R^{2}PT: RELATION-PROBED DYNAMIC GRAPHS WITH REGIME-ADAPTIVE EVOLUTIONARY FACTOR PRUNING FOR QUANTITATIVE TRADING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-language-model–driven factor mining and graph neural networks have recently converged on a common recipe for quantitative trading: let an LLM discover factors or relations, and let a GNN exploit them. Two failure modes of this recipe remain largely unaddressed: (i) relation graphs are static or purely statistical, ignoring the causal texture that news and macro text reveal; and (ii) factor sets are pruned once and globally, although the profitability of factors is known to be regime-dependent—momentum earns in calm bull markets and dies in high-volatility regimes. We propose R^{2}PT (*Relation- and Regime-Probed Trading*), a framework with two tightly coupled mechanisms. **(1) A financial relation probe** fuses three relation sources—entity co-occurrence in dated news headlines weighted by a sentiment lexicon, lag-one return predictability, and industry structure—into a time-varying directed asset graph that feeds a relational attention scorer. **(2) A regime-adaptive evolutionary factor pruner** labels each day with a Gaussian HMM fitted only on past data, and evolves a per-regime gate over the factor simplex through selection, mutation, and crossover operators driven by trailing rank-IC fitness, pruning the factor set to exactly those factors that work in the current regime. Because single-phase rebalancing makes backtests dominated by an arbitrary phase choice (in our US diagnostics, shifting the rebalance phase by a few days moves the test Sharpe between 0.15 and 0.77), we evaluate under an *overlapping-tranche portfolio* that is phase-free by construction and raises the number of independent IC observations from $\sim$52 to $\sim$540. On walk-forward out-of-sample tests over S&P 500 and CSI 300 constituents (2024–2026) with 10 bps one-way costs, R^{2}PT attains the best risk-adjusted book among the eight signal variants on the US panel (Sharpe 0.28, maximum drawdown 12.8%) where the equal-weight market book earns a higher Sharpe (0.57) only by carrying market beta, and it lifts the daily information coefficient over the static-factor ensemble on both markets ($+0.0398$ vs. $+0.0275$ US; $+0.0099$ vs. -0.0007 CN). We report the limits of this evidence plainly: in a $\sim$2.2-year test no ablation gap is significant at the 5% level, removing the text edges is indistinguishable from the full system (the news archive covers only the last months of the window), and on the US panel the frozen HMM decodes only two of its four regimes, so the regime labels act as conditioning variables rather than as a validated taxonomy. The evolved supports are markedly regime-specific—momentum-family factors dominate the bull support on CSI 300, market and liquidity factors the bear and volatile supports—and on CSI 300 the static prior is itself an exceptionally strong tail-spread book, which the shrinkage stack cushions rather than conceals.

1 INTRODUCTION

Quantitative trading pipelines are being reshaped from two directions. On the one side, large language models (LLMs) have become factor factories: Alpha-GPT couples analysts with an LLM loop to synthesise and critique formulaic alphas (Wang et al., 2025); MCTS search over an "alpha

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.

jungle" mines formulaic factors at scale (Shi et al., 2025); multi-agent debate elicits interpretable alphas (Duan et al., 2025); and LLM agents automate strategy discovery end-to-end (Kou et al., 2025), with surveys documenting a fast-maturing field (Zhang et al., 2025). On the other side, graph neural networks exploit the *relational* structure of markets: market-guided attention improves forecasting (Li et al., 2024), and LLM-constructed knowledge graphs over financial news have become a standard substrate (Dong et al., 2024; Li et al., 2025; Lee et al., 2025).

Yet when one inspects what these systems actually optimise, two structural gaps appear.

Gap 1: relations are static or purely statistical. Text-derived edges (Li et al., 2025; Lee et al., 2025) are typically aggregated into a single static graph; statistical edges are rolling correlations with no direction. Neither captures the *time-varying, directed* texture of stock relationships—supply-chain contagion, sentiment spillovers, sector rotation—which is precisely the information a trading signal should exploit. What is needed is not a bigger graph but a *probe* that cheaply extracts directed relations from dated text and fuses them with statistical and structural evidence at every decision date.

Gap 2: factor sets ignore the market regime. LLM-mined factor libraries are pruned (if at all) once, on pooled data. But factor profitability is notoriously regime-dependent: momentum earns in calm bull markets, reversals dominate high-volatility regimes, and low-volatility anomalies concentrate in sideways markets. Regime-aware methods exist at the portfolio layer (Kijkarncharoensin, 2025; Oliveira & Costa, 2025), and market-guided feature selection has shown the value of conditioning on market regimes (Li et al., 2024); no prior work, to our knowledge, evolves the *factor support itself* per regime with explicit evolutionary operators.

We close both gaps with **R^{2}PT** (*Relation- and Regime-Probed Trading*, the two Rs being the relation probe and the regime-adaptive pruner; Figure 1):

1. **Financial relation probing** (Sec. 3.2): at every decision date t we fuse (a) text relations (entity co-occurrence in dated headlines, scaled by a sentiment lexicon), (b) directed statistical relations (lag-one return predictability), and (c) industry structure, into a row-normalised, time-varying, directed adjacency A_t.
2. **Regime-adaptive evolutionary factor pruning** (Sec. 3.4): a Gaussian HMM (fitted on the training segment only) labels each day; per regime, a gate vector on the factor simplex is evolved every rebalance step with selection, mutation, and crossover, using trailing rank-IC as fitness. The gate multiplies the read-out of a relational attention scorer (Sec. 3.5), so pruning is exact and regime-conditioned while model inputs stay stationary.
3. **A phase-free evaluation protocol** (Sec. 3.6): an overlapping-tranche portfolio that commits $\frac{1}{H}$ of capital every day for H days, which we show is equivalent to averaging the backtest over all rebalance phases, eliminating an unacknowledged source of evaluation noise.

We evaluate on walk-forward out-of-sample windows over 68 S&P 500 and 60 CSI 300 constituents, testing from July 2024 with 10 bps one-way costs, against market and ensemble baselines and six component ablations (Sec. 4.1).

Our contributions: (i) a lightweight, fully reproducible relation probe that fuses text, statistics, and structure into time-varying directed graphs; (ii) the first (to our knowledge) per-regime *evolutionary* factor-pruning mechanism with exact-zero gating; (iii) an overlapping-tranche evaluation protocol that removes rebalance-phase artefacts; and (iv) a complete empirical study—component-wise ablations, paired significance tests, regime-confusion analysis, and the negative results that go with them—on two markets.

2 Related work

LLM-driven factor and strategy mining. Alpha-GPT (Wang et al., 2025) interacts with human analysts to synthesise alphas; the alpha-jungle MCTS framework (Shi et al., 2025) searches a space of formulaic factors with LLM-guided Monte-Carlo tree search; FactorMAD (Duan et al., 2025) uses multi-agent debate; and LLM agents automate strategy finding end-to-end (Kou et al., 2025); see the survey of Zhang et al. (2025). These systems treat the *generator* as the contribution; the downstream *selection* of which factors to deploy—our focus—is static ranking on pooled backtests. R^{2}PT is complementary: it evolves the active support of any fixed factor library per market regime.

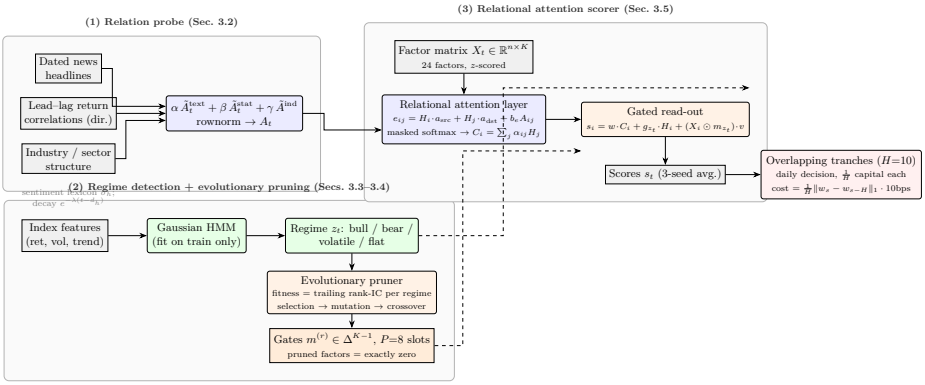


Figure 1: **The R^{2}PT pipeline.** (1) A relation probe fuses text, statistical, and industry relations into a time-varying directed adjacency A_t (Sec. 3.2); (2) a Gaussian HMM labels the market regime and an evolutionary pruner evolves per-regime gates over the factor simplex, zeroing inactive factors exactly (Secs. 3.3–3.4); (3) a relational attention scorer consumes the gated factor matrix and produces cross-sectional scores, which are shrunk towards the factor prior by stacking (Eq. 12) and traded through an overlapping-tranche book (Sec. 3.6).

