

7528419-an-auditable-public-data-u-s-equity-research-pipeline-point-in-time-factors-asse

SUMMARY: This paper examines a public-data U.S. equity research pipeline, focusing on the impact of various factors such as security identity, filing availability, and transaction costs on the performance of the research. The study uses a sample of 487 equities from 822 securities, providing daily market observations and 37,273 point-in-time fundamental observations. The analysis evaluates 715,447 daily market observations and 47 of 48 factor definitions, finding that multifactor specifications yield higher mean in-sample adjusted R-squared values (0.259–0.301) compared to the Capital Asset Pricing Model (CAPM) (0.063). The study also evaluates eight long-only portfolio methods and eight prediction models, finding that none of the models consistently show stable incremental value after accounting for transaction costs. The research contributes to the field by providing a connected empirical design that highlights the limitations of the public-data and short-window sample.

METHOD: The study uses a public-data U.S. equity research pipeline, with a sample of 487 equities from 822 securities. The sample provides 715,447 daily market observations and 37,273 point-in-time fundamental observations. The analysis evaluates 47 of 48 factor definitions and eight long-only portfolio methods, using eight prediction models in 17 purged monthly test folds. The study also evaluates the impact of transaction costs on the performance of the portfolio methods and prediction models.

KEY FINDINGS:

- Multifactor specifications yield higher mean in-sample adjusted R-squared values (0.259–0.301) compared to the Capital Asset Pricing Model (CAPM) (0.063).
- None of the eight long-only portfolio methods show stable incremental value after accounting for transaction costs.
- Eight prediction models are assessed in 17 purged monthly test folds using two characteristics.
- Non-constant models' bootstrap interval for mean rank correlation includes zero, and stable after-cost incremental value is not established.

IMPLICATION FOR TRADING: The research highlights the importance of considering various factors such as security identity, filing availability, and transaction costs in the development of equity research pipelines. The findings suggest that multifactor specifications can improve explanatory fit, but the impact of transaction costs on portfolio methods and prediction models is not consistent. These insights can be useful for practitioners in designing and evaluating equity research pipelines, as well as for understanding the limitations of public-data and short-window samples.

An Auditable Public-Data U.S. Equity Research Pipeline: Point-in-Time Factors, Asset Pricing, Portfolio Construction, and Walk-Forward Machine Learning

Amit Kumar Dudi

Department of Computational Intelligence

SRM Institute of Science and Technology

Kattankulathur, Chennai, India

ak2910@srmist.edu.in

Abstract—This paper studies a public-data U.S.-equity research pipeline in which security identity, filing availability, source transitions, transaction costs, and temporal validation are specified before performance is interpreted. The released sample screens 822 securities and admits 487 equities. It contains 715,447 daily market observations and 37,273 point-in-time fundamental observations. Forty-seven of 48 factor definitions are estimable. Five asset-pricing families show higher mean in-sample adjusted R-squared for multifactor specifications (0.259–0.301) than for CAPM (0.063). Eight long-only portfolio methods are evaluated under ex-ante one-way costs of 5, 10, and 20 basis points over a 32-month factor-portfolio window. Eight prediction models are assessed in 17 purged monthly test folds using two characteristics. Every non-constant model’s bootstrap interval for mean rank correlation includes zero, and stable after-cost incremental value is not established. The contribution is a connected empirical design that preserves negative and inconclusive findings while making the sample’s public-data and short-window limitations explicit.

Keywords—*empirical asset pricing; point-in-time data; factor models; portfolio construction; transaction costs; walk-forward validation; machine learning; reproducibility*

JEL classifications: C52, C55, G11, G12, G17

I. INTRODUCTION

A. Motivation

Empirical finance often begins after the hardest data decisions have already been made. Yet security identity, delisting coverage, accounting availability, corporate-action treatment, and source changes determine which observations a model sees. Portfolio and machine-learning layers add further choices about costs, sample splits, tuning, and interpretation. This study places those decisions inside one versioned workflow and asks what conclusions remain after the evidence is bounded explicitly.

B. Research problem

The project uses free and public sources rather than CRSP or Compustat. That choice expands reproducibility and access, but it also creates limitations that cannot be repaired by software alone. A fixed upstream ticker universe may omit firms that disappeared before the provider snapshot; a change from consolidated-tape to IEX-only bars alters volume comparability; and SEC facts become usable only after filing acceptance. The empirical problem is therefore both statistical and evidentiary.

C. Research questions

The analysis asks whether factor construction is feasible on the selected panel, whether multifactor models improve explanatory fit, how long-only allocations differ under fixed costs, whether costs reduce observed cumulative outcomes, and whether two-characteristic machine-learning models add stable out-of-sample information or economic value.

D. Contributions

The paper contributes a connected public-data study rather than a new factor theorem. It combines effective-dated identity, point-in-time accounting evidence, a broad factor registry, robust asset-pricing inference, eight constrained allocation methods, and purged walk-forward prediction. It also treats non-estimability, weak prediction, and incomplete hypothesis decisions as legitimate outcomes.

E. Paper structure

Sections II–V develop the literature, data, sample, and design. Sections VI–IX present factor, asset-pricing, portfolio, and machine-learning evidence. Sections X–XV evaluate robustness, hypotheses, implications, limitations, future work, and conclusions. Technical identities and extended tables appear in Appendices A–L.

II. LITERATURE REVIEW

A. Asset pricing and factor models

Markowitz [1] formalized portfolio choice under mean and variance, and Sharpe [2] connected systematic risk to expected return in the CAPM. Fama and French [3] expanded the empirical model with size and value factors; Carhart [4] added momentum; Fama and French [5] introduced profitability and investment factors; and Hou, Xue, and Zhang [6] developed a q-factor interpretation. These models motivate the five pre-specified families evaluated here, but higher explanatory fit is not treated as proof of persistent investable alpha.

B. Portfolio construction and estimation error

Mean-variance optimization is sensitive to expected-return and covariance estimation. DeMiguel, Garlappi, and Uppal [7] show why simple diversification can remain difficult to beat out of sample. Maillard, Roncalli, and Teiletche [8] develop equal-risk-contribution portfolios, López de Prado [9] proposes hierarchical risk parity, and Rockafellar and Uryasev [10] provide a tractable conditional-value-at-risk formulation. The present comparison therefore

spans naive, moment-based, diversification, risk-budgeting, hierarchical, and tail-risk approaches.

C. Statistical inference and multiple testing

White [11] and Newey and West [12] motivate covariance estimators robust to heteroskedasticity and, for HAC inference, serial dependence. Harvey, Liu, and Zhu [13] emphasize that a large factor search raises false-discovery risk. Because the released Phase-2 publication lacks the complete factor-level statistics needed for the original multiplicity rule, the corresponding hypothesis is left inconclusive rather than reconstructed after the fact.

D. Financial machine learning and temporal validation

Ridge regression [14], Elastic Net [15], Random Forest [16], and XGBoost [17] represent regularized linear and nonlinear prediction families. Gu, Kelly, and Xiu [18] document the potential of machine learning in asset pricing, while Bailey et al. [19] and López de Prado [20] stress temporal validation and backtest-overfitting controls. SHAP supplies fitted-model attributions [21], not causal effects.

