WHEN DO GRAPH NEURAL NETWORKS HELP IN CROSS-SECTIONAL STOCK RANKING? A MULTI-SEED, MULTI-UNIVERSE, COST-AWARE STUDY OF THE U.S. S&P 500

Anonymous authors
Paper under double-blind review

ABSTRACT

Papers often report that graph neural networks (GNNs) improve cross-sectional stock ranking by encoding relations among stocks as graph edges. Many reported gains, however, come from single-seed, single-split evaluations that do not adjust for multiple testing. We ask a narrower question: when, if at all, does the graph help? We evaluate an equal-budget tuned ladder from LightGBM to a heterogeneous relational GNN (L0–L7) over twelve expanding quarterly walk-forward folds and ten seeds on the U.S. S&P 500, with two pre-registered confirmatory families. The predictive family asks whether any tuned model beats tuned LightGBM. Hansen's SPA test does not reject in either feature universe ($p = 0.277$ and $p = 0.077$), and the benchmark comparisons are underpowered, so the result is a non-rejection and does not establish equality. Local DM contrasts place the observable signal in the non-graph MLP, and the best-supported contrast runs *against* the correlation graph. Adding a correlation-GAT to the MLP reduces IC ($\Delta\text{IC} = -0.012$); this penalty is BH-significant in *both* feature universes, and its sign holds when we drop any single quarter. The penalty is specific to that graph at its tuned operating point: in Universe C the sector, full-attention, HATS, and combined-edge arms all recover above the correlation-GAT, though Universe-C positives rest on a leak-selected basis. The MLP's gain over LightGBM ($\Delta\text{IC} = +0.015$) is significant only in the leak-selected universe and falls below its detectable-effect threshold, so we read it as suggestive. The edge-attribution family freezes the tuned correlation-GAT operating point and varies only the edge set. None of six edge contrasts survives Benjamini–Hochberg control, and all six are underpowered. A transaction-cost crosswalk (each IC contrast re-expressed as net Sharpe at 10 basis points) points the same way. The two Universe-C ladder contrasts we examine most closely (MLP−LightGBM and correlation-GAT−MLP) survive 10 bps; the tuned news-edge IC underperformance is near zero in net Sharpe, fragile across folds, and absent in the fixed-operating-point test. We therefore report conditional findings and failure modes.

1 INTRODUCTION

Cross-sectional stock ranking evaluates whether a model can predict the relative ordering of next-period returns across a fixed universe of stocks on each trading day; in our setting, this universe consists of the constituents of the U.S. S&P 500 index. The economic object is a dollar-neutral long-short portfolio: correct top- and bottom-decile rankings create a tradable signal, while transaction costs absorb near-chance rankings. We measure ranking quality with the *Information Coefficient* (IC), the daily cross-sectional Spearman rank correlation between predicted scores and realized forward returns. Recent GNN papers on stock ranking, including Feng et al. (2019), Kim et al. (2019), and Sawhney et al. (2021), report sizable IC gains from message passing over inter-stock relations such as sector membership, Wikidata links, and news co-mentions. Much of this evidence comes from one seed and one split, a design that leaves three concerns.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The first concern is seed dispersion. Even after tuning and averaging across seeds, per-arm 95% block-bootstrap confidence intervals (CIs) over test days are often comparable in width to the IC point estimates. For example, the Universe-C correlation-GAT yields a pooled IC of 0.0224 with a 95% interval of $[-0.0080, +0.0559]$, and retuning does not eliminate this variability. This is primarily a caveat about deployment precision; the confirmatory estimand remains the seed mean (§3.3). The second concern is regime concentration: a single high-dispersion quarter can disproportionately influence the annual estimate. The third is capacity confounding. A tuned edge ablation that compares networks of different sizes conflates the effect of the edge with the effect of increased model capacity; therefore, a causal interpretation requires holding the operating point fixed.

We address these concerns with one tuned ladder and two pre-registered confirmatory families. Each rung (arm) is one independently tuned model; L2, the correlation-GAT, is a graph attention network (GAT) over a trailing-window return-correlation graph, and L1 a multilayer perceptron (MLP) on the same features. The predictive family asks whether any tuned arm improves on tuned LightGBM, using Hansen's Superior Predictive Ability (SPA) test (Hansen, 2005) over nine candidates, with Diebold–Mariano / Harvey–Leybourne–Newbold (DM/HLN) pairwise tests (Diebold & Mariano, 1995; Harvey et al., 1997) under Benjamini–Hochberg (BH) false-discovery-rate control (Benjamini & Hochberg, 1995) as local evidence. The edge-attribution family freezes every tuned hyperparameter of the correlation-GAT and changes only the edge set, so the matched ΔIC estimates the edge effect at L2's operating point, not each edge set's retuned optimum (§3.3).

Two feature universes. Both universes contain the same fixed survivor snapshot of stocks (§4) and differ only in node features. **Universe B** is a 10-dimensional price-volume baseline. **Universe C** has 51 columns, 48 from Qlib's Alpha158 library (Yang et al., 2020) and three hand-crafted price-volume columns, selected from the project's earlier top-15 factor groups (the exploratory Plan-AAA ranking, Limitation L1).

Headline findings. Neither universe confirms a tuned arm as outperforming tuned LightGBM after selection adjustment. Hansen's SPA test does not reject in either universe (Universe B $p_{\text{consistent}} = 0.277$, Universe C $p_{\text{consistent}} = 0.077$), and the benchmark comparisons are underpowered because the observed gaps fall below the 80% minimum detectable effect.

The local ladder shows where any neural lift comes from, and its best-supported signal is negative for the correlation graph. Adding a correlation-GAT to the MLP's features *reduces* IC (L2$-$L1), BH-significantly in *both* universes, and the sign holds in every leave-one-quarter-out and leave-one-seed-out replicate. The positive direction is weaker: the tuned MLP beats LightGBM only in the leak-selected Universe C and below its detectable-effect threshold, so we read it as suggestive (Limitation L1). The edge-attribution family finds **0/6** BH-FDR rejections, with **6/6** contrasts underpowered (§5).

2 Related Work

Prior GNN rankers report single-split gains. Feng et al. (2019) learn to rank over industry and Wikidata relation graphs, and the HATS model of Kim et al. (2019) (75 Wikidata relation types) motivates our L7 rung, a HATS-*style* adaptation with three relation types, so L7$-$L2 is subject to relation starvation. Cheng & Li (2021) and Li et al. (2024) relax the graph toward dense learned attention, which our L6 rung proxies. Further variants span spatiotemporal hypergraphs (Sawhney et al., 2021), pattern-specific routing (Lin et al., 2021), MLP-only token mixing (Fan & Shen, 2024), and multi-relational dynamic graphs (Qian et al., 2024). None uses a multi-seed, multiplicity-controlled, cost-aware protocol (per-paper audit in Appendix B).

Empirical finance treats multiplicity first. We follow the empirical-asset-pricing line that treats multiple testing, dependence, and sample splitting as first-order concerns. Gu et al. (2020) set the recursive out-of-sample benchmark for machine-learning return prediction and find no gain beyond shallow networks. Hou et al. (2020) replicate 452 cross-sectional anomalies under one protocol: 65% fail the single-test hurdle ($|t| \geq 1.96$) and 82% fail their multiple-testing hurdle. Avramov et al. (2023) find that machine-learning long-short profits concentrate in hard-to-arbitrage stocks and shrink under cost screens, which motivates our cost layer. The purge-and-embargo validation and block-bootstrap methods of López de Prado (2018) shape our walk-forward design; Hansen

(2005) motivates the SPA test, Politis & Romano (1994) the stationary bootstrap, and Newey & West (1987) the heteroskedasticity-and-autocorrelation-consistent (HAC) standard errors.

The delta is the protocol. To our knowledge, this is the first S&P 500 stock-ranking study to combine ten seeds, 12 expanding walk-forward folds, a tuned ladder, two pre-registered confirmatory families, and disclosure of a tuned-configuration stability failure.

3 METHODOLOGY

3.1 WALK-FORWARD CROSS-VALIDATION

Our validation design is a 12-fold *expanding-window walk-forward*: in fold k, training ends at a quarter boundary, validation is the next quarter, and testing is the following quarter. The twelve test quarters run from 2023 Q1 through 2025 Q4, yielding $T = 749$ pooled test days. Because the label is a 21-trading-day forward return, we purge the final 21 days of both the training and the validation window to eliminate label overlap (López de Prado, 2018).

3.2 INFORMATION COEFFICIENT AND PORTFOLIO

We compute daily IC (§1) against 21-trading-day forward returns over stocks with non-missing features that day; the headline IC is its pooled mean across all test days in the 12 folds. The portfolio exercise buys the top and shorts the bottom decile of predicted scores, equal-weighted and dollar-neutral, rebalanced every 21 trading days. The long-short Sharpe ratio uses non-overlapping 21-day rebalance returns, annualized by $\sqrt{252/21}$. We compute it per fold-seed cell (about three rebalance returns each) and average the cells; paired contrasts bootstrap the 12 fold-level differences. The block = 21 d bootstrap handles daily-IC dependence separately (§3.3).

Net Sharpe subtracts one-way costs at each rebalance, $N = G - \tau_{\mathrm{L1}} c/10^4$ (net and gross rebalance returns, one-way cost c in basis points, bps), with $\tau_{\mathrm{L1}} = \sum_i |\Delta w_i|$ the realized per-rebalance $L1$ turnover (at most 4 under full rotation), measured per arm and fold. We report $c \in \{0, 5, 10, 15, 20, 30\}$ bps (headline 10 bps). Net Sharpe is descriptive and comes from the same tuned 12-fold confirmatory cells.

3.3 TWO CONFIRMATORY FAMILIES

The two confirmatory families stay statistically separate.

Family-1 (predictive). The benchmark is tuned LightGBM (L0), the candidates the other nine tuned rungs. We apply Hansen's SPA test with the consistent variant, loss $= -$daily IC, and a stationary bootstrap with block length 21 trading days. We also report local paired evidence from DM/HLN tests on the 20 pre-registered contrasts of Table 3 (five ladder and five edge-set contrasts, each in both universes). We correct by BH-FDR at $q = 0.05$ over the pooled 20-test family (per-universe correction gives the same rejections). The DM statistic combines Newey–West HAC standard errors (Newey & West, 1987) (pre-registered data-driven bandwidth; lag-21 check in Appendix C.8) with the HLN finite-sample correction for the 21-day horizon (formulas and one caveat in Appendix A.6).

Family-2 (edge attribution). Because a tuned-vs-tuned edge comparison confounds graph structure with effective capacity, the fixed-operating-point arms run the L3, L4, and L5 edge sets at L2's frozen tuned hyperparameter vector, so the matched ΔIC is an edge effect at that operating point.

One caveat: the intervention fixes the operating point and only indirectly the capacity. A denser edge set may be underfit under L2's hyperparameters, so part of the 0/6 non-significance could reflect underfitting instead of a null edge effect. Inference uses paired, fold-level, seed-averaged ΔIC over $n \approx 12$ fold blocks, with 5000 stationary block-bootstrap replications and BH-FDR control over six two-sided fold-block tests; the per-row confidence intervals are unadjusted and descriptive.

Power and inference unit. For each contrast we report a minimum detectable effect, MDE $= 2.8 \times \mathrm{SE}_{\mathrm{block}}$ (block-level standard error) with $n_{\mathrm{eff}} = T_{\mathrm{days}}/21 \approx 36$ non-overlapping blocks ($2.8 \approx z_{0.975} + z_{0.80}$: two-sided 5% test at 80% power; construction and caveats in Appendix C.5). The

MDE is a conservative block-level yardstick and does not measure the power of the daily DM test, so a contrast can reject and still sit below its MDE. Both families average the ten seeds' IC into one series before testing (per-day for Family-1, per-fold for Family-2), matching the pre-registered seed-mean estimand.

Pre-registration. We froze the protocol and tuned hyperparameters before the confirmatory run (hyperparameter file md5 `59ddd0a2`), and only the two families above enter confirmatory inference. We disclose the earlier analyses (horizon ablations, the loss-function horse race, the Plan-AAA factor ranking) as exploratory. IC is the only confirmatory metric, with no BH-FDR family for Sharpe. We fixed the plan in the project repository before the confirmatory run and did not lodge it with an external registry.

4 Data and Setup

The stock universe is a fixed survivor snapshot of the S&P 500: the 501 of the 503 share-class tickers listed at the snapshot date with continuous price, sector, and news histories from 2021 through 2026. Membership is not point-in-time (PIT): departed firms are absent (no delisting returns), and we exclude missing forward-return cells as NaN, so the estimand is this fixed survivor universe (Limitation L8).

The data frequency is daily close-to-close. The label, fixed across all experiments, is the cross-sectionally z-scored, market-excess transform of the 21-trading-day forward return $r_{T\to T+21} = p_{T+21}/p_T - 1$. Predictions for day T use features measured strictly at $T-1$, leaving a one-trading-day execution lag between the last feature timestamp and the beginning of the return window.

