled. This incremental design exists precisely to limit construction-induced turnover; Section 6 shows it is not sufficient.

3.2 Local layer

For subset S_k at date t, with $\widehat{\Sigma}_{t,k}$ an annualized Ledoit–Wolf covariance estimated from returns strictly before t (756-observation lookback, eigenvalue-floored), the local problem is

$$\begin{aligned} \max_{b_k} \quad & s_{t,S_k}^{\top} b_k - \epsilon_b \|b_k\|_2^2 \\ \text{s.t.} \quad & b_k^{\top} \widehat{\Sigma}_{t,k}\, b_k \le V_{\text{local}}^2, \qquad \mathbf{1}^{\top} b_k = 0, \qquad \|b_k\|_1 \le G_{\text{local}}, \end{aligned} \tag{1}$$

with $V_{\text{local}} = 10\%$ annualized, $G_{\text{local}} = 5$, and $\epsilon_b = 10^{-8}$. The volatility constraint is a cap that the objective normally makes binding; the gross constraint is a degeneracy guard. The global signal is used restricted to the subset — there is no within-subset re-ranking, since cash neutrality makes constant shifts irrelevant. There are no local name caps, turnover penalties, or factor-neutrality constraints: those belong to the final layer.

Because local covariance models can disagree about volatility, all raw local portfolios (*sleeves*) are normalized before recombination: each sleeve is projected through a common current-risk window, scaled exactly to V_{local}, and excluded if degenerate. Column scaling leaves the positive cone spanned by the basis unchanged but makes coefficients comparable across sleeves, which is a precondition for coefficient regularization.

3.3 Meta layer

Let $B_t \in \mathbb{R}^{N_t \times K_t}$ collect the normalized sleeves and C_t a basis-space covariance. The meta problem selects non-negative coefficients a:

$$\begin{aligned} \max_{a} \quad & s_t^\top B_t a - \lambda_1 \|B_t a - w_{t-1}\|_1 - \lambda_2 \|B_t a - w_{t-1}\|_2^2 - \mathrm{Dust}(B_t a - w_{t-1}) \\ \text{s.t.} \quad & a \geq 0, \qquad a^\top C_t\, a \leq V_{\text{final}}^2, \qquad -u_t \leq B_t a \leq u_t, \\ & \|B_t a\|_1 \leq G_{\text{final}}, \qquad \mathbf{1}^\top B_t a = 0, \end{aligned} \tag{2}$$

with $V_{\text{final}} = 8\%$ annualized, per-name cap $u_t = 3\%$, $G_{\text{final}} = 5$, and exact cash neutrality retained as a numerical invariant even though sleeves are individually neutral. All holding and trade constraints operate on $w = B_t a$, never on coefficients alone. The dust penalty suppresses economically meaningless small trades. A signed-coefficient challenger relaxes $a \geq 0$, adds $s_t^\top B_t a \geq 0$ and coefficient regularization, and measures the hedging value excluded by the positive cone without permitting a flip into an anti-signal portfolio.

3.4 Basis-space risk

Three covariance sources are blended in correlation space:

- **Current projected**: shrunk covariance of $Z_t^{\text{current}} = R_{[t-L,t)} B_t$, i.e. today's frozen basis back-projected through trailing returns. This estimates current exposures but looks through the basis's own selection sample, and is labeled accordingly.
- **Process**: shrunk covariance of genuinely prequential basis returns $z_{m,\tau+1}^{\text{process}} = B_{\tau,[:,m]}^\top r_{\tau+1}$, accumulated only for decisions already sealed — the risk of the *construction recipe*, including membership and estimator drift.
- **Structural**: a hierarchical factor model $(B^\top L) F (B^\top L)^\top + B^\top \mathrm{diag}(d)\, B$ fitted point-in-time, projected into basis space without forming a dense stock covariance.

Each source is converted to a unit-diagonal correlation matrix; a convex combination $\omega_c R^{\text{current}} + \omega_p R^{\text{process}} + \omega_s R^{\text{struct}}$ is re-scaled by one explicit variance diagonal, symmetrized, eigenvalue-floored, and verified positive semidefinite. Missing sources (e.g. insufficient process history) fall back with re-normalized weights and are never imputed from back-projections.

3.5 Transition layer

The final formulation adds a stock-space transition step: yesterday's realized portfolio is first repaired into today's constraint set (universe exits, name caps, neutrality), then moved toward today's RHPort target under L_1/L_2 trade penalties. This is a deliberate post-hoc smoothing scheme — the alternative to embedding dynamics in the objective — and its evaluation is one of the paper's main results.