III. DATA SOURCES AND SOURCE LIMITATIONS

TABLE I. DATA SOURCES AND PRINCIPAL LIMITATIONS

Source	Role	Temporal use	Principal limitation
HF Data Library	Market histories and candidate universe	Daily bars aggregated from intraday evidence	Fixed upstream universe before 2022; PiTrading/IEX source transition
SEC EDGAR	Filings, XBRL facts, ticker/CIK evidence	Actual acceptance timestamps	Taxonomy and concept heterogeneity
Alpha Vantage	Listing-status and lifecycle evidence	Historical lifecycle support; coverage may reach 2010	Not the start date of the empirical market panel
FRED DGS3MO	3-month constant-maturity Treasury yield reference	Frequency-matched risk-free series	Yield must be mapped to period returns
Kenneth French Data Library	Factor concepts and reference evidence	Daily/monthly where specified	Not a security master
SPY	Broad S&P 500 total-return proxy	Benchmark and 21-session relative ML target	ETF proxy rather than index database

A. HF Data Library

HF Data Library supplied the stock candidate set and market histories used locally. The official dataset record is Elkassabgi [22]. The provider's current documentation distinguishes pre-March-2022 consolidated-tape evidence from post-March-2022 IEX evidence and requests attribution to the original sources where noted.

B. SEC EDGAR

SEC [23] submissions, Company Facts, and tagged filing covers provide issuer facts and their availability times. Period end describes the accounting period; filing acceptance determines when a fact could enter research. Accession, taxonomy, concept, unit, and availability are retained.

C. Alpha Vantage lifecycle evidence

The Alpha Vantage [24] LISTING STATUS endpoint supports active and delisted-symbol evidence at historical dates. Some lifecycle evidence extends to 2010, but the released empirical market panel begins on 14 February 2018. These are distinct coverage statements.

D. FRED and Kenneth French

The three-month constant-maturity Treasury yield DGS3MO, produced by the Board of Governors of the Federal Reserve System and accessed through FRED [25], supplies the study's risk-free reference. Kenneth French's Data Library [26] supplies established factor conventions and comparison evidence; neither source replaces the project's security-level market or fundamental data.

E. Benchmark convention

SPY is the broad S&P 500 total-return proxy. Security-specific benchmark gaps are not compounded on different calendars; a common benchmark calendar is used for cross-portfolio comparison.

F. HF pre-2022 survivorship limitation

HF Data Library states that its universe was fixed around 2022–2023 and was not reconstructed point in time. Firms that delisted, failed, or were acquired before that snapshot can be absent. The limitation is most severe before 2022 and matters for long-horizon and cross-sectional asset-pricing work. The project's identity and lifecycle checks can reduce errors among supplied candidates, but they cannot restore securities missing from the upstream universe. The sample is therefore not CRSP-grade survivorship-complete [27].

G. March-2022 PiTrading/IEX structural break

The provider reports a source change in March 2022: earlier bars reflect consolidated-tape PiTrading evidence, whereas later bars derive from IEX. IEX represents approximately 2–3% of consolidated U.S. equity volume. Absolute volume and liquidity measures are therefore not fully comparable across the splice. Average dollar volume, turnover, Amihud illiquidity, and related characteristics require particular caution. Project screening rejects unresolved source splices but cannot eliminate the structural difference in accepted histories [27].

IV. SAMPLE CONSTRUCTION AND POINT-IN-TIME CONTROLS

A. Candidate universe

All 822 candidates passed through the same screening sequence before factor, portfolio, or prediction results were examined. The frozen sample accepts 487 equities and rejects 335: an acceptance rate of $487/822 = 59.25\%$ and a rejection rate of $335/822 = 40.75\%$.

TABLE II. UNIVERSE-SCREENING OUTCOMES

Disposition	Count	Share of candidates
Accepted	487	59.25%
Rejected	335	40.75%

B. Security identity

Listing identity is effective-dated and distinct from issuer identity. Ticker changes, share classes, venue changes, and issuer succession are not collapsed into a timeless present-day ticker.

C. Filing availability

Fundamental observations enter only after the recorded SEC availability time. The released fundamental evidence spans 23 February 2019 through 15 August 2026 by available_at, while the market panel spans 14 February 2018 through 4 August 2026.

D. Shares and split basis

Historical prices and shares must use compatible corporate-action bases. Candidates with unresolved share or split histories are excluded instead of forcing a mechanically complete market-cap series.

E. Lifecycle and terminal-event treatment

Terminal outcomes are bounded with available listing and issuer evidence. Unresolved exits remain exclusions. This control addresses supplied candidates only and does not fill the provider's pre-2022 universe omissions.

F. Final cross-sectional breadth

TABLE III. ANNUAL DAILY SECURITY BREADTH

Year	Minimum	Median	Mean	Maximum
2018	1	4	3.7	4
2019	4	14	17.8	46
2020	46	333	295.5	357
2021	357	398	393.3	404
2022	404	428	425.2	431
2023	431	463	457.6	466
2024	466	486	483.3	487
2025	487	487	487.0	487
2026	486	487	487.0	487

Annual cross-sectional breadth, 2018-2026

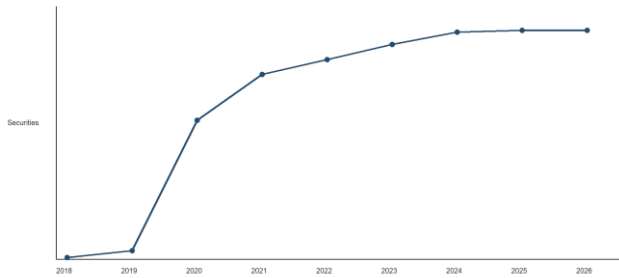


Fig. 1. Annual cross-sectional breadth, 2018–2026.

The early panel is intentionally sparse: only evidence available at the date is admitted. Breadth expands sharply in 2020 and reaches 487 by 2024. The 2026 row ends on 4 August. This pattern makes early cross-sectional estimates more fragile than later ones.

G. Descriptive statistics

TABLE IV. CORE DESCRIPTIVE STATISTICS FROM RELEASED PANELS

Variable	N	Mean	SD	P1	P25	Median	P75	P99
security return	715,447	0.0005868	0.02612	-0.07017	-0.01053	0.0005752	0.01157	0.0734
market capitalization	706,285	5.369e+10	1.803e+11	4.923e+08	1.015e+10	2.16e+10	5.017e+10	3.673e+11
average dollar volume	715,447	6.243e+07	3.062e+08	1.533e+05	4.897e+06	1.268e+07	4.232e+07	6.672e+08
book-to-market	585,078	0.3955	0.4406	-0.2019	0.1188	0.2926	0.5579	1.9823
12-minus-1 momentum	593,210	0.1250	0.4443	-0.6015	-0.1185	0.0691	0.2833	1.6455
rolling volatility	706,194	0.3899	0.2097	0.1574	0.2521	0.3295	0.4623	1.1676

Notes: Security return is expressed as a daily decimal return. Market capitalization is measured in U.S. dollars. Average dollar volume is measured in U.S. dollars per trading day. Book-to-market is a ratio. The 12-minus-1 momentum characteristic is expressed as a decimal cumulative-return characteristic. Rolling volatility is reported in the annualized decimal convention used by the released factor specification.

Market capitalization equals released price multiplied by released shares outstanding; average dollar volume equals price multiplied by volume. The characteristic rows are direct descriptive summaries of the long-form factor IDs book_to_market, 12-minus-1 momentum, and rolling_volatility. They do not represent new factor tests.

V. RESEARCH DESIGN AND HYPOTHESES

The workflow is one way: source evidence feeds point-in-time panels, factors, asset-pricing tests, portfolios, and finally

walk-forward prediction. Later phases may consume but do not rewrite earlier publications. Universe admission precedes analytical inspection, portfolio costs precede performance reading, and ML configuration precedes final fold evaluation.