All experiments use the same ten canonical seeds. Each ladder rung changes one nominal design axis relative to the rung below it (Table 1), and we tune each rung independently with 30 Optuna trials. Equal budget equalizes trial count, not search adequacy: neural arms have higher-dimensional spaces than LightGBM, a caveat for the "MLP beats GAT" interpretation. Universe B is the primary universe for the L2−L1 contrast because its basis was fixed a priori: ten standard price-volume features from the Alpha158 and GKX (Gu et al., 2020) families, all lagged one day and never ranked on this sample, so it carries no selection-stage leak of the kind that affects Universe C (Limitation L1; provenance and one caveat in Appendix C.1). We fixed the designation in the project decision log after the Universe-C leakage audit and before running the sweep; Appendix C.1 gives the dated record. A trials-sensitivity sweep that re-tunes L2 at $3\times$ the budget (90 trials, matching the MLP's search density) leaves the leak-free Universe-B penalty significant but drops Universe-C's below significance ($p = 0.059$, sign unchanged). The Universe-B penalty thus narrows but survives a tripled budget (Appendix C.1). The confirmatory claim does not depend on the designation: at the pre-registered budget the penalty is BH-significant in both universes, and the sweep is an unregistered sensitivity.

We construct four graph specifications. $\alpha 1$ is a 126-day Pearson correlation graph on daily returns, with an edge when $|\rho| > 0.6$. We compute snapshots on a 21-day grid, and each fold uses the last snapshot inside its purged training window, held fixed through training, validation, and test. $\alpha 2$ is static sector connectivity under the Global Industry Classification Standard (GICS). The news graph is point-in-time and fits no construction parameters. Two stocks share an edge on day T if an EODHD news article co-mentions them and its publication timestamp (not event date) falls in the seven calendar days before the NYSE close on $T-1$ (construction in Appendix A.5). $\alpha 4$ stacks the correlation, sector, and news edges.

The model classes are GAT (Veličković et al., 2018), GraphSAGE-Mean (Hamilton et al., 2017), MLP, and LightGBM; L7 follows an R-GCN-style relation-specific parameterization (Schlichtkrull et al., 2018), and we compute SPA with `arch` (Sheppard, 2024). All neural arms minimize mean-squared error (MSE) on the label, LightGBM uses its squared-error objective, and tuning selects on validation IC.

Pipeline positive control. A pre-confirmatory planted-signal control rules out a grossly nonfunctional graph path. With synthetic node features and a label planted on the frozen correlation topology (achievable IC 0.047), GAT recovers 82% and GraphSAGE-Mean 91% of the achievable IC while a feature-matched MLP recovers ≈ 0 (Appendix C.4).

Table 1: The tuned model ladder. Each rung changes one nominal design axis; every arm is independently retuned, so rung contrasts compare tuned operating points (§3.3).

Rung	Model / graph	Targets
L0	LightGBM (tuned), no graph	non-neural benchmark
L1	MLP, no graph	value of going neural
L2	GAT + α1 correlation	value of adding a graph
L2s	GraphSAGE-Mean + α1	aggregation control vs L2
L3	GAT + α1$\cup$news	marginal value of news edges
L4	GAT + α1$\cup\alpha$2 (sector)	marginal value of sector edges
L5	GAT + α4 (all edges)	edge-stacking complementarity
L5s	GraphSAGE-Mean + α4	aggregation control, full edges
L6	full self-attention, no mask	attention vs structure (dense)
L7	HATS-3R-adapt	relation-attention representative

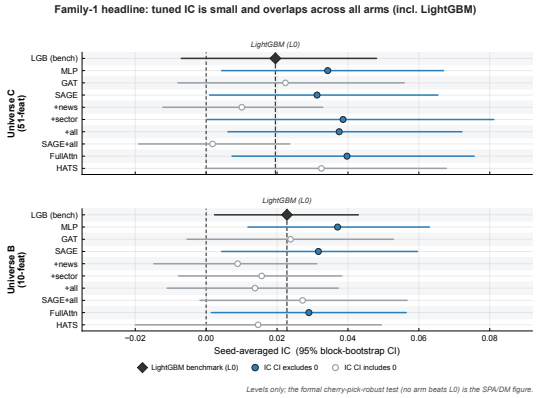

Figure 1: Per-arm pooled IC for the L0–L7 ladder. Intervals are 95% stationary block-bootstrap CIs (block = 21 d, 5000 reps); seeds = 10, folds = 12, benchmark = L0.

5 Results

5.1 Per-arm IC excludes zero for the MLP, not for the correlation-GAT

Figure 1 and Table 2 report seed-averaged pooled IC by arm, with 95% block-bootstrap confidence intervals over $T = 749$ test days. Intervals exclude zero most consistently for the non-graph MLP (L1, both universes) and the full-attention arm (L6), but for the pure correlation-GAT rungs (L2) in neither universe. C/L5s's pooled IC ≈ 0.002 is over defined cells only (27.5% collapse to undefined IC, §5.5); we exclude undefined days pointwise, and the arm stays in the $M = 9$ SPA candidate set.

5.2 No confirmed win over LightGBM after selection adjustment

Hansen's SPA test does not reject the null that no candidate beats LightGBM in either universe ($p_{\text{consistent}} = 0.2767$ for B and 0.0774 for C; $M = 9$, $T = 749$). In Universe C the SPA bracket is $p_{\text{lower}} = 0.055$, $p_{\text{consistent}} = p_{\text{upper}} = 0.077$, so even the lower bound stays above 5%, and the comparison is underpowered (observed gaps below the 80% MDE).

The local DM/HLN ladder (Figure 2, Table 3) gives a narrower view. We report each layer at its own error rate (SPA for the global benchmark question, the DM ladder as pre-registered local evidence at its own FDR level) and claim no joint family-wise rate. Two post-hoc sensitivities bound the layering: pooled BH over all 26 p-values rejects the same contrasts, and Benjamini–Yekutieli (Benjamini & Yekutieli, 2001) under arbitrary dependence keeps L2$-$L1 rejected in both universes while dropping C L1$-$L0, L3$-$L2 (both universes), and C L4$-$L2 (Appendix C.3).

Among the pre-registered contrasts, the strongest local evidence points to the non-graph MLP. In Universe C three graph or attention arms (L4, L5, L6) post higher pooled IC than the MLP (Table 2); the contrast set tests them only against L2, and those levels rest on the leak-selected ba-

Table 2: Per-arm pooled IC with 95% block-bootstrap CI; "e0": CI excludes zero. The C/L5s Univ-C value (0.0018) is conditional on its *defined* cells (27.5% collapse to undefined IC, §5.5).

Rung	Univ B IC	e0	Univ C IC	e0
L0 LightGBM	0.0228	y	0.0195	n
L1 MLP	0.0371	y	0.0343	y
L2 GAT(corr)	0.0238	n	0.0224	n
L2s SAGE	0.0317	y	0.0313	y
L3 +news	0.0089	n	0.0101	n
L4 +sector	0.0157	n	0.0386	y
L5 +both	0.0138	n	0.0375	y
L5s SAGE all	0.0272	n	0.0018	n
L6 full-attn	0.0290	y	0.0397	y
L7 HATS	0.0147	n	0.0325	n

Table 3: DM/HLN pairwise ladder: seed-averaged daily ΔIC and HLN p (*: BH-FDR reject at $q = 0.05$ over the 20-test family). Near-equal ΔIC can differ in p as the daily-loss variance differs by universe. The p-values use the Newey–West automatic lag; Tables 14 and 16 also report the lag-21 check, where C L1−L0 has $p = 0.063$. BH at lag 21 keeps 5 of the 11 rejections, including L2−L1 in both universes (Table 18).

Pair (A−B)	B ΔIC	B p	C ΔIC	C p
L1−L0	+0.0143	0.052	+0.0148	0.011^*
L2−L1	−0.0133	$< .001^*$	−0.0119	$< .001^*$
L6−L2	+0.0052	0.356	+0.0174	0.001^*
L7−L2	−0.0091	0.195	+0.0102	0.002^*
L2s−L2	+0.0079	0.118	+0.0089	$< .001^*$
L3−L2	−0.0149	0.006^*	−0.0123	0.009^*
L4−L2	−0.0081	0.227	+0.0163	0.012^*
L5−L2	−0.0100	0.110	+0.0152	$< .001^*$
L5−L4	−0.0019	0.172	−0.0011	0.831
L5−L3	+0.0049	0.066	+0.0275	$< .001^*$

sis. In Universe C the MLP beats LightGBM (L1−L0 = +0.0148), the correlation-GAT loses to the MLP (L2−L1 = −0.0119), and the tuned news-edge arm underperforms the correlation-GAT (L3−L2 = −0.0123; all BH-reject). The positive C L1−L0 contrast is only suggestive because the feature basis was leak-selected (Limitation L1; the magnitude-matched Universe-B contrast, +0.0143, does not survive BH). The leak-free universe offers a magnitude comparison: the two contrasts differ by 0.0005, an order of magnitude below the cross-seed range of the Universe-C contrast (Appendix C.2), while the p-values differ (0.052 versus 0.011). We do not identify what the leak contributed to either. The basis leak enters both arms through the same 51 columns, so it inflates levels and reaches the contrast only through differential exploitation, which the Universe-B anchor bounds at a small fraction if it is an adequate proxy. Two sensitivities bear on this reading: a test-informed 20-column subset gives +0.0134, and a basis re-selected on information through June 2022 gives −0.0024 with an interval containing zero. All three paired intervals contain zero, so neither run establishes a difference (Appendices C.6 and C.7). The negative correlation-GAT contrasts are less exposed: selection leakage can inflate apparent predictability but cannot explain why the correlation-GAT underperforms the MLP, and the L2−L1 penalty is BH-significant in Universe B as well (ΔIC = −0.0133).

Power is asymmetric. C L2−L1 exceeds its 80% MDE whereas C L1−L0 falls below its own, so the data identify a detectable correlation-GAT *penalty* while leaving any benchmark-beating *benefit* too small to detect. Under-search of the graph arms at the fixed 30-trial budget (§4) is an alternative reading. Two observations bound it: the same-budget, same-space arms L4 and L5 land *above* L2 in Universe C (+0.0163, +0.0152, both BH-reject), and the planted-signal control shows that the stack converts graph-borne signal when present.

Skill is episodic; the central contrasts are sign-stable. Pooled IC masks episodic predictability: most positive signal concentrates in a few quarters (especially 2024 Q4 and 2025 Q2), and roughly half of the test quarters are near zero or negative for almost every arm. None of the four crosswalk contrasts (L1−L0, L2−L1, L3−L2, L5−L3) changes sign in any leave-one-quarter-out (0/12, both universes) or leave-one-seed-out (0/10 each) replicate, with per-seed sign agreement between 8/10 and 10/10 (per-fold view and per-contrast counts in Appendices D and C). Sign stability is weaker than inferential stability: without 2025 Q2, C L1−L0 falls to +0.0090 and no longer rejects ($p = 0.125$; Table 15).

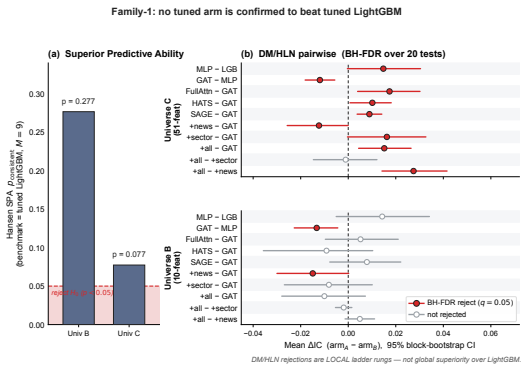

Figure 2: SPA and local DM/HLN evidence. Left: SPA $p_{\text{consistent}}$ by universe. Right: 20 pre-registered DM/HLN contrasts, with filled markers denoting BH-FDR rejection at $q = 0.05$.

Table 4: Gross/net crosswalk for the central Universe-C ladder contrasts and the edge contrast C L5−L3. "Carries?": the gross-IC result reproduces in net Sharpe; "NR" (not reproduced): the net ΔSharpe CI straddles zero.

Pair	Gross ΔIC (BH)	Net ΔSharpe@10 bps [CI]	Carries?
C L1−L0	$+0.0148$ (r)	$+1.17$ $[+0.36, +2.08]$	yes
C L2−L1	-0.0119 (r)	-0.72 $[-1.42, -0.10]$	yes
C L3−L2	-0.0123 (r)	$+0.08$ $[-0.77, +0.89]$	no (NR)
C L5−L3	$+0.0275$ (r)	$+0.88$ $[+0.12, +1.65]$	yes

5.3 The two load-bearing Universe-C contrasts survive transaction costs; the news-edge contrast does not

We re-express each pre-registered ranking contrast under a 10 bps net-of-cost convention (Table 4, Figure 3; descriptive, §3.3). The two central Universe-C ladder contrasts (C L1−L0 and C L2−L1) survive this economic layer, with paired net ΔSharpe confidence intervals excluding zero. The edge-stacking contrast C L5−L3 also reproduces in net Sharpe. Only the news-edge contrast C L3−L2 does not. In Universe B the net ΔSharpe signs match ($+0.43$ for L1−L0, -0.14 for L2−L1), but both paired intervals contain zero (Appendix C.9).

For L1−L0 the economic separation is larger than the IC gap suggests. At 10 bps, tuned LightGBM's mean per-cell net Sharpe is -0.22 (95% CI $[-0.82, +0.36]$) against the MLP's $+0.95$ ($[+0.18, +1.77]$). The object that matters is the *paired* per-fold net ΔSharpe (Table 4). Its point estimate is $+1.17$, and its stationary block-bootstrap CI over the 12 paired fold differences, $[+0.36, +2.08]$, excludes zero. It declines mildly with cost because the MLP turns over more (realized turnover τ_{L1} 2.90 versus 2.25 per rebalance in Universe C).

