

Hybrid prediction systems for long-horizon stock performance in equity investment

Andrii Gorbenko, *Carson-Newman University*, agorbenko@cn.edu

Seongyong Hong, *Carson-Newman University*, shong@cn.edu

Abstract

Traditional forecasting models often struggle with the noise and non-linearity of long-term market trends, leading to suboptimal strategic asset allocation. This research aims to develop a hybrid system that integrates diverse analytical techniques to enhance the accuracy of long-horizon predictions and provide more robust decision support for equity investors. Therefore, this study involves developing and testing a ranking stock-picking model that ranks U.S. stocks and predicts their outperformance against the SPDR S&P 500 Exchange-Traded Fund (SPY). The motivation for the research was its possible application for personal wealth building as well as school investment challenges. The stock scoring system is built upon three fundamentally different methods for assessing stock quality. First, profitability, leverage, liquidity, and operational efficiency are assessed via the Piotroski F-Score using SEC EDGAR financial statements. Second, intrinsic value is estimated by forecasting future free cash flows and discounting them to present value, then computing a margin of safety relative to the current price. Third, the Fama-French five-factor model estimates expected annual returns based on market beta, company size, value, profitability, and investment sensitivity. These three signals are combined in four architectures that combine the signals in progressively more selective ways: Parallel Composite Scoring, Factor-Adjusted DCF, Internally Consistent Hybrid Composite, and Conviction Gate Pipeline. We hold the three signals fixed and vary only how they are gated, ranked, and weighted, then evaluate each architecture on overlapping 36-month cohorts (2020–2026). The metric is the hit rate, the fraction of selected stocks that outperform the SPDR S&P 500 ETF (SPY) over the holding period. Architecture, not the construction of any single signal, accounts for most of the difference in performance: two architectures built from identical signals differ by 29 percentage points in average hit rate (27% versus 56%). Because this rests on only four overlapping cohorts, we treat the results as a feasibility and architecture-sensitivity study rather than evidence that one method predicts returns better than another. The contribution is methodological: it shows that how signals are combined is itself a major, and largely untested, driver of long-horizon stock selection.

Keywords: Discounted Cash Flow, Factor Models, Piotroski F-Score, Intrinsic Value, Portfolio Selection

Introduction

Predicting long-term stock performance is a tough task due to the natural ups and downs of financial markets. Traditional econometric models often get lost in "noise" over long periods. On the other hand, machine learning techniques alone can miss key economic principles. This study aims to create stronger hybrid prediction systems that blend the clear structure of financial theory with the powerful data handling of deep learning. By merging these approaches, this study hopes to improve accuracy in forecasts and

returns adjusted for risk for long-term investors. It seeks to connect computational skills with smart financial decision-making.

Outperforming a passive benchmark, the return you would earn by simply holding the whole index such as SPY instead of picking stocks, on a consistent basis is an exceedingly difficult task, yet academic research has demonstrated that certain firm characteristics are associated with persistent return differences. Evidence shows that value, profitability, and accounting quality demonstrate consistency and economic rationale (Fama & French, 1992, 2015; Piotroski, 2000). This research addresses one practical question: is it possible to develop a clear, replicable method using publicly available fundamentals and market data to periodically rank S&P 500 stocks expected to beat SPY over the next 36 months? A 36-month horizon is chosen deliberately. Daily or weekly forecasting is prone to significant noise and microstructure effects, whereas at the multi-year horizon, fundamental factors generating persistent abnormal returns have a chance to manifest.

Restricting the universe to S&P 500 proxies also provides a conservative test: academic literature consistently documents stronger factor effects in mid- and small-cap stocks. This research therefore combines three complementary approaches: the Piotroski F-Score, which scores a firm's financial health on nine points drawn from its statements; a discounted cash flow (DCF) engine with margin of safety, which values a firm from its projected future cash and gauges how far the price sits below that value; and a Fama-French five-factor expected return model, which predicts a stock's return from five common risk traits. All components are implemented in Python 3.12 with strict point-in-time data alignment, which restricts every calculation to the data available at the time, enforcing a 45-day SEC EDGAR filing lag, and each prediction window spans exactly 36 months, eliminating look-ahead bias, so the model never uses information that was unavailable on the decision date. Each of these three approaches is well documented on its own, but the literature says little about how to combine them, and it usually treats the combination step as an implementation detail rather than something worth studying. That gap is the starting point for this paper: with the signals held fixed, how much does the way they are combined change which stocks get selected? The four combination architectures, and the finding that the choice among them matters more than the choice of any single signal, are the paper's main contribution.