TABLE V. PRE-SPECIFIED RESEARCH HYPOTHESES

Hypothesis	Final pre-specified proposition
H1	At least one factor spread survives the complete pre-specified cost and multiplicity rule
H2	Multifactor models improve explanatory fit relative to CAPM and that improvement extends to stable out-of-sample economic evidence
H3	Estimated relations are stable through time under the original rule
H4	Higher transaction-cost assumptions reduce cumulative outcomes
H5	Constrained methods alter risk-return trade-offs and any apparent superiority is stable out of sample
H6	ML adds stable OOS predictive and after-cost value
H7	Explanation rankings are stable across methods

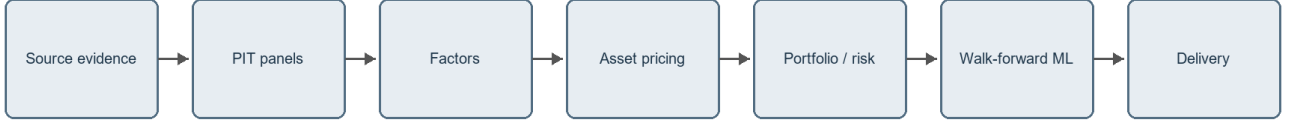


Fig. 2. One-way empirical research workflow.

The seven hypotheses permit supported, partially supported, inconclusive, and not-supported outcomes. H1 requires factor-level cost and multiplicity evidence; H3 requires the complete original temporal-stability decision. Missing required outputs are not replaced with post hoc tests.

VI. FACTOR CONSTRUCTION AND RESULTS

A. Factor families

The 48-definition registry is organized into seven configured families: Market, Size, Value, Momentum, Profitability, Investment/Leverage, and Risk/Liquidity. Values are winsorized and normalized according to the released configuration, with monthly portfolio formation where specified.

TABLE VI. FACTOR CONSTRUCTION SCALE AND ESTIMABILITY SUMMARY

Evidence	Released result
Configured definitions	48
Estimable definitions	47
Non-estimable definition	equity issuance
Long-form factor rows	34,341,456
Factor-portfolio observations	11,545
Sample dates	2018-02-14 to 2026-08-04

B. Portfolio construction

Cross-sectional ranks and high-minus-low portfolios translate characteristics into comparable evidence. Value-weighted terciles and the common benchmark calendar are implementation choices fixed in the release.

C. Missingness and estimability

Coverage differs across characteristics because market histories and accounting inputs become available at different dates. Estimability means a factor can be constructed under the released rule; it does not imply statistical significance.

D. Equity issuance limitation

Equity issuance is not estimable because the required issuance and shareholders' equity inputs cannot be reconciled

consistently from the released SEC evidence without unsupported imputation.

E. Available Phase-2 findings

The release establishes broad construction feasibility and one explicit non-estimability result. It does not persist complete factor IC, rank IC, standard error, t-statistic, and false-discovery evidence sufficient for H1. H1 therefore remains inconclusive.

VII. ASSET-PRICING MODELS AND RESULTS

A. CAPM

CAPM supplies the market-only benchmark [2]. Its mean adjusted R-squared is 0.063 across 123 diagnostic portfolios.

B. Fama-French 3

The three-factor specification adds size and value [3].

C. Carhart 4

Momentum augments the three-factor model following Carhart [4].

D. Fama-French 5

Profitability and investment extend the three-factor specification [5].

E. Configured q mapping

The project mapping uses log market capitalization, asset growth, and earnings yield as ME, IA, and ROE proxies. It is a study-specific implementation of q-motivated families, not a claim of exact replication of Hou, Xue, and Zhang [6]. No standalone custom model was estimated.

F. OLS and HAC inference

Coefficients are estimated by OLS. The reported intercept t-statistics use the released HAC/Newey-West covariance. The $p < .05$ column below is the share of the 123 diagnostic-portfolio intercept tests whose released HAC p-value is below

0.05. These in-sample intercepts are not interpreted as investable alpha. GRS and spanning statistics are unavailable in the released artifacts.

G. Empirical model comparison

TABLE VII. ASSET-PRICING MODEL COMPARISON AND HAC INTERCEPT INFERENCE

Model	N	Mean adj. R ²	Median adj. R ²	Mean monthly alpha (decimal)	Median monthly alpha (decimal)	Mean HAC t	p<.05	Mean market beta
CAPM	123	0.063	0.061	0.01187	0.01128	2.467	85.4%	1.402
Fama-French 3	123	0.259	0.247	0.00923	0.00937	2.146	58.5%	1.421
Carhart 4	123	0.277	0.261	0.00899	0.00910	2.048	52.8%	1.408
Fama-French 5	123	0.301	0.281	0.01056	0.01026	2.569	92.7%	1.284
Configured q mapping	123	0.296	0.295	0.00875	0.00876	2.090	58.5%	1.291

Notes: The $p < .05$ column reports the unadjusted share of the 123 portfolio-specific intercept tests whose released HAC p-value is below 0.05. It is descriptive and is not a multiple-testing-adjusted decision rule. Intercepts are not interpreted as evidence of persistent investable alpha.

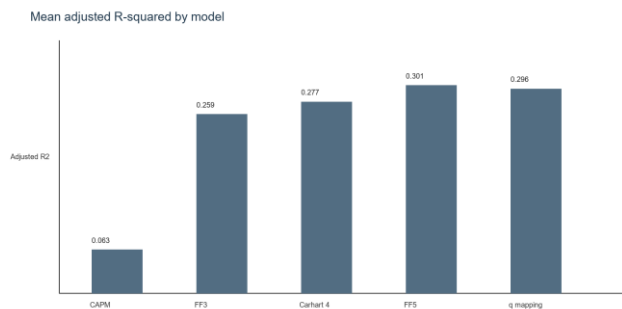


Fig. 3. Mean adjusted R-squared by asset-pricing model.

FF5 exceeds CAPM by $0.301 - 0.063 = 0.238$ in mean adjusted R-squared. The comparison supports improved in-

sample explanatory fit, not persistence or causal pricing success.

VIII. PORTFOLIO CONSTRUCTION, COSTS, AND RISK

A. Eight methods

The comparison includes equal weight, minimum variance, mean variance, maximum Sharpe, maximum diversification, risk parity, hierarchical risk parity, and CVaR.

B. Constraints

All methods are long-only, fully invested, and unlevered, with monthly rebalancing. No separate maximum-position or turnover cap was active in the released Phase-4 configuration. The tests use factor portfolios rather than direct investable securities.

C. Transaction costs

TABLE VIII. EX-ANTE ONE-WAY TRANSACTION-COST ASSUMPTIONS (BASIS POINTS)

Component	LOW	BASE	HIGH
Commission	0	0	0
Spread	2	5	10
Slippage	1	2	4
Fixed impact allowance	2	3	6
Total	5	10	20

These are assumptions fixed before result inspection, not observed broker executions. Nonlinear market impact was not independently estimated.

D. Performance

TABLE IX. BASE PORTFOLIO PERFORMANCE, 2024-01-31 TO 2026-08-04

Method	Annualized net return	Volatility	Sharpe	Maximum drawdown	Net cumulative
Equal weight	16.10%	11.14%	1.383	-8.57%	48.89%
Minimum variance	13.95%	10.00%	1.341	-6.69%	41.64%
Mean variance	12.72%	16.04%	0.818	-17.81%	37.60%
Maximum Sharpe	20.55%	12.30%	1.575	-7.51%	64.61%
Maximum diversification	17.31%	10.97%	1.499	-5.43%	53.08%
Risk parity	15.76%	10.79%	1.397	-8.33%	47.72%
Hierarchical risk parity	15.48%	10.56%	1.403	-8.13%	46.79%
CVaR	25.48%	14.21%	1.667	-8.55%	83.17%