The exception is C L3−L2: the news-edge rung underperforms the correlation-GAT in IC but matches it in net Sharpe (net ΔSharpe $+0.08$, CI $[-0.77, +0.89]$), so the news-edge IC underperformance is economically fragile and says nothing about news edges as such (§5.4). The IC loss is already absent from the Sharpe gap at 0 bps (Appendix C.9), so the divergence arises between full-cross-section IC and decile returns, and costs do not drive it. The crosswalk covers only Family-1 ranking pairs; Family-2 contrasts stay in IC.

5.4 At a fixed operating point, no edge set changes IC detectably

Family-1's news-edge result mixes the edge change with changes in the tuned model; Family-2 evaluates each edge set at the frozen L2 hyperparameter vector (Table 5, Figure 4). None of the six contrasts survives BH-FDR, and all six are underpowered (|matched ΔIC| <MDE@80%). Family-2's pre-registered unit is the fold block ($n \approx 12$), coarser than Family-1's daily unit, so a ΔIC that rejects in Family-1 can fail to reject here.

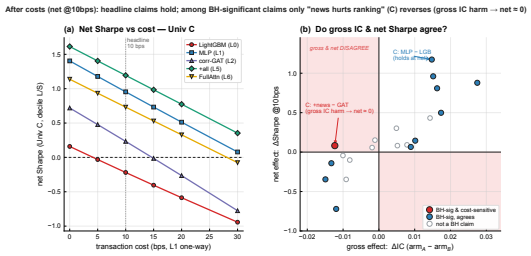

Figure 3: Net-of-cost crosswalk. Left: Universe-C cost ladder over 0–30 bps. Right: sign agreement between gross IC and net Sharpe across the 20 pre-registered pairs, with C L3−L2 annotated.

Table 5: Family-2 fixed-operating-point edge contrasts (matched ΔIC, BH-FDR over 6). UP = underpowered.

Univ	Edge	Matched ΔIC	BH	UP	Tuned ΔIC
B	news	+0.0015	n	y	−0.0149
B	sector	+0.0055	n	y	−0.0081
B	sector+news	+0.0042	n	y	−0.0100
C	news	+0.0011	n	y	−0.0123
C	sector	+0.0137	n	y	+0.0163
C	sector+news	+0.0141	n	y	+0.0152

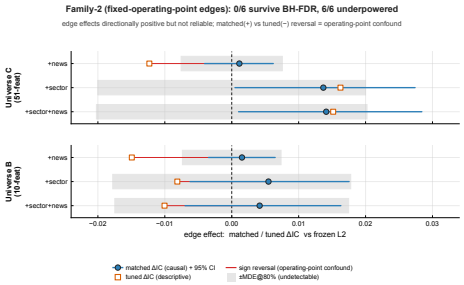

Figure 4: Matched-ΔIC versus tuned-ΔIC for the six edge contrasts. The ±MDE@80% band contains every matched estimate, and no contrast survives BH-FDR.

The sign reversals in Universe B are inconclusive: matched ΔIC is positive where tuned ΔIC was negative. In Universe C, the sector (+0.0137) and sector+news (+0.0141) contrasts have the same sign as their tuned counterparts, while the news contrast is small and positive (+0.0011) despite a negative tuned rung (−0.0123). The fixed-operating-point test therefore does not reproduce the tuned ladder's apparent news-edge harm. Given low power, this is a failure to detect harm; it does not show that the edge is harmless.

5.5 Disclosed failure modes

A tuned-configuration stability failure. The Universe-C GraphSAGE-Mean control arm with all edges (C/L5s) degenerates to constant predictions in 27.5% of its test fold-seeds, leaving Spearman IC undefined in those cells. The main conclusion is unchanged: under exclusion, zero-fill, or zero-skill handling, C/L5s mean IC remains approximately zero and the Universe-C SPA p-value remains between 0.077 and 0.080. Appendix D gives the cell-level breakdown and a plausible but unproven oversmoothing mechanism.

Exploratory failure modes. Appendix D details two exploratory analyses outside the confirmatory decision rules. One is a loss-function horse race in which ListMLE's advantage over MSE reversed sign across regimes; the other compares Universe-C's Plan-AAA feature-group ranking with a single-feature-IC proxy (Limitation L1).

6 Discussion

The evidence does not show that GNNs fail. Over this sample and protocol, with the graph path checked by a planted-signal control (§4), the SPA test does not reject the null of no win over tuned LightGBM (§5.2), and the strongest pre-registered neural contrast is the MLP's.

The one local DM contrast that holds in both universes under every correction is L2−L1 < 0: adding a correlation-GAT to the same node features reduces IC relative to the MLP. The C/L5s collapse hints at one mechanism: dense-edge aggregation plus high dropout smooth away cross-sectional signal.

The hint is weak, since we observe the collapse for GraphSAGE-Mean with all edges and not for L2's sparse-correlation GAT, so we bound the claim to the correlation graph at the tuned operating point. The other graph arms do not share the penalty: in Universe C, L4, L5, L6, and L7 all land above L2 with BH-rejected positive contrasts (Table 3).

The news-edge result shows why the two-family design matters: the tuned rung appears to harm ranking in Family-1, while the same edge change at the fixed L2 operating point is $+0.0011$ in Universe C, insignificant, and opposite-signed in Universe B. The tuned ladder answers a deployment question; the fixed-operating-point arms answer an edge-attribution question subject to underfitting.

Episodic signal also bounds the practical reading. Pooled IC and net Sharpe are averages over a regime mix in which a deployed strategy would spend roughly half its quarters flat or losing (§5.2, Appendix D), and twelve quarters cannot estimate how often favorable regimes recur. The practical conclusions concern model choice (prefer the MLP to the tuned correlation-GAT; expect no reliable win over LightGBM) and say nothing about attainable returns; a deployment case needs regime-conditional evaluation on a longer sample.

7 Limitations and Future Work

(L1) *Selection-stage leakage in Universe-C's basis.* Universe C's 51 columns come from PlanAAA top-15 feature groups ranked by neural-network permutation importance on quarters inside the evaluation window and under same-day open-high-low-close features. Only 5 of the 15 groups are also top-15 under a single-feature-IC proxy, and that proxy ranking is identical with and without the $T-1$ shift, so the disagreement reflects the importance measure and not the lag (Appendix D). We measure runtime features at $T-1$, but the basis *selection* leaked, so Universe-C positive results (MLP $>$ LightGBM, positive edge effects) are suggestive. A re-run of L0 and L1 on that 20-column subset (Appendix C.6) gives a similar point estimate but is itself conditional on the test-informed selection. A re-selection whose inputs stop in June 2022 (Appendix C.7) gives -0.0024 with a 95% interval $[-0.0256, +0.0178]$; the positive point estimate seen in Universe C does not reproduce there, but every paired interval contains zero, so the analysis does not establish that the contrasts differ. Neither run re-selects the graph or edge contrasts, and neither estimates how much leakage inflated Universe C. The Universe-B contrast is our closest leak-free comparison for L1$-$L0, and it bounds the inflation only under the assumption that Universe B is an adequate proxy; it says nothing about the Universe-C levels. Null and negative within-universe contrasts are less exposed (§5.2).

(L2) Predictive skill concentrates in a minority of the 12 test quarters (§6). **(L3)** We report the C/L5s collapse as a stability finding and do not repair it by retuning. **(L4)** The LightGBM comparison is underpowered: it gives no reliable evidence of superiority and no proof of equality. **(L5)** Net Sharpe is descriptive and sensitive to heavy tails and turnover, so we rely on fold-level paired ΔSharpe; the equal-weight decile portfolio ignores liquidity and short-borrow costs, so net Sharpe is an upper bound on realizable performance. **(L6)** L6 proxies the dense learned-attention family (Li et al., 2024; Cheng & Li, 2021); we include no learned-sparse architecture such as FinMamba (Chen et al., 2025), no dynamic multi-relational model (Qian et al., 2024), no all-MLP mixer (Fan & Shen, 2024), and no LSTM/Transformer time encoders. **(L7)** The evidence comes from one market and one prediction horizon.

(L8) *Survivorship and non-PIT composition.* The universe is a fixed end-window survivor snapshot (§4). A public-record audit finds 87 departed window members (14.8% of names, 8.2% of PIT stock-days) and an 8.1% look-ahead share from pre-addition inclusion, a 16.3% two-sided mismatch. Pairing reduces cross-arm imbalance but leaves the estimand non-PIT and the look-ahead bias, so per-arm *level* statements (Table 2; §5.3) overstate PIT-realizable levels (Appendix D). Graph arms also lose the omitted firms as neighbors, so the snapshot may not be neutral between graph and non-graph arms, in a direction we cannot sign. **(L9)** We fix the correlation-graph hyperparameters ($|\rho| > 0.6$, 126 days) without ablation. Each fold's graph dates from the end of training, so test-quarter edges are more than one quarter old, and we do not test a graph updated inside the test quarter. Next steps: a PIT rebuild with delisting returns, a longer sample, cross-market replication, intraday news rebalancing, and learned-sparse architectures.

AI USE STATEMENT

In this work, the authors used generative AI tools for the following tasks with required disclosure: implementing methods (experiment, evaluation, and analysis code), cleaning and reformatting the datasets, giving feedback on the research methodology and experimental design, and supporting the interpretation of results. The authors also used generative AI tools for tasks with recommended disclosure: editing the manuscript for readability and style, drafting parts of the paper (including its conversion to the ICLR template and first drafts of appendix sections), and writing the scripts that produce the figures. The authors did not use generative AI to propose or refine the research questions or hypotheses, or to identify related literature. Formulating or proving mathematical claims does not apply to this work, and the authors did not use generative AI to generate data; the synthetic node features of the planted-signal control come from a seeded random process in the experiment code. The authors reviewed and tested all AI-assisted code, checked every reported number against the experiment outputs with a claim-verification script, and revised all AI-assisted text. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses historical market data (prices, index membership, sector classifications, and news metadata) for a fixed universe of large-capitalization U.S. equities; it involves no human subjects, personal data, or user studies. The reported results are research findings on historical data under a specific evaluation protocol and do not constitute investment advice.

REPRODUCIBILITY STATEMENT

The project repository versions the code, confirmatory outputs, frozen tuned hyperparameters, the fold calendar, the Optuna grid, the ten seeds, and all analysis and figure scripts. We froze the protocol and tuned hyperparameters before the confirmatory run, with the hyperparameter file identified by md5 `59ddd0a2` (§3.3). A numeric-claim verifier checks the draft against underlying outputs, and the pre-registered protocol fixes the two-family decision rules. Experimental-setup details, including the DM/HLN estimator formulas, appear in Appendix A, and additional robustness analyses in Appendix C.

REFERENCES

Doron Avramov, Si Cheng, and Lior Metzker. Machine learning versus economic restrictions: Evidence from stock return predictability. *Management Science*, 69(5):2587–2619, 2023. doi: 10.1287/mnsc.2022.4449.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998.

Yifan Chen et al. FinMamba: Market-aware graph enhanced multi-level mamba for stock movement prediction. arXiv preprint arXiv:2502.06707, 2025.

Rui Cheng and Qing Li. Modeling the momentum spillover effect for stock prediction via attribute-driven graph attention networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 55–62, 2021.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. doi: 10.1080/07350015.1995.10524599.

Jinyong Fan and Yanyan Shen. StockMixer: A simple yet strong MLP-based architecture for stock price forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 8389–8397, 2024.

Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems*, 37(2):1–30, 2019. doi: 10.1145/3309547.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020. doi: 10.1093/rfs/hhaa009.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Peter Reinhard Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005. doi: 10.1198/073500105000000063.

David Harvey, Stephen Leybourne, and Paul Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291, 1997. doi: 10.1016/S0169-2070(96)00719-4.

Kewei Hou, Chen Xue, and Lu Zhang. Replicating anomalies. *The Review of Financial Studies*, 33(5):2019–2133, 2020. doi: 10.1093/rfs/hhy131.

Theis Ingerslev Jensen, Bryan T. Kelly, and Lasse Heje Pedersen. Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518, 2023. doi: 10.1111/jofi.13249.

Bryan T. Kelly, Semyon Malamud, and Kangying Zhou. The virtue of complexity in return prediction. *The Journal of Finance*, 79(1):459–503, 2024. doi: 10.1111/jofi.13298.

Raehyun Kim, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, and Jaewoo Kang. HATS: A hierarchical graph attention network for stock movement prediction. In *arXiv preprint arXiv:1908.07999*, 2019.

Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. MASTER: Market-guided stock transformer for stock price forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 162–170, 2024.

Hengxu Lin, Dong Zhou, Weiqing Liu, and Jiang Bian. Learning multiple stock trading patterns with temporal routing adaptor and optimal transport. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 1017–1026, 2021. doi: 10.1145/3447548.3467358.

Marcos López de Prado. *Advances in Financial Machine Learning*. John Wiley & Sons, 2018. ISBN 978-1-119-48208-6.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610.

Whitney K. Newey and Kenneth D. West. Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, 61(4):631–653, 1994. doi: 10.2307/2297912.

Manali Patel, Krupa Jariwala, and Chiranjoy Chattopadhyay. A systematic review on graph neural network-based methods for stock market forecasting. *ACM Computing Surveys*, 57(2):1–38, 2024. doi: 10.1145/3696411.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.

