

Jump-Residual State Learning for A-Share Stock Selection: A Residual Identification and Portfolio Validation Study

Hanzhi Sun

College of Science, Mathematics, China Agricultural University, Beijing, 100083, China
hanzhisun@cau.edu.cn

ABSTRACT

This study examines whether discontinuous price-movement information can improve cross-sectional stock selection in the Chinese A-share market after conventional return-prediction signals are controlled. A jump-residual state learning framework is proposed. The framework first estimates a non-jump machine-learning baseline, then learns the residual return component using robust jump indicators, idiosyncratic jump intensity, post-jump memory, graph-based peer confirmation and auditable semantic-event variables. This residual design tests a stricter incremental-value question than directly appending raw jump factors to a predictive model. Using a filtered HS300-100 universe, a five-day forward excess-return target, five-day rebalancing and transaction-cost-adjusted top-10 portfolios, JRS_Residual achieves a test-period cumulative return of 44.08% and a Sharpe ratio of 2.419, compared with CSI300 at 9.03%, EqualWeight at 21.44% and ML_Base at 24.87%. However, Newey-West RankIC and bootstrap tests indicate that statistical significance is mixed, and semantic augmentation is constrained by incomplete text-event coverage. The evidence therefore supports a cautious conclusion: residualized jump-state information exhibits sample-period out-of-sample value, while semantic enhancement remains a reproducible but not yet statistically confirmed extension.

Keywords: A-Share Market; Jump Risk; Residual Learning; Empirical Asset Pricing; Graph Risk; Semantic Events; Portfolio Construction

1. INTRODUCTION

Cross-sectional stock selection in the Chinese A-share market is a difficult empirical problem because stock returns are affected by conventional risk premia, retail-driven trading pressure, liquidity shocks, price-limit mechanisms, industry contagion and abrupt information arrivals. A factor that appears profitable in a small backtest may simply be a nonlinear transformation of known predictors, a proxy for market-wide volatility or a sample-specific artifact. Therefore, the central research question of this paper is not whether jump variables are correlated with future returns in isolation, but whether jump-state information contains incremental predictive value after conventional return-prediction signals are controlled.

This paper proposes a Jump-Residual State learning framework for A-share stock selection. The framework first estimates a non-jump baseline model using conventional price, turnover, volatility, liquidity and technical variables. It then defines the baseline residual as the component of the forward excess return not explained by the baseline model. Jump-state variables are trained to predict this residual component rather than the original return. This design turns the factor-discovery problem into an incremental-value test and reduces the risk that a complex model merely reuses known information.

The empirical setting uses a filtered HS300-100 stock universe from January 2024 to June 2026. The target variable is the five-day forward excess return over CSI300, and the portfolio is rebalanced every five trading days. All strategies share the same test period, transaction cost convention and top-10 long-only construction. The benchmark set includes CSI300, EqualWeight, LinearFactor, ML_Base, ML_Base_Plus_JumpRaw, JRS_Residual, JRS_Residual_GraphRisk and semantic-augmented JRS variants.

The main finding is economically meaningful but statistically cautious. JRS_Residual achieves the strongest test-period portfolio result among the compared models, while the raw jump augmentation performs poorly. This contrast is central to the paper: raw jump inclusion is not sufficient; jump information becomes useful only when it is organized as residual state information and combined through validation-based dynamic weighting. At the same time, significance tests are mixed, so the evidence should be interpreted as sample-period out-of-sample value rather than a fully established investment anomaly.

The contribution of this paper is threefold. First, it introduces a residual identification design for jump-state factors in A-share stock selection. Second, it adds graph confirmation and real semantic-event alignment without fabricating missing text data. Third, it reports recommendation audits, Newey-West RankIC tests, bootstrap confidence intervals and regime analysis, making the empirical conclusion more transparent and less dependent on a single backtest curve.

Figure 1 summarizes the complete research workflow, from the non-jump baseline and residual jump-state learner to graph confirmation, semantic alignment, portfolio formation and statistical validation.