4 Research protocol

Negative results are cheap to produce by accident — a leak inflates one arm, a re-used validation set deflates another — so the protocol is described at the same level of care as the method. Its design premise is that *future data are hostile*: correctness is enforced by construction and by adversarial tests, not by convention.

4.1 Sealed point-in-time inputs

Construction code never receives complete data panels. It receives sealed as-of snapshots exposing, for each logical source, only rows available strictly before the decision timestamp; every read is recorded in an availability audit, and a read of an unavailable value fails the run rather than returning data. The decision cutoff for weights at t is the previous session's close: no same-label return, price, volume, classification revision, or universe revision may enter construction at t, and PnL earned by w_t is evaluated against the return labeled $t+1$. Revisable sources carry two timestamps (effective and available). Development commands physically truncate inputs at a declared development end date, so holdout rows are absent from the process rather than merely filtered.

4.2 Sentinel tests

Leakage is detected by mandatory sentinel families rather than inspection: peekahead sentinels replace returns at or after the solve timestamp with extreme values and assert that weights and all sealed diagnostics are bit-identical; future-tail mutations are applied at full-refresh and incremental partition boundaries; cache-poisoning tests compare cold-cache against future-warmed-cache construction; checkpoint/resume must reproduce uninterrupted runs exactly; and prequential process returns are verified as $B_{t-1}^{\top} r_t$ on unique marker data. Universe canaries correlate future membership with subsequent winners and assert no clean decision changes before the canary's availability timestamp. At the time the project was closed the suite contained roughly 360 tests, and no promotion evidence could be marked as non-point-in-time.

4.3 Staged experiments with immutable selections

The experiment matrix (Section 5) was executed as a sequence of stages, each conditioned on an immutable selection from the previous stage, never as one Cartesian grid. Selection evidence binds candidate identity to executed artifacts: each stage's decision references artifact manifests, resolved configurations, and self-authenticating audit fingerprints (code SHA, configuration hash, data-vintage fingerprints), so a post-selection change of code or data invalidates the chain rather than silently contaminating it. Counterfactual mechanism questions — e.g. the turnover-penalty sweeps of Section 6 — were answered by sealed-artifact *replays* that recombine stored upstream decisions rather than re-running and re-tuning construction.

4.4 Frozen candidates and a single-consumption holdout

Before any holdout evaluation, a candidate must be frozen: code, resolved configuration, data vintage, cost assumptions, success metrics, and observed development metrics are hashed into a manifest. The holdout evaluator accepts exactly one frozen candidate, permits no optimizer or grid override, writes a consumed marker before evaluation, and refuses reruns; any subsequent change of code, report code, data vintage, or configuration requires a new candidate and a newly declared holdout. In the evaluated configuration the development window ran to the frozen data end, and the declared holdout began beyond it — a genuinely future, not historical, holdout.

4.5 Promotion gates and the stopping rule

Promotion to production required passing five gates: *leakage* (all sentinel families, availability audits, and fingerprint checks), *correctness* (tests, visible solver fallbacks, deterministic resume),

Table 1: Development-period summary of the (anonymized) return forecast.

Statistic	Value
Cross-sectional rank IC (mean)	[TODO: fill from grid artifacts]
IC t-statistic	[TODO: fill]
Forecast horizon (days, half-life)	[TODO: fill]
Coverage (mean names/day)	[TODO: fill]

mechanism (local conditioning improves without meta effective-rank collapse; results stable across partition seeds; the meta solver not merely selecting a few duplicated sleeves), *economic* (Sharpe and drawdown at least preserved versus flat benchmarks, improvement not explained by gross or concentration, turnover acceptable after realistic costs, gains persisting on the locked holdout), and *operational* (runtime, memory, and artifact reconstructibility at full production breadth).

The project was stopped at the economic gate, *without* consuming the declared holdout. By that point the development history had been used to select partitions, recombination modes, risk blends, and penalties; the remaining proposals — deeper hierarchies, more hedge directions, further transition tuning — were parameter perturbations of an already-searched space. Under any reasonable multiple-testing accounting (Bailey and López de Prado, 2014; Harvey and Liu, 2015), additional tuning against the same development results measures the research process, not the hypothesis. Stopping preserved the unconsumed holdout and the sealed replay infrastructure for any future variant that begins from a genuinely new premise.