Literature Review

There are many studies about cross-sectional return differences. In their famous article published in 1992, Fama and French showed that market beta, a measure of how strongly a stock moves with the overall market, explains less of the differences in returns than size and book-to-market ratios, a stock's accounting value divided by its market value, where high readings flag value stocks. The extension of the model made by these authors in 2015 added two new systematic factors, broad market-wide forces that move many stocks at once: profitability (RMW) and investment (CMA). Jegadeesh and Titman (1993) independently showed existence of the momentum effect, the tendency of recent winners to keep outperforming for a while, making an extension of the current approach plausible. According to McLean and Pontiff (2016), the signal predictability survives in practice, thus allowing using the documented signals in live trading strategies. In this line of work, signals are tested one at a time: the question is whether a given factor predicts returns, not how several validated factors should be combined.

Regarding the accounting fundamentals area, Piotroski demonstrated that a combination of nine profitability, liquidity, and quality signals reliably predicts future performance of value stocks. Another study by Novy-Marx (2010) confirmed the power of gross profitability in predicting stock returns. Bernard and Stober (1989) concluded that cash flows and accruals represent information of fundamentally different nature, with cash flows being superior in predicting fundamentals. Sloan (1996) confirmed that accruals component of profits is predictive of lower future returns, thus supporting the cash flow quality signal from F-Score. Bernard and Thomas (1990) demonstrated that earnings have not been completely impounded on stock prices, thus retaining their fundamental value. Discounted Cash Flow (DCF) valuation is another way of assessing company's value grounded on Williams (1938) framework and developed in Damodaran (2000). According to this method, company's intrinsic value equals the sum of discounted free cash flows and terminal growth value. The Gordon Growth formula (Gordon, 1959) is used for terminal value calculation. The intrinsic value to market price ratio is used as a scoring metric and as a margin of safety. For controlling risks, the methodology of Fabozzi, Markowitz, and Gupta (2008) will be followed, meaning equal-weighting portfolio construction to avoid calculating covariance matrix. According to Harvey, Liu, and Zhu (2016), many signals tested on a sample make significance threshold substantially higher than conventional ones.

The application of machine learning techniques to stock prediction in conjunction with fundamental analysis has been investigated recently. In their study, Phan and Chang (2024) use Long Short-Term Memory (LSTM), one-dimensional Convolutional Neural Networks (1D CNN), and Logistic Regression to forecast stock trends with the aid of fundamental variables in 269 publicly listed companies in stable market sectors in the U.S. between 2019 and 2023. The feature set used by the authors included the variables from financial statements as well as ratio-based calculations and the intrinsic values obtained through Discounted Cash Flow analysis. These elements are very similar to the ones used in current research. Importantly, Logistic Regression was found to produce higher test accuracy compared to the more complex models of deep learning. Particularly, Logistic Regression provided an average of 74.66% accuracy in predicting annual stock prices' movements and 72.85% in forecasting whether stock's intrinsic value is greater than its market price. The authors suggested that features used for the study are neither sequential nor spatial in nature, thus failing to leverage advantages of CNNs and LSTMs. This result is consistent with the findings in this study, which show that simple models based on fundamental analysis can be highly effective when forecasting stocks' behavior. The authors Gkolemis, Lee & Roudani (2024) devised a market neutral strategy based on a multi-factor model for the stocks of NYSE, using momentum indicators and other fundamental factors together with analyst recommendations. After starting with an initial pool of 162 factors, after performing multiple filtering operations such as correlation analysis, individual regression with economic sign check, and multivariate regression, the number of factors was reduced to 11 factors that showed some degree of predictability. Some factors selected by the researchers include the RSI indicator, book-to-market ratio, Return on Assets, Sales-to-Price ratio, and momentum factors, which are similar to the factors used in the Fama-French model in the current study. Three methods were tested for constructing portfolios: equally weighted, risk parity, and minimum variance beta-neutral portfolios, which showed that risk parity yielded the best results. Namely, the Sharpe Ratio, which measures return earned per unit of risk, was 0.81 compared to 0.57 for the S&P 500 index from 2016 to 2024, and the maximum drawdown, the largest peak-to-trough loss along the way, was -16.87% vs -33.93%. An important aspect noted by the authors regarding the methodology is related to the instability of factor coefficients over time, where the significance of factor signs became unstable in later test years. The work of Cohen et al. (2025) introduced a hybrid approach combining semantic augmentation of conventional machine learning-based predictions