Phase-4 BASE return-volatility comparison

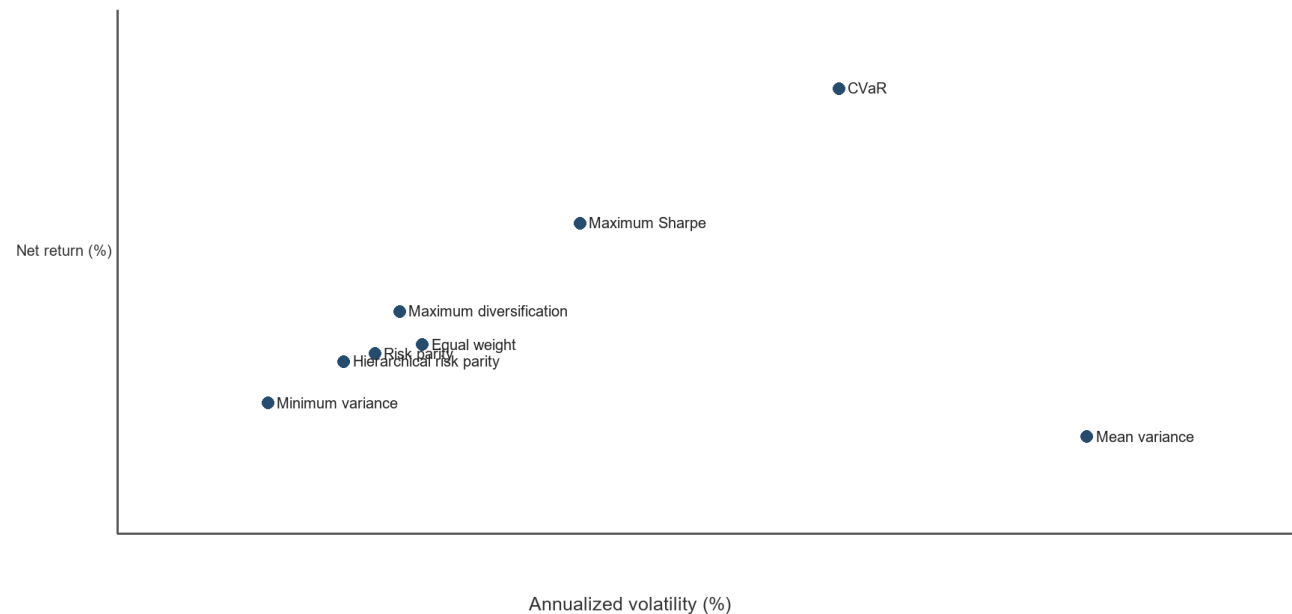


Fig. 4. Phase-4 BASE return-volatility comparison.

E. Cost sensitivity

TABLE X. CUMULATIVE-RETURN SENSITIVITY TO FROZEN COSTS

Method	Gross	LOW net	BASE net	HIGH net
Equal weight	49.12%	49.01%	48.89%	48.66%
Minimum variance	42.96%	42.30%	41.64%	40.34%
Mean variance	41.45%	39.52%	37.60%	33.85%
Maximum Sharpe	66.66%	65.63%	64.61%	62.59%
Maximum diversification	54.13%	53.60%	53.08%	52.03%
Risk parity	47.98%	47.85%	47.72%	47.47%
Hierarchical risk parity	47.42%	47.10%	46.79%	46.17%
CVaR	84.30%	83.73%	83.17%	82.06%

Cumulative-return cost sensitivity

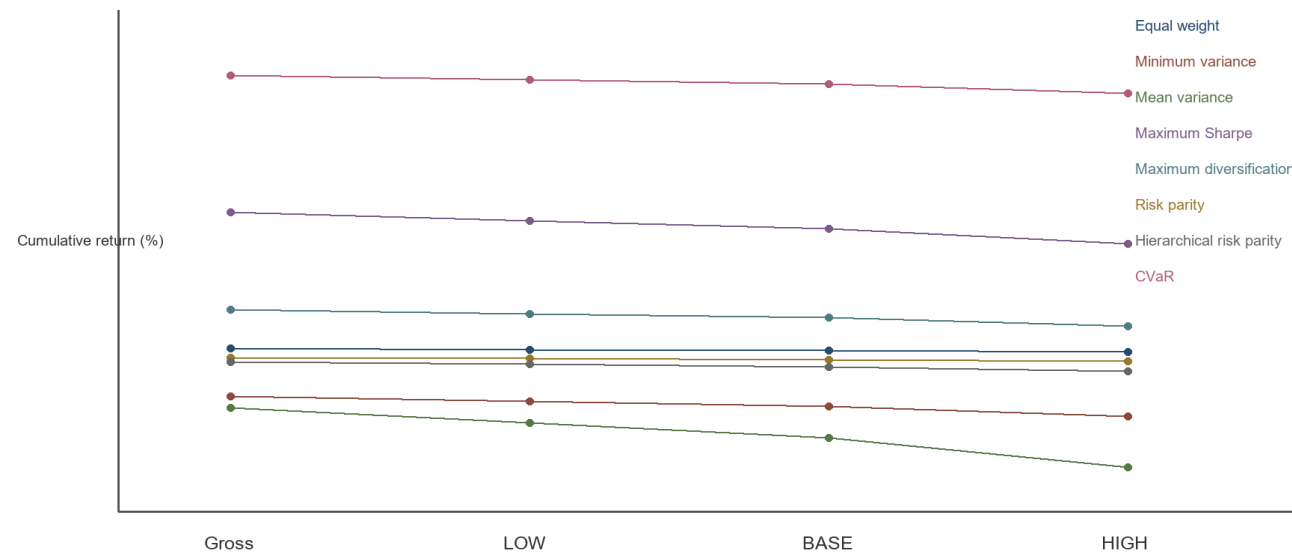


Fig. 5. Cumulative-return cost sensitivity by method.

Higher costs reduce every cumulative result. Equal weight's BASE drag is $49.12\% - 48.89\% = 0.23$ percentage points. Mean variance's HIGH drag is $41.45\% - 33.85\% = 7.60$ percentage points.

F. Risk metrics

TABLE XI. EXACT RELEASED BASE RISK METRICS

Method	Net cum.	Gross cum.	Vol.	Sharpe	Sortino	MDD	VaR95	CVaR95	Track err.	Info ratio	Beta
Equal Weight	48.89%	49.12%	11.14%	1.383	2.441	-8.57%	4.92%	5.38%	5.20%	-0.850	0.833
Minimum Variance	41.64%	42.96%	10.00%	1.341	2.136	-6.69%	4.55%	6.09%	10.03%	-0.641	0.498
Mean Variance	37.60%	41.45%	16.04%	0.818	1.153	-17.81%	4.53%	10.65%	13.55%	-0.495	0.753
Maximum Sharpe	64.61%	66.66%	12.30%	1.575	2.967	-7.51%	4.65%	5.85%	10.67%	-0.044	0.628
Maximum Diversification	53.08%	54.13%	10.97%	1.499	2.935	-5.43%	4.28%	4.94%	9.30%	-0.364	0.616
Risk Parity	47.72%	47.98%	10.79%	1.397	2.408	-8.33%	4.89%	5.39%	5.41%	-0.879	0.799
Hierarchical Risk Parity	46.79%	47.42%	10.56%	1.403	2.382	-8.13%	4.84%	5.43%	5.68%	-0.884	0.772
CVaR	83.17%	84.30%	14.21%	1.667	3.163	-8.55%	5.05%	7.58%	12.90%	0.298	0.622

Notes: VaR95 and CVaR95 are reported as positive loss magnitudes under the project's loss convention; maximum drawdown is reported as a negative peak-to-trough return. Tracking error and information ratio are benchmark-relative.

G. Scenarios

TABLE XII. RELEASED INSTANTANEOUS SCENARIO IMPACTS

Scenario family	Impact
Market	-0.244%
Rates	-0.073%
Volatility	-0.122%
Inflation	-0.061%
Liquidity	-0.037%

Scenario impacts are exposure mappings, not forecasts. A sixth project-specific custom mapping is retained in the authenticated scenario publication but is not interpreted in this summary because its shock definition is not stated in the manuscript; no generic interpretation is inferred.

