



Beyond pooling: The role of correlation structures for contemporaneous and lagged cross-asset stock prediction

Singh A

Submitted: December 16, 2025, Revised: version 1, March 25, 2026, version 2, June 14, 2026, version 3, June 18, 2026

Accepted: June 28, 2026

Abstract

We examined whether the correlation structure of the training universe affects cross-asset prediction of daily equity returns. Market-neutral residual returns were constructed for 75 S&P 500 stocks across five sectors by regressing daily stock returns against SPY, with market betas estimated using only training-fold data to prevent look-ahead bias. Four training regimes were defined for each target stock using quartiles of pairwise residual correlations: high-correlation, medium-correlation, anti-correlation, and same-sector, plus a pooled baseline. Two pipelines were evaluated: one based on contemporaneous residual correlation and another based on lagged predictive correlation, $\text{Corr}(\varepsilon_i(t), \varepsilon_{\text{tar}}^{\text{res}}(t+1))$, estimated exclusively within each training fold. Ridge regression, Random Forest, and XGBoost were evaluated using three-fold walk-forward TimeSeriesSplit across 10 target stocks and 10 random seeds (9,000 total model fits). BCa bootstrap confidence intervals clustered by target stock showed that no pairwise regime differences were significant under contemporaneous construction, whereas under lagged construction the medium-correlation regime significantly outperformed the high- and anti-correlation regimes despite all regimes producing near-zero or negative out-of-sample R^2 . This pattern is consistent with unstable training-fold predictive correlations that fail to generalize out of sample rather than robust lead-lag structure. Together, both pipelines converge on the same conclusion: correlation structure does not reliably improve cross-asset forecasting performance for daily large-cap U.S. equities, consistent with the Efficient Market Hypothesis.

Keywords

Machine Learning, Cross-asset prediction, Stock return forecasting, Financial time series, Residual returns, Market neutralization, Lead-lag effects, Correlation structure, Time series forecasting, Efficient Market Hypothesis

Aayush Singh, MOT Charter High School, 1275 Cedar Lane Rd, Middletown, DE 19709, USA.

aayush.singh2023@gmail.com

1. Introduction

Cross-asset ML for stock return prediction has been studied extensively over the past decade. Several studies have shown that pooling training data across many stocks produces a better out-of-sample R^2 than single-asset models (1, 2), and large-scale work by Gu et al. (3) demonstrated that modern ML methods applied to hundreds of firm characteristics achieve out-of-sample R^2 of $\sim 0.3\text{--}0.4\%$ at the monthly frequency, which may be small in absolute terms but still is economically significant. With this in consideration, one question has received little attention: does the composition of the training universe matter, or does any sufficiently large training universe produce comparable performance?

The lead-lag literature suggests that training universe composition should matter. Lo and MacKinlay (4) documented that large-cap stock returns systematically lead small-cap returns, suggesting that cross-asset information transfer is real and structured. Cohen and Frazzini (5) showed that supplier-customer firm pairs exhibit return predictability matched with slow diffusion of fundamental news. Hou (6) found that intraindustry information diffusion drives the lead-lag effect between large and small firms, with neglected industries showing stronger effects. Hong, Torous, and Valkanov (7) demonstrated that 14 out of 34 U.S. industries forecast the aggregate market up to two months in advance. The literature implies that stocks with economic or informational links to the target should predict better.

The theoretical mechanism behind these empirical patterns is provided by Hong and

Stein (8), whose gradual information diffusion model showed that when information spreads slowly across the investor population, prices underreact and stocks sharing information sources exhibit lead-lag co-movement. Chordia and Swaminathan (9) showed that trading volume governs the speed of this diffusion, with low-volume stocks lagging behind high-volume peers. Both papers imply that correlation structure in residual returns, after removing common market sensitivity provided by SPY, reflects genuine economic linkages between assets.

To the author's knowledge, no prior work has directly tested how training universe composition affects cross-asset ML performance. Most work trains on the full cross-section (3, 10, 11) or individual assets; neither varies training universe composition in a controlled way. We systematically varied the correlation structure of the training universe while holding sample size constant to determine whether training-universe composition influences cross-asset predictive performance.

1.1 Study objectives and overview

This study investigates whether the composition of the training universe influences cross-asset predictive performance when training set size is held constant. Specifically, we test whether selecting training stocks based on their relationship to the target stock—high correlation, medium correlation, anti-correlation, or sector membership—produces different out-of-sample predictive performance than training on an equally sized randomly selected stock universe.

To address this question, we evaluate two complementary experimental pipelines. The first defines training regimes using contemporaneous residual correlations, whereas the second defines regimes using one-day-ahead predictive correlations, $\text{Corr}(\varepsilon_i(t), \varepsilon_{\text{tar}}^{\text{pe}}(t+1))$, estimated exclusively within each walk-forward training fold to eliminate information leakage. Comparing these two approaches allows us to determine whether training-universe composition based on contemporaneous co-movement or predictive lead-lag relationships provides a more effective basis for cross-asset forecasting. The overall experimental design therefore evaluates two independent dimensions of training-universe design: [1] training-set composition, defined by the correlation regime from which stocks are selected, and [2] the temporal structure used to define those regimes, namely contemporaneous versus lagged predictive correlations.

2. Methods

2.1 Stock universe and data collection

We collected daily open, high, low, close, and volume (OHLCV) data using the YFinance Python package (14) for 76 assets: 75 S&P 500 stocks across five GICS sectors plus SPY as the market benchmark. The five sectors and constituent stocks are: Information Technology (AAPL, MSFT, NVDA, AVGO, ORCL, AMD, ADBE, CRM, INTC, CSCO, QCOM, TXN, IBM, NOW, AMAT), Financials (JPM, BAC, WFC, GS, MS, BLK, C, SCHW, AXP, USB, PNC, TFC, COF, BK, MCO), Healthcare (LLY, UNH, JNJ, ABBV, MRK, TMO, ABT, DHR, PFE, BMY, AMGN, GILD, SYK, ISRG, CVS), Energy (XOM, CVX, COP, EOG, SLB, MPC,

PSX, VLO, WMB, OXY, OKE, DVN, BKR, FANG, HAL), and Consumer Discretionary (AMZN, TSLA, HD, MCD, NKE, LOW, SBUX, TJX, BKNG, CMG, GM, F, ORLY, DHI, LEN). All stocks are current or recent S&P 500 assets, selected to represent major liquid large-cap equities across economically distinct sectors.

Data were collected using `period='max'` to retrieve the full available history for each stock. After downloading, the dataset was trimmed to the latest common start date across all 76 assets and the earliest common end date, ensuring all correlation computations and residual estimates use identical calendar dates for all stocks. This avoided the structural distortion that would arise if some stocks contributed data from time periods not available for others, where apparent correlations would partly reflect regime differences rather than genuine cross-asset relationships (15).

2.2 Market neutralization via OLS residuals

To isolate stock-specific return components from systematic market movements, we estimated market betas for each stock using ordinary least squares (OLS) linear regression of daily stock returns against SPY daily returns. This single-factor market model, introduced by Sharpe (16) and formalized in the Fama-MacBeth (17) framework, decomposes each stock's return into a systematic component proportional to the market return and an idiosyncratic residual. Formally, for stock i on day t , $r_{i,t} = \alpha_i + \beta_i \times r_{\text{SPY},t} + \varepsilon_{i,t}$ (eq1), where β_i was estimated by OLS using only the training rows of each walk-forward fold. The residuals ε represent the portion of stock i 's return that is

unexplained by the market on that day. These residuals served as our primary return series throughout the study.

Without market neutralization, correlations between large-cap stocks would be dominated by their shared sensitivity to SPY, making most stock pairs appear highly correlated regardless of any genuine economic relationship (18). By removing the market component, residual correlations reflect real cross-asset informational linkages, which can actually generate lead-lag predictability as in the theoretical frameworks of Hong and Stein (8) and Chordia and Swaminathan (9). The Fama-French three-factor model would further remove size and value premiums, but our 75 stocks were all large-cap and our target was short-term daily returns, so the single-factor model is appropriate and more parsimonious.

2.3 Correlation matrix and regime construction

We ran two separate pipelines that differed only in how training regimes were constructed. Both pipelines shared the same model architectures, cross-validation scheme, and bootstrap evaluation procedure; the only difference was the correlation measure used to assign stocks to regimes.