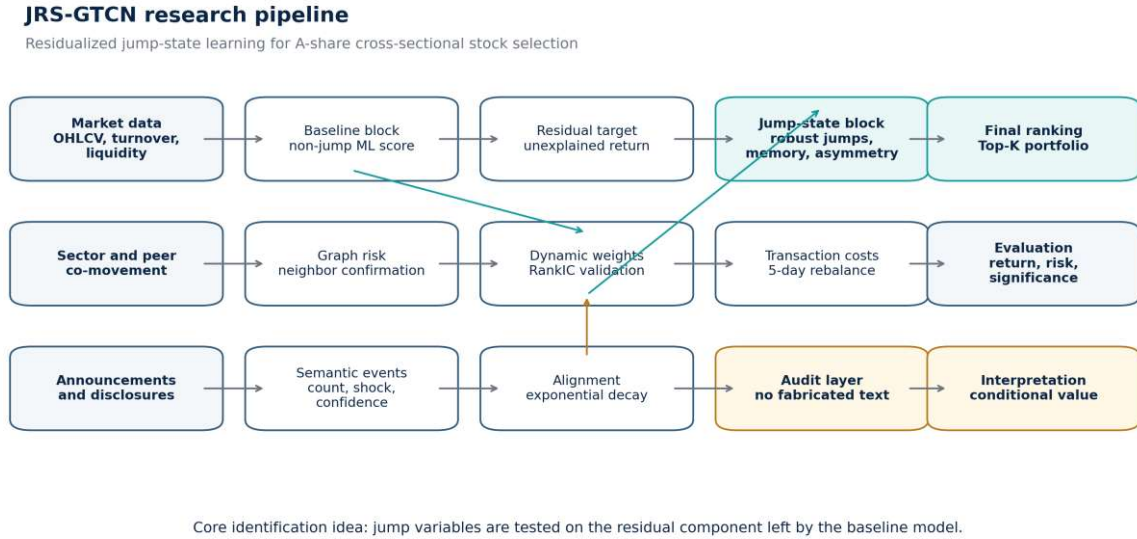


Figure 1. Research framework for residualized jump-state stock selection.

2. LITERATURE REVIEW AND RESEARCH HYPOTHESES

The first related literature is empirical asset pricing and factor-based stock selection. Classic studies document common risk factors, momentum, reversal and other forms of return predictability [1-3]. Later work shows that many anomalies can overlap, weaken after publication or fail to survive stronger multiple-testing standards [8-9]. This literature implies that a new factor block should not be evaluated only by its standalone correlation with returns. It should be tested for incremental explanatory power relative to established predictors. The residual design in this paper follows this principle by asking whether jump-state variables explain the part of future excess returns that a non-jump baseline leaves unexplained.

The second literature concerns volatility, realized variation and jumps. GARCH-type models describe time-varying conditional volatility [4], while realized-volatility studies and jump tests distinguish continuous volatility from discontinuous price moves [5-7]. In equity markets, jumps may reflect firm-specific news, investor attention, liquidity imbalance or market-wide stress. These interpretations have different implications for expected returns. A positive jump can be followed by continuation if it reflects information discovery, but it can also be followed by reversal if it reflects temporary overreaction. This ambiguity motivates a state-based approach rather than a single raw jump variable.

The third literature applies machine learning to return prediction. Flexible models can capture nonlinear interactions, heterogeneous effects and high-dimensional predictor structures [10-11]. However, model flexibility also makes overfitting easier. A machine-learning model that performs well in the training sample may not necessarily provide interpretable incremental alpha. This paper therefore uses machine learning as both a baseline predictor and a residual learner, and evaluates the economic meaning of the residual component through out-of-sample portfolios and statistical tests.

The fourth literature studies connectedness, industry transmission and graph-based financial learning. Stocks are not independent observations. They are linked through sectors, supply chains, common institutional ownership, factor exposures and correlated investor attention. Graph methods provide one way to use this dependence structure [12]. In the present paper, graph risk is used as transparent peer confirmation rather than as an opaque end-to-end neural architecture. This choice is consistent with the objective of improving interpretability and auditability.

The fifth literature examines financial text, media tone and investor sentiment. Textual information may reach investors before prices fully adjust, and media sentiment can influence trading pressure ^[13-15]. Recent evidence also shows that high-dimensional text representations can be used for return prediction, although their economic value depends on disciplined out-of-sample evaluation ^[16]. Nevertheless, text-based factors are sensitive to coverage, event timing, source quality and language processing. This paper therefore includes semantic factors only when real announcements and disclosure events are available. The semantic module is evaluated as an auditable extension rather than treated as validated alpha by assumption.