Hao Qian et al. MDGNN: Multi-relational dynamic graph neural network for comprehensive and dynamic stock investment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 14642–14650, 2024.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. Stock selection via spatiotemporal hypergraph attention network: A learning to rank approach. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 497–504, 2021.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web (ESWC)*, pp. 593–607, 2018.

Kevin Sheppard. arch: Autoregressive conditional heteroskedasticity (arch) and other tools for financial econometrics. Python package, `https://github.com/bashtage/arch`, 2024.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Sheng Xiang, Dawei Cheng, Chencheng Shang, Ying Zhang, and Yuqi Liang. Temporal and heterogeneous graph neural network for financial time series prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 3584–3593, 2022. doi: 10.1145/3511808.3557089.

Xiao Yang, Weiqing Liu, Dong Zhou, Jiang Bian, and Tie-Yan Liu. Qlib: An ai-oriented quantitative investment platform. *arXiv:2009.11189*, 2020.

A Experimental setup details

A.1 Walk-forward fold calendar

Our validation design is a 12-fold *expanding-window walk-forward* (Table 6). In fold k, training ends at a quarter boundary, validation is the next quarter, and testing is the following quarter. The twelve test quarters run from 2023 Q1 through 2025 Q4, yielding $T = 749$ pooled test days. Because the label is a 21-trading-day forward return, we purge the final 21 days of both the training and the validation window to eliminate label overlap (López de Prado, 2018). A secondary sliding 252-day specification is a robustness exercise only, opening no additional inferential family. Price, sector, and news histories span 2021 through 2026.

Table 6: Walk-forward fold calendar. Twelve expanding-window folds; each test quarter follows a one-quarter validation set, and we purge the final 21 days of both the training and the validation window. Pooled test span is $T = 749$ trading days.

Fold	Test quarter	Test days	Fold	Test quarter	Test days
0	2023 Q1	62	6	2024 Q3	64
1	2023 Q2	62	7	2024 Q4	64
2	2023 Q3	63	8	2025 Q1	60
3	2023 Q4	63	9	2025 Q2	62
4	2024 Q1	61	10	2025 Q3	64
5	2024 Q2	63	11	2025 Q4	61

A.2 Hyperparameter search space

We tune every ladder rung independently with Optuna, $N = 30$ trials per arm (arms are the ladder rungs defined in the main text) under an equal compute budget and six hyperparameter dimensions per family (Table 7). Search centers are the pilot defaults, so the pilot is the $N = 1$ center sample and remains comparable. Fixed, untuned settings for the neural arms: epochs $= 100$, patience $= 15$, gradient accumulation $= 32$; HATS relations $= 3$ (`linear_shared`; the three relation types are correlation, GICS sector, and news co-occurrence, feeding a 21-day ranking head). The LightGBM grid is six-dimensional to match the neural dimension count and preserve an equal tuning budget for the benchmark.

Tuning uses one split for every study: training through 2022-06-30 and validation on the second half of 2022, before the first test quarter. Each study runs its trials with one tuning seed, re-runs the top five configurations with three tuning seeds, and selects the highest mean validation IC. The tuning seeds are disjoint from the ten evaluation seeds, and the tuning correlation graph is the snapshot frozen at the tuning training end.

Table 7: Hyperparameter search grid (Optuna, $N = 30$ trials per arm). Top panel: neural arms (MLP / GAT / L6 full-attention / L7 HATS). Bottom panel: LightGBM (L0).

Dimension	Centre	Search range
Neural arms		
learning rate	1×10^{-3}	log-uniform $[1 \times 10^{-4}, 1 \times 10^{-2}]$
weight decay	1×10^{-4}	log-uniform $[1 \times 10^{-5}, 1 \times 10^{-3}]$
dropout	0.3	$\{0.1, 0.2, 0.3, 0.5\}$
hidden channels	64	$\{32, 64, 128\}$
num layers	2	$\{1, 2, 3\}$
attention heads	4	$\{2, 4, 8\}$
LightGBM (L0)		
num leaves	31	$\{15, 31, 63, 127\}$
learning rate	0.05	log-uniform $[0.01, 0.1]$
min data in leaf	20	$\{10, 20, 50, 100\}$
n estimators	100	early-stopped (not gridded)
`lambda_l1`	1×10^{-8}	log-uniform $[1 \times 10^{-8}, 1.0]$
`lambda_l2`	1×10^{-8}	log-uniform $[1 \times 10^{-8}, 1.0]$

A.3 Frozen tuned hyperparameters

We froze the protocol and the tuned hyperparameters before the confirmatory run. The hyperparameter file has md5 `59ddd0a2` (`artifacts/storya_v21_tune/frozen_hparams.json`, committed to the repository before launch). Table 8 reports the discrete winning hyperparameters of each of the 20 tuning studies verbatim from the frozen file. The frozen file records the continuous winners (learning rate and weight decay) at full floating-point precision, and we do not re-round them here. For L0 (LightGBM), the frozen discrete winners are `num_leaves` $= 15$ with `min_data_in_leaf` $= 50$ in Universe B and `num_leaves` $= 15$ with `min_data_in_leaf` $= 100$ in Universe C. The frozen file records `n_trials` $= 25$ for study C/L5s and 29 for C/L6; all other studies record 30.

Both studies ran the full 30-trial budget, but five C/L5s trials and one C/L6 trial returned an undefined validation objective, and the archiving step counts only trials with a finite objective. The objective is undefined when no validation day has a defined rank IC, which happens when a configuration produces constant cross-sectional predictions. The five failed C/L5s configurations all used three layers, 128 hidden channels, and learning rates between 6.2×10^{-3} and 1.0×10^{-2}, with dropout 0.5 in four of them and 0.3 in the fifth. The C/L5s winner (three layers, dropout 0.5, 32 hidden channels, and a learning rate inside that range) shares the depth, dropout, and learning-rate range of the failed trials and differs only in width. The tuning stage therefore already showed a precursor of the test-period collapse reported in §D.2.

Table 8: Frozen tuned discrete hyperparameters per neural tuning study (verbatim from `frozen_hparams.json`, md5 `59ddd0a2`). Arm definitions follow the model ladder in the main text; "–" marks a dimension that is not part of the arm's recorded winner. † rows are the 90-trial L2 re-tunes from the M14 search-budget sensitivity (§C.1; `frozen_hparams_m14.json`); all other arms stay at 30 trials.

Univ.	Arm	Drop.	Hid.	Lay.	Heads	Trials	Univ.	Arm	Drop.	Hid.	Lay.	Heads	Trials
B	L1	0.3	128	3	–	30	C	L1	0.1	128	1	–	30
B	L2	0.3	64	2	2	30	C	L2	0.1	128	1	4	30
B	L2s	0.3	128	1	–	30	C	L2s	0.1	128	1	–	30
B	L3	0.2	32	2	4	30	C	L3	0.1	32	2	2	30
B	L4	0.1	32	2	4	30	C	L4	0.2	128	1	4	30
B	L5	0.1	32	2	4	30	C	L5	0.1	128	1	4	30
B	L5s	0.1	64	1	–	30	C	L5s	0.5	32	3	–	25
B	L6	0.1	32	2	8	30	C	L6	0.1	64	1	4	29
B	L7	0.3	128	1	2	30	C	L7	0.1	128	1	4	30
B	L2†	0.1	32	1	2	90	C	L2†	0.1	128	1	8	90

An equal budget equalizes the trial count only. Neural arms search higher-dimensional spaces than LightGBM, so search adequacy differs across arms, a caveat for the "MLP beats GAT" interpretation. A trials-sensitivity sweep probes this caveat. Re-tuning L2 at $3\times$ the budget (90 trials, matching the MLP's search density) leaves the leak-free Universe-B penalty significant (L2$-$L1 $= -0.0090$, HLN $p = 0.002$, rejected under Benjamini–Hochberg (BH) false-discovery-rate (FDR) control) but drops Universe-C's below significance (-0.0069, $p = 0.059$, sign unchanged). The gap narrows without closing, so the primary leak-free result is not an equal-budget artifact.

A.4 Cell budget accounting

The confirmatory main table is 2160 cells for the nine non-HATS rungs (9 arms $\times$ 2 universes $\times$ 12 folds $\times$ 10 seeds), plus 240 L7/HATS cells, plus 720 fixed-capacity Family-2 cells (L3fc/L4fc/L5fc $\times$ $2 \times 12 \times 10$), for 3120 cells in total. Each of these three stores is internally complete (0 failed, 0 duplicate `cell_id`), and each pre-specified family consumes its own store: Family-1 reads the main table plus L7, and Family-2 reads the fixed-capacity store. The fixed-capacity arms reuse the main per-arm `cell_id` numbering by design, so we keep the three stores as separate key spaces. The L7 contingency rule is an automatic health gate that demotes HATS out of the family if more than 20% of its cells diverge or collapse. It did not fire (diverge fraction 0, collapse fraction 0; source `family1_summary.md`). L7 stays in the family, and the Superior Predictive Ability (SPA) test runs at $M = 9$.

The ten canonical seeds, used in all experiments, are $\{7, 34, 86, 99, 123, 456, 789, 1024, 2024, 2026\}$. Universe B carries 10 features and Universe C carries 51, all measured strictly at $T - 1$.

A.5 News-graph construction

The news graph fits no construction parameters. Ticker tags are the provider's per-article annotations, an edge requires only co-mention in a multi-ticker article, and two temporal parameters are fixed a priori. The article's publication timestamp must fall before the NYSE close on $T - 1$ and within the preceding seven calendar days (about five trading days); edges expire after that window. The construction estimates nothing from full-sample or future data.

A.6 DM/HLN estimator details

Local paired evidence comes from Diebold–Mariano tests (Diebold & Mariano, 1995) with the Harvey–Leybourne–Newbold small-sample correction (Harvey et al., 1997) on 20 pre-registered contrasts: five ladder contrasts, $\{$L1$-$L0, L2$-$L1, L6$-$L2, L7$-$L2, L2s$-$L2$\}$, and five edge-set contrasts, $\{$L3$-$L2, L4$-$L2, L5$-$L2, L5$-$L4, L5$-$L3$\}$, each in both universes. The main correction is BH-FDR at $q = 0.05$ over the pooled 20-test family (Benjamini & Hochberg, 1995). A per-universe BH correction gives the same rejection decisions.

The DM statistic uses Newey–West heteroskedasticity-and-autocorrelation-consistent (HAC) standard errors (Newey & West, 1987) with the data-driven bandwidth of Newey & West (1994), $L = \lfloor 4(T/100)^{2/9} \rfloor$ ($L = 6$ at $T = 749$), to absorb serial correlation in the daily loss differential. The HLN factor, $\sqrt{(T + 1 - 2h + h(h-1)/T)/T}$ with $h = 21$, is a finite-sample correction to the DM variance estimate at an h-step horizon, and we compare the corrected statistic with a t_{T-1} distribution. Harvey et al. (1997) derive the factor for a rectangular window of $h - 1$ lags. We pair it with a Bartlett-kernel HAC estimate, so the combination is a hybrid, and the lag-21 column checks its sensitivity.

The automatic bandwidth ($L = 6$) is shorter than the 21-day label overlap, which induces moving-average dependence up to lag 20 in the daily loss differential. We froze the automatic lag before the confirmatory run, and it is the pre-registered primary specification. The fixed lag-21 HAC check (column `HLN_p_t_lag21`) is a sensitivity, and Table 18 reports it for all 20 contrasts. At lag 21, L2$-$L1 stays below 0.05 in both universes, and C L1$-$L0 does not.

Table 9: Evaluation-protocol audit of the eight most closely related deep/GNN cross-sectional ranking systems and this work. Cells reflect each paper's *reported* protocol; "n.r." marks a property the paper does not report, and the table records reporting differences only, so a gap does not imply that the method could not satisfy the property. Universe counts, where shown in parentheses, are numbers of stocks. Costs: no audited paper models transaction costs; RSR and THGNN explicitly dismiss them, and AD-GAT reports no portfolio backtest at all. Seeds: AD-GAT's entry is the mean of the top five of 30 initializations *selected on validation*; TRA and MASTER additionally report standard deviations over their five seeds. Graph-off: "yes"/"no"/"partial" indicates whether the paper reports a no-graph variant of its own framework (details in §B.2); "n/a" marks the non-graph TRA. The final row abbreviates our protocol: 12-fold expanding walk-forward with a 21-day purge; ten seeds plus a leave-one-seed-out (LOSO) check; BH-FDR at $q = 0.05$ with Benjamini–Yekutieli and pooled-26 sensitivities; a 0–30 bps cost grid with per-arm, per-fold turnover; and the feature-matched L2−L1 graph-off contrast as the headline, complemented by the fixed-operating-point edge-attribution family. Full citations for the audited systems appear in the opening paragraph of this appendix.