2.3.1 Contemporaneous pipeline

We computed pairwise Pearson correlations between all residual return series, producing a 75×75 correlation matrix (Figure 1). Before computing quartile thresholds for regime assignment, we excluded the self-correlation of 1.0 from each stock's correlation vector, as including it would artificially inflate the 75th percentile threshold and misclassify borderline pairs. For each target stock, we defined three correlation-based training regimes using quartile thresholds applied to that stock's row of the correlation matrix:

[1] High-correlation regime: all stocks whose residual correlation with the target exceeded the 75th percentile of that target's distribution.

[2] Medium-correlation regime: all stocks whose residual correlation fell between the 25th and 75th percentiles.

[3] Anti-correlation regime: all stocks whose residual correlation fell below the 25th percentile. Across all 10 target stocks, the 25th percentile of residual correlations was itself negative (ranging from -0.003 to -0.167), meaning this regime consisted entirely of stocks with negative residual correlation to the target.

[4] Same-sector regime: contained all stocks in the same GICS sector as the target.

[5] All regime: randomly drawn sample of the full 74-stock universe (leave one out applied), used as a pooled baseline.

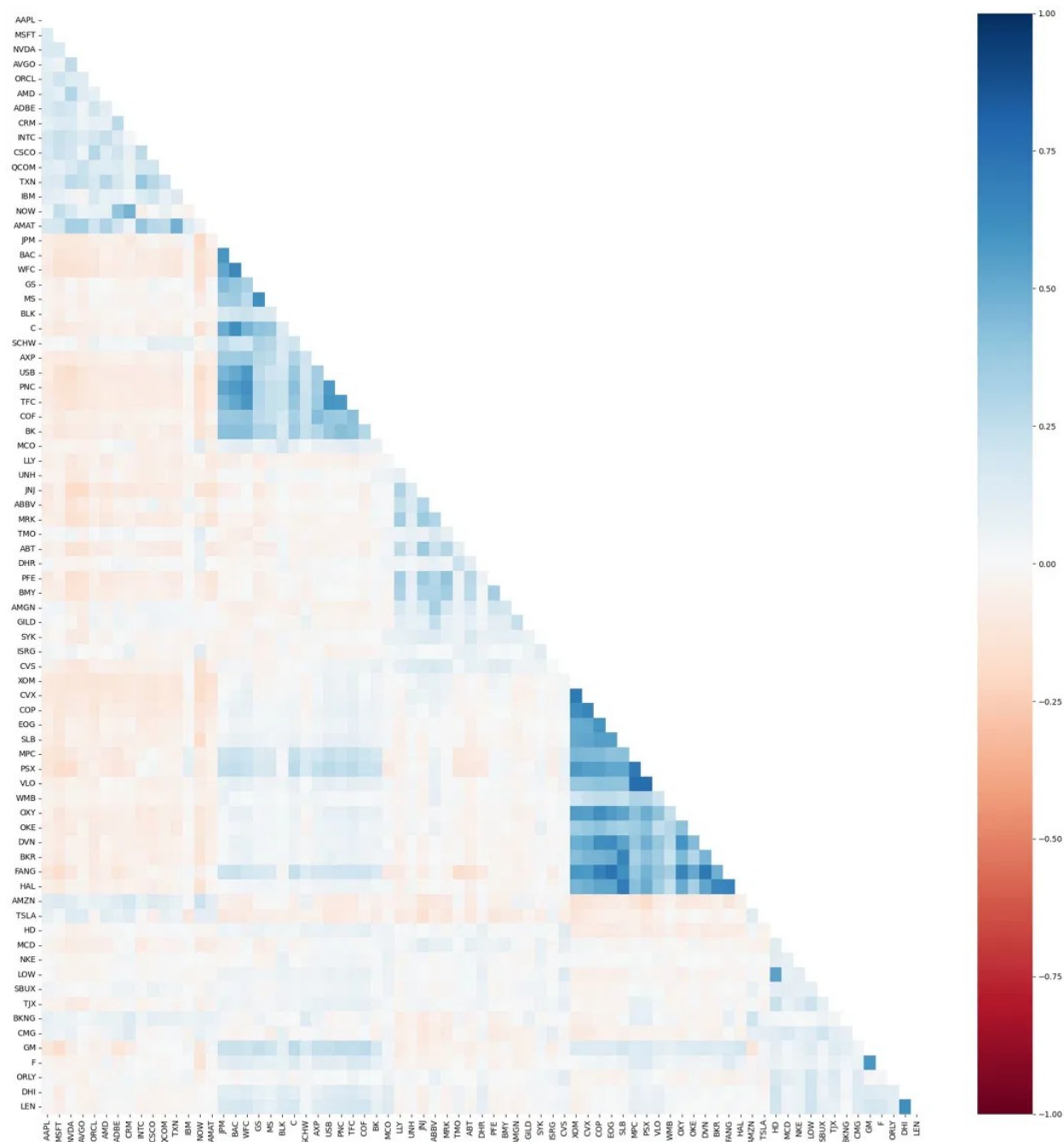


Figure 1. Correlation matrix with respect to the residuals of the relevant stock universe, with blue shades representing values with higher correlation and red shades representing values with lower/anti correlation. This is computed on training data only.

Figure 1 illustrates the pairwise residual residualization. Clear block-like regions of correlation matrix for the 75-stock universe after stronger positive correlation are visible within market neutralization using SPY several groups of stocks, whereas correlations

between many stocks from different sectors are substantially weaker and, in some cases, slightly negative. The market-neutralization procedure removes much of the broad market co-movement, resulting in a wider range of residual correlation values than would be expected using raw returns. These residual correlations provide the basis for assigning stocks to the high-correlation, medium-correlation, and anti-correlation training regimes used throughout the subsequent analyses. Because the correlation matrix is computed exclusively from training-fold data, the regime construction is free from look-ahead bias.

2.3.2 Lagged pipeline

Rather than grouping stocks by same-day co-movement, we grouped them by one-day-ahead predictive correlation, $\text{Corr}(\epsilon_i(t), \epsilon_{\text{tar}}^{\text{oe}}(t+1))$, estimated exclusively within each training fold to prevent leakage. The five regime categories were identical to the contemporaneous pipeline, with quartile thresholds now applied to this lagged correlation measure. This method tested whether stocks that historically led the target by one day produced better out-of-sample predictions than stocks selected by same-day co-movement.

Because quartile thresholds were computed per target stock, the high-correlation regime for a technology stock may contain healthcare or financial stocks if they happen to co-move strongly with the target after market neutralization, which is an intentional choice.

To control for sample size across regimes, we set the training set size for each regime equal to the minimum regime size across all five regimes

for that target stock, then randomly sampled that number of stocks from each regime without replacement. The same-sector regime was the binding constraint for most target stocks (typically 14 stocks, all sector peers minus the target itself). The all regime was also randomly sampled to the same size. This controlled for training set size, ensuring that any observed performance differences were attributable to training-set composition rather than the number of training stocks.

2.4 Feature construction and target variable

Features for each model were the current-day (day t) residual returns of all stocks in the training regime. The target variable was the next-day (day $t+1$) residual return of the target stock. Lo and MacKinlay (4) motivated this setup directly, and Cohen and Frazzini (5) demonstrated that this method extended across economic supply chain links. Using residuals rather than raw returns prevented SPY's return from acting as an artificial predictor of future returns and reduced the likelihood that measured cross-asset relationships simply reflected shared market beta rather than stock-specific interactions.

All feature matrices were standardized to zero mean and unit variance using the StandardScaler fitted on the training data before PCA, ensuring that high-variance features did not disproportionately influence the principal components. Standardization was fit exclusively on the training fold of each cross-validation split and applied as a transform to the test fold. After standardization, principal component analysis (PCA) was performed with $n_components=0.95$, retaining the minimum

number of orthogonal components that explained 95% of feature variance. PCA was similarly fit on the training fold only. This dimensionality reduction addressed multicollinearity among training stock residuals, especially high in the high-correlation regime, and has been used in related ML factor model work by Connor and Korajczyk (19) and Lettau and Pelger (20).

2.5 Model architectures and hyperparameters

Three model families were evaluated: Ridge regression, Random Forest, and XGBoost. Ridge regression adds an L2 penalty on coefficient magnitudes to a standard linear regression, providing a principled way to handle the collinearity that remains even after PCA. We used RidgeCV with an alpha grid of [0.1, 1.0, 10.0, 100.0, 1000.0], which selected the regularization strength within each training fold using cross-validation on training data only, and also eliminated the arbitrary choice of a fixed alpha.