IX. WALK-FORWARD MACHINE LEARNING

A. Features and target

The final feature set contains book-to-market and 12-minus-1 momentum. The target is the security's 21-trading-session total return minus SPY's return over the same horizon.

B. Temporal split

The design has 60 training months, 12 validation months, and one test month, with monthly retraining. Purging removes overlapping label horizons and a one-month embargo separates validation from testing. The released decision panel

contains 27,640 observations across 91 monthly dates; 17 complete out-of-sample folds cover 28 February 2025 through 30 June 2026.

C. Models

The eight rows are zero, historical mean, factor composite, linear, Ridge, Elastic Net, Random Forest, and XGBoost. Constant predictions make rank correlation undefined, so both constant baselines report IC as N/A rather than zero.

D. Tuning/search

Ridge alpha is selected from [0.1, 1.0, 10.0]. Elastic Net uses alpha [0.0001, 0.001] and l1_ratio [0.25, 0.5, 0.75]. Random Forest uses n_estimators [50, 100] and max_depth [3, 6]. XGBoost uses n_estimators [50, 100], max_depth [2, 3], and learning_rate [0.05, 0.1]. Validation Spearman information coefficient (IC) is the model-selection criterion. Linear preprocessing is fit on training data only; seed 17 and one thread are fixed.

E. Bootstrap methodology

The 17 fold-level Spearman IC observations are resampled with replacement 1,000 times to construct 95% bootstrap confidence intervals using random seed 17. The small number of complete folds is reflected in the width of these intervals.

F. Predictive results

TABLE XIII. EXACT EIGHT-MODEL PREDICTIVE RESULTS

Model	Mean IC	95% bootstrap CI	Positive IC folds	Positive rate
Zero baseline	N/A	N/A	N/A	N/A
Historical mean	N/A	N/A	N/A	N/A
Factor composite	0.017412	[-0.026401, 0.110608]	10/17	58.8%
Linear	0.017412	[-0.026401, 0.110608]	10/17	58.8%
Ridge	0.017414	[-0.026410, 0.110637]	10/17	58.8%
Elastic Net	0.027535	[-0.024701, 0.151649]	11/17	64.7%
Random Forest	0.009558	[-0.074074, 0.131357]	10/17	58.8%
XGBoost	-0.009069	[-0.049268, 0.055085]	8/17	47.1%

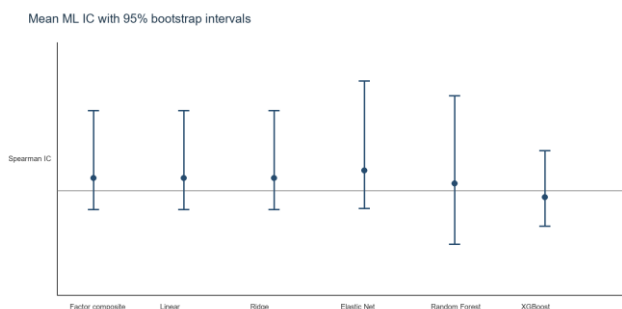


Fig. 6. Mean ML information coefficients with 95% bootstrap intervals.

Every non-constant model interval includes zero. Stable predictive information is not established.

G. Economic results

TABLE XIV. EXACT EIGHT-MODEL AFTER-COST ECONOMIC FOLD RESULTS

Model	Mean net fold return (%)	Median net fold return (%)	Positive economic folds
Zero baseline	-0.1415%	-0.1376%	8/17
Historical mean	-0.1415%	-0.1376%	8/17
Factor composite	-0.0973%	-0.0849%	7/17
Linear	-0.0973%	-0.0849%	7/17
Ridge	-0.0973%	-0.0849%	7/17
Elastic Net	-0.1039%	-0.0333%	7/17
Random Forest	+0.0074%	+0.0942%	9/17
XGBoost	-0.0044%	+0.1838%	11/17

Notes: Mean and median net fold returns are reported in percentage terms. The frozen model outputs are decimal returns and are multiplied by 100 for display. Positive economic folds report the number of positive fold-level outcomes among the 17 complete out-of-sample folds.

Random Forest has a small positive mean and 9 positive folds, while XGBoost has a positive median but a negative mean. Neither pattern establishes stable after-cost incremental value.

H. Explanations

Linear-family coefficients rank book-to-market above momentum. Random Forest permutation importance ranks book-to-market first while its SHAP summary ranks momentum first. XGBoost ranks book-to-market first under both methods. These are fitted-model attribution measures, not causal estimates; the inconsistency supports H7 as not supported.

X. ROBUSTNESS AND RESEARCH-INTEGRITY CHECKS

The released design includes a common benchmark calendar, LOW/BASE/HIGH cost sensitivity, multiple portfolio methods, six scenario mappings, constant and model-based ML challengers, purged temporal folds, bootstrap intervals, feature ablations, drift diagnostics, and point-in-time joins. The released implementation also passed software-integrity, restart, and publication-authentication checks; technical details are reported in Appendix J.

XI. HYPOTHESIS ASSESSMENT

TABLE XV. FINAL HYPOTHESIS ASSESSMENT

Hypothesis	Classification	Basis
H1	Inconclusive	Complete factor-level cost and multiplicity decision is not persisted.
H2	Partially supported	Multifactor models improve in-sample adjusted R-squared; stable out-of-sample economic improvement is not established.
H3	Inconclusive	Rolling evidence exists, but the complete original temporal-stability decision is absent.
H4	Supported	Every higher cost setting reduces cumulative results.
H5	Partially supported	Methods alter observed risk-return profiles; durable superiority is not established.
H6	Not supported	All non-constant IC intervals include zero and economic gains are unstable.
H7	Not supported	Attribution rankings vary across methods and are noncausal.

XII. DISCUSSION

A. Factor and asset-pricing evidence

The selected public-data panel supports broad characteristic construction, but feasibility is not significance. Multifactor models explain more in-sample variation than CAPM on the project's diagnostic portfolios. The result is

consistent with established factor structures, although the sample and test assets limit external validity.

B. Portfolio evidence

Allocation rules produce materially different observed volatility, drawdown, benchmark tracking, and tail risk. CVaR has the highest observed cumulative return in the short window, while minimum variance has the lowest volatility;

neither is labelled the best strategy because persistence has not been established.

C. Transaction costs

The monotonic cost results are the most direct economic finding. They show that even simple pre-specified allowances change cumulative comparisons, especially for turnover-sensitive allocations. Because the costs are assumptions rather than executions, they measure sensitivity rather than realized implementation shortfall.

D. Why the negative ML result matters

The ML evidence resists a common temptation to tune until a positive narrative emerges. With only two features and 17 folds, estimates are imprecise; all IC intervals include zero. Preserving that outcome distinguishes an empirical test from a model-selection exercise optimized for presentation.

E. Implications for public-data research

Public sources can support useful, traceable empirical work when availability and source changes are explicit. They cannot substitute for absent upstream securities or standardized proprietary histories. The appropriate contribution is therefore source-bounded evidence, not equivalence to CRSP, Compustat, or historical-index databases.

XIII. LIMITATIONS

- The source-selected universe is not CRSP-, Compustat-, historical S&P 500-, or Russell-replication evidence.
- HF Data Library warns that its pre-2022 fixed universe omits provider-absent delisted firms; project controls cannot reconstruct them [27].
- The March-2022 PiTrading-to-IEX transition creates a volume and liquidity break; IEX represents approximately 2–3% of consolidated volume [27].
- Cross-sectional breadth is sparse in 2018–2019, and SEC concepts remain heterogeneous despite point-in-time availability.
- Equity issuance is not estimable, and unresolved identity, terminal-event, share, and splice candidates are excluded.
- Phase 4 covers only 32 months and uses factor-portfolio test assets. Costs are assumed rather than observed, and nonlinear market impact is not estimated.
- Phase 5 has two features and 17 complete folds; all aggregate IC intervals include zero. Explanation methods are noncausal.
- H1 and H3 remain incomplete under their original decision rules.