Work	Universe	Splits	Seeds	Mult.	Costs	Graph-off
RSR	NASDAQ/NYSE	single (2017)	5	none	none	yes
STHAN-SR	NASDAQ/NYSE/TSE	single (×3)	5	none	none	yes
AD-GAT	S&P 500 (198)	single (70d)	top-5/30	none	none	no
TRA	CSI800	single, purged	5	none	none	n/a
THGNN	S&P 500/CSI300	single (2020)	5	none	none	no
MASTER	CSI300/CSI800	single (10q)	5	none	none	no
StockMixer	NASDAQ/NYSE/S&P	single (×3)	3	none	none	partial
MDGNN	CSI100/CSI300	7 rolling folds	n.r.	none	none	no
This work	S&P 500 (501)	12-fold WF	10+LOSO	BH+BY	0–30 bps	matched L2−L1

B Extended related work and evaluation-protocol audit

This appendix audits the evaluation protocols *reported* by the eight most closely related deep-learning systems for cross-sectional stock ranking, published between 2019 and 2024 at TOIS, AAAI, KDD, and CIKM. The systems are RSR (Feng et al., 2019), STHAN-SR (Sawhney et al., 2021), AD-GAT (Cheng & Li, 2021), TRA (Lin et al., 2021), THGNN (Xiang et al., 2022), MASTER (Li et al., 2024), StockMixer (Fan & Shen, 2024), and MDGNN (Qian et al., 2024). It extends the related-work discussion in the main text. TRA has no graph, and we include it as the closest published Qlib-based ranking baseline. HATS (Kim et al., 2019) and FinMamba (Chen et al., 2025) motivate rungs of our ladder, and we exclude them from the audit because neither appears at an archival venue. We compiled every entry from the papers as published. An entry records what a paper *reports*: "n.r." marks a property the paper does not report, and we treat every gap as a reporting difference. A gap is neither a claim that the property is impossible for the method nor evidence that this protocol could not evaluate earlier methods.

B.1 Protocol comparison matrix

Table 9 yields three aggregate facts. First, seven of the eight papers evaluate on a single chronological split; MDGNN alone uses seven rolling folds, and only TRA adds purge gaps. Second, none of the eight applies any multiple-testing correction, although the papers typically compare on the order of 8–15 models across 3–6 metrics and 2–3 markets. Four of the eight report no hypothesis test at all, and the remaining four use uncorrected t or Wilcoxon tests against one or two reference models. Third, none of the eight models transaction costs; RSR and THGNN dismiss them explicitly. Seed counts cluster at five (minimum three; one paper does not report seeds). No audited paper reports a leave-one-seed-out check, a pre-registered analysis plan, a power analysis, or a planted-signal positive control.

B.2 Per-paper positioning

THGNN shares our edge type, not our protocol (Xiang et al., 2022). Among the eight systems, THGNN has the cleanest graph construction: it regenerates its edges daily from a trailing 20-day

return correlation with a $|\rho| \geq 0.6$ threshold, and the construction, as described, uses no future or full-sample information. THGNN and our $\alpha 1$ graph share the edge type, a thresholded trailing return correlation with a 0.6 cutoff. They differ in window (20 versus 126 days) and in update rule: THGNN regenerates edges daily, and we hold one Pearson-correlation snapshot per fold, the last inside the training window, through validation and test. Our headline feature-matched contrast tests the edge type that THGNN relies on. THGNN's evaluation, however, uses a single test year (calendar 2020) per market with five repeated runs, no significance tests, and explicitly ignored costs. All of its reported ablations retain the graph, so the framework has no internal no-graph variant. The two studies also differ in label, horizon, and period, so the connection positions our contrast relative to THGNN and does not replicate it.

RSR (Feng et al., 2019) and STHAN-SR (Sawhney et al., 2021) use look-ahead graphs and show mixed graph-off evidence. Both works attach relation graphs built from a single Wikidata snapshot, mined at collection time and held fixed through the test period (in STHAN-SR's case including a TSE test extending to August 2020). Neither paper discusses the resulting look-ahead exposure in graph construction, and RSR also restricts its universe to stocks with near-complete full-sample histories, a survivorship-style filter. Our news graph is strictly point-in-time (publication-timestamp gated) and our correlation graph is trailing-only. Both papers' own ablations already contain mixed graph-off evidence. In RSR, the no-graph Rank_LSTM attains a NASDAQ investment return ratio of 0.68, above the GCN variant (0.24) and both industry-relation RSR variants (0.20 and 0.23); only the Wikidata-relation RSR_I variant (1.19) exceeds it. In STHAN-SR, the plain hypergraph convolution without attention (Sharpe 0.93) trails the LSTM-only variant (0.95), while the full model reaches 1.42. The internal evidence of these two papers is thus consistent with a configuration-conditional graph benefit; we measure that conditionality under controlled inference.

AD-GAT's selection step inflates its estimand (Cheng & Li, 2021). AD-GAT evaluates on an S&P 500 subset of 198 stocks, retained by requiring no missing data and at least 100 news articles over the full sample period; this coverage filter conditions on future information. Its single split leaves one 70-day test window. It trains 30 initializations and reports the mean of the top five selected on validation, with direction accuracy and AUC only (no portfolio backtest) and uncorrected t-tests. Relative to a deployment estimand, this run-selection protocol yields an upward-biased estimate of test performance; our seed-mean estimand with a leave-one-seed-out check follows the opposite convention. We record this as a difference in reported estimands; in the paper's own terms it is not an error.

The complexity counterpoint is compatible with our result. The virtue-of-complexity results of Kelly et al. (2024) are the strongest published counterweight to any "simple beats complex" reading of our findings. Their random-feature ridge models with up to $P = 12{,}000$ features deliver an out-of-sample market-timing Sharpe gain of approximately 0.47 per year ($t \approx 3$) even while out-of-sample R^2 remains negative (e.g., -3.8% alongside a Sharpe ratio of 0.46). Their setting, however, is single-asset market timing in the heavily shrunk $P > T$ regime. Our equal-budget tuned ladder compares standard GNN architectures at tuned operating points (hidden sizes 32–64, at most two layers), far from that regime. The two sets of results are therefore compatible: our negative graph contrast applies to the correlation graph at the tuned operating point only and says nothing about over-parameterized models under ridge shrinkage. Their argument that R^2-type fit measures are incomplete guides to economic value is also consistent with our choice to report net Sharpe alongside Information Coefficient (IC), with IC as the only confirmatory metric and the net-Sharpe layer descriptive.

Our multiplicity control errs in the stricter direction. Studying 153 factors across 93 countries, Jensen et al. (2023) show that the multiple-testing framework alone can move a headline replication rate by roughly 45 percentage points. The rate runs from 35% under an HXZ-style construction, to 55.6% under their capped value-weighted construction, to 82.4% under a CAPM-alpha test object, with 75.6% under frequentist Benjamini–Yekutieli control and 82.4% under their hierarchical Bayesian model. Their hierarchical empirical-Bayes treatment of correlated signals has higher power than BH/BY and can rescue borderline findings. Our design takes the conservative frequentist direction: pre-registered BH families with a Benjamini–Yekutieli sensitivity valid under arbitrary dependence. Under that sensitivity, 7 of our 11 BH-rejected ladder contrasts survive: the graph-

penalty contrast L2−L1 still rejects in both universes, while C L1−L0, L3−L2 in both universes, and C L4−L2 drop. A joint Bayesian treatment of the correlated ladder contrasts, in their spirit, is a natural extension that we leave to future work; because our choice errs in the stricter direction, our surviving results stand under it.

B.3 The matrix supports a bounded caution about reported GNN gains

Read as a whole, the matrix supports a bounded caution. None of the eight audited papers corrects for multiplicity while comparing on the order of 8–15 models across 3–6 metrics and 2–3 markets, so the reported margins carry unquantified selection effects. None models costs while relying on daily top-k rotation portfolios that imply near-100% daily turnover, so gross performance gaps need not survive frictions. Where internal graph-off evidence exists (RSR, STHAN-SR), it already points to a configuration-conditional benefit. None of this shows that any specific published gain is spurious; the audited papers follow the reporting norms of their venues and periods. The matrix does show that the published record, as reported, cannot bound how much of the headline gains would survive a multi-seed, multiplicity-controlled, cost-aware re-evaluation. Two independent observations are at least consistent with our findings. StockMixer, an MLP-family architecture, reportedly outperforms GNN hybrids on the field's own benchmark datasets (Fan & Shen, 2024), converging with our (suggestive) non-graph MLP result, though under the same single-split, three-seed, cost-free protocol. The field's systematic review (Patel et al., 2024) catalogs architectures without tabulating the prevalence of significance testing, cost modeling, or seed reporting, so the survey does not quantify the rigor gap this audit measures. We apply the same standard to our own results: the SPA outcome is a non-rejection, which is not evidence that the models are equal, and the graph-penalty finding applies to the correlation graph at the tuned operating point only.

C Additional robustness analyses

The analyses in this appendix are post-hoc sensitivities and one pre-confirmatory pipeline control. They supplement the two pre-registered confirmatory families; they are not new confirmatory families and do not replace the pre-registered decision rules.

C.1 Trials-sensitivity of the correlation-GAT penalty

The equal budget could favor the MLP. The equal 30-trial tuning budget equalizes trial count and leaves search adequacy unequal. The correlation-GAT's categorical hyperparameter space is three times the MLP's (the extra attention-heads dimension yields 108 versus 36 combinations), so the finding that the tuned correlation-GAT (L2) underperforms the tuned MLP (L1) could in principle be a search-budget artifact. A validation-to-test diagnostic cannot separate under-search from inherent fragility, so we ran a direct sensitivity.

We re-tune only L2 at the MLP's search density. We re-tuned only L2, at 90 trials (matching the MLP's search density), in both universes, under the same nominal protocol; this grants a one-sided advantage to the graph arm. Because the tuner restarts its Tree-structured Parzen Estimator (TPE) sampler per process, the 90-trial search is an independent draw and does not contain the confirmatory 30-trial search as a subset. The 90-trial winner accordingly differs from the 30-trial winner (in Universe B: dropout $0.3 \to 0.1$, hidden width $64 \to 32$, two layers $\to$ one). We merged the winning configuration into the frozen hyperparameter file, re-evaluated the 240 L2 model-fold-seed cells, and recomputed L2−L1 against the unchanged confirmatory arms; all other arms remain at 30 trials.

More search narrows the gap without closing it. More search finds better *validation* configurations, as expected. The selected configuration's validation IC rises from 0.05429 to 0.05595 in Universe B and from 0.05503 to 0.06304 in Universe C, and the 90-trial Universe-B winner is *simpler* (hidden width 32, one layer, dropout 0.1, versus 64, two layers, 0.3). Pooled *test* IC also rises but stays below the MLP. L2 moves from 0.0238 to 0.0281 in Universe B and from 0.0224 to 0.0274 in Universe C, against MLP (L1) pooled ICs of 0.0371 and 0.0343. The graph-versus-MLP gap narrows by roughly 30–40% but does not close. Table 10 reports the paired DM/HLN tests (Diebold & Mariano, 1995; Harvey et al., 1997).

Table 10: Trials-sensitivity of the L2−L1 contrast: confirmatory 30-trial tuning versus an independent 90-trial L2-only retune (3× budget, density-matched to the MLP). Only L2 receives the additional budget. BH denotes Benjamini–Hochberg FDR control at $q = 0.05$ (Benjamini & Hochberg, 1995).

Universe	L2 tuning budget	ΔIC (L2−L1)	HLN p	BH ($q = 0.05$)
B (leak-free, primary)	30 trials (confirmatory)	−0.0133	3.97×10^{-4}	reject
B (leak-free, primary)	90 trials (sensitivity)	−0.0090	0.0021	reject
C (leak-selected)	30 trials (confirmatory)	−0.0119	6.9×10^{-6}	reject
C (leak-selected)	90 trials (sensitivity)	−0.0069	0.059	not rejected

The leak-free penalty survives the larger budget. The headline conclusion, that the tuned correlation-GAT underperforms the tuned MLP, survives BH significance at three times the GAT budget in the leak-free Universe B. This addresses the specific equal-30-trial, categorical-search-density objection in the primary leak-free universe. It does *not* prove exhaustive tuning fairness or rule out every under-tuning explanation, because density-matching under adaptive TPE search is a heuristic and carries no optimizer-fairness guarantee. The objection has partial merit: extra search shrinks the gap by roughly 30–40%, and in the leak-selected Universe C the significance does not survive ($p = 6.9 \times 10^{-6} \to 0.059$), although the ΔIC stays negative and short of the nominal two-sided level. We therefore describe the contrast as BH-significant at both budgets in the leak-free Universe B and as search-budget-sensitive (direction unchanged, near-nominal) in the leak-selected Universe C. We treat the 90-trial run as a sensitivity. The project decision log designated Universe B the primary leak-free universe on 2026-06-26 and 2026-06-30, after the Universe-C leakage audit of 2026-05-27 and before the 90-trial run of 2026-07-02 and 2026-07-03; the designation was a project decision and was not written into a separate sensitivity plan. Universe B's ten features (5-, 10-, and 21-day mean and standard deviation of returns, 12-month momentum, maximum daily return, dollar volume, and 5-day price-volume correlation) were assembled in April 2026 from standard Alpha158 and GKX-style factors, before the confirmatory design, and were never ranked on the confirmatory sample. One single-feature diagnostic on a 63-day window of an earlier fold layout inside the confirmatory test span informed the choice to add volume features; every feature is lagged one day.