Random Forest (21) builds an ensemble of decision trees trained on bootstrap samples of the training data with random feature subsets, reducing variance through averaging. We used `n_estimators=100` and `max_depth=10` to limit overfitting on noisy daily return labels. XGBoost (22) builds trees sequentially, each correcting the residual errors of the previous, with `n_estimators=100`, `max_depth=6`, and `learning_rate=0.05`. For both tree-based models, hyperparameters were held constant across all regimes and stocks to ensure that performance differences reflected training universe composition rather than model tuning. Both models used the random seed of the current

outer seed loop iteration to ensure reproducibility while varying model stochasticity across seeds.

We note that Krauss et al. (23) found gradient-boosted trees and random forests to be among the strongest performers for equity prediction and statistical arbitrage on the S&P 500, and Fischer and Krauss (24) showed LSTMs can outperform simpler methods with sufficient data. We chose Ridge, RF, and XGBoost for consistency and interpretability across regimes. Specialized time-series models like ARIMA and GARCH were excluded, because the goal was to compare residual effects across consistent architectures, not to maximize absolute forecasting accuracy with a possible loss of interpretability.

2.6 Walk-forward cross-validation

All models were evaluated using three-fold walk-forward TimeSeriesSplit, which maintains strict chronological ordering, where each fold trains on all data before a given cutoff and tests on the subsequent period. Standard k-fold cross-validation was not appropriate for time series data because random shuffling creates look-ahead bias (25). StandardScaler and PCA were fit on each training fold separately and applied to the corresponding test fold. R^2 was computed on each test fold and averaged across the three folds to produce a single regime-level performance estimate per target stock per model per seed.

2.7 Multi-seed experimental design

Because training regimes were populated by random sampling from each correlation bucket, a single random seed may produce unusually

informative or uninformative stock selections, biasing results toward or away from a particular regime. To address this, we repeated the entire procedure across 10 random seeds: [42, 22, 27, 15, 7, 13, 99, 31, 55, 8]. For each seed, the full pipeline executed, where regimes were sampled, features were constructed, models were trained across three folds, and R^2 was recorded. This yielded 100 observations per regime per pipeline (10 seeds $\times$ 10 target stocks) and 9,000 total model fits across both pipelines (10 seeds $\times$ 10 stocks $\times$ 5 regimes $\times$ 3 models $\times$ 3 folds $\times$ 2 pipelines). Cross-seed variance captured both sampling variance (which stocks were drawn) and model stochasticity (XGBoost and RF internal randomness). Ridge has no internal randomness, so its cross-seed variance reflected sampling variance only.

2.8 Bootstrap confidence intervals

To formally test whether performance differences across regimes were statistically significant, we applied the bias-corrected and accelerated (BCa) bootstrap (26) via `scipy.stats.bootstrap` with 1,000 resamples at the 95% confidence level. The BCa method corrected for both bias and skewness in the bootstrap distribution, which is appropriate for the small-to-moderate sample sizes typical of financial ML evaluation (27).

Each pipeline produced 100 Ridge R^2 estimates per regime, one for each (seed, target) combination. These 100 values were not independent: the 10 seeds for a given target shared the same residual series and the same fold-wise regime assignment, and differed only in which stocks were randomly drawn into each regime. Treating them as 100 independent

observations would understate the true uncertainty and artificially narrow the confidence intervals. To avoid this, we clustered the bootstrap by target stock. For each regime we first averaged the 10 seed-level R^2 values within each target, yielding 10 target-level estimates, then resampled those 10 target-level values with replacement.

We bootstrapped two quantities: [1] the mean R^2 per regime, to characterize absolute performance with uncertainty, and [2] the pairwise difference in mean R^2 between regimes, to test whether one regime outperformed another. A pairwise difference whose 95% interval excluded zero indicated that the regimes differed at the 95% level. We note that even a statistically significant difference can be economically negligible.

3. Results

3.1 Overall pattern across regimes and models

All five training regimes produced near-zero or negative R^2 across all models and target stocks. This matches prior work (3, 28), where daily large-cap returns are close to a random walk. Ridge regression consistently showed the least negative R^2 across all regimes and stocks, followed by Random Forest, with XGBoost most negative. As Gu et al. (3) and Welch and Goyal (28) also found, simpler and regularized models perform better when signal-to-noise is low. Figure 2 presents the mean Ridge R^2 for each regime under both the contemporaneous and lagged pipelines, and Figure 3 shows the corresponding per-stock breakdown across all target stocks and regimes.

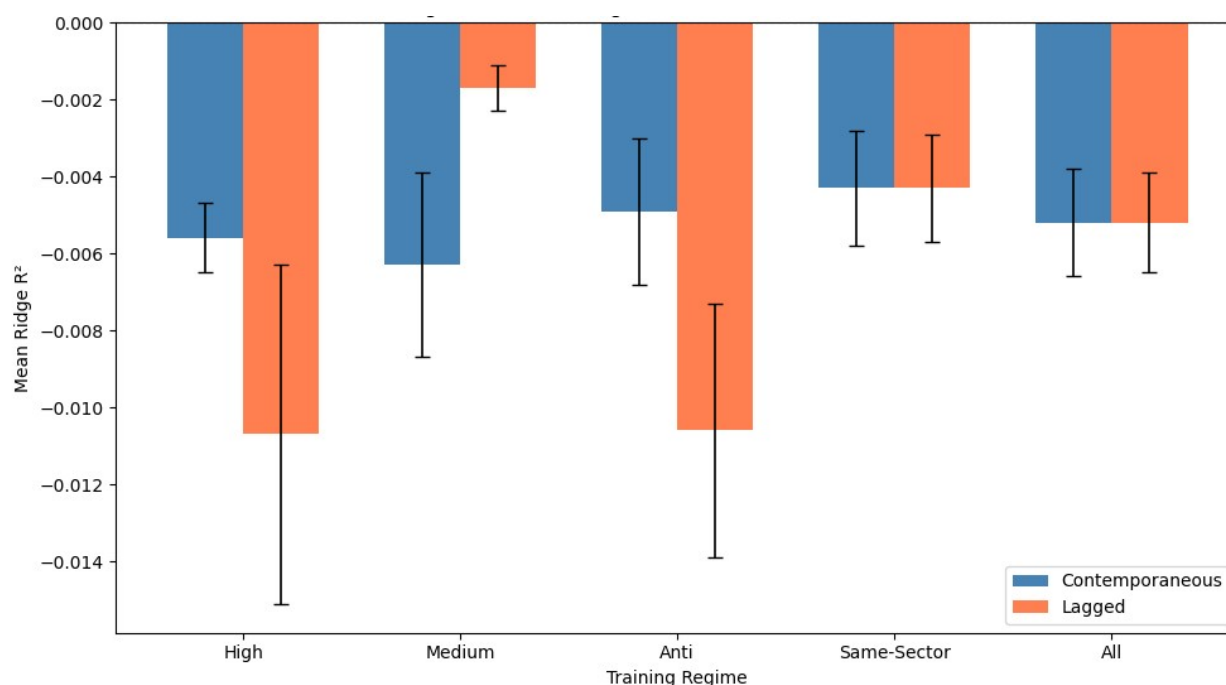


Figure 2. Mean Ridge R^2 by training regime for the contemporaneous and lagged pipelines, averaged across 10 target stocks. Error bars represent 95% BCa bootstrap confidence intervals, clustered by target stock (10 target-level means computed from 100 runs per regime). Under contemporaneous construction, all five regimes perform indistinguishably. Under lagged construction, the medium-correlation regime is substantially less negative than all others, while high and anti-correlation regimes perform worst.

Figure 2 summarizes the mean Ridge regression performance across all five training regimes for both the contemporaneous and lagged pipelines, with uncertainty quantified using 95% BCa bootstrap confidence intervals. Under the contemporaneous pipeline, the mean R^2 values for the five regimes are closely clustered and their confidence intervals overlap substantially, indicating little difference in predictive performance as a function of training-universe composition. In contrast, the lagged pipeline exhibits a distinct pattern in which the medium-correlation regime achieves the least negative mean R^2 , whereas both the high-correlation and anti-correlation regimes perform worst. The same-sector and pooled baselines occupy

intermediate positions. This divergence between the contemporaneous and lagged pipelines motivated the formal pairwise statistical comparisons reported subsequently in the tables.