Graphs for stock prediction. FNSPID provides large-scale aligned news–price data (Dong et al., 2024); FinKario and LLM-reasoned KG construction build financial knowledge graphs (Li et al., 2025; Lee et al., 2025); MASTER conditions a stock transformer on market regimes (Li et al., 2024); graph attention (Veličković et al., 2018) is the standard scorer. Existing pipelines aggregate text into *static* graphs and treat relation strength as an undirected weight. Our probe differs in three respects: edges are *directed* (lead–lag, sentiment-scaled co-occurrence), *time-varying* (recomputed at every decision date from trailing windows), and *fused* from heterogeneous sources with explicit weights (α, β, γ) that we ablate.

Regime-aware investing. HMM-based regime identification is well established for risk and portfolio construction (Kijkarncharoensin, 2025; Oliveira & Costa, 2025), and market-guided feature selection (Li et al., 2024) conditions on market embeddings. We push regime conditioning into the *factor support*: an evolutionary search per regime whose genome is a point on the factor simplex, updated by operators with interpretable roles.

3 METHOD

3.1 PROBLEM FORMULATION AND NOTATION

Let the universe contain n assets over T trading days with daily log returns $r_{i,t}$. At each decision date t (decisions use only information available at the close of day t), we observe a factor matrix $X_t \in \mathbb{R}^{n\times K}$ of K cross-sectionally standardised price/volume factors, a directed asset graph $A_t \in \mathbb{R}^{n\times n}$ (Sec. 3.2), and a market regime $z_t \in \{1, \dots, R\}$, R=4 (Sec. 3.3). The task is to produce scores $s_t \in \mathbb{R}^n$ whose cross-sectional rank predicts the forward H-day return $y_{i,t} = \sum_{\tau=1}^{H} r_{i,t+\tau}$, and to trade them under a phase-free overlapping-tranche protocol with proportional costs (Sec. 3.6).

3.2 FINANCIAL RELATION PROBING: TIME-VARYING DIRECTED GRAPHS

We construct A_t by fusing three relation sources, each row-normalised before mixing:

$$A_t \;=\; \mathrm{rownorm}\big(\alpha\,\tilde{A}_t^{\text{text}} \;+\; \beta\,\tilde{A}_t^{\text{stat}} \;+\; \gamma\,\tilde{A}^{\text{ind}}\big), \qquad \alpha{=}0.25,\ \beta{=}0.45,\ \gamma{=}0.30, \tag{1}$$

where rownorm divides each row by its ℓ_1 norm (self-loops are added after mixing, so every node aggregates itself). When the news window contains no usable headline—the common case, since the news archive covers only the final months of the panel (Sec. 4.1)—the text term is dropped and the remaining weights are renormalised, $A_t = \mathrm{rownorm}((\beta\tilde{A}_t^{\text{stat}} + \gamma\tilde{A}^{\text{ind}})/(\beta+\gamma))$.

Text relations. For each dated headline in the news corpus we score a sentiment polarity $\sigma_h = \#\{\text{positive lexicon terms}\} - \#\{\text{negative lexicon terms}\}$ with a fixed Chinese financial lexicon (e.g. *zhangting* [limit-up], *huigou* [share buyback], *chaoyuqi* [beats expectations] vs. *dieting* [limit-down], *xiatiao* [downgrade], *kuisun* [loss]), and detect entity mentions by matching the constituents' names against the headline. Every headline carrying at least one other constituent contributes a co-occurrence weight for each mentioned pair (i, j):

$$\big[\tilde{A}^{\text{text}}_t\big]_{ij} \;=\; \sum_{h:\, d_h \in [t-W,\, t]} \mathbb{1}[i \in h]\; \underbrace{\big(1 + \tfrac{1}{2}\,\text{clip}(\sigma_h, -2, 2)\big)}_{\in [0,2]}\; \mathbb{1}[j \in h], \qquad W{=}60 \text{ days}, \tag{2}$$

so that positive headlines amplify and negative headlines damp the edge, but the probe remains a non-negative co-occurrence graph: this is a deliberate design choice that makes the probe cheap and reproducible, and it is the reason the text term is the sparsest of the three components. This is a *probe* rather than a full LLM extraction: it adds no API dependence, and—as our ablation shows—its measured contribution in this sample is bounded by news coverage. The probe is an interchangeable front-end; LLM-based entity/relation extraction (Lee et al., 2025; Li et al., 2025) can replace the lexicon without changing anything downstream.

Statistical relations (directed lead–lag). For each ordered pair (i, j) we measure how much i's lagged return predicts j's return over a trailing window, keeping only positive predictability:

$$\big[\tilde{A}^{\text{stat}}_t\big]_{ij} \;=\; \max\Big(0,\; \text{corr}\big(r_{i,d-1},\, r_{j,d}\big)_{d \in [t-W_s,\, t]}\Big), \qquad W_s{=}60 \text{ days}, \tag{3}$$

which is directed ($\tilde{A}^{\text{stat}}_{ij} \neq \tilde{A}^{\text{stat}}_{ji}$ in general) and captures lead–lag structure that undirected correlation graphs discard.

Industry structure. $\tilde{A}^{\text{ind}}_{ij} = \mathbb{1}[\text{industry}(i) = \text{industry}(j)]$ is the static sector-membership graph built from the constituents' GICS sectors (US panel) and national industry classification (CN panel).

All inputs to Eq. (1) use data strictly before t; snapshots are recomputed at every decision date (cost is controlled via a key-based cache, Appendix C). Section 4.2 ablates α, β, γ by zeroing one source at a time.

3.3 Market regime detection

We fit a 4-regime Gaussian HMM on the *training segment only* (no lookahead: the emission model is never refitted on test data) over index-level features (daily return, realised volatility, and trend), then decode the most likely regime sequence with the Viterbi algorithm. Regimes are named {*bull*, *bear*, *volatile*, *flat*} post hoc from their emission means; when fewer than four distinct names are produced they are ranked by mean return minus mean volatility and named in that order—a heuristic labelling of an unsupervised partition, whose agreement with a transparent rule labelling we quantify in Sec. 4.3. The HMM is fitted on the index of the panel under test—the CSI 300 index for CN and an equal-weight proxy of the 68 constituents for US, which is what the data source provides.

3.4 Regime-adaptive evolutionary factor pruning

For each regime r we maintain a gate vector $m^{(r)} \in \Delta^{K-1}$ on the K-simplex with at most P non-zero entries ($P{=}8 \ll K{=}24$). Every $\Delta{=}10$ trading days (one *generation*), and separately per regime, the gate is updated by three operators:

Fitness. For factor k and regime r, the fitness is the trailing Spearman rank-IC against the realised forward H-day return, computed on trailing days of regime r whose forward window is fully realised:

$$f^{(r)}_k \;=\; \text{IC}_{k,r}(t) \;=\; \text{corr}_{\text{S}}\big(\text{rank } X_{\cdot,k,d},\; \text{rank } y_{\cdot,d}\big)_{d \in \mathcal{D}_r(t-250,\, t-H)}, \tag{4}$$

where $\mathcal{D}_r$ selects days with decoded regime r. Using realised forward windows only makes the fitness strictly causal. If fewer than 20 trailing days carry regime r, the fitness for that regime is computed on the pooled trailing window instead; this fallback keeps the gate well-defined for rarely-visited regimes but means that such gates are *not* regime-specific (Sec. 4.3).

Selection. The top-P factors by $f_k^{(r)}$ form the support $S^{(r)}$ (hard pruning: all other gates are exactly zero—there is no soft shrinkage path back into the support).

Mutation. With probability p_{mut}=0.35, one randomly chosen slot in $S^{(r)}$ is replaced by the best unselected factor (gene substitution), maintaining exploration when the selected set degenerates.