Hypothesis 1. Raw jump factors do not necessarily improve stock selection once conventional predictors are considered.

Hypothesis 2. Residualized jump-state factors provide stronger incremental value than raw jump augmentation.

Hypothesis 3. Graph-based peer confirmation improves stability when jump states contain common information across related stocks.

Hypothesis 4. Semantic-event factors may add conditional value, but the effect depends on event coverage and alignment quality.

3. DATA, VARIABLES AND EXPERIMENTAL DESIGN

3.1 Universe Construction and Sample Split

The empirical universe is constructed from HS300 candidate stocks. Stocks with insufficient observations, poor liquidity or excessive missing values are removed. The final experiment uses a 100-stock A-share universe. The market data start on 2024-01-02 and end on 2026-06-22. The training period ends on 2025-03-31, the validation period ends on 2025-09-30 and the remaining observations form the out-of-sample test period.

The sample split is deliberately chronological. No future information is used to train or tune earlier predictions. All model variants use the same split, the same rebalancing frequency and the same transaction-cost setting. This design makes the baseline comparison meaningful because differences in performance are attributable to the signal construction rather than to different evaluation conventions.

The reproducibility controls are embedded in the main experimental design. They include a fixed chronological split, a common transaction-cost assumption across all baselines, an identical top-K rule for comparable portfolios, real event-source records only, no fabricated semantic observations, duplicate date-stock audits, Newey-West RankIC testing, bootstrap confidence intervals and regime-level robustness comparisons.

Figure 2 displays the chronological training, validation and out-of-sample test windows together with the common five-day prediction and rebalancing design.

Experimental Timeline and Walk-Forward Evaluation



Prediction horizon = 5 trading days; Rebalancing frequency = 5 trading days; Top-K = 10

Figure 2. Chronological training, validation and out-of-sample testing design.

3.2 Return Target and Evaluation Unit

For stock i on trading date t , the target is the five-day forward excess return over CSI300:

$$y_{i,t}^{(5)} = r_{i,t:t+5} - r_{m,t:t+5} \quad (1)$$

The ranking score is evaluated by daily cross-sectional IC and RankIC:

$$IC_t = \text{corr}(S_{i,t}, y_{i,t}^{(5)}), \text{RankIC}_t = \text{corr}_{\text{Spearman}}(S_{i,t}, y_{i,t}^{(5)}) \quad (2)$$

Because the final investment rule selects the top-ranked stocks, RankIC and Precision@10 are more aligned with the portfolio objective than mean squared prediction error alone.

3.3 Semantic Event Data

The semantic module uses real public announcements and disclosure events. The current text snapshot contains 2,564 events covering 19 stocks from 2025-10-01 to 2026-06-22. Event-level variables are converted into stock-date features such as event counts, weighted counts, semantic shock, risk-keyword intensity and semantic confidence. No synthetic news or placeholder text is used.

Figure 3 reports the ranking evidence for the leading semantic-event factors and the associated recommendation-quality audit, making the limited coverage and the duplicate-control checks visible in the main text.

Semantic factors and recommendation-quality audit



Figure 3. Semantic-event factor ranking and recommendation audit snapshot.

4. METHODOLOGY

4.1 Baseline Model and Residual Target

Let $X_{i,t}^B$ denote the baseline non-jump predictor vector, including momentum, volatility, turnover, liquidity and technical indicators. The baseline learner estimates:

$$\hat{y}_{i,t}^B = f_B(X_{i,t}^B) \quad (3)$$

The residual target is then defined as:

$$\epsilon_{i,t}^B = y_{i,t}^{(5)} - \hat{y}_{i,t}^B \quad (4)$$

The key methodological choice is to train jump-state variables on epsilon rather than on y itself. If jump variables only duplicate conventional signals, they should have limited ability to explain this residual component. If they contain incremental information, they should improve the residual prediction and the final ranking.

Figure 4 illustrates this residual identification logic and clarifies why the jump learner is trained on the baseline forecast error rather than directly on the original return target.