Figure 3 presents the Ridge regression results separately for each target stock and training regime, allowing visualization of the consistency of performance across individual equities. Under the contemporaneous pipeline, no training regime consistently produces superior performance across the ten target stocks, with R^2 values remaining uniformly close to zero regardless of regime assignment. In contrast, the lagged pipeline demonstrates that the medium-correlation regime generally

produces less negative R^2 values across most target stocks, whereas the high-correlation and anti-correlation regimes frequently yield the poorest performance. Although considerable heterogeneity remains among individual stocks,

the heatmaps indicate that the aggregate differences observed in Figure 2 are not driven by a single outlying stock but instead reflect a broadly consistent pattern across the evaluated target universe.

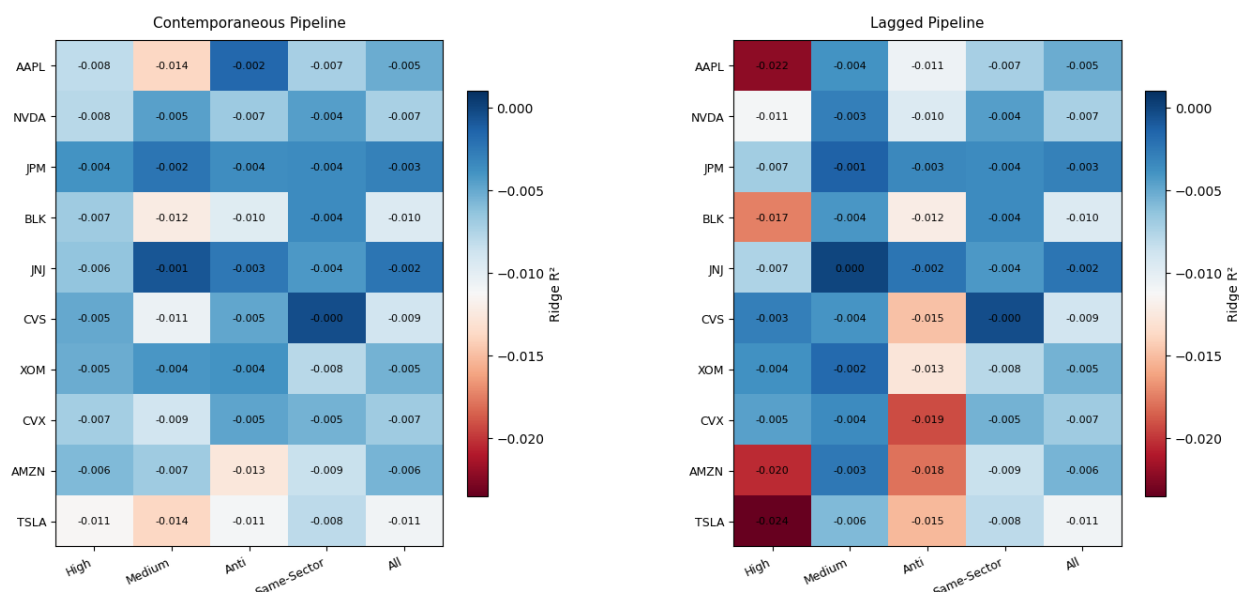


Figure 3. Per-stock Ridge R^2 heatmap for each training regime under the contemporaneous (left) and lagged (right) pipelines. Blue shades represent values closer to zero (better performance), and red shades represent more negative values (worse performance). The lagged pipeline shows how medium-regime has an advantage across most target stocks, not present in the contemporaneous pipeline.

3.2 Contemporaneous pipeline

Tables 1 through 5 summarize predictive performance under the contemporaneous pipeline, in which training regimes are defined according to same-day residual correlations between stocks. Mean out-of-sample R^2 values are reported for Ridge regression, XGBoost, and Random Forest across the 10 target stocks, averaged over 10 random seeds where applicable. Across all five training regimes, Ridge regression consistently achieves the least negative R^2 values, followed by Random Forest and XGBoost, although all three models produce predominantly negative out-of-sample

R^2 values. Tables 6 and 7 provide BCa bootstrap confidence intervals for the mean Ridge R^2 of each regime and for pairwise differences between regimes, respectively. While small numerical differences are observed among the high-correlation, medium-correlation, anti-correlation, same-sector, and pooled training regimes, the bootstrap analysis indicates substantial overlap in their confidence intervals, and none of the pairwise comparisons reaches statistical significance. These findings suggest that, under contemporaneous regime construction, varying the correlation structure of

the training universe produces little measurable effect on predictive performance.

Table 1. Contemporaneous high-correlation regime results (Mean $R^2 \pm$ SD across 10 seeds)

Stock	Ridge R^2	XGBoost R^2	Random Forest R^2
AAPL	-0.0081 ± 0.0017	-0.1773 ± 0.0399	-0.0793 ± 0.0159
NVDA	-0.0079 ± 0.0023	-0.1878 ± 0.0736	-0.0931 ± 0.0240
JPM	-0.0039 ± 0.0010	-0.1104 ± 0.0178	-0.0714 ± 0.0169
BLK	-0.0069 ± 0.0016	-0.1159 ± 0.0218	-0.0615 ± 0.0165
JNJ	-0.0064 ± 0.0010	-0.0839 ± 0.0139	-0.0508 ± 0.0133
CVS	-0.0050 ± 0.0017	-0.0556 ± 0.0128	-0.0336 ± 0.0109
XOM	-0.0052 ± 0.0038	-0.0692 ± 0.0150	-0.0296 ± 0.0050
CVX	-0.0071 ± 0.0009	-0.0811 ± 0.0121	-0.0530 ± 0.0126
AMZN	-0.0058 ± 0.0015	-0.2288 ± 0.0728	-0.1790 ± 0.0352
TSLA	-0.0114 ± 0.0013	-0.1120 ± 0.0159	-0.0555 ± 0.0139

Values are reported as mean $\pm$ standard deviation across all 10 random seeds, except for the same-sector regime where standard deviation is zero because sector membership is deterministic and does not vary across seeds. Ridge regression with RidgeCV alpha selection. XGBoost uses gradient boosting with learning_rate=0.05. Random Forest uses max_depth=10.

Table 2. Contemporaneous medium-correlation regime results (Mean $R^2 \pm$ SD across 10 seeds)

Stock	Ridge R^2	XGBoost R^2	Random Forest R^2
AAPL	-0.0138 ± 0.0028	-0.1593 ± 0.0238	-0.0753 ± 0.0086
NVDA	-0.0046 ± 0.0017	-0.2755 ± 0.1259	-0.1390 ± 0.0875
JPM	-0.0023 ± 0.0019	-0.0817 ± 0.0114	-0.0527 ± 0.0131
BLK	-0.0123 ± 0.0025	-0.1046 ± 0.0593	-0.0495 ± 0.0225
JNJ	-0.0008 ± 0.0031	-0.0598 ± 0.0120	-0.0265 ± 0.0087
CVS	-0.0105 ± 0.0021	-0.0777 ± 0.0226	-0.0550 ± 0.0207
XOM	-0.0041 ± 0.0016	-0.0753 ± 0.0146	-0.0449 ± 0.0098
CVX	-0.0089 ± 0.0025	-0.1276 ± 0.0269	-0.1011 ± 0.0404
AMZN	-0.0069 ± 0.0054	-0.2533 ± 0.0484	-0.1348 ± 0.0402
TSLA	-0.0138 ± 0.0063	-0.1381 ± 0.0257	-0.0750 ± 0.0259

Medium-correlation regime contains stocks with residual correlations between the 25th and 75th percentiles of the target stock's correlation distribution. Values are mean $\pm$ SD across 10 random seeds.