C.2 Per-seed sign agreement and leave-one-seed-out

The confirmatory tests average the ten seeds into one series before testing, so cross-seed dispersion enters as a deployment caveat and stays outside family-wise inference. As a sensitivity, we reconstruct per-seed pooled ΔIC for the four BH-rejected ladder contrasts (L1−L0, L2−L1, L3−L2, L5−L3), restricted to their BH-rejected instances (six universe–contrast pairs in total). We build each per-seed value from fold-level mean ICs weighted by test-day counts (equivalent to the day-level pooled mean). Table 11 reports, for each contrast, the number of seeds whose individual pooled ΔIC has the same sign as the seed-mean estimate, the number of leave-one-seed-out (LOSO) sign flips, and the per-seed range. No BH-rejected contrast changes sign when we remove any single seed (0/10 each). Per-seed signs agree in 8/10 for the two Universe-B penalties, 9/10 for their Universe-C counterparts, and 10/10 for C L1−L0 and C L5−L3.

Table 11: Per-seed sign agreement and leave-one-seed-out (LOSO) for the six BH-rejected universe–contrast pairs. "Same sign" counts seeds whose per-seed pooled ΔIC matches the sign of the seed-mean estimate; "LOSO flips" counts seeds whose removal flips the sign of the pooled ΔIC.

Universe	Contrast	Pooled ΔIC	Same sign	LOSO flips	Per-seed min	Per-seed max
B	L2−L1	−0.01327	8/10	0/10	−0.03116	0.01361
B	L3−L2	−0.01493	8/10	0/10	−0.03997	0.02573
C	L1−L0	0.01477	10/10	0/10	0.00708	0.02825
C	L2−L1	−0.01193	9/10	0/10	−0.03663	0.00133
C	L3−L2	−0.01231	9/10	0/10	−0.02452	0.0093
C	L5−L3	0.02747	10/10	0/10	0.00419	0.04504

C.3 POOLED BH AND BENJAMINI–YEKUTIELI SENSITIVITY

The main analysis controls the 20 pre-registered DM/HLN contrasts as one BH-FDR family at $q = 0.05$ (Benjamini & Hochberg, 1995) and the six edge-attribution (Family-2) tests as a separate family. Two post-hoc sensitivities bound this layering choice. First, pooling the 20 DM and six Family-2 p-values into a single 26-test BH family rejects the same 11 DM contrasts and no Family-2 contrast, so the rejection decisions match the pre-registered per-family decisions. Second, Benjamini–Yekutieli (Benjamini & Yekutieli, 2001), valid under arbitrary dependence ($q/c(m)$ with $m = 20$, $c(m) = 3.598$), keeps 7 of the 11 rejections: L2−L1 in *both* universes, and C L2s−L2, C L5−L2, C L5−L3, C L6−L2, C L7−L2. The contrasts that drop (C L1−L0, L3−L2 in both universes, and C L4−L2) are the ones the main text already labels suggestive or not reproduced in net Sharpe. The headline graph penalty (L2−L1 < 0) therefore still rejects in both universes under the most conservative dependence assumption. Table 12 lists the full decision grid.

Table 12: BH versus Benjamini–Yekutieli (BY) rejection decisions for the 20 pre-registered DM/HLN contrasts ($q = 0.05$; BY uses $q/c(m)$, $m = 20$, $c(m) = 3.598$). • = reject, ∘ = not rejected.

	Universe B		Universe C	
Contrast	BH	BY	BH	BY
L1−L0	∘	∘	•	∘
L2−L1	•	•	•	•
L6−L2	∘	∘	•	•
L7−L2	∘	∘	•	•
L2s−L2	∘	∘	•	•
L3−L2	•	∘	•	∘
L4−L2	∘	∘	•	∘
L5−L2	∘	∘	•	•
L5−L4	∘	∘	∘	∘
L5−L3	∘	∘	•	•

C.4 PLANTED-SIGNAL POSITIVE CONTROL: DETAIL

The control plants a label on the real graph. Before the confirmatory run, a planted-signal control tested whether the graph path works at all. It uses the real frozen correlation-graph topology, synthetic node features $X \sim \mathcal{N}(0, 1)$, and a label planted *on* the graph: for each day, the label is β times the normalized-adjacency neighbor mean of one observed feature, plus noise. By construction, a working message-passing GNN must recover this signal and a feature-matched non-graph MLP cannot; we exclude isolated nodes from evaluation. We calibrate the coefficient β so that the measured oracle IC (the Spearman correlation of the true signal with the label on test days, over connected stocks) is approximately 0.05. The pre-registered pass rule requires a GNN test IC of at least $0.7\times$ the measured achievable IC *and* the GNN significantly above the MLP (seed-averaged-per-fold paired HLN, $h = 21$, BH-FDR). The run comprises 60 cells: {GAT, GraphSAGE-Mean, MLP} $\times$ 4 seeds $\times 5$ folds in Universe B.

Both GNNs pass the pre-registered rule. Against a measured achievable IC of 0.0469, GraphSAGE-Mean attains test IC 0.0427 (91% recovery) and GAT 0.0384 (82%), while the MLP control attains 0.0024 ($\approx$0). Both GNNs are significantly above the MLP (BH-FDR, HLN $p \approx 0$), and both pass the pre-registered rule (Table 13). The recovery is consistent at the cell level: the GAT and GraphSAGE-Mean ICs are positive in all 20 fold-seed cells each (per-cell ranges $[0.0248, 0.0536]$ and $[0.0353, 0.0577]$), whereas the MLP is positive in only 12 of 20 (range $[-0.0106, 0.0205]$). The verdict is also insensitive to the pass threshold: at $0.7\times$ both GNNs pass; at $0.8\times$ (target 0.0375) both pass; at $0.9\times$ (target 0.0422) GraphSAGE-Mean passes and GAT is marginally below.

The control certifies gross operation only. The control establishes that the graph path converts graph-borne signal into cross-sectional IC, ruling out the *gross* broken-pipeline failure modes: edges not wired into learnable parameters, gross training failure, a representational bottleneck, or

Table 13: Planted-signal positive control (60 cells: 3 models $\times$ 4 seeds $\times$ 5 folds, Universe B). Recovery is relative to the measured achievable IC of 0.0469.

Model	Test IC	Recovery	Positive cells	Per-cell range
GraphSAGE-Mean	0.0427	91%	20/20	$[0.0353, 0.0577]$
GAT	0.0384	82%	20/20	$[0.0248, 0.0536]$
MLP (control)	0.0024	$\approx$0	12/20	$[-0.0106, 0.0205]$

a lookahead-broken aggregation. It does *not* certify that the architecture or hyperparameters are optimal for real financial features; a synthetic-feature positive control cannot exclude a subtle misspecification on real features. The licensed claim is therefore that the graph pipeline is operational and that the null is not a broken-pipeline artifact. The control does not show that the null is definitively a task property. The control ran on the pilot pipeline that the confirmatory runners reuse import-only, so it certifies gross operation only.

C.5 Minimum detectable effect: construction and caveats

For each contrast, we report a minimum detectable effect (MDE) to distinguish statistical non-rejection from evidence of a precisely estimated zero. We define MDE $= 2.8 \times \mathrm{SE}_{\mathrm{block}}$, with $n_{\mathrm{eff}} = T_{\mathrm{days}}/21$. The multiplier $2.8 \approx z_{0.975} + z_{0.80}$ is the standard two-sided 5% test, 80% power factor for a z-test under an approximately normal block-bootstrap sampling distribution. This MDE rests on a coarse non-overlapping block SE over $n_{\mathrm{eff}} \approx 36$ blocks and is a deliberately conservative power yardstick. It differs from the HAC daily-loss SE behind the DM/HLN p-values. "Underpowered" therefore means the gap is small relative to block-level dispersion, and it says nothing about the resolution of the daily DM test.

Three sample-size notions enter the analysis: daily-IC inference uses $T = 749$ pooled test days, the IC/Sharpe MDE uses $n_{\mathrm{eff}} = T/21 \approx 36$ non-overlapping blocks, and Family-2 uses about 12 fold blocks. Family-2's fold-block unit is coarser than Family-1's daily unit, so the two families' p-values are not comparable at equal ΔIC.

This yardstick separates the two directions of the headline result. The correlation-graph penalty exceeds its 80% MDE in Universe C ($|-0.0119| > 0.0089$), a detectable effect. The MLP's gain over LightGBM falls below its own MDE (C L1$-$L0 $= +0.0148 < 0.0220$) and is therefore underpowered: the data give no reliable evidence of superiority and do not prove equality.

C.6 Feature-subset sensitivity (post-hoc, test-informed)

We re-tuned and re-evaluated L0 (LightGBM) and L1 (MLP) on the 20 Universe-C columns belonging to the five Plan-AAA groups that are also top-15 under a single-feature-IC proxy (ROC30, KMID, KUP, CNTP20 and CORR60 groups), under the frozen protocol (12-fold expanding walk-forward, 21-day purge, 10 seeds, 30-trial equal-budget tuning per arm). This is a sensitivity analysis, not a confirmatory test: both rankings that define the subset were scored on quarters inside the evaluation window, so the subset cannot resolve or bound the selection leakage discussed in Limitation L1, and no multiple-testing family is opened (nominal, unadjusted p-values). On the subset the MLP$-$LightGBM contrast is $+0.0134$ (95% 21-day block-bootstrap CI $[+0.0008, +0.0283]$; nominal HLN $p = 0.008$ at the NW automatic lag and 0.054 at lag $=$ horizon; 10/10 seeds same sign), versus $+0.0148$ in Universe C and $+0.0143$ in Universe B. The paired daily change C $-$ subset is $+0.0013$, and its paired interval, $[-0.0159, +0.0189]$, contains zero; the run does not establish equivalence. As in C and B, the contrast is below the design's approximate MDE ($\approx$0.020) and concentrated in 2025 Q2 (53% of the pooled value; excluding that quarter, $+0.007$ $[-0.003, +0.018]$). Both arms' tuning-window validation IC on the subset was negative for every finalist (L0 -0.012, L1 -0.045), so the tuned configurations are protocol-consistent but not validated optima, and the subset MLP is $\approx 14\times$ smaller than the Universe-C MLP. A replicate on a different backend (Mac MPS vs CUDA) gives $+0.0132$ $[+0.0002, +0.0279]$. Universe B remains the leak-free feature-basis anchor.

Three caveats govern the reading. The subset's smaller p-value comes from a smaller block SE (0.0070; C 0.0079, B 0.0098) and not from a larger effect. Its percentile interval excludes zero only at the boundary: the symmetric normal margin, $1.96 \times \text{SE} = 0.0138$, exceeds the point estimate 0.0134. At lag 21 none of the three contrasts reaches the nominal 0.05 level. The paired B − subset change is $+0.0008$ $[-0.0186, +0.0200]$, and the paired MDEs (0.0245 for C − subset, 0.0276 for B − subset) exceed the Universe-C contrast itself, so the non-rejection does not establish equivalence. The subset MLP has 2,337 parameters against 31,745 in Universe C; per-seed counts measure seed stability on the same data and are not independent replications. Table 14 places the subset beside the confirmatory Universe-C and Universe-B rows, and Table 15 gives the quarter-concentration and paired-change detail.

Table 14: Feature-subset sensitivity of the L1−L0 contrast (post-hoc, test-informed): the 20-column Universe-C subset (C5) beside the confirmatory Universe-C and Universe-B rows. ΔIC is the seed-averaged daily MLP−LightGBM difference; the CI is a 21-day block bootstrap; HLN p-values are nominal and unadjusted, at the implementation-default Newey–West automatic lag and at lag 21 (the label horizon); MDE is $2.8 \times \text{SE}_{\text{block}}$ (§C.5); "same sign" counts seeds whose per-seed contrast matches the sign of the seed mean. All three contrasts fall below their MDE.

Universe	ΔIC	95% CI	p (auto)	p (lag 21)	MDE	Same sign
C5 (post-hoc)	$+0.0134$	$[+0.0008, +0.0283]$	0.008	0.054	0.0197	10/10
C (confirmatory)	$+0.0148$	$[-0.0004, +0.0304]$	0.011	0.063	0.0220	10/10
B (confirmatory)	$+0.0143$	$[-0.0051, +0.0341]$	0.052	0.181	0.0275	10/10

Table 15: Quarter concentration and paired change for the feature-subset sensitivity. Top: the fold-9 (2025 Q2) contribution and the contrast with that quarter excluded. Bottom: paired daily change of a confirmatory contrast minus the subset contrast (positive when the confirmatory universe's L1−L0 exceeds the subset's); "paired MDE" is $2.8\times$ the paired block SE. p-values are nominal HLN at the Newey–West automatic lag and at lag 21.

Universe	Fold 9 (share)	Ex fold 9	95% CI	p (auto)	p (lag 21)
C5	$+0.0862$ (53%)	$+0.0069$	$[-0.0032, +0.0179]$	0.132	0.231
C	$+0.0785$ (44%)	$+0.0090$	$[-0.0052, +0.0228]$	0.125	0.248
B	$+0.0529$ (31%)	$+0.0108$	$[-0.0089, +0.0300]$	0.143	0.312
Paired change	**Paired MDE**	**Mean**	**95% CI**	p **(auto)**	p **(lag 21)**
C − C5	0.0245	$+0.0013$	$[-0.0159, +0.0189]$	0.838	0.882
B − C5	0.0276	$+0.0008$	$[-0.0186, +0.0200]$	0.915	0.938

C.7 Pre-evaluation feature re-selection sensitivity

We conducted a retrospective sensitivity analysis using feature re-selection whose scoring and grouping inputs were restricted to information through June 2022, conditional on the study's fixed stock panel, and re-tuned LightGBM and MLP under the existing evaluation protocol. The selector scored 168 columns (10 hand-crafted and 158 Alpha158, all lagged one day) by single-feature daily Spearman IC on 231 feature days from 2021-07-01 to 2022-05-31, whose labels end on 2022-06-30, the tuning training boundary. A column enters scoring only at a coverage of at least 0.50. The selector then scored each of the 61 Plan-AAA groups by the mean absolute time-mean IC of its members and kept the top 15 groups, whose member union is 48 columns (5 hand-crafted and 43 Alpha158). These 48 columns share 22 with Universe C's 51 and 4 with the 20-column C5 subset. Two qualifications apply. First, only the inputs are pre-evaluation. The rule form (marginal |IC| scoring, the 61-group partition, the top-15 cut, and the coverage threshold) comes from the single-feature-IC proxy of §D.4 and the test-informed Plan-AAA construction; we kept it for comparability and did not re-optimize it. Second, the top-15 cut is a near tie: rank 15 scores 0.03198 and rank 16 scores 0.03150, and all 60 rankable groups lie within one label-overlap-inflated standard error of the cut. The 48 columns are one draw from many near-tied sets.