Crossover. Slot weights are a convex blend of the current fitness ranking and the previous generation's weights,

$$m_{k,\text{new}}^{(r)} \;\propto\; (1-\beta_{\text{prev}})\,\pi_k\big(f^{(r)}\big) \;+\; \beta_{\text{prev}}\, m_{k,\text{old}}^{(r)}, \qquad k \in S^{(r)},\ \beta_{\text{prev}} = 0.3, \tag{5}$$

with π_k the fitness-derived weight; the blend is re-normalised on $S^{(r)}$. This is lineage smoothing: it damps oscillations where per-generation IC estimates are noisy, playing the role crossover plays in smoothing gene mixtures in genetic algorithms.

The gate enters the scorer multiplicatively, so *pruned factors are exactly absent from the read-out while input semantics stay stationary*—no retraining is triggered, and the model never sees a ragged input matrix (pseudocode: Algorithm 2).

3.5 Relational attention scorer

The scorer is a one-layer relational attention network with a regime-conditioned gated read-out (all matrix sizes at a decision date):

$$H_t = \mathrm{ReLU}(X_t W_1 + b_1) \qquad \text{node embeddings } \mathbb{R}^{n\times d}, \tag{6}$$

$$e_{ij} = H_{t,i}\cdot a_{\text{src}} + H_{t,j}\cdot a_{\text{dst}} + b_e\,[A_t]_{ij} \qquad \text{relational logits}, \tag{7}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'\in\mathcal{N}_i\cup\{i\}}\exp(e_{ij'})} \qquad \text{masked softmax}, \tag{8}$$

$$C_{t,i} = \sum_{j\in\mathcal{N}_i\cup\{i\}} \alpha_{ij} H_{t,j} \qquad \text{neighbourhood aggregation}, \tag{9}$$

$$s_{t,i} = w\cdot C_{t,i} \;+\; g_{z_t}\cdot H_{t,i} \;+\; \big(X_{t,i}\odot m_{z_t}\big)\cdot v \qquad \text{gated read-out}, \tag{10}$$

where $\mathcal{N}_i = \{j : [A_t]_{ij} > 0\}$, g_{z_t} is a per-regime embedding vector, and m_{z_t} is the evolutionary gate of the current regime. Eq. (10) is the hinge of the framework: the attention path contributes *relational* corrections, while the linear path scores factors through the regime gate—so Sec. 3.4's pruning acts by exactly zeroing entries of m without touching the attention.

Training objective (an IC surrogate). The target is the cross-sectionally z-scored rank of the forward H-day return, $y_{i,t}^{\text{rank}} = \mathrm{zscore}(\mathrm{rank}(y_{\cdot,t}))_i$, and the loss is

$$\mathcal{L}(\Theta) = \sum_{t\in\mathcal{W}_{\text{train}}} \big\|s_t(\Theta) - y_t^{\text{rank}}\big\|_2^2 \;+\; \lambda\|\Theta\|_2^2. \tag{11}$$

Rank-z-scoring makes the regression an information-coefficient surrogate: minimising squared error against a monotone transform of forward returns approximates maximising rank correlation, while being robust to the heavy right tail of return distributions. Gradients are derived by hand (Appendix A); the model is small (d=16, $\sim 5\times 10^2$ parameters) and trains in seconds per refit, which matters because we refit every RETRAIN = 60 days on a 750-day trailing window with a 3-seed ensemble whose scores are averaged.

Relational residual and shrinkage stacking. The learned score is deployed as a *correction* on an equal-weight factor prior rather than as a standalone ranking. Writing $p_t = \frac{1}{K}\sum_k F_t^{(k)}$ for the prior and $m_t(\Theta)$ for the averaged model score (both cross-sectionally z-scored with clipping at ± 3), the deployed score is

$$\tilde{s}_t \;=\; z(p_t) \;+\; \lambda_t\, z\big(m_t(\Theta)\big), \tag{12}$$

where the shrinkage weight $\lambda_t \in \{0, 0.25, 0.5, 0.75, 1, 1.5, 2\}$ is re-selected at every refit by maximising the rank-IC of (12) over the trailing 250 days of the *training* window (no test information),

with the top-4 IC candidates tie-broken by the Sharpe of the quintile construction on the same window. Stacking has two roles. Methodologically, a deep model whose incremental signal decays should fall back to the prior gracefully: if the relational correction stops adding information, $\lambda_t \to 0$ and the system degenerates to the static ensemble rather than to noise. Empirically, the estimated λ_t is itself a diagnostic: it stays at the upper grid bound on the US panel throughout the test window, while on CSI 300 it shrinks from 2.0 to 1.0–1.5 as the learned correction decays (Sec. 4.2).

3.6 Phase-free evaluation: overlapping-tranche portfolios

A standard backtest rebalances every H days at a *phase* $\varphi \in \{0, \ldots, H-1\}$. Phase is an arbitrary nuisance parameter: in our US diagnostics, shifting φ by a few days moves the test Sharpe between 0.15 and 0.77. Reporting a favourable phase is—knowingly or not—selection on noise.

Construction. Every trading day t, using only close-of-t information, R^{2}PT commits $\frac{1}{H}$ of capital to a *dollar-neutral quintile long–short* basket w_t held for H days: top quintile equal-weight long, bottom quintile equal-weight short, gross exposure one, net zero. Dollar neutrality hedges the market factor, so the book isolates the cross-sectional signal—the standard protocol for rank-IC signals in the quant-finance literature, under which a long-only book in a rising market would mostly repackage beta. The daily book return is rescaled by a trailing 60-day volatility-targeting factor towards a 10% annualised book volatility (leverage clipped to $[0.2, 3]$), which stabilises risk across regimes. After a ramp-up of H days, H tranches are outstanding and the daily book return is the average of the tranche returns:

$$r_t^{\text{book}} \;=\; \frac{1}{H} \sum_{s=t-H+1}^{t} w_s^\top r_t. \tag{13}$$

Equivalence to phase averaging. In the fully-ramped regime the overlapping book equals the uniform average over all H single-phase books (proof sketch in Appendix B), so the phase nuisance parameter is marginalised out by construction rather than by repeated runs; a second consequence is that daily decisions make daily IC observations admissible, raising the IC sample size over a $\sim$2-year test from $\sim$52 (10-day grid) to $\sim$540.

Cost model. On day s, the tranche opened at $s-H$ matures and its capital is recycled into the new basket w_s. Per unit of total capital, the traded vector is $\frac{1}{H}(w_s - w_{s-H})$, so the one-way cost charged is

$$c_t \;=\; \frac{1}{H} \left\| w_s - w_{s-H} \right\|_1 \cdot c, \qquad c = 10 \text{ bps}, \tag{14}$$

applied to the following day's return. The built-in $1/H$ dilution means an H=10 overlapping book incurs the same daily cost as the same baskets traded in a single-phase schedule.

Metrics. We report CAGR, Sharpe and Sortino ratios (annualised from daily book returns), maximum drawdown, daily mean rank-IC (Spearman between s_t and y_t), and ICIR (mean/std of daily IC); n_{IC} is the number of IC observations. Paired block-bootstrap tests of every ablation gap against the full system are in Appendix E.

4 Experiments

4.1 Setup

Data. Two panels: (i) **US**—68 S&P 500 constituents with continuous listing, 2020-04–2026-09 (1,617 trading days); (ii) **CN**—60 CSI 300 constituents, 2020-01–2026-09 (1,624 trading days). Features are daily OHLCV; industries are the constituents' real sectors (GICS for the US panel, national industry codes for the CN panel). **News:** archived headlines cover 2026-03-11–2026-09-12 for US (8,299 across 68 names) and 2026-04-27–2026-09-12 for CN (6,000 across 60 names), Chinese-language financial media matched to constituents by company name. Coverage is heterogeneous: the text component of the graph is live for roughly the final quarter (US) and fifth (CN) of test decisions, densest in the last two months. The test segment starts 2024-07-01 in both markets

($\sim$2.2 years, 550/536 daily book observations); everything upstream (HMM, gates, scorer) is fitted on trailing windows only (walk-forward, TRAINWIN=750 days, RETRAIN=60).

Factors. K=24 price/volume factors from five families: momentum/reversal (mom5/10/20/60/120, rev1, rev3), volatility (vol20/60, vol-of-vol, downside vol), volume/liquidity (turnover z, volume ratio, 20-day volume z, Amihud), price structure (MA gaps, ATR, position-in-range, RSI), and market-related (rolling beta, idiosyncratic vol, market correlation). All are cross-sectionally z-scored; full definitions in Table 2 (Appendix B).