Table 3. Contemporaneous anti-correlation regime results (Mean $R^2 \pm SD$ across 10 seeds)

Stock	Ridge R^2	XGBoost R^2	Random Forest R^2
AAPL	-0.0016 ± 0.0020	-0.1714 ± 0.0190	-0.0823 ± 0.0192
NVDA	-0.0068 ± 0.0017	-0.3089 ± 0.0981	-0.1250 ± 0.0399
JPM	-0.0036 ± 0.0017	-0.0865 ± 0.0152	-0.0334 ± 0.0077
BLK	-0.0098 ± 0.0013	-0.0695 ± 0.0100	-0.0275 ± 0.0075
JNJ	-0.0025 ± 0.0008	-0.0625 ± 0.0135	-0.0299 ± 0.0079
CVS	-0.0049 ± 0.0008	-0.0438 ± 0.0086	-0.0278 ± 0.0085
XOM	-0.0039 ± 0.0014	-0.0706 ± 0.0101	-0.0352 ± 0.0076
CVX	-0.0048 ± 0.0007	-0.1070 ± 0.0346	-0.0631 ± 0.0261
AMZN	-0.0126 ± 0.0030	-0.2503 ± 0.0533	-0.1504 ± 0.0354
TSLA	-0.0110 ± 0.0017	-0.1526 ± 0.0285	-0.1051 ± 0.0200

Anti-correlation regime contains stocks with residual correlations below the 25th percentile, which is negative for all 10 target stocks (ranging from -0.003 to -0.167). Values are mean $\pm$ SD across 10 random seeds.

Table 4. Contemporaneous same-sector regime results (Single estimate, deterministic membership)

Stock	Ridge R^2	XGBoost R^2	Random Forest R^2
AAPL	-0.0072	-0.2118	-0.0687 ± 0.0055
NVDA	-0.0045	-0.1669	-0.1096 ± 0.0089
JPM	-0.0035	-0.0963	-0.0721 ± 0.0066
BLK	-0.0035	-0.1126	-0.0525 ± 0.0056
JNJ	-0.0042	-0.0690	-0.0305 ± 0.0028
CVS	-0.0004	-0.0415	-0.0209 ± 0.0026
XOM	-0.0079	-0.0445	-0.0225 ± 0.0017
CVX	-0.0055	-0.0531	-0.0369 ± 0.0038
AMZN	-0.0085	-0.4605	-0.2321 ± 0.0124
TSLA	-0.0080	-0.1230	-0.0593 ± 0.0046

Same-sector regime contains all stocks in the same GICS sector as the target, subsampled to match the minimum regime size. Same-sector regime $SD = 0$ for Ridge and XGBoost because sector membership is deterministic. RF varies slightly across seeds due to internal randomness.

Table 5. Contemporaneous all-stocks regime results (pooled baseline) (Mean $R^2 \pm$ SD across 10 seeds)

Stock	Ridge R^2	XGBoost R^2	Random Forest R^2
AAPL	-0.0052 ± 0.0029	-0.1653 ± 0.0392	-0.0913 ± 0.0363
NVDA	-0.0073 ± 0.0023	-0.3160 ± 0.2418	-0.1900 ± 0.0851
JPM	-0.0030 ± 0.0032	-0.0994 ± 0.0263	-0.0612 ± 0.0161
BLK	-0.0096 ± 0.0030	-0.1280 ± 0.0518	-0.0734 ± 0.0248
JNJ	-0.0023 ± 0.0029	-0.0669 ± 0.0138	-0.0342 ± 0.0058
CVS	-0.0089 ± 0.0040	-0.0582 ± 0.0097	-0.0421 ± 0.0207
XOM	-0.0055 ± 0.0025	-0.0894 ± 0.0166	-0.0516 ± 0.0116
CVX	-0.0069 ± 0.0026	-0.1090 ± 0.0351	-0.0553 ± 0.0106
AMZN	-0.0056 ± 0.0049	-0.2012 ± 0.0494	-0.1703 ± 0.0629
TSLA	-0.0108 ± 0.0038	-0.1281 ± 0.0242	-0.0597 ± 0.0132

The all-stocks baseline samples randomly from the full 74-stock universe (excluding the target), subsampled to match the minimum regime size. Values are mean $\pm$ SD across 10 random seeds.

Table 6. Contemporaneous bootstrap 95% CIs for mean ridge R^2 per regime

Regime	Mean R^2	95% CI Lower	95% CI Upper
Same-Sector	-0.0053	-0.0066	-0.0037
Anti-Correlation	-0.0061	-0.0087	-0.0042
All-Stocks	-0.0065	-0.0082	-0.0049
High-Correlation	-0.0068	-0.0084	-0.0058
Medium-Correlation	-0.0078	-0.0106	-0.0052

BCa bootstrap with 1,000 resamples. Unit of resampling is the target stock (10 target-level means per regime). Regimes ordered from least negative (best) to most negative.

Table 7. Contemporaneous pairwise R^2 differences between regimes

Comparison	CI Lower	CI Upper	Significant?
High vs Anti	-0.0025	0.0019	No
High vs Same-Sector	-0.0028	0.0002	No
High vs Medium	-0.0014	0.0031	No
High vs All	-0.0015	0.0011	No
Medium vs Anti	-0.0054	0.0008	No
Same-Sector vs Anti	-0.0016	0.0029	No

BCa bootstrap 95% CIs for the difference in mean Ridge R^2 between regime pairs. Positive values indicate the first-listed regime outperforms the second. No comparison reaches significance under contemporaneous regime construction.

3.3 Lagged pipeline

Tables 8 through 14 report the equivalent results for the lagged pipeline, where regimes are

defined by $\text{Corr}(\varepsilon_i(t), \varepsilon_{\text{tar}}^{\text{reg}}(t+1))$ estimated

within each training fold.

Table 8. Lagged high-correlation regime results (Mean $R^2 \pm$ SD across 10 seeds)

Stock	Ridge R^2	XGBoost R^2	Random Forest R^2
AAPL	-0.0222 ± 0.0017	-0.1886 ± 0.0280	-0.1064 ± 0.0144
NVDA	-0.0110 ± 0.0025	-0.3707 ± 0.1363	-0.1695 ± 0.0433
JPM	-0.0070 ± 0.0011	-0.0821 ± 0.0098	-0.0392 ± 0.0098
BLK	-0.0174 ± 0.0024	-0.0984 ± 0.0210	-0.0669 ± 0.0217
JNJ	-0.0075 ± 0.0015	-0.1189 ± 0.0463	-0.0761 ± 0.0365
CVS	-0.0029 ± 0.0016	-0.0504 ± 0.0070	-0.0296 ± 0.0055
XOM	-0.0038 ± 0.0011	-0.0784 ± 0.0145	-0.0409 ± 0.0081
CVX	-0.0046 ± 0.0018	-0.0871 ± 0.0171	-0.0425 ± 0.0132
AMZN	-0.0204 ± 0.0052	-0.3278 ± 0.0590	-0.2914 ± 0.1135
TSLA	-0.0235 ± 0.0060	-0.1346 ± 0.0248	-0.0789 ± 0.0145

Table 9. Lagged medium-correlation regime results (Mean $R^2 \pm$ SD across 10 seeds)

Stock	Ridge R^2	XGBoost R^2	Random Forest R^2
AAPL	-0.0039 ± 0.0010	-0.1390 ± 0.0231	-0.0658 ± 0.0216
NVDA	-0.0029 ± 0.0010	-0.3136 ± 0.1343	-0.1248 ± 0.0398
JPM	-0.0013 ± 0.0009	-0.0858 ± 0.0172	-0.0576 ± 0.0200
BLK	-0.0041 ± 0.0012	-0.1009 ± 0.0255	-0.0578 ± 0.0403
JNJ	$+0.0000 \pm 0.0009$	-0.0661 ± 0.0116	-0.0335 ± 0.0082
CVS	-0.0040 ± 0.0015	-0.0667 ± 0.0209	-0.0531 ± 0.0290
XOM	-0.0017 ± 0.0007	-0.0737 ± 0.0097	-0.0341 ± 0.0066
CVX	-0.0035 ± 0.0009	-0.1025 ± 0.0163	-0.0585 ± 0.0088
AMZN	-0.0025 ± 0.0018	-0.2087 ± 0.0567	-0.1163 ± 0.0437
TSLA	-0.0058 ± 0.0009	-0.0953 ± 0.0203	-0.0394 ± 0.0096

Tables 8 through 12 summarize model performance under the lagged pipeline, in which training regimes are defined using one-day-ahead predictive correlations, $\text{Corr}(\varepsilon_i(t), \varepsilon_{\text{target}}(t+1))$, estimated exclusively within each walk-forward training fold. As in the contemporaneous pipeline, Tables 8 through 12 report mean R^2 values across the 10 target stocks for Ridge regression, XGBoost, and Random Forest, averaged over 10 random seeds where applicable. Table 13 presents BCa bootstrap confidence intervals for the mean Ridge R^2 of each training regime, while Table 14 reports bootstrap confidence intervals for pairwise differences between regimes to determine whether observed performance differences are statistically significant. Unlike the contemporaneous pipeline, the lagged pipeline

reveals a clear separation among training regimes. The medium-correlation regime consistently achieves the least negative Ridge R^2 , whereas the high-correlation and anti-correlation regimes perform worst. The bootstrap analysis confirms that several of these differences are statistically significant, although all regimes continue to produce near-zero or negative out-of-sample R^2 values, indicating that the observed improvements are modest in absolute magnitude.