We re-tuned both arms at 30 trials per arm under the frozen protocol (top-5 finalists at three seeds; training through 2022-06-30, validation on the second half of 2022). The L0 champion's validation

IC is $+0.0311$ and the L1 champion's $+0.0135$, every finalist in both arms is positive, and the L1 champion has one layer of 128 units (31,361 parameters). All 240 cells (ids 3600 to 3839) completed under the frozen hyperparameters (md5 `a8fdfb8f`). LightGBM is deterministic under frozen parameters, so its three tuning seeds coincide; these validation values are selection metrics and do not constitute independent validation. As in §C.6, this is a post-hoc sensitivity analysis: we open no multiple-testing family and report nominal p-values.

Table 16 places C-pre beside the confirmatory Universe-C and Universe-B rows and the post-hoc C5 subset, and Table 17 reports the paired daily contrast of each of those bases against C-pre on the same 749 days. The C-pre contrast is sensitive to fold 9 (2025 Q2), where LightGBM's IC is 0.204 and the MLP's 0.066 ($\Delta\text{IC} = -0.139$ from unrounded values), the most negative fold contribution, and the opposite sign from Universe C and C5 in that quarter. Excluding that fold gives $+0.0099$ $[-0.0069, +0.0265]$ ($p = 0.198$ at the automatic lag, 0.348 at lag 21). The full-period result is primary; the exclusion is a diagnostic whose series splices across the removed quarter.

On C-pre, the seed-averaged daily MLP$-$LightGBM contrast was -0.0024 (95% CI $[-0.0256, +0.0178]$; nominal HLN $p = 0.786$ / 0.847). The interval contains zero; this does not establish absence of a contrast, and $|\Delta\text{IC}|$ lies below the C-pre MDE of 0.0313. The positive pooled point estimate observed in C, C5 and B was not reproduced on C-pre. However, all three paired difference intervals contain zero, so the analysis does not establish that the underlying contrasts differ across these bases. The paired MDE for C $-$ C-pre (0.0433) exceeds both the observed change and the Universe-C contrast itself, so this non-rejection does not establish that the contrasts agree. Both models had low pooled IC point estimates on C-pre (LightGBM 0.0057, MLP 0.0033, intervals include 0), lower than on C and B; this descriptive pattern does not establish negligible predictive content or identify the cause. C-pre does not estimate how much leakage inflated C: it restricts only the selection and grouping inputs to information through 2022-06-30, conditional on the fixed stock panel and the existing protocol, and Universe B remains the leak-free anchor. We did not re-select the basis for the graph and edge contrasts and did not re-run them on C-pre.

Table 16: Pre-evaluation re-selection sensitivity of the L1$-$L0 contrast: the 48-column C-pre basis beside the confirmatory Universe-C (51 columns) and Universe-B (price-volume) rows and the post-hoc 20-column C5 subset. ΔIC is the seed-averaged daily MLP$-$LightGBM difference; the CI is a 21-day block bootstrap; HLN p-values are nominal and unadjusted, at the Newey–West automatic lag and at lag 21; IC L0 and IC L1 are the pooled per-arm point estimates (the C-pre per-arm intervals are $[-0.0230, +0.0376]$ for LightGBM and $[-0.0181, +0.0245]$ for the MLP); MDE is $2.8 \times \text{SE}_{\text{block}}$ (§C.5); "same sign" counts seeds whose per-seed contrast matches the sign of the seed mean; "LOSO" counts leave-one-seed-out replicates whose sign differs from the seed mean. All four contrasts fall below their MDE.

Basis	ΔIC	95% CI	p auto / lag 21	IC L0	IC L1	MDE	Same sign	LOSO
C-pre	-0.0024	$[-0.0256, +0.0178]$	0.786 / 0.847	0.0057	0.0033	0.0313	5/10	1/10
C	$+0.0148$	$[-0.0004, +0.0304]$	0.011 / 0.063	0.0195	0.0343	0.0220	10/10	0/10
B	$+0.0143$	$[-0.0051, +0.0341]$	0.052 / 0.181	0.0228	0.0371	0.0275	10/10	0/10
C5	$+0.0134$	$[+0.0008, +0.0283]$	0.008 / 0.054	0.0203	0.0337	0.0197	10/10	0/10

Table 17: Paired daily contrasts against C-pre on the same 749 days: the L1$-$L0 contrast of each comparison basis minus the C-pre contrast (positive when the comparison basis's MLP$-$LightGBM gap exceeds C-pre's); "paired MDE" is $2.8\times$ the paired block SE; p-values are nominal HLN at the Newey–West automatic lag and at lag 21. Every interval contains zero, and every mean lies below its paired MDE.

Paired change	Mean	95% CI	p (auto)	p (lag 21)	Paired MDE
C $-$ C-pre	$+0.0172$	$[-0.0111, +0.0494]$	0.102	0.271	0.0433
B $-$ C-pre	$+0.0167$	$[-0.0133, +0.0509]$	0.131	0.309	0.0447
C5 $-$ C-pre	$+0.0158$	$[-0.0095, +0.0481]$	0.131	0.292	0.0414

C.8 HAC LAG SENSITIVITY FOR ALL 20 CONTRASTS

The 21-day label overlap induces dependence in the daily loss differential beyond the automatic bandwidth ($L = 6$; Appendix A.6). Table 18 repeats the HLN test at a fixed lag of 21 for all 20 pre-registered contrasts. The automatic lag stays the pre-registered primary specification. The lag-21 column is a sensitivity from the same run, and the BH pass over it is post-hoc. Every p-value rises at lag 21. BH at $q = 0.05$ keeps 5 of the 11 rejections: L2−L1 in both universes, C L2s−L2, C L5−L2, and C L5−L3. It drops C L1−L0, L3−L2 in both universes, C L4−L2, C L6−L2, and C L7−L2. Three of the dropped contrasts (B L3−L2, C L6−L2, C L7−L2) stay below an unadjusted 0.05 and miss the BH step-up thresholds. L2−L1 is the only contrast that rejects in both universes at lag 21.

Table 18: HAC lag sensitivity for the 20 pre-registered DM/HLN contrasts. ΔIC and p (auto) repeat Table 3; p (lag 21) uses a fixed Bartlett bandwidth of 21. *: BH-FDR reject at $q = 0.05$ over the 20-test family, applied separately at each lag. The automatic lag is the pre-registered primary specification.

	Universe B			Universe C		
Pair (A−B)	ΔIC	p (auto)	p (lag 21)	ΔIC	p (auto)	p (lag 21)
L1−L0	+0.0143	0.052	0.181	+0.0148	0.011*	0.063
L2−L1	−0.0133	< .001*	0.011*	−0.0119	< .001*	< .001*
L6−L2	+0.0052	0.356	0.510	+0.0174	0.001*	0.018
L7−L2	−0.0091	0.195	0.403	+0.0102	0.002*	0.028
L2s−L2	+0.0079	0.118	0.305	+0.0089	< .001*	0.004*
L3−L2	−0.0149	0.006*	0.049	−0.0123	0.009*	0.070
L4−L2	−0.0081	0.227	0.392	+0.0163	0.012*	0.070
L5−L2	−0.0100	0.110	0.260	+0.0152	< .001*	0.009*
L5−L4	−0.0019	0.172	0.301	−0.0011	0.831	0.877
L5−L3	+0.0049	0.066	0.155	+0.0275	< .001*	< .001*

C.9 UNIVERSE-B COST LAYER AND THE L3−L2 COST PATH

The main-text crosswalk (Table 4) covers Universe C. Table 19 adds the same four contrasts in Universe B, from the same confirmatory cells and the same 12 paired fold differences. At 10 bps the Universe-B net ΔSharpe has the Universe-C sign for L1−L0, L2−L1, and L5−L3, and all four Universe-B intervals contain zero. We read these rows as descriptive: the signs agree, and the intervals are too wide to separate any of the four contrasts from zero. Mean realized turnover per rebalance in Universe B is 2.01 (L0), 2.61 (L1), 2.75 (L2), and 3.03 (L3).

The last row follows C L3−L2 across the cost grid. Its ΔSharpe is already positive at 0 bps, with an interval that contains zero, so the IC loss is absent from decile returns before any cost. The divergence therefore arises between full-cross-section IC and decile returns, and costs do not create it. Mean realized turnover is 2.91 for C/L2 and 3.19 for C/L3.

Table 19: Universe-B cost layer for the four crosswalk contrasts, and the C L3−L2 cost path. Cells give paired ΔSharpe at 0, 10, and 30 bps one-way cost, with a 95% stationary block-bootstrap CI over the 12 fold-level differences. "(r)" / "(n)": gross ΔIC BH-rejected / not rejected (Table 3). Descriptive, as in §5.3.

Pair	Gross ΔIC (BH)	ΔSharpe@0 bps [CI]	@10 bps [CI]	@30 bps [CI]
B L1−L0	+0.0143 (n)	+0.43 [−0.16, +0.99]	+0.43 [−0.14, +0.97]	+0.43 [−0.17, +1.01]
B L2−L1	−0.0133 (r)	−0.12 [−0.52, +0.33]	−0.14 [−0.53, +0.29]	−0.18 [−0.58, +0.25]
B L3−L2	−0.0149 (r)	−0.20 [−1.25, +0.82]	−0.35 [−1.37, +0.63]	−0.63 [−1.60, +0.28]
B L5−L3	+0.0049 (n)	+0.23 [−0.28, +0.81]	+0.30 [−0.22, +0.90]	+0.45 [−0.12, +1.09]
C L3−L2	−0.0123 (r)	+0.15 [−0.61, +0.82]	+0.08 [−0.77, +0.89]	+0.01 [−1.12, +1.21]

D SURVIVORSHIP AUDIT, STABILITY FAILURE, AND EXPLORATORY DIAGNOSTICS

The diagnostics in this appendix qualify the headline results and leave the BH-rejected contrasts and the SPA conclusion in place. The estimand is a survivor snapshot (§D.1), the dense-edge control arm's tuned configuration collapses (§D.2), the signal concentrates in a few test quarters (§D.3), and two exploratory failure modes sit outside the confirmatory families (§D.4).

D.1 SURVIVORSHIP AND COMPOSITION AUDIT

(L8) *Survivorship and non-point-in-time composition.* The universe is a fixed end-window survivor snapshot; it is not point-in-time (PIT) membership. A public-record constituent audit, in place of official S&P DJI or CRSP history, identifies two approximate deviations. First, 87 window members left before the snapshot date and are absent, equal to 14.8% of ever-member names and 8.2% of PIT member stock-days. Because reported results are within-universe paired contrasts on the same snapshot, this reduces mechanical cross-arm imbalance but does not make the estimand PIT. Pairing protects only cross-arm *differences*. Graph arms also lose the omitted firms as neighbors, so the snapshot may not be neutral between graph and non-graph arms, in a direction we cannot sign. Pairing does not protect the few *level* statements (per-arm IC excluding zero, Table 2; per-arm net Sharpe, §5.3), and the survivor-and-look-ahead universe inflates them, so we read them as within-sample descriptive levels that overstate PIT-realizable levels. Attrition is non-uniform, and omitted names include 2023 regional-bank failures and acquisition targets. Second, the snapshot over-includes current constituents before index-addition dates, creating 8.1% look-ahead stock-days and a two-sided composition mismatch of 16.3%. Paired contrasts do not remove this sampling bias. The mismatch has a liquidity bound (among 76 mid-window additions, the smallest market capitalization is \$6.6 B and none is below \$5 B) but no distress bound: SVB, Signature, and First Republic remain omitted.

Table 20 itemizes the audit over the window 2021-01-29 to 2026-01-28 (1,255 trading days), for the fixed price ∩ sector snapshot universe of 501 tickers. We reconstructed the audit from public records (the Wikipedia S&P 500 change history) only; because we did not consult official S&P Dow Jones or CRSP membership files, it may miss ticker-change, share-class, or correction events. Removed names are absent from the universe entirely (never in the price file); this differs from the within-universe exclusion of missing forward-return cells. On the audit's reading, both sides of the 16.3% mismatch sit at the index's size boundary: cap-change removals at the bottom edge, and look-ahead names freshly graduated into the index, all within S&P 500 liquidity scale. The residual is a boundary biased-sampling effect and involves no small-cap contamination; the distress-side omissions noted above remain.