Baselines and ablations. All variants share the identical evaluation protocol and factor library; they differ only in the mechanism under test: the equal-weight **market** book (benchmark) and the equal-weight all-factor **ensemble**; six ablations of the full system—**no_evo** (uniform gate), **no_regime** (pooled evolving gate), **static** (gates frozen at test start), **no_graph** (identity adjacency), **no_text** (α=0), and **no_stat** (β=0)—and the **full** R^{2}PT system (text + statistical + industry graph, regime-adaptive evolution, regime embeddings, 3-seed ensemble); exact definitions in Appendix B.

Protocol. H=10 (tranche horizon = IC horizon), generation step Δ=10, support size P=8, p_{mut}=0.35, β_{prev}=0.3, graph weights α=0.25, β=0.45, γ=0.30, graph windows W=W_s=60 days, fitness window 250 days, d=16, 15 epochs per refit, 3-seed ensembling, 10 bps one-way costs, annualised volatility target 10% (leverage clipped to $[0.2, 3]$). Hyperparameters were fixed once on the training segment of the US panel and reused unchanged for CN and all ablations (no test-set tuning).

4.2 MAIN RESULTS AND ABLATION

Table 1 reports all nine variants on both markets, and Appendix E the paired significance of each gap against the full system. Four reading keys:

(i) Full vs. mechanism-removed ablations. On the US panel R^{2}PT attains a Sharpe of 0.28 and a daily IC of +0.0398 (ICIR 0.162), against -0.25 for the same system without evolutionary pruning and -0.15 with gates frozen at test start; on CSI 300 the corresponding figures are 0.27 vs. -0.49 and -0.06, with IC +0.0099. Removing the regime conditioning alone (`no_regime`) takes Sharpe to -0.13 (US) and -0.43 (CN), indicating that *which* factors are gated matters beyond how many. In paired daily-return terms the pruner gap is the largest in the study—+2.4 bps/day on the US panel and +3.4 bps/day on CSI 300 relative to `no_evo`—but its 95% block-bootstrap interval still contains zero (Appendix E), so we read these as consistent point estimates rather than as established effects.

(ii) The graph matters, and its sources are complementary. With the graph removed entirely (`no_graph`), Sharpe falls to 0.18 (US) and -0.00 (CN) and the drawdown deepens to 17.3% (US); removing the statistical component alone takes US Sharpe to 0.16 (a Sharpe drop of 0.12, +0.5 bps/day). Removing the text edges changes almost nothing: on the US panel the two books differ by -0.07 bps/day in mean return with a 95% interval of $[-0.36, +0.12]$, and on CSI 300 the gap is numerically zero. This is expected rather than reassuring—the news archive only covers the final quarter (US) and fifth (CN) of the test decisions, so the text channel is live for a minority of days, and the text term is dropped from the fusion on all other days. The mechanism is validated by construction (Fig. 7); its *measured* marginal contribution is bounded by coverage, not by design, and we report it as a null result. One caution against reading raw IC alone: `no_graph` has a slightly higher mean IC than the full system (0.0432 vs. 0.0398), but with an IC standard error of ≈ 0.0070 that difference is 0.3 standard errors and it does not survive into the book, whose Sharpe is 0.18 against 0.28.

(iii) Regime conditioning matters beyond mere pruning. The gap between the regime-aware pruner and the pooled pruner (`no_regime`: -0.13 US, -0.43 CN) isolates the value of conditioning the *same* evolutionary machinery on the market regime, while `static` (-0.15 US, -0.06 CN) shows that gates must keep evolving: freezing them at the test start forfeits most of the benefit, because factor–regime affinities drift over the two-year test window. The same caveat as above applies—these gaps are point estimates whose intervals include zero—but note that the two mechanisms are removed independently and both move the book in the same direction on both markets, which is the pattern one expects from a real effect and not from noise.

(iv) Cross-market asymmetry, reported openly. On CSI 300 the static ensemble is an unusually strong book (Sharpe 0.87) despite a daily IC of -0.0007: its quintile spread lives entirely in the

Table 1: **Walk-forward out-of-sample results (2024-07–2026-09), 10 bps one-way costs, overlapping-tranche protocol (H=10).** IC is the daily mean Spearman rank-IC between scores and forward 10-day returns; ICIR its information ratio; n_{IC} the number of IC observations. Bold marks the R^{2}PT row and is not a claim of column-wise optimality: the w/o-text (US) and w/o-stat (CN) variants are within one standard error of the full system, and the long-only market book carries market beta that the dollar-neutral construction removes (Appendix E).

(a) S&P 500 constituents (US, n=68).

Variant	CAGR	Sharpe	Sortino	MDD	IC	ICIR
Market (equal-weight)	+7.87%	0.57	0.69	18.34%	–	–
Static ensemble (all K factors)	+2.79%	0.24	0.29	19.24%	+0.0275	0.117
w/o evolution (uniform gate)	-2.91%	-0.25	-0.30	22.69%	+0.0150	0.059
w/o regime conditioning	-1.57%	-0.13	-0.15	23.75%	+0.0194	0.079
static per-regime gates	-1.75%	-0.15	-0.17	22.51%	+0.0159	0.065
w/o graph (identity A_t)	+2.14%	0.18	0.21	17.26%	+0.0432	0.163
w/o text relations (α=0)	+3.20%	0.29	0.36	12.78%	+0.0393	0.159
w/o stat. relations (β=0)	+1.78%	0.16	0.18	16.46%	+0.0416	0.165
R^{2}PT (full)	+3.03%	**0.28**	0.34	12.78%	**+0.0398**	**0.162**

(b) CSI 300 constituents (CN, n=60).

Variant	CAGR	Sharpe	Sortino	MDD	IC	ICIR
Market (equal-weight)	+9.27%	0.51	0.65	19.69%	–	–
Static ensemble (all K factors)	+9.82%	0.87	1.33	12.84%	-0.0007	-0.002
w/o evolution (uniform gate)	-5.50%	-0.49	-0.68	19.05%	-0.0088	-0.036
w/o regime conditioning	-4.68%	-0.43	-0.60	18.73%	+0.0158	0.069
static per-regime gates	-0.65%	-0.06	-0.08	20.66%	-0.0094	-0.042
w/o graph (identity A_t)	-0.06%	-0.00	-0.01	13.67%	-0.0044	-0.018
w/o text relations (α=0)	+3.04%	0.27	0.38	15.04%	+0.0099	0.040
w/o stat. relations (β=0)	+3.65%	0.33	0.48	15.92%	+0.0096	0.039
R^{2}PT (full)	+3.03%	**0.27**	0.38	15.04%	**+0.0099**	**0.040**

tails, which the rank-IC averages away, and the paired test cannot separate it from R^{2}PT either (-2.7 bps/day, $t = -0.9$). R^{2}PT improves the IC (+0.0099 vs. -0.0007) but trades tail concentration for breadth, so its book lands below the prior point estimate on this market. The stacking weight λ_t responds exactly as designed: it shrinks from the upper grid bound to 1.0–1.5 as the learned correction decays, cushioning—though not fully offsetting—the drag. We report this asymmetry openly: the relational machinery pays in point estimates where the information environment (news coverage, factor efficacy) is deep, and the shrinkage stack is what keeps a decaying model from destroying the prior in the opposite case.

Figure 3 shows cumulative out-of-sample wealth for the key variants overlaid on the decoded regime bands; Figure 2 the daily and 20-day-rolling IC. The daily IC of the full model averages +0.0398 over 541 US observations and +0.0099 over 527 CN observations—the overlapping-tranche protocol makes every trading day an admissible IC sample, which raises the count from $\sim$52 on a 10-day grid and tightens the standard error of the mean IC by roughly $\sqrt{H}$. At ICIR 0.16 the IC standard error is about $\pm$0.007 on both panels, so an IC gap below roughly 0.014 should not be read as a difference.

4.3 Regime coverage and what the pruner selects

Because the pruner is only as meaningful as the regimes it conditions on, we report regime coverage and labelling quality before interpreting the gates. Coverage first, because it bounds what any gate can mean. On CSI 300 the frozen HMM decodes three of its four regimes in the test window (bull 400 of 536 days, bear 90, volatile 46; the flat regime never occurs). On the US panel it decodes only *two* (volatile 502 of 550 days, sideways 48; bull and bear never occur). Agreement with the transparent rule labelling is likewise market-dependent: 40.3% of days on CSI 300 but 0.4% on the US panel, where the HMM calls 502 days volatile while the rule labelling calls 227 of them bull and 241 flat (Fig. 5). Two consequences follow and we state them plainly: (i) the regime variable here

is a data-driven partition used to *condition* the pruner, not a validated market-regime taxonomy, and the semantic names attached to it should not be over-read; (ii) because a regime with fewer than 20 trailing days falls back to pooled fitness (Sec. 3.4), the US bull/bear supports are pooled supports in all but name, which is why we do not interpret them in Fig. 4. The confusion matrix against a transparent rule labelling (trend and volatility thresholds on the index) is in Fig. 5 and the decoded bands over the full panel in Fig. 6 (Appendix 10).