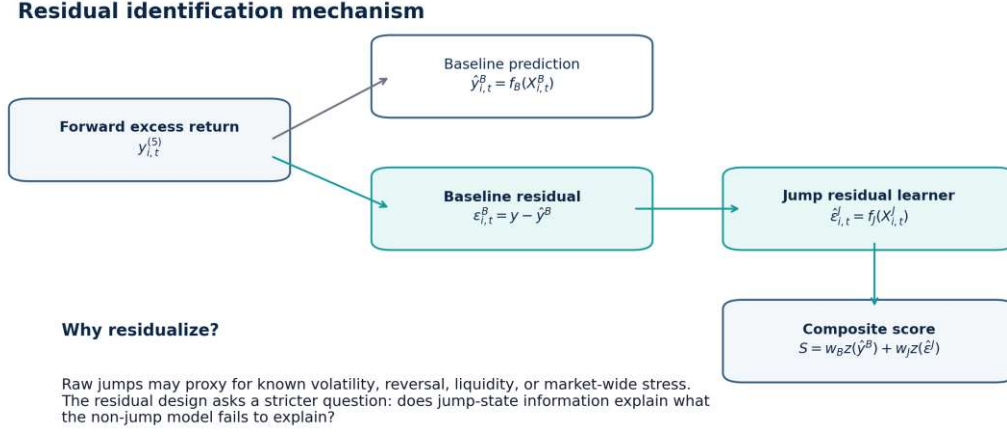


Figure 4. Residual identification mechanism for testing incremental jump-state value.

4.2 Robust Jump-State Construction

Daily jumps are measured using robust rolling statistics. Let $r_{i,t}$ be the daily return, $m_{i,t}^{(L)}$ the rolling median over window L and $MAD_{i,t}^{(L)}$ the median absolute deviation. The robust jump score is:

$$z_{i,t}^R = [r_{i,t} - m_{i,t}^{(L)}] / [1.4826 MAD_{i,t}^{(L)} + \epsilon] \quad (5)$$

A jump flag is assigned when $|z_{i,t}^R|$ exceeds a threshold. Additional state variables include positive and negative jump counts, jump asymmetry, time since last jump, volume-confirmed jumps, idiosyncratic jump intensity and post-jump memory. This transformation converts isolated jumps into a state description that can distinguish recent shocks from persistent jump regimes.

4.3 Residual Jump Learner and Dynamic Fusion

Let $X_{i,t}^J$ denote the jump-state feature vector. The residual jump learner estimates:

$$\hat{\epsilon}_{i,t}^J = f_J(X_{i,t}^J) \quad (6)$$

The final JRS score is a dynamically weighted standardized combination of the baseline score and the residual jump score:

$$S_{i,t}^{JRS} = w_t^B z(\hat{y}_{i,t}^B) + w_t^J z(\hat{\epsilon}_{i,t}^J) \quad (7)$$

Weights are determined from validation-period ranking performance and constrained to avoid unstable signal dominance. This mechanism is not intended to create a new complex architecture; it is a disciplined way to combine a conventional return predictor with an incremental residual signal.

4.4 Graph Confirmation and Semantic Alignment

For graph confirmation, stocks are connected by sector membership, return correlation and jump co-movement. The edge weight between stocks i and j is:

$$a_{ij,t} = 0.4I(\text{sector}_i = \text{sector}_j) + 0.4\rho_{ij,t}^r + 0.2\rho_{ij,t}^J \quad (8)$$

The graph-risk score aggregates peer jump pressure and helps distinguish isolated noise from broader sector or correlation-cluster stress. For semantic alignment, event variables are accumulated with exponential time decay:

$$A_{i,t}(\gamma) = \sum_{e \in E_i, \tau_e \leq t} \exp[-\gamma(t - \tau_e)] q_e \quad (9)$$

Here q_e is an event quality or risk score and γ controls how quickly older events lose relevance. The semantic features are deliberately treated as conditional auxiliary variables because current coverage is incomplete.