5 Experimental design

5.1 Data and forecast

Experiments use a liquid US equity universe of roughly 3,000 names under a rolling, point-in-time membership rule, over a development period of approximately eleven years ending in mid-2026, at daily rebalancing. The return forecast is a proprietary cross-sectional model, z-scored in the cross-section and unbiased with respect to cash neutrality; it is held *fixed* across every arm, so all performance differences are attributable to construction. Its strength is summarized by rank information statistics in Table 1 rather than by its composition. Sector and industry strata use current classifications, declared as an explicit research assumption rather than a point-in-time claim; this affects only stratum balancing, not returns or universe membership, and is held identical across arms. Costs are assessed with linear cost sweeps applied to exactly reconstructed stock-level turnover.

5.2 Benchmarks

The comparison target is deliberately strong: a flat stock-space optimizer solving the analogue of problem (2) directly in stock weights — same forecast, volatility target (8% annualized), name caps (3%), gross cap, exact cash neutrality, and turnover penalties — equipped with (i) a full shrunk stock covariance and (ii) the production hierarchical factor covariance. Any hierarchy premium must therefore be incremental to competent flat construction, not to $1/N$.

5.3 Staged matrix

Stages ran sequentially, each conditioned on an immutable selection from its predecessor (Section 4); the full Cartesian product was never instantiated.

Stage A — synthetic failure modes. Known covariance and alpha models (independent assets; strong market factor; positive block factors; weak localized factors; negatively correlated pairs; covariance regime shift; signal concentrated in one cluster; signal requiring a cross-cluster hedge), plus future-return and future-membership canaries. Oracle risk, realized out-of-sample risk, basis regret, and weight stability are compared before any real data are touched.

Stage B — recombination. With partitioning fixed (subset size 50, one low-correlation partition, three-year local risk, one-year current meta risk): equal-weight bag of sleeves; partition-level meta; direct sleeve-level meta; versus the two flat benchmarks.

Stage C — partition construction and subset size. Low-correlation versus correlation-clustered versus stratified-random partitions; subset sizes $\{25, 50, 100, 200\}$, judged on local and meta conditioning plus development-period economics.

Stage D — oversampling. P repeated decorrelated partitions ($P \in \{1, 2, 3\}$ for direct meta; up to 25 for bag/partition-meta), tracking basis rank, pair repetition, and marginal value per added partition.

Stage E — risk estimator. Current-only versus fixed current/structural blends (0.75/0.25, 0.5/0.5), structural-only control, and current/process/structural blends once process history is seasoned; blend weights coarse and fixed, never adapted on the evaluation sample.

Stage F — lookbacks. Local two versus three years; meta six months versus one year; surviving configurations only.

Stage G — feedback and turnover. Zero versus one cross-subset feedback sweep; positive versus signed coefficients; dust off versus a small pre-registered grid; turnover penalties selected on net performance and trade-size distributions, not gross Sharpe.

Stage H — stability and holdout. For finalists: repeated deterministic partition seeds, small input perturbations, membership/basis/ coefficient/weight stability, paired block bootstrap on performance differences, and a single locked-holdout evaluation. *Stage H's holdout evaluation was never opened* (Section 4).

Later additions — signed alpha/hedge separation, two-stage feasibility/economic solves, forecast-vintage averaging, and the stock-space transition layer — entered as sealed-artifact replays of stage outputs rather than as new grid dimensions.

6 Results

We organize the empirical account around five recurring findings. Performance is reported relative to the flat hierarchical-factor benchmark (ratios and differences), consistent with the disclosure constraints of a proprietary setting; mechanism diagnostics are reported in full.

Table 2: Local and meta conditioning by partition scheme and subset size (development period). [TODO: fill from Stage C diagnostics]

	Local condition number		Meta effective rank	
Partition	median	p95	median	p95
Low-correlation	[TODO:]	[TODO:]	[TODO:]	[TODO:]
Correlation-clustered	[TODO:]	[TODO:]	[TODO:]	[TODO:]
Stratified random	[TODO:]	[TODO:]	[TODO:]	[TODO:]

[TODO: All tables and figures in this section are placeholders pending extraction from the sealed grid artifacts (quickgrid task prefixes); the qualitative statements are from the closed research record.]