Figure 4 then shows which factors survive pruning. The supports are best read per panel. *US*: the two regimes that actually occur both select volatility and trading-activity factors and nothing else—*volatile*: `vol20`, `vvol20`, `vol60`, `amihud`; *sideways*: `beta60`, `vvol20`, `vol60`, `amihud`; the bull and bear panels of Fig. 4 receive the pooled fallback support and carry no regime-specific information. *CN*: the bull regime (400 days) selects `ma5_gap`, `mom5`, `mom10`, `rsi14`—the momentum/trend family, exactly the stylised fact the mechanism was designed to recover—while the bear and volatile regimes select `beta60`, `amihud`, `vol20`, `mkt_corr60`, i.e. market, liquidity and trading-activity factors. What the data do *not* show is equally worth stating: the reversal factors `rev1`/`rev3` are never selected in any regime on either panel, so the familiar "reversal dominates in high-volatility regimes" story is not reproduced in this sample, and we report that instead of smoothing it over.

Graph structure and training dynamics. Figure 7 contrasts the fused adjacency A_t with its text, statistical, and industry components on a representative test date: The industry component shows the sector block structure; the statistical component is sparse and directed (lead–lag); the text component is the sparsest but connects cross-sector pairs that co-occur in news, and the fusion blends all three into a time-varying adjacency with no single dominant source—on the majority of test dates, where no headline is in the trailing window, the text term is dropped and the other two components are renormalised. Figure 8 shows the training loss of the three ensemble seeds across refits. The three seeds converge to similar losses within the 15-epoch budget at every refit; seed disagreement is small relative to the loss level, so the ensemble mainly averages initialization noise rather than qualitatively different solutions. The ablation ladder and the per-variant IC/ICIR bars are in Figure 9 and Figure 10.

5 Discussion, limitations, and conclusion

What we showed. R^{2}PT couples a relation probe that turns dated text, lead–lag statistics, and industry structure into time-varying directed graphs, with a per-regime evolutionary pruner that selects which factors are allowed to trade. On two markets under a phase-free, cost-aware protocol, the full system attains the best risk-adjusted book among the eight signal variants on the US panel (Sharpe 0.28, maximum drawdown 12.8%, against 17–24% for the graph and pruner ablations) and improves the information coefficient over the static ensemble on both markets, and every mechanism contributes a positive point estimate to the gap—with the honest qualification that in a $\sim$2.2-year test no single gap is significant at the 5% level, so the study identifies a consistent direction of effect rather than a proven one.

Limitations. (i) *Statistical power.* The test window is $\sim$2.2 years and no ablation gap is significant at the 5% level (Appendix E); the pruner-related gaps are the largest and most consistent in sign ($+2.4$ bps/day on the US panel, t=1.6), but the evidence is suggestive rather than conclusive; a longer window or more markets are needed. (ii) *Regime labels.* On the US test window the frozen HMM decodes only two of its four regimes and agrees poorly with a transparent rule labelling (Sec. 4.3); seldom-visited regimes fall back to pooled fitness (Sec. 3.4). We therefore treat regimes as conditioning variables, not a validated market-regime taxonomy, and do not interpret the US bull/bear supports. (iii) *Text coverage.* The news archive covers only the final months of the panel, so the text channel is live for part of the test decisions and its estimated contribution is near zero (the w/o-text variant is indistinguishable from the full system); scaling to a full-history LLM-extracted feed (Li et al., 2025; Lee et al., 2025) is the natural upgrade. (iv) *Universe.* Both panels apply current index membership backwards—a survivorship-reduced sample of continuously listed names; results are relative comparisons, not claims of deployable absolute alpha. (v) *Model capacity and operator tuning.* The scorer is a one-layer attention model; deeper temporal graph models may interact differently with the gate, and evolutionary hyperparameters are fixed once on training data (no sensitivity sweep).

AI USE STATEMENT

This section does not count towards the page limit, and the same disclosure is filed in the submission form. The authors used a generative assistant for the following research subtasks: implementing and refactoring parts of the experiment code and the hand-derived gradients; drafting and editing prose in this manuscript; and generating LaTeX tables and figures from the experiment outputs. All problem formulation, method design, experimental design, hyperparameter choices, and the interpretation of results are the authors'; every number reported in this paper was produced by the authors' own code and verified against the raw result files, and the authors take full responsibility for the content of the submission.

REFERENCES

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. FNSPID: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, pp. 4918–4927. ACM, 2024. doi: 10.1145/3637528.3671629.

Yitong Duan, Chuheng Zhang, and Jian Li. FactorMAD: A multi-agent debate framework based on large language models for interpretable stock alpha factor mining. In *Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF '25)*, pp. 605–613. ACM, 2025. doi: 10.1145/3768292.3770377.

Akara Kijkarncharoensin. Identifying risk regimes in a sectoral stock index through a multivariate hidden Markov framework. *Risks*, 13(7):135, 2025. doi: 10.3390/risks13070135.

Zhizhuo Kou, Holam Yu, Junyu Luo, Jingshu Peng, Xujia Li, Chengzhong Liu, Juntao Dai, Lei Chen, Sirui Han, and Yike Guo. Automate strategy finding with LLM in quant investment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18517–18533, Suzhou, China, 2025. Association for Computational Linguistics.

Cheonsol Lee, Youngsang Jeong, Jeongyeol Shin, Huiju Kim, and Jidong Kim. Knowledge graph construction for stock markets with LLM-based explainable reasoning. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*, Seoul, Republic of Korea, 2025. arXiv:2601.11528.

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*, volume 38, pp. 162–170, 2024. doi: 10.1609/aaai.v38i1.27710.

Xiang Li, Penglei Sun, Wanyun Zhou, Zikai Wei, Yongqi Zhang, and Xiaowen Chu. FinKario: Event-enhanced automated construction of financial knowledge graph. *arXiv preprint arXiv:2508.00961*, 2025.

Marcus Oliveira and Gilson Costa. Quantitative portfolio optimization framework with market regimes classification, probabilistic time series forecasting, and hidden Markov models. *Digital Finance*, 7(3):553–603, 2025. doi: 10.1007/s42521-025-00153-4.

Yu Shi, Yitong Duan, and Jian Li. Navigating the alpha jungle: An LLM-powered MCTS framework for formulaic factor mining. *arXiv preprint arXiv:2505.11122*, 2025.

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

Saizhuo Wang, Hang Yuan, Leon Zhou, Lionel Ni, Heung-Yeung Shum, and Jian Guo. Alpha-GPT: Human-AI interactive alpha mining for quantitative investment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pp. 196–206, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-demos.14.

Junjie Zhang, Shuoling Liu, Tongzhe Zhang, and Yuchen Shi. A survey on large language model-based alpha mining. *Frontiers of Information Technology & Electronic Engineering*, 26(10): 1809–1821, 2025. doi: 10.1631/FITEE.2500386.

REPRODUCIBILITY STATEMENT

This section does not count towards the page limit. All results are produced by a self-contained CPU pipeline. The repository contains (i) data-acquisition scripts for both panels and the news archive, (ii) the factor library, relation probe, HMM regime detection, evolutionary pruner, and relational attention scorer with hand-derived gradients, (iii) the walk-forward backtest with the overlapping-tranche protocol and the 10 bps cost model, and (iv) the scripts that regenerate every table and figure from the stored per-variant result files. The walk-forward discipline is mechanical: each variant runs the same loop, and no object is ever fitted on data at or after the decision date (HMM and gates use trailing windows with fully realised forward returns, and λ_t is selected on training-window days only). Every hyperparameter is listed in Sec. 4.1; compute is minutes per variant on a single CPU core. The per-variant daily book returns used in the significance tests of Appendix E are released with the code, so the gaps in Table 1 can be re-tested directly.