Table 10. Lagged anti-correlation regime results (Mean $R^2 \pm$ SD across 10 seeds)

Stock	Ridge R^2	XGBoost R^2	Random Forest R^2
AAPL	-0.0106 ± 0.0022	-0.1669 ± 0.0272	-0.1133 ± 0.0345
NVDA	-0.0096 ± 0.0024	-0.1429 ± 0.0692	-0.0565 ± 0.0198
JPM	-0.0034 ± 0.0019	-0.0955 ± 0.0109	-0.0647 ± 0.0249
BLK	-0.0122 ± 0.0012	-0.0825 ± 0.0204	-0.0488 ± 0.0219
JNJ	-0.0023 ± 0.0017	-0.0498 ± 0.0072	-0.0211 ± 0.0100
CVS	-0.0148 ± 0.0025	-0.0611 ± 0.0085	-0.0539 ± 0.0165
XOM	-0.0127 ± 0.0013	-0.0892 ± 0.0147	-0.0521 ± 0.0081
CVX	-0.0192 ± 0.0016	-0.1329 ± 0.0136	-0.1347 ± 0.0370
AMZN	-0.0179 ± 0.0025	-0.3178 ± 0.0910	-0.2014 ± 0.0674
TSLA	-0.0152 ± 0.0029	-0.1667 ± 0.0351	-0.0846 ± 0.0167

Table 11. Lagged same-sector regime results (Single estimate, deterministic membership)

Stock	Ridge R^2	XGBoost R^2	Random Forest R^2
AAPL	-0.0072	-0.2080	-0.0703 ± 0.0059
NVDA	-0.0045	-0.1620	-0.1109 ± 0.0100
JPM	-0.0035	-0.0874	-0.0711 ± 0.0066
BLK	-0.0035	-0.1100	-0.0521 ± 0.0053
JNJ	-0.0042	-0.0685	-0.0303 ± 0.0024
CVS	-0.0004	-0.0504	-0.0213 ± 0.0029
XOM	-0.0079	-0.0513	-0.0227 ± 0.0020
CVX	-0.0055	-0.0516	-0.0362 ± 0.0033
AMZN	-0.0085	-0.4657	-0.2340 ± 0.0119
TSLA	-0.0080	-0.1125	-0.0596 ± 0.0057

Same-sector regime is identical across both pipelines because sector membership does not depend on the correlation construction method.

Table 12. Lagged all-stocks regime results (pooled baseline) (Mean $R^2 \pm$ SD across 10 seeds)

Stock	Ridge R^2	XGBoost R^2	Random Forest R^2
AAPL	-0.0052 ± 0.0029	-0.1645 ± 0.0427	-0.0900 ± 0.0352
NVDA	-0.0073 ± 0.0023	-0.3061 ± 0.2362	-0.1894 ± 0.0834
JPM	-0.0030 ± 0.0032	-0.1024 ± 0.0258	-0.0603 ± 0.0163
BLK	-0.0096 ± 0.0030	-0.1256 ± 0.0495	-0.0731 ± 0.0241
JNJ	-0.0023 ± 0.0029	-0.0693 ± 0.0120	-0.0335 ± 0.0055
CVS	-0.0089 ± 0.0040	-0.0576 ± 0.0123	-0.0425 ± 0.0213
XOM	-0.0055 ± 0.0025	-0.0877 ± 0.0135	-0.0519 ± 0.0121
CVX	-0.0069 ± 0.0026	-0.1052 ± 0.0338	-0.0575 ± 0.0128
AMZN	-0.0056 ± 0.0049	-0.2071 ± 0.0515	-0.1693 ± 0.0637
TSLA	-0.0108 ± 0.0038	-0.1250 ± 0.0255	-0.0603 ± 0.0144

The all-stocks baseline samples randomly from the full 74-stock universe (excluding the target), subsampled to match the minimum regime size. Values are mean $\pm$ SD across 10 random seeds.

Table 13. Lagged bootstrap 95% CIs for mean ridge R^2 per regime

Regime	Mean R^2	95% CI Lower	95% CI Upper
Medium-Correlation	-0.0030	-0.0040	-0.0020
Same-Sector	-0.0053	-0.0066	-0.0036
All-Stocks	-0.0065	-0.0081	-0.0049
Anti-Correlation	-0.0118	-0.0148	-0.0082
High-Correlation	-0.0120	-0.0176	-0.0076

BCa bootstrap with 1,000 resamples. Unit of resampling is the target stock (10 target-level means per regime). Regimes ordered from least negative (best) to most negative. The medium-correlation regime is substantially less negative than all others.

Table 14. Lagged pairwise R^2 differences between regimes

Comparison	CI Lower	CI Upper	Significant?
High vs Anti	-0.0045	0.0055	No
High vs Same-Sector	-0.0110	-0.0025	Yes
High vs Medium	-0.0134	-0.0047	Yes
High vs All	-0.0099	-0.0012	Yes
Medium vs Anti	0.0059	0.0114	Yes
Same-Sector vs Anti	0.0034	0.0096	Yes

BCa bootstrap 95% CIs for the difference in mean Ridge R^2 between regime pairs. Positive values indicate the first-listed regime outperforms the second.

4. Discussion

4.1 Contemporaneous correlation structure does not determine predictive performance

Under contemporaneous regime construction, no pairwise comparison between training regimes reached statistical significance. The high-correlation, medium-correlation, anti-correlation, same-sector, and pooled all-stocks regimes all produced indistinguishable out-of-sample Ridge R^2 , with BCa bootstrap confidence intervals for every pairwise difference crossing zero. When training stocks were selected by how much they co-moved with the target on the same day, the composition of the training universe did not affect next-day prediction accuracy.

With 100 (seed, ticker) observations per regime and BCa intervals that are reasonably tight, the experiment was sensitive enough to detect differences of the magnitude that would be economically relevant. Instead, the findings reflect the underlying structure of large-cap U.S. equity markets, where, after market neutralization, daily residual returns contain little exploitable cross-asset predictive signal for correlation-based training-universe selection to improve out-of-sample forecasting. This interpretation is consistent with the semi-strong form of the Efficient Market Hypothesis (12, 13), which predicts that historically observable cross-asset relationships among highly followed S&P 500 stocks should be rapidly incorporated into prices, leaving little persistent predictive signal available for out-of-sample forecasting (30).

The criterion used to assign stocks to training regimes (same-day residual correlation) and the criterion used to evaluate model performance (next-day prediction accuracy) represent distinct statistical relationships. Consequently, the experimental design does not contain a tautological linkage between regime assignment and model evaluation. High contemporaneous residual co-movement does not necessarily imply predictive information one day later. Therefore, a null result under the contemporaneous pipeline was not predetermined. It remained plausible that stocks exhibiting strong same-day residual co-movement would also share transferable lead-lag information, as suggested by the information diffusion frameworks of Hong and Stein (8) and Chordia and Swaminathan (9).

4.2 Lagged correlation structure is significant but spurious

The lagged pipeline tells a different story, though it leads to the same conclusion. When regimes were defined by one-day-ahead predictive correlation $\text{Corr}(\varepsilon_i(t), \varepsilon_{\text{tar}}^{\text{op}}(t+1))$, estimated within each training fold, the medium-correlation regime outperformed all others significantly. BCa confidence intervals for medium vs. high (CI: [0.0047, 0.0134]) and medium vs. anti-correlation (CI: [0.0059, 0.0114]) both excluded zero. The high and anti-correlation regimes performed the worst, with mean Ridge R^2 of -0.0120 and -0.0118 respectively, while the medium regime reached -0.0030 .