6.1 Construction mechanics improved

The hierarchical premise held at the level it was stated. Local correlation matrices in subsets of 25–100 names were well conditioned at the 756-day lookback; the local solvers were reliable (in the reference construction, all subset solves reached optimality); and the low-correlation partition delivered materially better local conditioning than correlation-clustered or random assignment at equal subset size. Final portfolios satisfied their constraint contracts exactly — gross near target, net cash exposure at numerical zero, name weights at or under cap, forecast volatility on target.

6.2 The basis moved through time

The central mechanical failure is that the sleeve basis is itself a time-varying object. Quarterly refreshes rotate subset membership; daily incremental updates move entrants and rebalance capacities; local re-solves re-orient sleeves as covariances update. As a result, yesterday's realized stock portfolio — still reasonable *in stock space*, with stable exposures — was frequently outside today's positive sleeve cone $\{B_t a : a \geq 0\}$. Reaching today's feasible set then forces trades that carry no forecast content. Coefficient warm starts, stable sleeve identities, and incremental membership all reduced but could not remove this mechanical turnover, because the cone rotation is intrinsic to re-solving local problems on new information.

6.3 Richer bases were mostly redundant

Oversampled partitions (Stage D) and additional exported directions — factor-response sleeves and internal-hedge directions — were motivated as enlarging the reachable set. In practice their marginal contribution saturated quickly: added directions were largely inside the span of the existing basis, so they contributed little usable rank while degrading the conditioning of the basis-space covariance and, with it, solver reliability. This is the random-matrix expectation (Laloux et al., 1999; Bun et al., 2017) — beyond the strong common directions, sample-estimated directions are mostly noise, and adding more of them adds noise rank — realized at the sleeve level rather than the stock level.

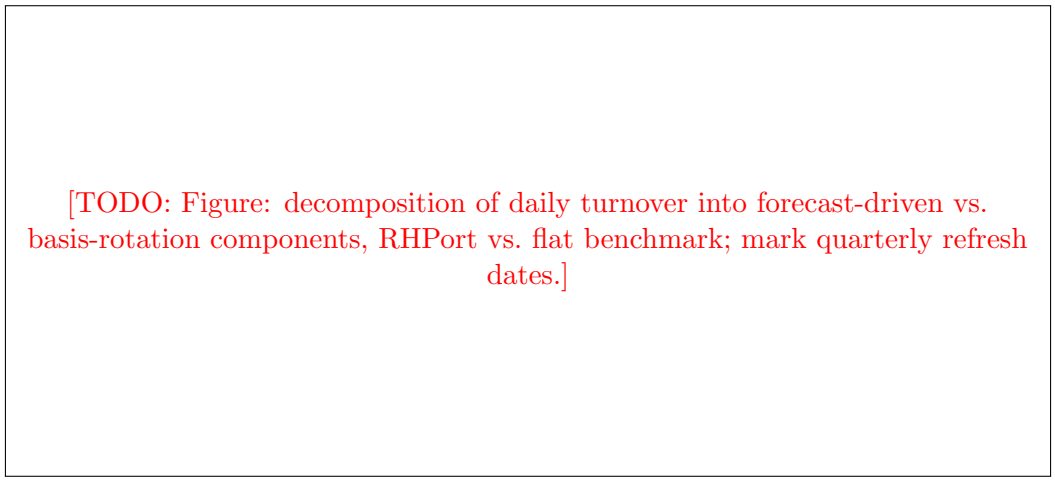

Figure 1: Turnover decomposition. Basis rotation generates turnover with no counterpart in the flat construction.

6.4 Cleaner contracts did not mean better portfolios

Separating positive hierarchical alpha allocation from signed internal-node hedging, and splitting the solve into a feasibility stage followed by an economic stage, materially improved the *model contract*: infeasibility became attributable, solver failures became diagnosable, and fallbacks became fail-closed. None of it changed the economic solution appreciably. We report this as a caution about a seductive failure mode of construction research: architectural refinements that improve the engineering readout are easily mistaken for progress on the investment problem.

6.5 Turnover was a symptom; smoothing had a narrow useful band

The stock-space transition layer (Section 3.5) was evaluated by sealed replays over a pre-registered penalty sweep. At moderate penalties it reduced realized turnover without materially changing gross performance — evidence that much of the excess turnover was indeed mechanical rather than informational. But the useful band was narrow: as penalties increased, the layer began suppressing the trades that carried the forecast before it had removed all mechanical rotation, and net performance deteriorated. A repair layer cannot distinguish rotation from information, exactly as Gârleanu and Pedersen (2013) imply for post-hoc smoothing relative to trading-toward-an-aim inside the objective.