A HAND-DERIVED GRADIENTS

All gradients are derived by hand and implemented in vectorised numpy. For one decision date with $X \in \mathbb{R}^{n\times K}$, $A \in \mathbb{R}^{n\times n}$, target $y \in \mathbb{R}^n$, write $E = (e_{ij})$ for the logits of Eq. (7), $P = (\alpha_{ij})$ for the masked softmax, C for the aggregation, and s for the read-out. With $\delta = 2(s - y)$ and the mask $M = \mathbb{1}[A_{ij} > 0 \text{ or } i = j]$:

Read-out.

$$\frac{\partial \mathcal{L}}{\partial v} = (X \odot m)^\top \delta, \quad \frac{\partial \mathcal{L}}{\partial w} = C^\top \delta, \quad \frac{\partial \mathcal{L}}{\partial g_z} = H^\top \delta, \quad \frac{\partial \mathcal{L}}{\partial X} \mathrel{+}= \delta v^\top \odot m, \tag{15}$$

and m is treated as a constant (it is evolved, not learnt).

Aggregation. $\partial \mathcal{L}/\partial C = \delta w^\top$, hence $\partial \mathcal{L}/\partial H_j \mathrel{+}= \sum_i \alpha_{ij}\,(\delta w^\top)_i$ (transposed aggregation $\widehat{H} = P^\top(\delta w^\top)$) plus the direct term $g_z \delta^\top$ from the skip path.

Masked softmax. With $\partial \mathcal{L}/\partial P = G$ (from the aggregation), the logit gradient is the standard softmax Jacobian with mask M:

$$\frac{\partial \mathcal{L}}{\partial e_{ij}} = M_{ij}\,\alpha_{ij}\Big(G_{ij} - \textstyle\sum_{j'} G_{ij'}\alpha_{ij'}\Big). \tag{16}$$

Logits and embedding. $\partial \mathcal{L}/\partial a_{\mathrm{src}} = \sum_{ij}(\partial \mathcal{L}/\partial e_{ij})H_i$, $\partial \mathcal{L}/\partial a_{\mathrm{dst}} = \sum_{ij}(\partial \mathcal{L}/\partial e_{ij})H_j$, $\partial \mathcal{L}/\partial b_e = \sum_{ij}(\partial \mathcal{L}/\partial e_{ij})A_{ij}$, and

$$\frac{\partial \mathcal{L}}{\partial H} \mathrel{+}= \big(\tfrac{\partial \mathcal{L}}{\partial e}\mathbf{1}\big)a_{\mathrm{src}}^\top + \big(\tfrac{\partial \mathcal{L}}{\partial e}\big)^\top a_{\mathrm{dst}}, \qquad \frac{\partial \mathcal{L}}{\partial W_1} = X^\top\big(\tfrac{\partial \mathcal{L}}{\partial H} \odot \mathbb{1}[HW_1{+}b_1 > 0]\big). \tag{17}$$

All parameters are updated with Adam (lr $= 0.02$, $\ell_2 = 10^{-4}$, 15 epochs) over the training window; a full refit takes seconds on CPU, which is what makes the walk-forward protocol with 60-day refits and 3-seed ensembling affordable.

B FACTOR DEFINITIONS

Equivalence to phase averaging (proof sketch). Write the single-phase strategy at phase φ as holding $w_t^{(\varphi)} = w_{\varphi+kH}$ for $t \in [\varphi + kH, \varphi + (k{+}1)H)$, with book return $r_t^{(\varphi)} = w_t^{(\varphi)\top} r_t$. In the fully-ramped regime, tranche $s = t - H + 1 + \varphi$ realises exactly the single-phase return of phase φ at day t, hence $r_t^{\mathrm{book}} = \frac{1}{H}\sum_{\varphi=0}^{H-1} r_t^{(\varphi)}$: the overlapping book is the *uniform average over all H phases*, so the phase nuisance parameter is marginalised out by construction rather than by repeated runs. A second consequence is statistical: daily decisions make daily IC observations admissible, raising the IC sample size over a $\sim$2-year test from $\sim$52 (10-day grid) to $\sim$540.

Table 2: The $K{=}24$ factor library (all cross-sectionally z-scored at each date).

Family	Factors	Definition (trailing window w)
Momentum / reversal	mom5, mom10, mom20, mom60, mom120	$\log(P_t/P_{t-w})$
	rev1, rev3	$-r_t$; $-\frac{1}{3}\sum_{k\le 3} r_{t-k+1}$
Volatility	vol20, vol60	std of daily returns
	vvol20	20-day mean of 5-day return std (vol-of-vol)
	dvol20	std of negative returns (downside vol)
Volume / liquidity	turn_z	cross-sectional z of $\log(1+\text{volume})$
	vratio	5-day / 20-day mean volume ratio
	vol_z20	z of 20-day mean volume
	amihud	Amihud: $\text{mean}(\lvert r\rvert/\log(1+\text{vol}))$
Price structure	ma5/20/60_gap	$P_t/\,\text{MA}_w - 1$
	atr20	mean daily (high$-$low)/close
	pos_range	mean $(P-\text{low})/(\text{high}-\text{low})$ (5d)
	rsi14	RSI(14)
Market-related	beta60	rolling β vs. index (60d)
	idio_vol60	std of β-hedged residual
	mkt_corr60	rolling correlation with index

Variants. All variants share the identical evaluation protocol and factor library; they differ only in the mechanism under test:

market	equal-weight universe book (benchmark)
ensemble	all K factors, equal weights, same long–short construction
no_evo	uniform gate: all factors, no evolutionary pruning
no_regime	evolving gate, but pooled across regimes (regime-agnostic pruning)
static	per-regime gates frozen at test start (no temporal evolution)
no_graph	identity adjacency: no relational information
no_text	graph = statistical + industry only ($\alpha{=}0$)
no_stat	graph = text only where news exists, otherwise industry only ($\beta{=}0$)
full (R^2PT)	text + statistical + industry graph, regime-adaptive evolution, regime embeddings, 3-seed ensemble

C Implementation details and complexity

Graph snapshots. A_t is recomputed at every decision date for the variant under test via a key-based cache (`snapshot_key`): each ablation needs exactly one adjacency flavour, so per-variant cost is $1\times$ a full graph computation rather than $4\times$ (computing all component variants). Over a $\sim$550-day test segment this is $\sim$550 snapshot computations per variant.

Evolutionary gates advance every $\Delta = 10$ days; between steps the gate of the most recent generation is used. Fitness windows use only realised forward returns, so at date t the newest usable label is $t - H$.

Ensembling. Three seeds per refit; their scores are averaged before ranking. Seed variance across training curves is shown in Figure 8.

Compute. Everything runs on CPU in minutes per variant: the scorer has $\sim 5{\times}10^2$ parameters; the dominant cost is the daily graph snapshots.

D Full walk-forward algorithm

Algorithm 1 runs the complete loop for one market; every variant in Table 1 is the same loop with the mechanism under test replaced.

Algorithm 1: R^{2}PT walk-forward loop (one market)

Input: panel (prices, volumes, headlines, sectors); test start t_0; horizon H; generation step Δ; refit step RETRAIN

Fit HMM on $[t_0 - \textsc{TrainWin}, t_0)$; decode regimes on the full panel with the frozen model;
Initialise gates $m^{(r)}$ uniformly on top-P factors per regime;
for $t = t_0, \ldots, T-1$ ***(daily)*** **do**
    if $t \equiv 0 \pmod{\Delta}$ **then**
        update gates $m^{(r)}$ by Algorithm 2 (fitness on realised days only);
    else
        $m \leftarrow$ gate of the most recent generation;
    if $t \equiv 0 \pmod{\textsc{Retrain}}$ **then**
        refit the 3-seed scorer ensemble on $[t - \textsc{TrainWin}, t)$;
        re-select shrinkage λ_t on the trailing 250 training days (Eq. 12);
    $A_t \leftarrow$ relation-probe snapshot (Eq. 1, data $< t$ only);
    $s_t \leftarrow \text{scorer}(X_t, A_t, z_t, m_{z_t})$ (Eqs. 6–10);
    $\tilde{s}_t \leftarrow z(p_t) + \lambda_t\, z(s_t)$ (Eq. 12);
    $w_t \leftarrow$ dollar-neutral quintile long–short basket of $\tilde{s}_t$;
    rescale by the trailing volatility-target factor (10% annualised);
    charge cost $\frac{1}{H}\|w_t - w_{t-H}\|_1 \cdot c$ (Eq. 14);