4.5 Portfolio Construction and Inference

At each rebalancing date, the model ranks the stock universe by $S_{i,t}$ and selects the top $K = 10$ stocks. The implemented strategy is a long-only equal-weight rule, not a mean-variance optimizer. Let H_t denote the selected set. Portfolio weights are defined explicitly as:

$$w_{i,t} = 1/K \quad \text{if } i \in H_t, \quad w_{i,t} = 0 \quad \text{otherwise,} \quad K = 10 \quad (10)$$

The portfolio is rebalanced every five trading days. A one-way transaction-cost rate of $c = 0.001$ is applied consistently to all comparable strategies through portfolio turnover. The net portfolio return is:

$$R_{p,t+1}^{\text{net}} = \sum_i w_{i,t} r_{i,t+1} - c \sum_i |w_{i,t} - w_{i,t-}| \quad (11)$$

where $w_{i,t-}$ is the pre-rebalancing weight after market movements. Because the implemented allocation is equal weighted, no covariance estimator, risk-aversion parameter or optimized stock-specific weight is used. Statistical inference uses Newey-West RankIC t-statistics, paired return tests, bootstrap confidence intervals and rebalancing-period win rates.

5. EMPIRICAL RESULTS

5.1 Portfolio Performance

Table 1. Test-period portfolio performance.

Model	Cum. return	Ann. return	Sharpe	Max drawdown
CSI300	9.03%	13.67%	0.841	-7.78%
Equal Weight	21.44%	33.37%	1.679	-8.94%
Linear Factor	22.17%	34.57%	1.134	-12.68%
ML Base	24.87%	39.00%	1.286	-17.92%
ML + Raw Jump	1.76%	2.62%	0.094	-14.35%
JRS Residual	44.08%	71.84%	2.419	-13.35%
JRS + Graph	35.75%	57.32%	2.046	-12.42%
SA-JRS Residual	42.25%	68.61%	2.280	-11.59%

Table 1 shows that JRS_Residual is the strongest model in the current out-of-sample period. It outperforms CSI300, EqualWeight, LinearFactor and ML_Base in cumulative return and Sharpe ratio. The poor performance of ML_Base_Plus_JumpRaw supports Hypothesis 1: raw jump variables do not automatically improve prediction. The performance of JRS_Residual supports Hypothesis 2: residualized jump-state learning is more effective than naive jump augmentation.

Figure 5 provides a visual comparison of cumulative return, annualized return, Sharpe ratio and maximum drawdown. The horizontal layout preserves the full model names without label overlap.