6.6 The return profile was weak and unstable

Against the flat benchmarks, no surviving configuration — across recombination modes, partitions, subset sizes, oversampling, risk blends, lookbacks, coefficient schemes, and transition penalties — delivered a convincing net improvement. An attractive early sub-period did not persist, and forecast-horizon rescaling and causal forecast-vintage averaging did not repair it. Given the number of arms examined, the early-period result is exactly what selection over a development history produces under the null (Bailey and López de Prado, 2014).

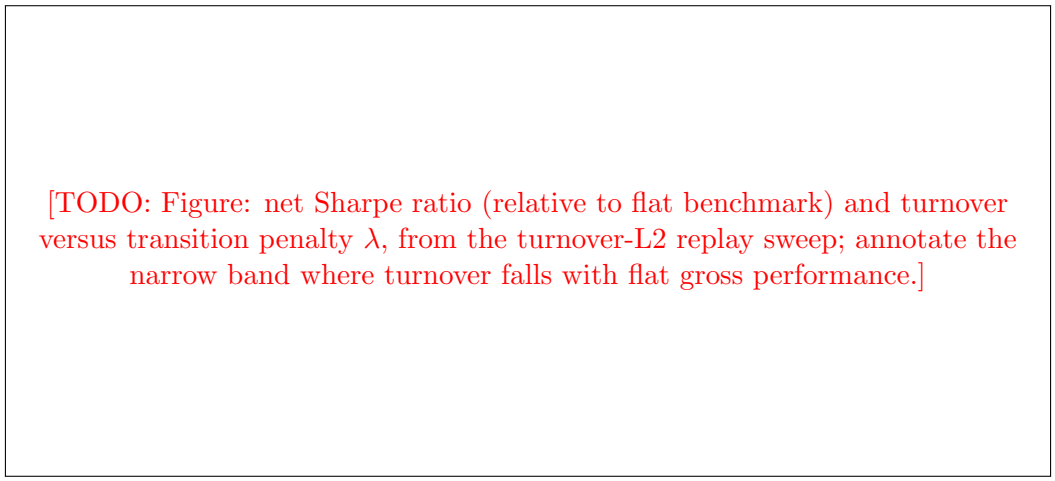

Figure 2: Transition-layer penalty sweep: turnover is reduced cheaply only in a narrow band; beyond it, smoothing consumes the signal.

Table 3: Development-period economics relative to the flat hierarchical-factor benchmark (differences; positive favors RHPort). [TODO: fill from Stage B/G artifacts and replay sweeps]

Configuration	ΔSharpe (net)	Turnover ratio	ΔMaxDD
Equal bag	[TODO:]	[TODO:]	[TODO:]
Partition meta	[TODO:]	[TODO:]	[TODO:]
Direct sleeve meta	[TODO:]	[TODO:]	[TODO:]
Direct + signed challenger	[TODO:]	[TODO:]	[TODO:]
Direct + transition layer (best λ)	[TODO:]	[TODO:]	[TODO:]

6.7 Summary

The hierarchy did what its premise promised locally — better-conditioned small covariances, reliable small solves — and still failed globally, because the costs of hierarchical delivery (basis rotation, redundant rank, repair-layer losses) consumed gains the fixed forecast was not strong enough to fund. The construction redistributed the existing signal; it could not amplify it.

7 Discussion

7.1 The fundamental-law reading

Write the expected information ratio as $\mathrm{IR} \approx \mathrm{IC} \cdot \sqrt{\mathrm{BR}} \cdot \mathrm{TC}$ (Grinold, 1989; Clarke et al., 2002). Holding the forecast fixed fixes IC and BR; every RHPort design choice acts on $\mathrm{TC} \leq 1$, and net of costs, on TC *minus* the cost of delivery. The experiment is therefore a direct measurement of how much transfer-coefficient headroom a competent flat optimizer leaves on the table at this scale: our answer is *not enough to fund a hierarchy.* The flat benchmark with a hierarchical factor covariance already achieves high effective transfer at low mechanical turnover; the hierarchical delivery mechanism spent its budget rotating its own basis.

This reading sharpens where the hypothesis went wrong. The estimation-risk premise — small covariances are better estimated — was *true* and is visible in the conditioning diagnostics.