Output: daily book returns (Eq. 13), IC series, metrics

E PAIRED SIGNIFICANCE OF THE ABLATION GAPS

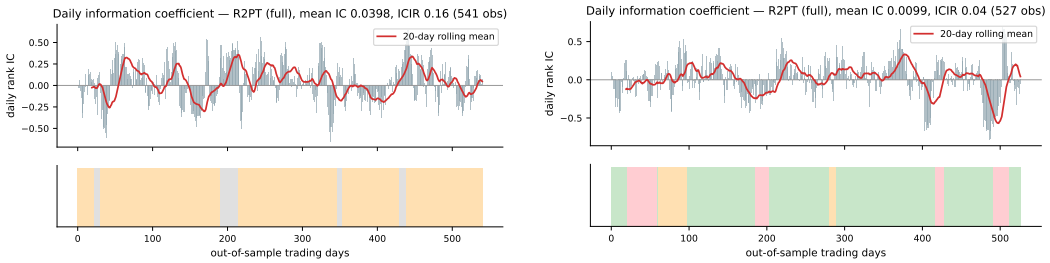
Table 3: **Paired significance of the ablation gaps.** For each variant we difference the daily book returns against the full system on the days both cover, and report the Sharpe gap, the mean daily return gap in basis points, its t-statistic, and a stationary block-bootstrap 95% confidence interval (block length 20 days, 4,000 resamples). No gap is significant at the 5% level in a $\sim$2.2-year test; the pruner-related gaps are the largest and the most consistent in sign.

(a) S&P 500 (US).

Variant (removed)	**Sharpe**	**Δ Sharpe**	**Δ bps/day**	t	**95% CI (bps/day)**
w/o evolution	-0.25	+0.52	+2.36	+1.60	[-0.79, +5.67]
w/o regime cond.	-0.13	+0.41	+1.83	+1.18	[-1.13, +5.09]
static gates	-0.15	+0.43	+1.90	+1.32	[-1.09, +5.20]
w/o graph	0.18	+0.09	+0.35	+0.34	[-1.63, +2.70]
w/o text (α=0)	0.29	-0.01	-0.07	-0.42	[-0.36, +0.12]
w/o stat. (β=0)	0.16	+0.12	+0.50	+0.46	[-1.76, +2.56]
Static ensemble	0.24	+0.03	+0.09	+0.05	[-3.82, +3.93]
Market (equal-weight)	0.57	-0.29	-1.92	-0.49	[-7.93, +4.13]

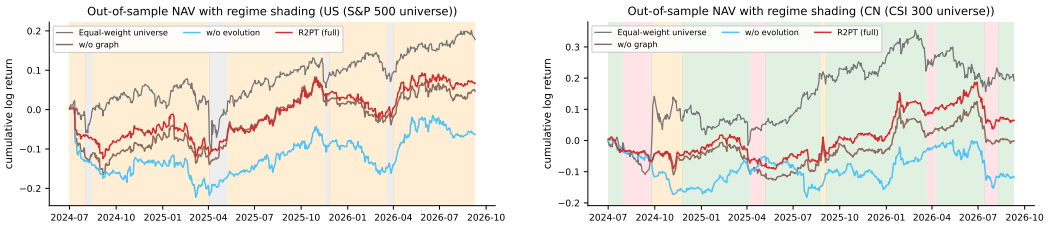
(b) CSI 300 (CN).

Variant (removed)	**Sharpe**	**Δ Sharpe**	**Δ bps/day**	t	**95% CI (bps/day)**
w/o evolution	-0.49	+0.75	+3.39	+0.97	[-3.13, +10.11]
w/o regime cond.	-0.43	+0.69	+3.06	+0.78	[-3.72, +10.25]
static gates	-0.06	+0.32	+1.46	+0.41	[-4.65, +8.04]
w/o graph	-0.00	+0.27	+1.22	+1.20	[-0.67, +3.04]
w/o text (α=0)	0.27	-0.00	-0.00	-0.70	[-0.01, +0.00]
w/o stat. (β=0)	0.32	-0.06	-0.25	-0.18	[-3.37, +2.40]
Static ensemble	0.87	-0.61	-2.69	-0.90	[-8.43, +2.37]
Market (equal-weight)	0.51	-0.25	-2.48	-0.49	[-12.96, +7.17]



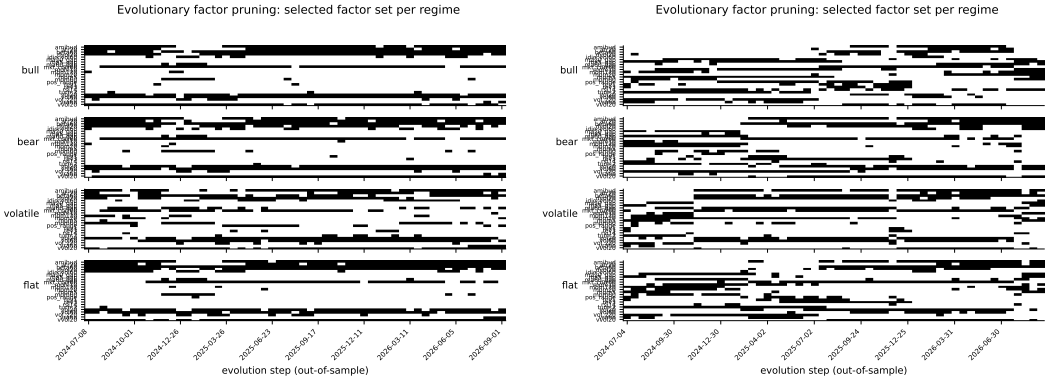
(a) S&P 500 (US). (b) CSI 300 (CN).

Figure 2: **Daily rank-IC** (dots) and 20-day rolling mean (line) of R^{2}PT scores against forward 10-day returns. IC is noisy daily but persistently positive on rolling windows on both panels.



(a) S&P 500 (US). (b) CSI 300 (CN).

Figure 3: **Out-of-sample cumulative wealth** (log scale) of the market book, the static ensemble, and R^{2}PT, with decoded regimes shaded (green=bull, red=bear, orange=high-volatility, grey=sideways). R^{2}PT's drawdowns cluster at regime transitions on CSI 300, where the pruner rotates the support between momentum and market/liquidity factors; on the US panel the book spends almost the whole window inside one decoded state.



(a) S&P 500 (US). (b) CSI 300 (CN).

Figure 4: **Evolved factor gates per regime** over the test segment (rows: $K{=}24$ factors ordered by family; columns: generations every 10 days; brightness: gate weight, white = pruned to exactly zero).

F ADDITIONAL FIGURES

Algorithm 2: Regime-adaptive evolutionary factor pruning (one generation at date t)

Input: factors $X_d \in \mathbb{R}^{n \times K}$, forward returns y_d for realised days; decoded regimes z_d; gates $m^{(r)}$, $r = 1..R$

for $r \in \{1, \ldots, R\}$ **do**
- $\mathcal{D}_r \leftarrow \{d \in [t-250, t-H] : z_d = r,\ y_d \text{ realised}\}$;
- **if** $|\mathcal{D}_r| < 20$ **then**
  - $\mathcal{D}_r \leftarrow \{d \in [t-250, t-H] : y_d \text{ realised}\}$; `// pooled fallback`
- **for** $k \in \{1, \ldots, K\}$ **do**
  - $f_k^{(r)} \leftarrow \mathrm{corr_S}(\mathrm{rank}\, X_{\cdot,k,\mathcal{D}_r},\ \mathrm{rank}\, y_{\mathcal{D}_r})$;
- $S^{(r)} \leftarrow$ top-P factors by $f^{(r)}$;
- **if** $\mathrm{rand}() < p_{\mathrm{mut}}$ **then**
  - replace a random slot of $S^{(r)}$ by $\arg\max_{k \notin S^{(r)}} f_k^{(r)}$;
- $m^{(r)} \leftarrow \mathrm{norm}\left((1-\beta_{\mathrm{prev}})\,\pi(f^{(r)}) + \beta_{\mathrm{prev}}\, m^{(r)}\right)$ restricted to $S^{(r)}$;

Output: gates $m^{(r)}$ (used until the next generation)