D.2 TUNED-CONFIGURATION STABILITY FAILURE

The C/L5s stability failure leaves the Universe-C SPA conclusion (Hansen, 2005) unchanged under every handling of the degenerate cells; this subsection gives the cell-level breakdown, a candidate mechanism, and that sensitivity.

The Universe-C GraphSAGE-Mean control arm with all edges (C/L5s) degenerates to constant predictions in 27.5% of its test fold-seeds: 25 fully collapsed and 8 partially collapsed cells out of 120. This is not a training crash: affected cells report convergence but plateau at no signal, leaving Spearman IC mathematically undefined in the affected cells. A partially collapsed cell has undefined IC on some of its test days. We compute IC per seed and day, treat undefined seed-days as missing, and average the defined seeds, so the daily series averages fewer seeds on those days. The 33 affected cells come from seven of the ten seeds; three seeds never collapse in any fold, every fold keeps at least four seeds with a complete daily IC series, and fold 11 keeps all ten. The tuning study had already produced five undefined-objective trials at the same depth, dropout, and learning-rate range as the frozen configuration (§A.3).

A plausible but unproven mechanism is oversmoothing: GraphSAGE-Mean aggregation over a dense edge set, with tuned dropout 0.5, can smooth away cross-sectional variation. The failure occurs for GraphSAGE-Mean with all edges, whereas L2 is a GAT on the sparse correlation graph.

Table 20: Constituent-audit summary underlying Limitation L8. The audit is a public-record (Wikipedia-reconstructed) approximation of official S&P DJI or CRSP history. "PIT" = point-in-time; stock-days are equal-weight exposure proxies over the 1,255-day window.

Names and stock-days	
Fixed snapshot universe (price ∩ sector)	501 names
S&P 500 constituent changes in window	97 (87 distinct removed tickers)
PIT superset (snapshot ∪ removed-in-window)	588 names
Removed-in-window names absent from snapshot	87 / 588 (14.8%)
Reconstructed PIT membership stock-days	629,580
Missing stock-days (survivorship side)	51,841 (8.2%)
Look-ahead stock-days (pre-addition inclusion)	51,016 (8.1%)
Two-sided composition mismatch	(51,841 + 51,016) / 629,580 (16.3%)
Snapshot stock-days ($501 \times 1{,}255$)	628,755
Snapshot-name PIT stock-days	577,739
Exit reason among the 87 removed names	
Market-capitalization change (shrank out)	54
M&A / corporate action	28
Other / unspecified	5
Look-ahead side: 76 mid-window additions	
Minimum market capitalization	\$6.6 B
Median market capitalization	\$31.8 B
Additions below \$5 B	0

We treat undefined IC as missing instead of as a measured zero, and we do not retune only this losing arm. The main conclusion is unchanged: under exclusion, zero-fill, or zero-skill handling, C/L5s mean IC remains approximately zero and the Universe-C SPA p-value remains between 0.077 and 0.080.

D.3 Regime concentration: a per-fold view

Figure 5 disaggregates the pooled IC of the tuned ladder into its 12 quarterly test folds. A few quarters, especially 2024 Q4 and 2025 Q2, hold most of the positive signal. Roughly half of the test quarters are near zero or negative for almost every arm, so the pooled estimate is an average over a nonstationary sequence. Single-quarter Sharpe ratios are even more fragile because they use few non-overlapping rebalance returns. This concentration affects the pooled IC *levels* and leaves the sign of the four BH-rejected ladder contrasts unchanged. Sign stability is weaker than inferential stability: without 2025 Q2, C L1−L0 no longer rejects (Table 15). In a leave-one-quarter-out check, none of these contrasts (Benjamini & Hochberg, 1995) (L1−L0, L2−L1, L3−L2, L5−L3) changes sign when we remove any single quarter (0/12 in both universes). A leave-one-seed-out check mirrors this along the seed axis: no BH-rejected contrast changes sign when we remove any single seed (0/10 each). Per-seed signs agree in 8/10 for the two Universe-B penalties, 9/10 for their Universe-C counterparts, and 10/10 for C L1−L0 and C L5−L3.

Table 21 gives a complementary, purely descriptive per-quarter view from the untuned anchor runs (cell-mean IC averaged over four model classes and ten seeds). Signal recurs across roughly five strong quarters (2023 Q4, 2024 Q1, 2024 Q4, 2025 Q2, 2025 Q4), but about half of the quarters (2023 Q1–Q3, 2024 Q2, 2025 Q1) are approximately zero or negative. This is consistent with the reading that cross-sectional ranking pays off only in high-dispersion regimes. This table is descriptive and predates the tuned confirmatory ladder; we show it to document that the concentration pattern also appears outside the tuned configuration.

D.4 Exploratory failure modes

The analyses in this subsection sit outside the confirmatory families. In an earlier loss-function horse race (we later fixed MSE for the confirmatory ladder), ListMLE showed a regime-shift inversion: per-cell mean IC was -0.0458, compared with $+0.0113$ for MSE. A plausible explanation is that

Table 21: Per-quarter mean IC in the untuned anchor runs, averaged over four model classes and ten seeds (descriptive; not part of the confirmatory families).

23Q1	23Q2	23Q3	23Q4	24Q1	24Q2	24Q3	24Q4	25Q1	25Q2	25Q3	25Q4
−0.006	−0.029	−0.010	+0.056	+0.074	+0.002	+0.012	+0.084	−0.009	+0.095	+0.023	+0.034

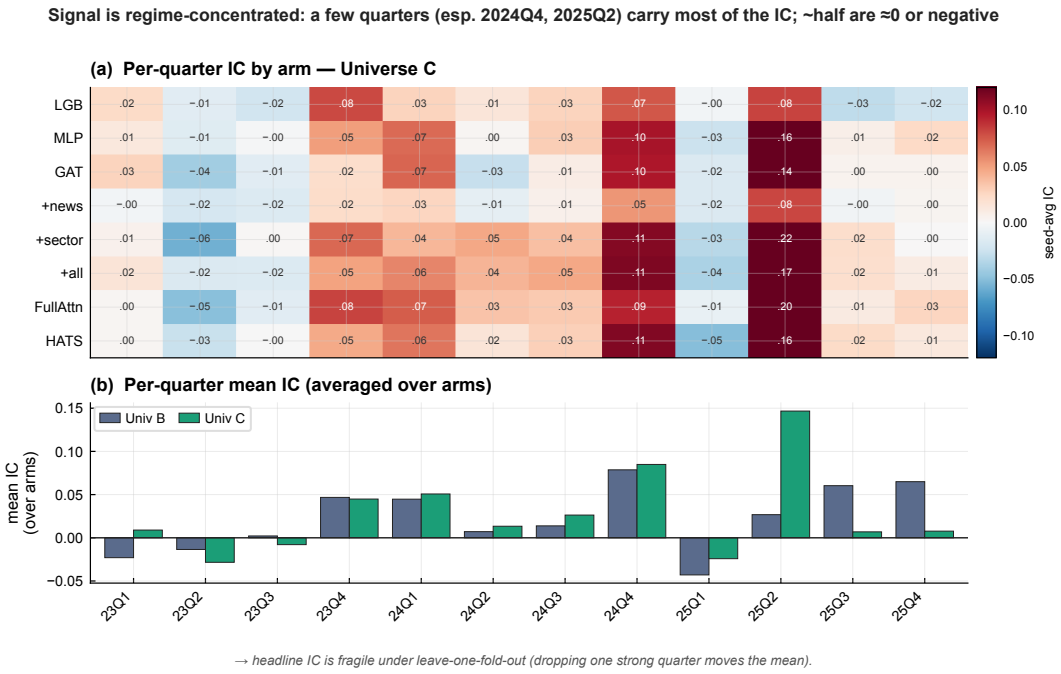

Figure 5: Per-fold IC by arm (descriptive companion to the pooled contrasts). (a) Heatmap of seed-averaged IC for the tuned ladder arms across the 12 walk-forward quarters in Universe C; (b) per-quarter mean IC averaged over arms for Universes B and C. We exclude degenerate C/L5s cells, matching the Family-1 handling (§D.2). The per-fold view explains why pooled IC levels are fragile under leave-one-quarter-out, whereas the BH-rejected pairwise contrasts do not change sign in any leave-one-quarter-out replicate.

the softmax likelihood overweights the in-sample rank order, which can invert under regime shift. Figure 6 visualizes this inversion.

That horse race used the same 501-stock panel and label, three-column and 158-column (Alpha158) feature sets, an MLP and GraphSAGE-Mean at fixed hyperparameters, and five expanding folds testing 2024 Q2 to 2025 Q2 with a 21-day purge. ListMLE is the per-day Plackett–Luce log-likelihood over the full cross-section, the pairwise arm a hinge loss; a 153-run pilot on one fold set per-loss learning rate and dropout. Stage 1 ran 3 losses, 2 models, 2 feature sets, 5 folds, 10 seeds: 600 cells. The pre-registered analysis (mixed model for ΔIC, per-fold block bootstrap for ΔSharpe, Bonferroni co-primary $\alpha/2 = 0.025$, BH-FDR, fold-4 leave-one-fold-out check) rejected 0 of 8 co-primary contrasts; pairwise per-cell mean IC was $+0.0007$. On fold 4, ListMLE's seed-averaged IC lay between -0.28 and -0.36 for every model and feature set; its fold-2 pilot validation IC ($+0.118$) became -0.053 at test. Follow-ups (rolling two-year training, a lagged-dispersion hypothesis) left the conclusion unchanged. That fold-4 window is 2025 Q2, fold 9 of the 12-fold design, which dominates the Universe-C, C5, and C-pre L1−L0 contrasts (§C.6, §C.7, §D.3); we draw no causal link.

A separate issue is Universe-C's feature-basis leakage. Its 51 columns (48 from Alpha158 and three hand-crafted) come from the Plan-AAA top-15 feature groups, ranked by neural-network permutation importance and under same-day open-high-low-close (OHLC) features. Only 5 of the 15 groups are also top-15 under a single-feature-IC proxy, and that proxy ranking is identical with and without

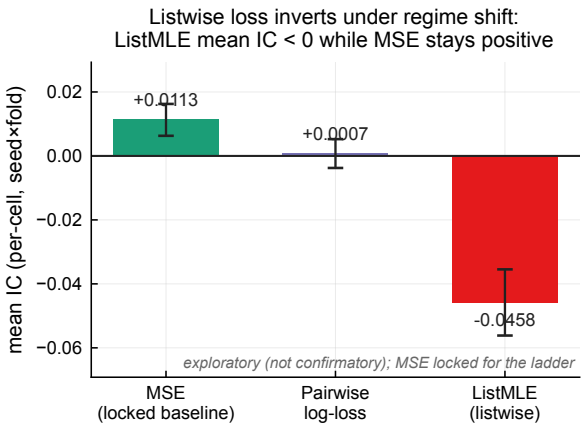

Figure 6: ListMLE inversion in the loss-function horse race (exploratory, outside the confirmatory families). Mean per-cell IC by training loss, with ± 1 standard error across cells. ListMLE's per-cell mean IC is -0.0458, against $+0.0113$ for MSE, the loss later fixed for the confirmatory ladder. A plausible explanation is that the softmax likelihood overweights the in-sample rank order, which can invert under regime shift.

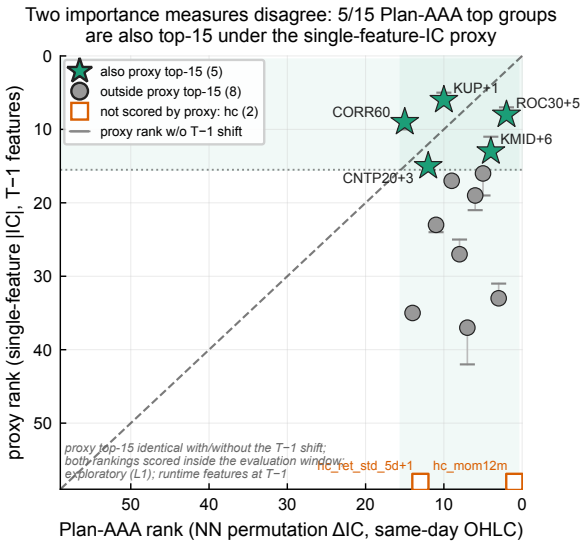

Figure 7: Plan-AAA feature-group ranking against a single-feature-IC proxy (exploratory, outside the confirmatory families). Original Plan-AAA rank (permutation importance, same-day OHLC) versus the proxy rank (single-feature $|\mathrm{IC}|$ computed on $T-1$ features); the proxy ranking is identical with and without the $T-1$ shift. Only 5 of the 15 groups that define Universe C's basis are also top-15 under the proxy, so the disagreement reflects the importance measure and not the lag; both rankings score quarters inside the evaluation window. The leak is in the basis *selection*, so Universe-C positive results are suggestive (Limitation L1; pre-evaluation re-selection sensitivity in §C.7).

the $T-1$ shift, so the disagreement reflects the importance measure and not the lag (Figure 7). Both rankings were scored inside the evaluation window: the proxy on the last 313 valid label days (2024-09-27 to 2025-12-26) and the Plan-AAA ranking on the five test quarters 2024-04-01 to 2025-06-30 (folds 5 to 9). Universe-C's positive results (MLP $>$ LightGBM, positive edge effects) are suggestive, while null and negative within-universe contrasts are less exposed (Limitation L1); a post-hoc sensitivity on the resulting 20-column subset appears in §C.6 and a pre-evaluation re-selection sensitivity in §C.7.