The error was expecting an estimation-quality improvement on the risk side to convert into net return. Risk-side improvements raise TC only insofar as the flat construction was being damaged by covariance error, and modern shrinkage plus factor structure had already removed most of that damage.

7.2 Relation to the HRP evidence

Our result is consistent with, not contrary to, the hierarchical risk parity literature read carefully: reported HRP gains concentrate in out-of-sample *risk* behavior — volatility, drawdown, weight stability — in long-only allocation without a return forecast (López de Prado, 2016; Raffinot, 2017). We asked the hierarchical premise a harder question: to deliver a fixed alpha through production constraints, net of turnover, against a strong flat baseline. The premise survives as risk engineering; it fails as alpha engineering. Studies that benchmark hierarchical methods against $1/N$ or against unshrunk mean–variance systematically overstate what practitioners should expect.

7.3 Turnover as a design boundary, not a tuning parameter

The transition-layer result generalizes: when a construction's continuity is extrinsic — restored by penalizing trades after targets are formed — smoothing faces an information/rotation identification problem it cannot win. Continuity must be *intrinsic*: either the parameterization itself is persistent (stable bases by construction, not by matching), or dynamics enter the objective as in Gârleanu and Pedersen (2013) and Boyd et al. (2017), where the optimizer trades toward an aim that already reflects the cost of future rebalancing. A hierarchical construction whose local problems re-solve freely each day cannot be repaired into low turnover from outside.

7.4 Stopping without the holdout

We stopped at the economic gate with the holdout unconsumed. The argument is worth making explicit, because the alternative — "one more sweep, then the holdout" — is the modal failure of proprietary research. After Stages B–G plus replay sweeps, the development history had adjudicated every proposed refinement; remaining ideas were perturbations within an already-searched neighborhood. Consuming a single-use holdout to adjudicate among arms pre-selected on development data yields a biased estimate whose bias grows with the search (Bailey and López de Prado, 2014; Harvey and Liu, 2015; Arnott et al., 2019); preserving it costs nothing and keeps a genuinely out-of-sample test available for a future variant that begins from a new premise. The sealed-replay infrastructure makes that future test cheap.

7.5 Boundary conditions and reopening criteria

The negative result is a statement about a regime, not a theorem. We pre-registered, at closure, the conditions under which the architecture is worth revisiting:

1. **A materially stronger, independently validated forecast.** The fundamental-law ceiling scales with IC; a forecast strong enough to fund delivery costs changes the arithmetic that killed this instance.

2. **Intrinsically continuous hierarchical construction.** A formulation whose basis persistence is a property of the parameterization — not a post-processing repair — removes the dominant mechanical cost.

3. **A constraint the flat optimizer cannot express.** Hierarchy earns complexity only when it enlarges the feasible set in a way the stock-space problem cannot (e.g. genuinely local constraint structure), not when it merely re-parameterizes it.

Absent one of these, further tuning of hierarchy depth, hedge directions, transition penalties, or numerical tolerances against the existing development results would fit the development history rather than test a hypothesis.

7.6 Threats to validity

Three limitations qualify the result. First, sector/industry strata used current classifications (declared, and identical across arms); this touches only stratum balancing, but a fully point-in-time classification test was not run. Second, costs were assessed by linear sweeps on exact turnover rather than an impact model; since RHPort's handicap is *higher* turnover, concave impact would penalize it further, making the linear treatment conservative in RHPort's favor. Third, the forecast is one proprietary signal family at one horizon; a materially different signal could interact differently with hierarchical delivery — which is precisely reopening criterion 1.

8 Conclusion

We built, hardened, and systematically evaluated a hierarchical long–short portfolio construction at production scale, holding the return forecast fixed and benchmarking against a competent flat optimizer net of turnover. The hierarchical premise was locally correct — small covariances are better conditioned and small solves are more reliable — and globally insufficient: basis rotation manufactured turnover with no informational counterpart, richer bases added noise rank, post-hoc smoothing could not separate rotation from information, and the fixed forecast was not strong enough to fund the difference. Portfolio construction, in the fundamental-law arithmetic, is a transfer-coefficient technology bounded above by one; it can stop a forecast from being wasted, but it cannot be a substitute for the forecast.

The findings we believe travel beyond this instance are: benchmark hierarchical constructions against strong flat optimizers, not naive diversification; treat turnover as a design boundary requiring intrinsic continuity or in-objective dynamics, not a post-hoc penalty; distinguish improvements to the model contract from improvements to the portfolio; and make negative results credible by sealing inputs, staging experiments, freezing candidates, and being willing to stop with the holdout unconsumed.