We interpreted this pattern as being more consistent with unstable training-fold predictive

correlations than with genuine, persistent lead-lag structure. The lagged construction selects stocks by estimating $\text{Corr}(\varepsilon_i(t), \varepsilon_{\text{tar}}^{\text{op}}(t+1))$ within each training fold, which means that it actively optimizes for the training-fold version of the prediction task. Stocks assigned to the high predictive correlation regime are those that happened to lead the target most strongly during that particular fold. Out of sample, this relationship does not hold. The stocks selected as strong predictors in-sample performed no better than randomly selected stocks, and in fact performed worse, suggesting that the selection process simply identified noise.

This interpretation is consistent with the gradual information diffusion framework of Hong and Stein (8) and Chordia and Swaminathan (9). Those papers predict that genuine lead-lag relationships arise from slow diffusion of information across investor populations, a structural feature that should be stable across time periods. If the lagged correlations we estimated within training folds reflected real information diffusion links, we would expect them to generalize out of sample. They did not, which suggests that the estimated correlations were driven by transient co-movements rather than the durable economic linkages that the theory describes. The cross-predictability documented by Lo and MacKinlay (4), Cohen and Frazzini (5), and Hou (6) may be largely arbitrated away in the large-cap stocks examined here, since all three papers note that their effects are strongest in neglected, low-coverage stocks, whereas our universe consisted entirely of highly covered S&P 500 names.

An alternative explanation for the superior performance of the medium-correlation regime is that moderately correlated training stocks provide greater informational diversity than highly correlated stocks. Highly correlated assets may contain largely redundant predictive information, whereas moderately correlated assets could capture complementary aspects of market dynamics while still remaining economically related to the target. Because the present study was not designed to distinguish redundancy from instability in estimated predictive correlations, both explanations remain plausible. Future work could explicitly quantify feature redundancy within each training regime to determine whether informational diversity contributes to the observed medium-regime advantage.

4.3 Convergence on a null result

The two pipelines reached the same substantive conclusion by different routes. The contemporaneous pipeline produced a clean null, where no regime outperformed any other. The lagged pipeline produced a significant result that we attributed to in-sample overfitting of the regime construction criterion itself. Together, they suggested that correlation structure, whether measured by same-day co-movement or historical predictive lead, simply does not reliably determine cross-asset forecasting performance for large-cap equities at the daily frequency.

4.4 Consistency across models

The near-zero or negative R^2 pattern was consistent across Ridge regression, Random Forest, and XGBoost, though magnitudes differed. Across both pipelines, XGBoost

consistently produced the most negative R^2 values, followed by Random Forest, whereas Ridge regression achieved the highest average performance across all training regimes. This pattern is consistent with the possibility that highly flexible nonlinear models derive little benefit when the underlying predictive signal is weak, whereas the simpler linear structure and L2 regularization of Ridge regression may provide greater robustness in a low signal-to-noise environment. This ordering matches findings in Gu et al. (3) and Krauss et al. (23), who also found that model complexity did not translate to better performance when the predictive signal was weak.

4.5 Same-sector vs. cross-sector correlation

Under contemporaneous construction, same-sector training was statistically indistinguishable from every other regime, including from high-correlation training. This is notable because sector peers are the most natural candidates for informational relatedness, as they share customers, suppliers, regulatory environments, and macroeconomic exposures. Yet a healthcare stock that happens to co-move with a technology target in residual return space is no more or less useful than a sector peer. The signal, to the extent it exists at all, travels through residual correlation structure rather than industry classification.

This finding contrasts with the intuition of Moskowitz and Grinblatt (29), who demonstrated that industry momentum explains a substantial portion of individual stock momentum. If GICS sector boundaries captured the primary source of informational relatedness between stocks, same-sector training should

have outperformed. It did not, which suggested that the economic linkages relevant to daily cross-asset prediction were not well-described by revenue-based industry classification.

4.6 Market efficiency and negative R^2

All regimes produced negative out-of-sample R^2 values across both pipelines, indicating that every model performed worse than predicting the mean residual return, which is $\sim$ zero after market neutralization. For near-random-walk residual returns at the daily frequency, this is the expected outcome under market efficiency (12, 13). The daily frequency is well-established as the hardest prediction horizon for large-cap equities (28), as the signal-to-noise ratio is too low for any training universe composition to overcome it. The finding that no correlation-based selection strategy produced positive R^2 confirmed that for S&P 500 stocks, cross-asset predictive structure in daily residual returns is either absent or too small to be reliably captured by the models and sample sizes in this study.

4.7 Implications for cross-asset ML model design

For daily large-cap equity prediction, the composition of the training universe along correlation dimensions did not meaningfully affect performance. Although several pairwise comparisons in the lagged pipeline reached statistical significance, the corresponding effect sizes remained extremely small. All regimes continued to exhibit near-zero or negative out-of-sample R^2 values, indicating that the statistically detectable differences did not translate into practically useful forecasting performance. This distinction between statistical and practical significance is

particularly important in financial prediction, where large sample sizes can reveal reliable but economically negligible differences. The absolute magnitude of any regime difference, even the statistically significant ones in the lagged pipeline, was economically negligible, with the best-case improvement on the order of 0.009 R^2 units against a backdrop of consistently negative performance.

More broadly, these findings illustrate an important principle in applied machine learning. Increasing the sophistication of training-data selection does not necessarily translate into improved predictive performance. Even when training universes are constructed using theoretically motivated correlation criteria, predictive performance may remain unchanged if the underlying signal is weak or unstable. This emphasizes that improvements in data selection strategies cannot compensate for a fundamentally low signal-to-noise ratio.

4.8 Limitations

This study has several important limitations. The stock universe consisted entirely of current or recent S&P 500 constituents, introducing survivorship bias (31). Incorporating historical constituents through a database such as CRSP would address this limitation. Although 10 target stocks were sufficient for estimating aggregate effects, they represent a relatively small sample, and per-stock results exhibit substantial heterogeneity. Consequently, the aggregate analyses may mask meaningful stock-level variation, and any stock-specific conclusions should be considered exploratory. Hyperparameters for Random Forest and XGBoost were held constant across all regimes

and target stocks; although regime-specific tuning might improve performance, it would also introduce an important confounder by making it impossible to distinguish the effects of training-universe composition from those of model optimization. The lead-lag feature set used here (current-day residual returns of training stocks) was deliberately simple and directly motivated by the underlying theory, but omitted fundamental variables, macroeconomic indicators, order-flow information, and higher-frequency features routinely incorporated into professional forecasting models. Furthermore, we evaluated only daily prediction. Lead-lag effects have been reported to be stronger at intraday and weekly frequencies (4, 6), where training-universe composition may have a greater influence on predictive performance. We also examined only a single lag structure; multi-day lookback windows or rolling correlation features may capture additional cross-asset relationships. Finally, although the anti-correlation regime consisted of stocks selected for negative residual correlation with the target, the magnitudes of these correlations were generally small (25th-percentile values ranging from approximately -0.003 to -0.167 across target stocks). Consequently, this regime likely represents weak idiosyncratic divergence rather than strong economically meaningful inverse relationships between assets. True anti-correlation arising from mechanically opposing economic exposures (e.g., commodity producers versus commodity-intensive consumers) is uncommon among large-cap S&P 500 equities and is therefore not represented in the present universe. Among these limitations, the relatively small number of target stocks is likely the most consequential, as between-stock

variability dominates the aggregate estimates and is reflected in the width of the bootstrap confidence intervals.

4.9 Perspective

Replicating this experimental design using a larger set of target stocks, particularly mid-cap or less-followed equities, would help determine whether training-universe composition has a greater influence in markets where cross-asset information diffusion is slower (5, 6). Intraday or weekly data could reveal whether the correlation-structure effect strengthens at frequencies where lead-lag relationships are more pronounced (4, 6). Finally, extending the regime framework to include rolling or time-varying correlation windows would help assess whether the relative ordering of regimes is stable across different market conditions, including periods of elevated volatility or of correlation breakdown (18).

The present findings should not be interpreted as evidence that lead-lag relationships are absent from financial markets. Rather, they indicate that correlation-based selection of training stocks does not reliably improve daily prediction of large-cap U.S. equities under the modeling framework evaluated here. Lead-lag effects documented in previous literature may still emerge under alternative prediction horizons, different asset classes, less efficient markets, higher-frequency data, or models incorporating richer sources of information.