XIV. FUTURE RESEARCH

Future work should use a point-in-time, delisting-inclusive security master; independently reconcile consolidated volume across the 2022 source boundary; extend the out-of-sample portfolio window; incorporate observed execution data where lawfully available; and add pre-specified characteristics only after the present evidence is frozen. A larger fold count would

permit more informative ML uncertainty estimates. None of these extensions is inferred from the current results.

XV. CONCLUSION

A connected public-data pipeline can support broad factor construction, robust asset-pricing comparison, constrained portfolio analysis, and temporally ordered prediction while preserving uncertainty. In this sample, multifactor models improve in-sample explanatory fit, portfolio methods produce different observed risk profiles, and higher assumed costs reduce every cumulative result. The analysis does not establish stable factor significance under H1, temporal stability under H3, persistent portfolio superiority, or stable machine-learning value. Those boundaries are part of the result.

DATA AND CODE AVAILABILITY

Project repository:
<https://github.com/amitdudi04/institutional-multi-factor-asset-pricing-platform>

Empirical tag: project-complete-public-data-v1

Empirical commit:
5a3e930665756fa7aaedecb5f9ab90792bcf849

Code and configuration are public. Third-party data are not redistributed, and exact reproduction requires lawful access to the source evidence. The repository's MIT licence does not supersede source licences. Publication identifiers and checksums support artifact-level verification.

FUNDING

This research received no external funding.

COMPETING INTERESTS

The author declares no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

AI-assisted tools were used for limited technical and editorial support, including coding assistance, debugging, document formatting, and language refinement. The research questions, data-source selection, empirical methodology, model specifications, validation procedures, analysis, interpretation of results, and conclusions were determined and verified by the author. The author reviewed the generated outputs and takes full responsibility for the final manuscript.

ACKNOWLEDGMENT

This research was conducted under the academic supervision of Mitul Sudhirkumar Nagar, Assistant Professor, Department of Computational Intelligence, SRM Institute of Science and Technology.

REFERENCES

- [1] H. Markowitz, "Portfolio selection," *Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952. <https://doi.org/10.1111/j.1540-6261.1952.tb01525.x>.
- [2] W. F. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk," *Journal of Finance*, vol. 19, no. 3, pp. 425–442, 1964. <https://doi.org/10.1111/j.1540-6261.1964.tb02865.x>.
- [3] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 33, no. 1, pp. 3–56, 1993. [https://doi.org/10.1016/0304-405X\(93\)90023-5](https://doi.org/10.1016/0304-405X(93)90023-5).
- [4] M. M. Carhart, "On persistence in mutual fund performance," *Journal of Finance*, vol. 52, no. 1, pp. 57–82, 1997. <https://doi.org/10.1111/j.1540-6261.1997.tb03808.x>.

- [5] E. F. Fama and K. R. French, "A five-factor asset pricing model," *Journal of Financial Economics*, vol. 116, no. 1, pp. 1–22, 2015. <https://doi.org/10.1016/j.jfineco.2014.10.010>.
- [6] K. Hou, C. Xue, and L. Zhang, "Digesting anomalies: An investment approach," *Review of Financial Studies*, vol. 28, no. 3, pp. 650–705, 2015. <https://doi.org/10.1093/rfs/hhu068>.
- [7] V. DeMiguel, L. Garlappi, and R. Uppal, "Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?," *Review of Financial Studies*, vol. 22, no. 5, pp. 1915–1953, 2009. <https://doi.org/10.1093/rfs/hhm075>.
- [8] S. Maillard, T. Roncalli, and J. Teiletche, "The properties of equally weighted risk contribution portfolios," *Journal of Portfolio Management*, vol. 36, no. 4, pp. 60–70, 2010. <https://doi.org/10.3905/jpm.2010.36.4.060>.
- [9] M. López de Prado, "Building diversified portfolios that outperform out of sample," *Journal of Portfolio Management*, vol. 42, no. 4, pp. 59–69, 2016. <https://doi.org/10.3905/jpm.2016.42.4.059>.
- [10] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *Journal of Risk*, vol. 2, no. 3, pp. 21–41, 2000. doi:10.21314/JOR.2000.038.
- [11] H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, vol. 48, no. 4, pp. 817–838, 1980. <https://doi.org/10.2307/1912934>.
- [12] W. K. Newey and K. D. West, "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, vol. 55, no. 3, pp. 703–708, 1987. <https://doi.org/10.2307/1913610>.
- [13] C. R. Harvey, Y. Liu, and H. Zhu, "... and the cross-section of expected returns," *Review of Financial Studies*, vol. 29, no. 1, pp. 5–68, 2016. <https://doi.org/10.1093/rfs/hhv059>.
- [14] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970. <https://doi.org/10.1080/00401706.1970.10488634>.
- [15] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.
- [16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. <https://doi.org/10.1023/A:1010933404324>.
- [17] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD 2016*, pp. 785–794, 2016. <https://doi.org/10.1145/2939672.2939785>.
- [18] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *Review of Financial Studies*, vol. 33, no. 5, pp. 2223–2273, 2020. <https://doi.org/10.1093/rfs/hhaa009>.
- [19] D. H. Bailey, J. M. Borwein, M. López de Prado, and Q. J. Zhu, "Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance," *Notices of the AMS*, vol. 61, no. 5, pp. 458–471, 2014.
- [20] M. López de Prado, *Advances in Financial Machine Learning*. Wiley, 2018.
- [21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [22] A. Elkassabgi, "HF Data Library: High-Frequency U.S. Equity Data (Version 1.0)," [Dataset], Zenodo, 2026. <https://doi.org/10.5281/zenodo.19501605>.
- [23] U.S. Securities and Exchange Commission, "EDGAR application programming interfaces," 2026. [Online]. Available: <https://www.sec.gov/search-filings/edgar-application-programming-interfaces>.
- [24] Alpha Vantage, "API documentation: LISTING_STATUS," 2026. [Online]. Available: <https://www.alphavantage.co/documentation/>.
- [25] Board of Governors of the Federal Reserve System (US), "Market Yield on U.S. Treasury Securities at 3-Month Constant Maturity, Quoted on an Investment Basis [DGS3MO]," FRED, Federal Reserve Bank of St. Louis, 2026. Retrieved August 26, 2026, from <https://fred.stlouisfed.org/series/DGS3MO>.
- [26] K. R. French, "Data Library," Dartmouth College. Retrieved August 26, 2026. Kenneth R. French - Data Library.
- [27] HF Data Library, "Documentation: Data sources, source transition, and universe construction & survivorship," Retrieved August 26, 2026. 1-Minute Stock Data: Cleaning Methodology - HF Data.