Out-of-sample portfolio performance across baselines

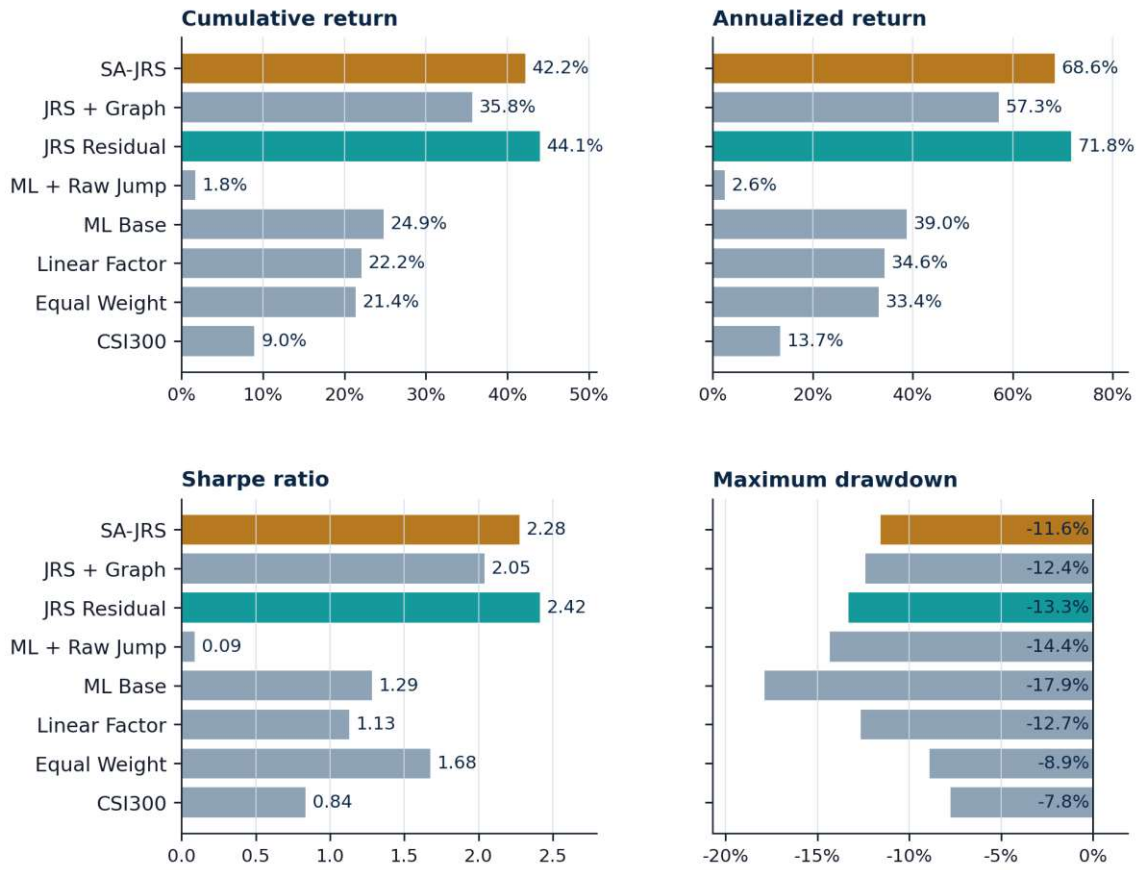


Figure 5. Test-period return and risk comparison across major baselines.

5.2 Prediction and Ranking Evidence

Table 2. Test-period prediction and ranking metrics.

Model	IC	RankIC	ICIR	Top10 hit
Linear Factor	-0.0002	-0.0089	-0.001	0.337
ML Base	-0.0063	-0.0215	-0.044	0.299
ML + Raw Jump	-0.0331	-0.0268	-0.234	0.288
JRS Residual	0.0283	0.0255	0.180	0.320
JRS + Graph	0.0181	0.0229	0.117	0.321
SA-JRS Residual	0.0275	0.0175	0.177	0.356
SA-JRS Gated	0.0149	0.0003	0.100	0.327

Table 2 reports the test-period prediction and ranking evidence. The metrics clarify why the portfolio evidence should be interpreted carefully. JRS_Residual has a positive test-period RankIC, while several baselines have weaker or negative ranking statistics. However, the magnitude of RankIC is modest, which is typical in cross-sectional equity prediction but still requires formal significance testing.

6. STATISTICAL SIGNIFICANCE AND ROBUSTNESS

6.1 Newey-West and Bootstrap Tests

Table 3. Newey-West RankIC significance tests.

Model	Mean RankIC	t-stat	p-value	N
ML Base	-0.0215	-1.158	0.249	165
ML + Raw Jump	-0.0268	-1.472	0.143	165
JRS Residual	0.0255	1.628	0.105	165
JRS + Graph	0.0229	1.545	0.124	165
SA-JRS Residual	0.0175	1.197	0.233	165
SA-JRS Gated	0.0003	0.017	0.986	165

Table 4. Bootstrap confidence intervals for rebalancing excess returns.

Baseline	Mean excess	CI low	CI high	p-value	Win rate
CSI300	0.66%	-0.35%	1.87%	0.210	0.559
Equal Weight	0.32%	-0.59%	1.32%	0.508	0.559
ML Base	0.16%	-0.61%	0.88%	0.669	0.500
JRS Residual	-0.24%	-1.05%	0.50%	0.547	0.529

Table 3 reports the Newey-West RankIC tests, while Table 4 reports bootstrap confidence intervals for rebalancing-period excess returns. The statistical evidence is weaker than the economic backtest. JRS_Residual has a positive Newey-West RankIC statistic, but it does not pass a strict 5% significance threshold in the current sample. Bootstrap confidence intervals for semantic-gated excess returns also include zero. These results do not invalidate the strategy, but they prevent overclaiming. The correct conclusion is that residualized jump states show sample-period out-of-sample value, while stronger data and longer samples are needed for an SCI-level causal or anomaly claim.

Figure 6 visualizes the RankIC estimates and bootstrap intervals, emphasizing that economically positive point estimates are not uniformly statistically significant.