The engineering survives the hypothesis: sealed as-of data access, sentinel families, sparse parameter-free solver graphs, resumable checkpoints, and sealed-artifact replays now serve as general research infrastructure. The holdout remains locked, awaiting a hypothesis that earns it.

References

Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3(2):5–39, 2001.

Robert Almgren, Chee Thum, Emmanuel Hauptmann, and Hong Li. Direct estimation of equity market impact. *Risk*, 18(7):58–62, 2005.

Rob Arnott, Campbell R. Harvey, and Harry Markowitz. A backtesting protocol in the era of machine learning. *Journal of Financial Data Science*, 1(1):64–74, 2019.

David H. Bailey and Marcos López de Prado. The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5):94–107, 2014.

David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. The probability of backtest overfitting. *Journal of Computational Finance*, 20(4):39–69, 2017.

Michael J. Best and Robert R. Grauer. On the sensitivity of mean-variance-efficient portfolios to changes in asset means: Some analytical and computational results. *Review of Financial Studies*, 4(2):315–342, 1991.

Fischer Black and Robert Litterman. Global portfolio optimization. *Financial Analysts Journal*, 48(5):28–43, 1992.

Stephen Boyd, Enzo Busseti, Steven Diamond, Ronald N. Kahn, Kwangmoo Koh, Peter Nystrup, and Jan Speth. Multi-period trading via convex optimization. *Foundations and Trends in Optimization*, 3(1):1–76, 2017.

Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

Vijay K. Chopra and William T. Ziemba. The effect of errors in means, variances, and covariances on optimal portfolio choice. *Journal of Portfolio Management*, 19(2):6–11, 1993.

Roger Clarke, Harindra de Silva, and Steven Thorley. Portfolio constraints and the fundamental law of active management. *Financial Analysts Journal*, 58(5):48–66, 2002.

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy? *Review of Financial Studies*, 22(5):1915–1953, 2009.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

Andrea Frazzini, Ronen Israel, and Tobias J. Moskowitz. Trading costs. *SSRN Working Paper 3229719*, 2018.

Nicolae Gârleanu and Lasse Heje Pedersen. Dynamic trading with predictable returns and transaction costs. *Journal of Finance*, 68(6):2309–2340, 2013.

Richard C. Grinold. The fundamental law of active management. *Journal of Portfolio Management*, 15(3):30–37, 1989.

Richard C. Grinold and Ronald N. Kahn. *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk*. McGraw-Hill, 2nd edition, 2000.

Campbell R. Harvey and Yan Liu. Backtesting. *Journal of Portfolio Management*, 42(1):13–28, 2015.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016.

J. D. Jobson and Bob Korkie. Estimation for Markowitz efficient portfolios. *Journal of the American Statistical Association*, 75(371):544–554, 1980.

Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters. Noise dressing of financial correlation matrices. *Physical Review Letters*, 83(7):1467–1470, 1999.

Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, 2003.

Olivier Ledoit and Michael Wolf. Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, 30(4):110–119, 2004.

Olivier Ledoit and Michael Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *Review of Financial Studies*, 30(12):4349–4388, 2017.

Marcos López de Prado. Building diversified portfolios that outperform out of sample. *Journal of Portfolio Management*, 42(4):59–69, 2016.

Marcos López de Prado. *Advances in Financial Machine Learning*. Wiley, 2018.

Marcos López de Prado. A robust estimator of the efficient frontier. *SSRN Working Paper 3469961*, 2019.

Rosario N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B*, 11(1):193–197, 1999.

R. David McLean and Jeffrey Pontiff. Does academic research destroy stock return predictability? *Journal of Finance*, 71(1):5–32, 2016.

Robert C. Merton. On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, 8(4):323–361, 1980.

Richard O. Michaud. The Markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1):31–42, 1989.

Robert Novy-Marx and Mihail Velikov. A taxonomy of anomalies and their trading costs. *Review of Financial Studies*, 29(1):104–147, 2016.

Thomas Raffinot. Hierarchical clustering-based asset allocation. *Journal of Portfolio Management*, 44(2):89–99, 2017.

Stephen A. Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360, 1976.

Michele Tumminello, Fabrizio Lillo, and Rosario N. Mantegna. Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior & Organization*, 75(1):40–58, 2010.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.