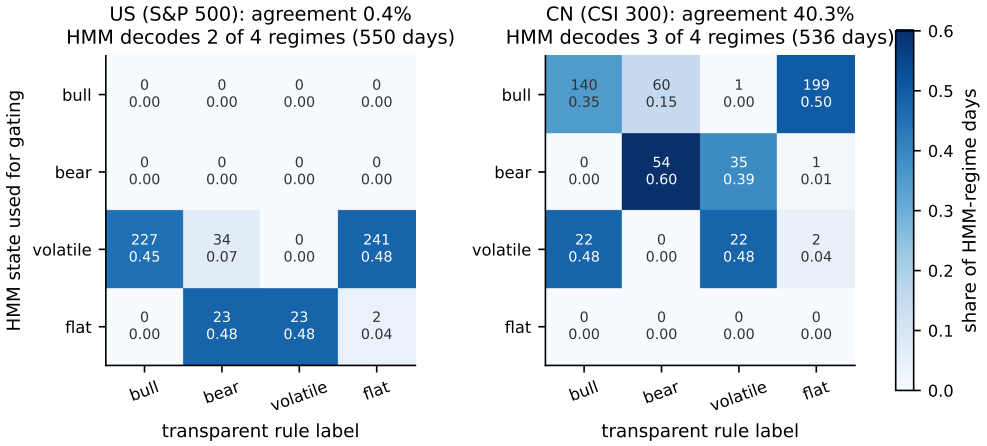
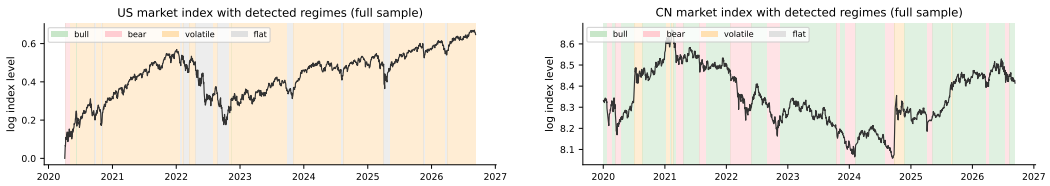


Figure 5: **Regime labelling: HMM regimes used for gating vs. transparent rule labels**, test segment. Cell entries are day counts and the share of the row. Regimes that the frozen HMM never decodes in the window appear as empty rows of zeros. The two labellings agree on 40.3% of CN days but on only 0.4% of US days, where the HMM calls 502 of 550 days *volatile* while the rule labelling calls 227 of them *bull* and 241 *flat*: the regime variable is a data-driven partition used for conditioning, not a validated market taxonomy.



(a) S&P 500 (US). (b) CSI 300 (CN).

Figure 6: **Decoded market regimes** over the full panel (index level, top; regime band, bottom). The HMM is fitted on the training segment only. For the US panel the index is an equal-weight proxy of the 68 constituents, since the data source provides no aligned S&P 500 series.

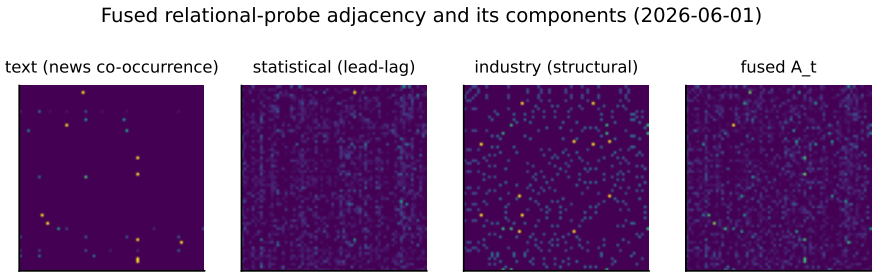
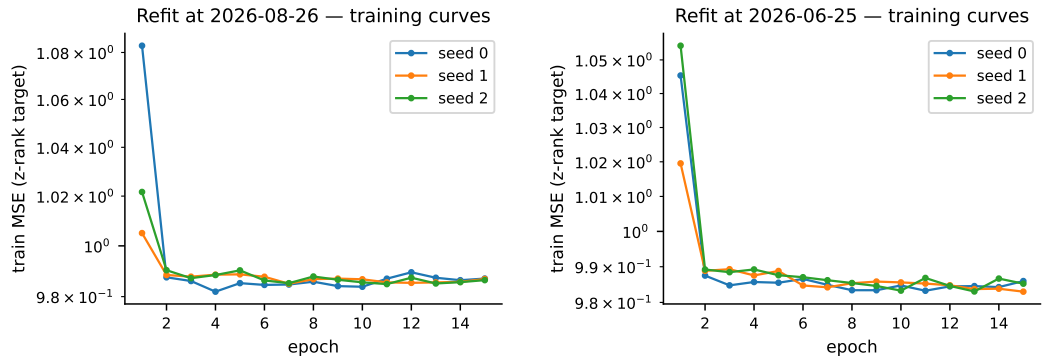
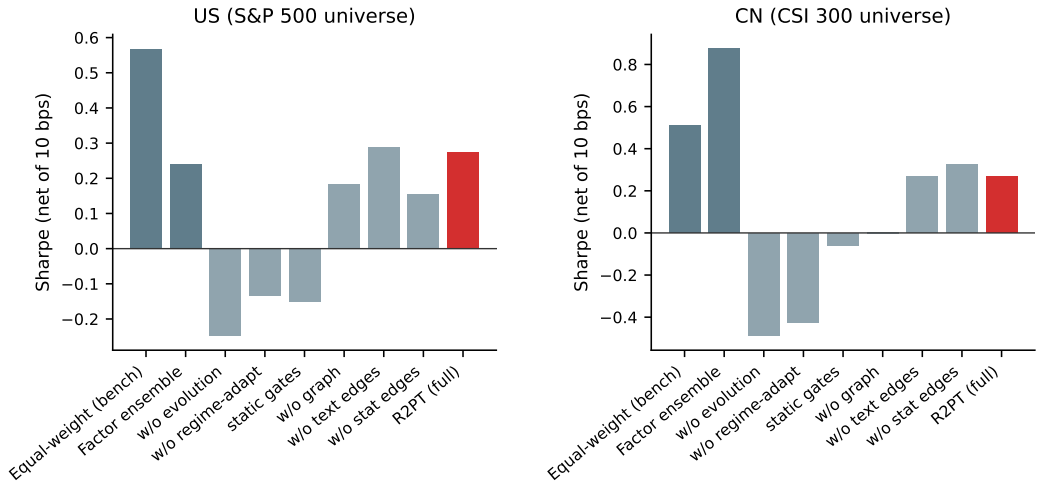


Figure 7: **Relation probe output on a US test date**: fused adjacency A_t (left) and its three row-normalised components—text, statistical lead–lag, industry. The fused graph is directed and time-varying; component heatmaps are ordered by sector to make the industry block structure visible. On dates without news coverage the text term is dropped and the remaining two components are renormalised.



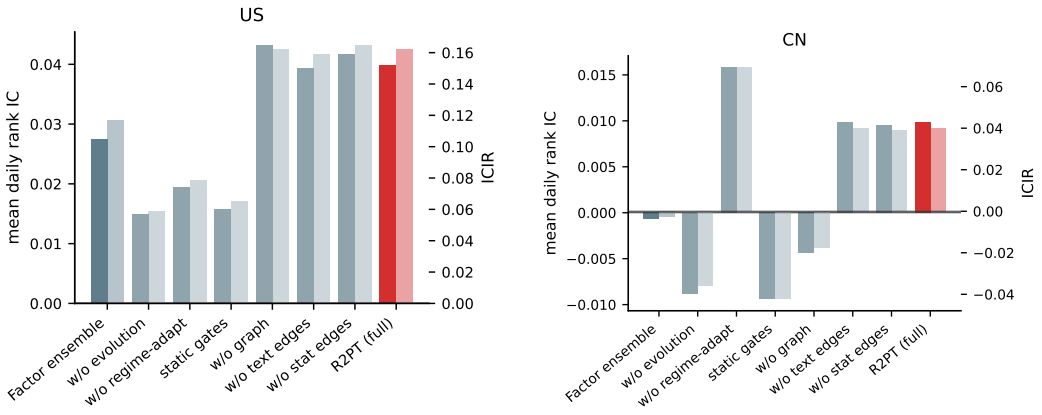
(a) S&P 500 (US). (b) CSI 300 (CN).

Figure 8: **Training loss of the three ensemble seeds** across walk-forward refits (final refit shown).



(a) S&P 500 (US). (b) CSI 300 (CN).

Figure 9: **Ablation ladder**: out-of-sample Sharpe by variant.



(a) S&P 500 (US). (b) CSI 300 (CN).

Figure 10: **IC and ICIR by variant** with n_{IC} daily observations.