Statistical evidence: economically positive but not uniformly significant

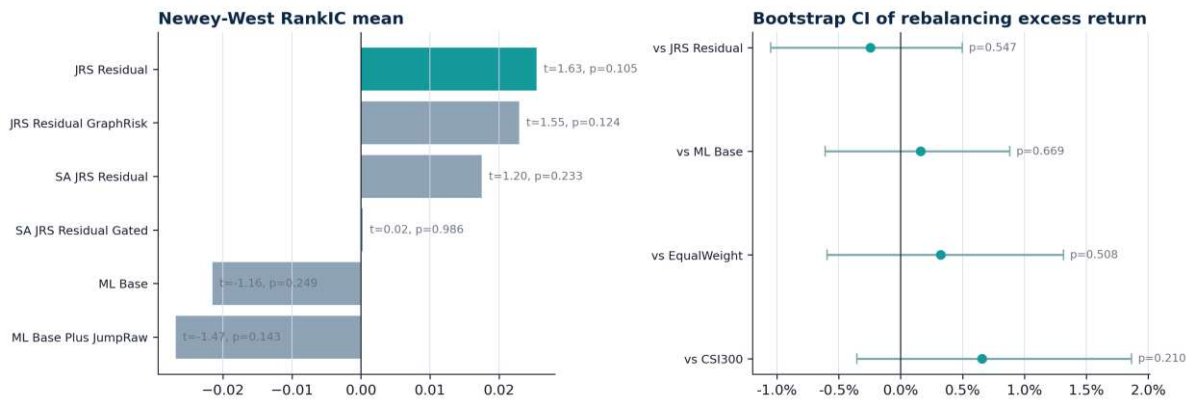


Figure 6. RankIC significance and bootstrap confidence intervals.

6.2 Regime Analysis

Regime analysis examines whether the value of jump-state information is conditional on market states. The evidence suggests that performance varies across high-volatility, low-volatility, positive-market, negative-market, high-jump and semantic-shock regimes. This is economically plausible because jump factors should be most relevant when discontinuous information arrival or market segmentation is more important. Figure 7 compares regime-specific Sharpe ratios across the main model variants.



Figure 7. Regime-dependent Sharpe ratios across model variants.

7. REPRODUCIBILITY AND RECOMMENDATION AUDIT

7.1 Reproducibility Controls

The empirical workflow uses a fixed chronological sample split, a common five-day rebalancing cycle, the same one-way transaction-cost rate across comparable strategies, and an identical Top-10 equal-weight allocation rule. The pipeline also records source-level semantic-event provenance, does not replace missing text with synthetic events, and applies Newey-West, paired-return and bootstrap procedures to the same out-of-sample observations. These controls were previously summarized in an appendix; they are reported here because they are part of the main empirical design.

7.2 Latest Recommendation Output

A practical stock-selection model must produce clean recommendation tables. The latest recommendation audit verifies that the generated table contains no duplicate date-stock observations and that the selected Top10 list contains exactly 10 distinct stocks. This audit prevents accidental concentration caused by duplicated predictions, repeated replacement-pool rows or concatenated model outputs.

Table 5. Latest audited Top10 recommendation list.

No.	Code	Name	JRS score	Main positive factors
1	688521	VeriSilicon	3.792	ML baseline; Residual jump
2	601456	Guolian Minsheng	2.345	ML baseline; Residual jump
3	688082	ACM Research Shanghai	1.965	ML baseline; Residual jump; Peer confirmation
4	301308	Longsys Electronics	1.894	Residual jump
5	600026	COSCO Shipping Energy	0.924	ML baseline; Residual jump; Peer confirmation
6	300450	Lead Intelligent Equipment	0.863	ML baseline; Residual jump; Peer confirmation

Table 5. (continued) Latest audited Top10 recommendation list.

No.	Code	Name	JRS score	Main positive factors
7	600160	Juhua Co.	1.261	Residual jump
8	601059	Cinda Securities	0.919	ML baseline; Residual jump
9	603260	Hoshine Silicon	0.401	ML baseline; Residual jump; Peer confirmation
10	000999	China Resources Sanjiu	0.616	ML baseline; Residual jump

Table 5 reports the latest audited Top10 list using English company names. The recommendation table should not be interpreted as investment advice. It is an output of a controlled research backtest and should be used for academic validation, model diagnostics and further robustness testing.