APPENDIX A. COMPLETE 48-FACTOR DEFINITION CATALOG

ID	Family	Inputs	Conceptual implementation	Estimability
market_return	Market	market_return	approved broad-market total return	ESTIMABLE
excess_return	Market	market_return, risk_free	market total return - risk-free return	ESTIMABLE
risk_free_rate	Market	risk_free	frequency-matched approved risk-free simple return	ESTIMABLE
rolling_market_beta	Market	return, market_return	rolling $\text{cov}(\text{return}, \text{market}) / \text{var}(\text{market})$	ESTIMABLE
market_cap	Size	price, shares_outstanding	price * shares_outstanding	ESTIMABLE
log_market_cap	Size	price, shares_outstanding	$\ln(\text{market capitalization})$	ESTIMABLE
size_small	Size	market_cap, exchange	market cap $\leq$ NYSE 30th percentile	ESTIMABLE
size_mid	Size	market_cap, exchange	NYSE 30th < market cap < NYSE 70th	ESTIMABLE
size_large	Size	market_cap, exchange	market cap $\geq$ NYSE 70th percentile	ESTIMABLE
book_to_market	Value	book_equity, market_cap	book equity / market capitalization	ESTIMABLE
price_to_book	Value	book_equity, market_cap	market capitalization / book equity	ESTIMABLE
earnings_yield	Value	net_income, market_cap	net income / market capitalization	ESTIMABLE
cash_flow_yield	Value	operating_cash_flow, market_cap	operating cash flow / market capitalization	ESTIMABLE
dividend_yield	Value	dividends, market_cap	dividends / market capitalization	ESTIMABLE
momentum_1m	Momentum	return	compound return over 21 sessions	ESTIMABLE
momentum_3m	Momentum	return	compound return over 63 sessions	ESTIMABLE
momentum_6m	Momentum	return	compound return over 126 sessions	ESTIMABLE

ID	Family	Inputs	Conceptual implementation	Estimability
momentum_12_1m	Momentum	return	compound return months 12 through 2	ESTIMABLE
momentum_12m	Momentum	return	compound return over 252 sessions	ESTIMABLE
momentum_24m	Momentum	return	compound return over 504 sessions	ESTIMABLE
residual_momentum	Momentum	return, market_return	compound residual return after rolling market beta	ESTIMABLE
roe	Profitability	net_income, shareholder equity	net income / shareholder equity	ESTIMABLE
roa	Profitability	net income, total assets	net income / total assets	ESTIMABLE
gross_profitability	Profitability	gross profit, total assets	gross profit / total assets	ESTIMABLE
operating_margin	Profitability	operating income, revenue	operating income / revenue	ESTIMABLE
net_margin	Profitability	net income, revenue	net income / revenue	ESTIMABLE
asset_turnover	Profitability	revenue, average assets	revenue / average assets	ESTIMABLE
accruals	Profitability	total accruals, total assets	total accruals / total assets	ESTIMABLE
leverage	Investment / leverage	total debt, total assets	total debt / total assets	ESTIMABLE
debt_to_equity	Investment / leverage	total debt, shareholder equity	total debt / shareholder equity	ESTIMABLE
interest_coverage	Investment / leverage	operating income, interest expense	operating income / interest expense	ESTIMABLE
asset_growth	Investment / leverage	total assets, prior total assets	total assets / prior total assets - 1	ESTIMABLE
capex_growth	Investment / leverage	capex, prior capex	capital expenditure / prior capital expenditure - 1	ESTIMABLE
equity_issuance	Investment / leverage	net equity issuance, shareholder equity	net equity issuance / shareholder equity	NOT ESTIMABLE
working_capital_growth	Investment / leverage	working capital, prior working capital	working capital / prior working capital - 1	ESTIMABLE
rolling_volatility	Risk / liquidity	return	annualized rolling standard deviation	ESTIMABLE
downside_volatility	Risk / liquidity	return	annualized rolling standard deviation of negative returns	ESTIMABLE
semi_variance	Risk / liquidity	return	annualized mean squared negative return	ESTIMABLE
beta	Risk / liquidity	return, market_return	rolling market beta	ESTIMABLE
idiosyncratic_volatility	Risk / liquidity	return, market_return	annualized volatility of market residual	ESTIMABLE
average_dollar_volume	Risk / liquidity	price, volume	rolling mean price * volume	ESTIMABLE
turnover	Risk / liquidity	volume, shares outstanding	volume / shares outstanding	ESTIMABLE
amihud_illiquidity	Risk / liquidity	return, price, volume	abs(return) / dollar volume	ESTIMABLE
bid_ask_proxy	Risk / liquidity	high, low	(high-low) / midpoint	ESTIMABLE
rolling_beta	Risk / liquidity	return, market_return	rolling market beta	ESTIMABLE
rolling_correlation	Risk / liquidity	return, market_return	rolling correlation with market	ESTIMABLE
tracking_error	Risk / liquidity	return, benchmark_return	annualized volatility of return-benchmark	ESTIMABLE
downside_beta	Risk / liquidity	return, market_return	cov(return, market)/var(market) on negative-market dates	ESTIMABLE

APPENDIX B. DETAILED UNIVERSE SCREENING AND ANNUAL BREADTH

Screening retained 487 and rejected 335 candidates after attribution, splice, identity, share-basis, terminal-event, asset-type, venue, and fundamental-evidence checks. Annual breadth is reproduced in Table III; the complete daily evidence remains in the frozen market publication.

Universe rejection reasons

Rejection reason	Count	Share of 822 candidates
Attribution	114	13.87%
Source splice	56	6.81%
Identity	68	8.27%

Rejection reason	Count	Share of 822 candidates
Shares	70	8.52%
Terminal event	16	1.95%
Split basis	7	0.85%
Asset type	2	0.24%
Non-US / venue	1	0.12%
Fundamentals	1	0.12%
Total rejected	335	40.75%

The large number of exclusions is a deliberate consequence of the study design: a candidate was removed when attribution, source continuity, effective-dated identity, share basis, terminal-event treatment, instrument eligibility, venue, or point-in-time fundamental evidence could not be defended from the available public sources.

Stage	Start	End	Frequency
Phase 1 market	2018-02-14	2026-08-04	Daily
Phase 1 fundamentals (available at)	2019-02-23	2026-08-15	Event time
Phase 2 factors	2018-02-14	2026-08-04	Daily/monthly formation

Stage	Start	End	Frequency
Phase 3 asset pricing	2018-03-29	2026-08-04	Monthly
Phase 4 portfolios	2024-01-31	2026-08-04	Monthly
Phase 5 decisions	2025-02-28	2026-06-30	Monthly

APPENDIX C. ADDITIONAL DESCRIPTIVE STATISTICS

Factor ID	N	Mean	SD	P1	P25	Median	P75	P99
book_to_mar ket	585,078	0.395473	0.440632	-0.201907	0.118791	0.292552	0.557899	1.982315
momentum_1 2_1m	593,210	0.124992	0.444258	-0.601507	-0.118453	0.069121	0.283342	1.645525
rolling_volati lity	706,194	0.389935	0.209722	0.157417	0.252073	0.329544	0.462254	1.167571

APPENDIX D. ASSET-PRICING INFERENCE NOTES

Table VII reports the persisted cross-model summary fields. The five empirical model specifications are defined in Appendix K, Equations (3)–(7). Classical, HC0–HC3, and HAC/Newey–West covariance outputs and rolling coefficients remain in the released Phase-3 artifacts. GRS and spanning statistics are not available in the frozen release and are not reconstructed here.

APPENDIX E. PORTFOLIO AND RISK RESULT COVERAGE

Tables IX–XII report the uniformly persisted cross-method performance, transaction-cost, risk, and scenario evidence used in the manuscript. HHI, turnover, and risk-contribution summaries are omitted because they are not uniformly persisted across all methods. No missing cross-method statistic is imputed.

APPENDIX F. TRANSACTION-COST AND WORKED CALCULATIONS

Calculation	Result
Acceptance	487 / 822 = 59.25%
Rejection	335 / 822 = 40.75%
FF5 fit improvement	0.301 - 0.063 = 0.238
Equal-weight BASE drag	49.12% - 48.89% = 0.23 percentage points
Mean-variance HIGH drag	41.45% - 33.85% = 7.60 percentage points

APPENDIX G. EIGHT-MODEL ML CONFIGURATION AND HYPERPARAMETER GRID

Tables XIII–XIV retain one row per model. Frozen grids are Ridge alpha [0.1, 1.0, 10.0]; Elastic Net alpha [0.0001, 0.001] and l1 ratio [0.25, 0.5, 0.75]; Random Forest

n_estimators [50, 100] and max_depth [3, 6]; XGBoost n_estimators [50, 100], max_depth [2, 3], learning_rate [0.05, 0.1].