5. Conclusion

We tested whether the correlation structure of the training universe affects cross-asset predictive performance for large-cap U.S.

equities using two matched experimental pipelines that differed only in how the training regimes were defined: contemporaneous residual correlation versus lagged predictive correlation. Under contemporaneous regime construction, no pairwise comparison between the high-correlation, medium-correlation, anti-correlation, same-sector, and pooled baselines reached statistical significance. When training-universe composition was defined by same-day residual co-movement, it did not meaningfully influence next-day predictive performance.

The inclusion of a second, independently constructed lagged pipeline strengthened this conclusion by evaluating a more direct test of cross-asset information transfer. Rather than selecting training stocks according to contemporaneous co-movement, the lagged pipeline grouped stocks using one-day predictive correlations estimated exclusively within each training fold. Although this approach produced statistically significant differences among some regimes, these improvements remained economically small and did not yield positive out-of-sample predictive performance. The convergence of two independent experimental designs toward the same practical conclusion increases confidence that the observed findings are robust rather than artifacts of a particular regime construction strategy.

The lagged pipeline shows that the medium-correlation regime outperforms both the high- and anti-correlation regimes, but we interpret this pattern as reflecting instability in historically estimated predictive correlations rather than genuine, persistent lead-lag

structure. Predictive correlations estimated within individual training folds identify stocks that appear to lead the target during the estimation period, but these relationships fail to generalize out of sample. The apparent predictive signal is therefore likely to reflect transient or sample-specific structure rather than stable cross-asset information transfer.

More broadly, the study illustrates an important methodological principle for financial machine learning. Increasing the sophistication of training-universe selection does not necessarily improve predictive performance when the underlying predictive signal is weak. Correlation-based filtering, even when motivated by established theories of information diffusion and implemented under strict leakage controls, cannot compensate for the limited predictability of daily large-cap equity returns. These findings therefore emphasize the importance of evaluating both statistical significance and practical forecasting performance when proposing new training strategies for financial machine-learning models. Although the principal finding is a negative result, it establishes an important empirical boundary condition for cross-asset machine learning by demonstrating that neither contemporaneous nor fold-wise lagged correlation-based training-universe selection provides a practically meaningful advantage for daily prediction of large-cap U.S. equities under controlled experimental conditions.

Both pipelines support the same conclusion. Across all regimes, models, and both construction methods, out-of-sample R^2 is near-zero or negative. The best-case improvement

across any regime comparison is on the order of 0.009 R^2 units, far too small to constitute exploitable alpha. For daily residual returns of highly covered S&P 500 stocks, cross-asset predictive structure is either absent or too weak to be captured by the models and sample sizes examined here, consistent with the semi-strong form of the Efficient Market Hypothesis (12, 13). Training-universe composition is therefore not the primary constraint on predictive performance. Instead, improvements are more likely to arise from richer feature representations and other aspects of the forecasting pipeline than from correlation-based selection of the training universe.

Finally, the conclusions of this study should be interpreted within the scope of the experimental design. The results apply specifically to daily prediction of market-neutralized residual returns for large-cap U.S. equities using correlation-based training-universe selection. They should not be interpreted as evidence that lead-lag relationships or cross-asset information transfer are absent from financial markets more generally. Instead, the present findings indicate that, under the modeling framework and prediction horizon examined here, correlation structure alone does not provide a reliable basis for improving cross-asset forecasting performance. Future work extending this framework to alternative asset classes, smaller-cap stocks, international markets, and higher-frequency data will help determine the broader generality of these observations.

Acknowledgements

I would like to acknowledge Shrey Tayal for his helpful comments on the paper, and Shireesh

Apte from JHSS for his editorial feedback. All experiments were conducted independently by the author.

6. References

1. Cao, L. (2022). AI in finance: Challenges, techniques, and opportunities. *ACM Computing Surveys*, 55(3), 1–38. <https://doi.org/10.1145/3502289>
2. Fieberg, C., Metko, D., Poddig, T., & Loy, T. (2023). Machine learning techniques for cross sectional equity returns' prediction. *OR Spectrum*, 45(1), 289–323. <https://doi.org/10.1007/s00291-022-00693-w>
3. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273. <https://doi.org/10.1093/rfs/hhaa009>
4. Lo, A. W., & MacKinlay, A. C. (1990). When are contrarian profits due to stock market overreaction? *The Review of Financial Studies*, 3(2), 175–205. <https://doi.org/10.1093/rfs/3.2.175>
5. Cohen, L., & Frazzini, A. (2008). Economic links and predictable returns. *The Journal of Finance*, 63(4), 1977–2011. <https://doi.org/10.1111/j.1540-6261.2008.01379.x>
6. Hou, K. (2007). Industry information diffusion and the lead-lag effect in stock returns. *The Review of Financial Studies*, 20(4), 1113–1138. <https://doi.org/10.1093/revfin/hhm003>
7. Hong, H., Torous, W., & Valkanov, R. (2007). Do industries lead stock markets? *Journal of Financial Economics*, 83(2), 367–396. <https://doi.org/10.1016/j.jfineco.2005.09.010>
8. Hong, H., & Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance*, 54(6), 2143–2184. <https://doi.org/10.1111/0022-1082.00184>
9. Chordia, T., & Swaminathan, B. (2000). Trading volume and cross-autocorrelations in stock returns. *The Journal of Finance*, 55(2), 913–935. <https://doi.org/10.1111/0022-1082.00231>
10. Cakici, N., Fieberg, C., Metko, D., & Zaremba, A. (2023). Machine learning goes global: Cross-sectional return predictability in international stock markets. *Journal of Economic Dynamics and Control*, 155, 104725. <https://doi.org/10.1016/j.jedc.2023.104725>

11. Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, 135(2), 271–292. <https://doi.org/10.1016/j.jfineco.2019.06.008>
12. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417. <https://doi.org/10.1111/j.1540-6261.1970.tb00518.x>
13. Fama, E. F. (1991). Efficient capital markets: II. *The Journal of Finance*, 46(5), 1575–1617. <https://doi.org/10.1111/j.1540-6261.1991.tb04636.x>
14. Yahoo Finance. (2025). YFinance Python package (Version 0.2.65) [Computer software]. <https://pypi.org/project/yfinance/>
15. Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56. [https://doi.org/10.1016/0304-405X\(93\)90023-5](https://doi.org/10.1016/0304-405X(93)90023-5)
16. Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3), 425–442. <https://doi.org/10.1111/j.1540-6261.1964.tb02865.x>
17. Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3), 607–636. <https://doi.org/10.1086/260061>
18. Ang, A., & Bekaert, G. (2002). International asset allocation with regime shifts. *The Review of Financial Studies*, 15(4), 1137–1187. <https://doi.org/10.1093/rfs/15.4.1137>
19. Connor, G., & Korajczyk, R. A. (1988). Risk and return in an equilibrium APT: Application of a new test methodology. *Journal of Financial Economics*, 21(2), 255–289. [https://doi.org/10.1016/0304-405X\(88\)90062-1](https://doi.org/10.1016/0304-405X(88)90062-1)
20. Lettau, M., & Pelger, M. (2020). Factors that fit the time series and cross-section of stock returns. *The Review of Financial Studies*, 33(5), 2274–2325. <https://doi.org/10.1093/rfs/hhaa020>
21. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
22. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>

23. Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2), 689–702. <https://doi.org/10.1016/j.ejor.2016.10.031>
24. Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669. <https://doi.org/10.1016/j.ejor.2017.11.054>
25. Bailey, D. H., Borwein, J., López de Prado, M., & Zhu, Q. J. (2016). The probability of backtest overfitting. *Journal of Computational Finance*, 20(4), 39–69. <https://doi.org/10.21314/jcf.2016.322>
26. Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), 171–185. <https://doi.org/10.1080/01621459.1987.10478410>
27. White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097–1126. <https://doi.org/10.1111/1468-0262.00152>
28. Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4), 1455–1508. <https://doi.org/10.1093/rfs/hhm014>
29. Moskowitz, T. J., & Grinblatt, M. (1999). Do industries explain momentum? *The Journal of Finance*, 54(4), 1249–1290. <https://doi.org/10.1111/0022-1082.00146>
30. Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1), 65–91. <https://doi.org/10.1111/j.1540-6261.1993.tb04702.x>
31. Jiang, W. (2021). Applications of deep learning in stock market prediction: Recent progress. *Expert Systems with Applications*, 184, 115537. <https://doi.org/10.1016/j.eswa.2021.115537>