8. DISCUSSION: PUBLICATION READINESS AND REMAINING GAPS

The current enhanced manuscript is substantially closer to a journal-style empirical finance paper than the initial project report. It now contains a clearer research question, a residual identification strategy, explicit hypotheses, mathematical definitions, statistical inference, regime analysis, recommendation audit and a more cautious conclusion. These elements are necessary for SCI-oriented writing because reviewers usually ask not only whether the strategy earns returns, but also why the signal is incremental and how robust the evidence is.

Nevertheless, the paper is not yet at the level of a mature SCI submission. The main limitation is the short out-of-sample period. The current test window is useful for a course-level or preliminary research paper, but a top empirical-finance journal would expect a longer sample, multiple market cycles and stronger stress-period coverage. The second limitation is universe size. A 100-stock HS300 subset is acceptable for controlled model development, but a stronger paper should test all HS300 constituents or a broader A-share universe.

The third limitation is semantic coverage. The semantic pipeline is reproducible and auditable, but only 19 stocks currently have real text-event coverage. Therefore, semantic augmentation cannot yet be presented as a validated alpha source. The fourth limitation is model simplicity. The paper intentionally avoids adding complex new architectures, which improves interpretability, but future work may compare the transparent residual approach with more advanced temporal graph or cross-modal models once the data foundation is stronger.

The strongest defensible claim is therefore not that JRS-GTCN has discovered a stable investment law. The defensible claim is that residualizing jump-state information against a non-jump baseline is a promising research design, and that in the current HS300-100 experiment it improves out-of-sample economic performance relative to several baselines while statistical validation remains incomplete.

9. CONCLUSION

This paper develops and evaluates a jump-residual state learning framework for A-share stock selection. The core idea is to test jump-state information as an incremental residual predictor rather than to append raw jump variables directly to a baseline model. This design is both economically motivated and empirically useful: raw jump inclusion performs poorly, while the residualized JRS model produces the strongest test-period portfolio result in the current experiment.

The empirical results support a cautious conclusion. JRS_Residual outperforms CSI300, EqualWeight, LinearFactor and ML_Base in the current out-of-sample backtest. However, Newey-West RankIC tests and bootstrap confidence intervals do not provide uniformly strong statistical evidence. Semantic augmentation is operational and auditable, but not yet sufficiently covered to prove stable incremental value. Therefore, the paper should present JRS_Residual as a promising sample-period result and SA-JRS as a reproducible extension requiring more data.

Future work should expand the universe, extend the sample to cover more market cycles, improve semantic coverage, add external validation periods and test whether the residual jump mechanism remains effective under alternative rebalancing frequencies, transaction costs and portfolio constraints. These steps are necessary before the model can support a stronger SCI-level claim of stable factor discovery.

REFERENCES

- [1] Fama, E.F., French, K.R. (1993) Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33: 3-56
- [2] Jegadeesh, N., Titman, S. (1993) Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48: 65-91
- [3] Lo, A.W., MacKinlay, A.C. (1988) Stock market prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies*, 1: 41-66
- [4] Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31: 307-327
- [5] Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P. (2003) Modeling and forecasting realized volatility. *Econometrica*, 71: 579-625
- [6] Barndorff-Nielsen, O.E., Shephard, N. (2004) Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2: 1-37
- [7] Ait-Sahalia, Y., Jacod, J. (2009) Testing for jumps in a discretely observed process. *The Annals of Statistics*, 37: 184-222
- [8] Hou, K., Xue, C., Zhang, L. (2015) Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28: 650-705
- [9] Harvey, C.R., Liu, Y., Zhu, H. (2016) ... and the cross-section of expected returns. *The Review of Financial Studies*, 29: 5-68
- [10] Gu, S., Kelly, B., Xiu, D. (2020) Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33: 2223-2273
- [11] Kelly, B.T., Pruitt, S., Su, Y. (2019) Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134: 501-524
- [12] Kipf, T.N., Welling, M. (2017) Semi-supervised classification with graph convolutional networks. In: *International Conference on Learning Representations*
- [13] Baker, M., Wurgler, J. (2006) Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61: 1645-1680
- [14] Tetlock, P.C. (2007) Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62: 1139-1168
- [15] Loughran, T., McDonald, B. (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66: 35-65
- [16] Ke, Z.T., Kelly, B.T., Xiu, D. (2019) Predicting returns with text data. Working paper