APPENDIX H. REPRODUCIBILITY IDENTIFIERS AND

ARTIFACT LINEAGE

Stage / evidence	Reproducibility identifier
Phase 1 market data	factor_market_input-89848f8e3c34d53d0432f6e0
Phase 1 point-in-time fundamentals	factor_fundamental_input-962c8ada3e74e32b00568cd7
Phase 2 factor construction	factor-5a73616f131fca2102fa2ca4141ad64f
Phase 3 asset-pricing analysis	asset-pricing-c0a72fae5349f1048e18557250da1927
Phase 4 portfolio and risk analysis	Model-level identifiers retained in the repository reproducibility guide
Phase 5 walk-forward ML analysis	Model-level identifiers retained in the repository reproducibility guide
Phase 6 research reports	Report-level identifiers retained in the repository reproducibility guide
Frozen empirical Git tag	project-complete-public-data-v1
Frozen empirical Git commit	5a3e930665756fa7aaedecb5f9ab90792bcf849

The complete portfolio-method, machine-learning-model, and report-level artifact identifiers are retained in the public repository's versioned reproducibility guide.

APPENDIX I. DETAILED H1–H7 EVIDENCE MATRIX

Hypothesis	Required evidence	Available evidence	Classification
H1	At least one factor spread survives the complete pre-specified cost and multiplicity rule	Complete factor-level cost and multiplicity decision is not persisted.	Inconclusive
H2	Multifactor models improve explanatory fit relative to CAPM and that improvement extends to stable out-of-sample economic evidence.	Multifactor models improve in-sample adjusted R-squared; stable out-of-sample economic improvement is not established.	Partially supported
H3	Estimated relations are stable through time under the original rule	Rolling evidence exists, but the complete original temporal-stability decision is absent.	Inconclusive
H4	Higher transaction-cost assumptions reduce cumulative outcomes	Every higher cost setting reduces cumulative results.	Supported
H5	Constrained portfolio methods alter risk-return trade-offs and any apparent superiority is stable out of sample.	Methods alter observed risk-return profiles; durable superiority is not established.	Partially supported

Hypothesis	Required evidence	Available evidence	Classification
H6	ML adds stable OOS predictive and after-cost value	All non-constant IC intervals include zero and economic gains are unstable.	Not supported
H7	Explanation rankings are stable across methods	Attribution rankings vary across methods and are noncausal.	Not supported

APPENDIX J. SOFTWARE ASSURANCE AND RUNTIME VERIFICATION

The latest post-remediation assurance records 301 passing tests, 90.52% branch-aware coverage, and 143 retained publications at the final restart check. Lint, formatting, strict type checks, lock verification, catalog validation, restart checks, and publication authentication passed. Runtime remediation addressed discovery timing, concurrent DuckDB lock collisions, and stale publication ordering without changing Phase 1–5 empirical results. Docker and GitHub billing constraints are operational matters, not empirical limitations.

APPENDIX K. FORMULA REFERENCE

Equations (1)–(26) summarize the empirical transformations referenced in the study. Symbols follow the definitions in the corresponding methods sections; these formulas do not introduce additional empirical estimates.

$$Rank(i, t) = rank[X(i, t)] / N(t) \quad (1)$$

$$F(t) = R(high, t) - R(low, t) \quad (2)$$

$$R(p, t) - R(f, t) = \alpha(p) + \beta(M, p)[R(M, t) - R(f, t)] + \epsilon(p, t) \quad (3)$$

$$R(p, t) - R(f, t) = \alpha + \beta(M)MKT + \beta(S)SMB + \beta(H)HML + \epsilon(p, t) \quad (4)$$

$$Carhart4 = FF3 + \beta(MOM)MOM \quad (5)$$

$$FF5 = \alpha + \beta(M)MKT + \beta(S)SMB + \beta(H)HML + \beta(R)RMW + \beta(C)CMA + \epsilon(p, t) \quad (6)$$

$$R_{p, t} - R_{f, t} = \alpha_p + \beta_M MKT_t + \beta_{ME} ME_t + \beta_{IA} IA_t + \beta_{ROE} ROE^{\text{proxy}}_t + \epsilon_{p, t} \quad (7)$$

where ME = log(market capitalization), IA = asset growth, and ROE^{proxy} = earnings yield. These are study-specific mappings and are not presented as an exact replication of Hou, Xue, and Zhang [6].

$$\beta = (X^T X)^{-1} X^T y \quad (8)$$

$$R_{adj}^2 = 1 - (1 - R^2) \left(\frac{n - 1}{n - k - 1} \right) \quad (9)$$

$$\min_w w' \Sigma w \quad (10)$$

$$\max_w w' \mu - \left(\frac{\lambda}{2} \right) w' \Sigma w \quad (11)$$

$$\max_w \left(\frac{w' \mu - R_f}{\sqrt{w' \Sigma w}} \right) \quad (12)$$

$$DR(w) = \frac{w' \sigma}{\sqrt{w' \Sigma w}} \quad (13)$$

$$\text{Risk contribution: } RC_i = \frac{w_{i(\Sigma w)_i}}{\sqrt{w^T \Sigma w}} \quad (14)$$

$$Turnover_t = \frac{1}{2} \sum_i |w_{i, t}(target) - w_{i, t}(pre-trade)| \quad (15)$$

$$Cost(t) = Turnover(t) \times c(t) \quad (16)$$

$$R(net, t) = R(gross, t) - Cost(t) \quad (17)$$

$$\text{Value at Risk: } VaR_\alpha(L) = \inf\{\ell : P(L \leq \ell) \geq \alpha\} \quad (18)$$

$$\text{Conditional Value at Risk: } CVaR_\alpha(L) = E[L | L \geq VaR_\alpha(L)] \quad (19)$$

Here, L denotes portfolio loss, and VaR and CVaR follow the loss-sign convention used in the empirical analysis.

$$\sigma_{annual} = \sigma_{period} \sqrt{K} \quad (20)$$

$$Sharpe = \left(\frac{E[R_p] - R_f}{\sigma(R_p - R_f)} \right) \sqrt{K} \quad (21)$$

$$MDD = \min(t) \left\{ \frac{V(t)}{\max(s \leq t) V(s)} - 1 \right\} \quad (22)$$

$$Y(i, t) = R(i, t \text{ to } t + 21) - R(SPY, t \text{ to } t + 21) \quad (23)$$

$$Ridge: \min_{\beta} \{ \sum_i (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_2^2 \} \quad (24)$$

$$Elastic \ Net: \min_{\beta} \{ \sum_i (y_i - x_i^T \beta)^2 + \lambda [a \|\beta\|_1 + (1 - a) \|\beta\|_2^2] \} \quad (25)$$

where λ controls the overall regularization strength and a determines the relative contribution of the L1 and L2 penalties.

$$IC_t = \rho_s(\hat{y}_{i,t}, y_{i,t}) \quad (26)$$

The information coefficient is the cross-sectional Spearman rank correlation between model predictions and realized 21-session benchmark-relative returns at each test date.

APPENDIX L. NUMERICAL PROVENANCE STATEMENT

Numerical claims in this manuscript were checked against the frozen empirical publications and the project's controlled evidence ledger. The retained base ledger contains 225 controlled records, of which 221 are marked VERIFIED and four are explicitly marked NOT AVAILABLE. The descriptive statistics for book-to-market, 12-minus-1 momentum, and rolling volatility were subsequently recomputed directly from the frozen long-form factor publication; the corrected values reported in Table IV and Appendix C supersede the earlier descriptive-statistic entries. Manuscript numerical expressions are classified as direct, derived, methodological constants, or reference values.