with LLM scores provided by ChatGPT-4o for fundamental, technical, and entropy-based predictions on NASDAQ-100 stocks in 2020-2025. Based on findings, an appropriate weighting scheme of LLM-semantic and ML-algorithmic signals varies significantly depending on both prediction types and investment horizons involved. For instance, the purely ML-based technical predictor generated cumulative returns of 1,978% with monthly rebalancing, while the hybrid weighting (ML weight $w = 0.15$) worked best for the fundamental strategy and $w = 0.70$ was optimal for the entropy one. Such differences imply that semantic content derived by LLMs provides an additional value in qualitative and long-term areas of interest like fundamental analysis. Specifically, quarterly rebalancing in the fundamental strategy yielded the highest returns in cases where LLM semantic signals alone were considered ($w = 0$). Hence, interpretation of earnings releases, management expectations, and macroeconomic information may become more efficient than numerical scoring of news over longer investment periods. These recent studies tune the predictive model itself, whether an LSTM, CNN, logistic regression, or LLM augmentation, while keeping the signal set largely fixed. The reverse question, how to combine established signals well, is the one this study takes up, and it has received far less attention.

Methodology

The three methods were chosen because they measure different, mostly non-overlapping aspects of stock quality. The Piotroski F-Score captures accounting quality and whether the fundamentals are improving; the DCF margin of safety captures valuation against estimated intrinsic worth; and the Fama-French five-factor model captures expected return from systematic risk exposures. A stock can score well on one of these and badly on another, and that is exactly when combining them helps. Figure 1 shows how the data sources, signals, and architectures fit together.

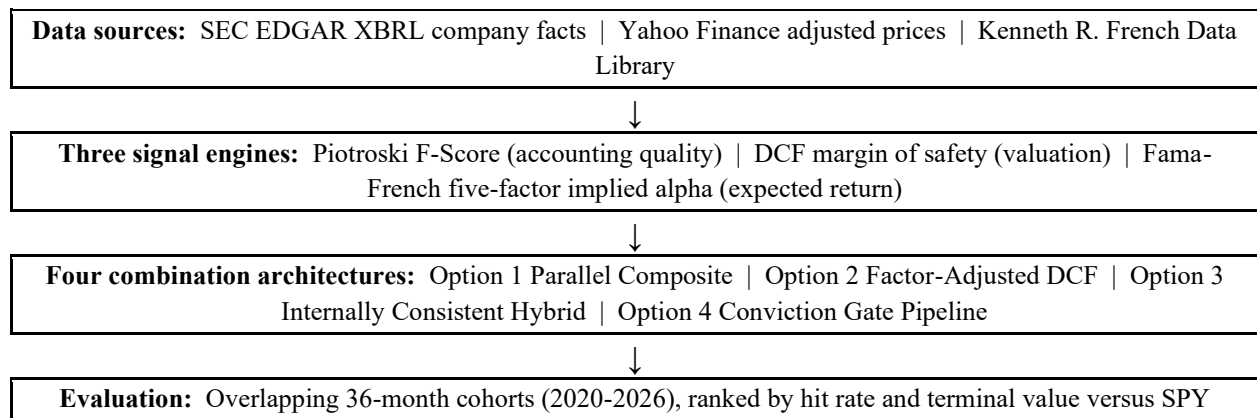


Figure 1. Conceptual Research Framework

Three data sources were used to create the process. Financial statement items (operating cash flow, capital expenditures, net income, total assets, long-term debt, current assets and liabilities, shares outstanding, revenue, and cost of goods sold) are sourced from SEC EDGAR XBRL company facts. Adjusted closing prices can be downloaded via Yahoo Finance. The monthly Fama-French five-factor returns, and the risk-free rate are obtained from the Kenneth R. French Data Library. This library is freely available for academic

purposes. All parts of the program were developed and executed in Python 3.12 on a MacBook Air with an Apple M5 processor. A strict two-layer point-in-time policy was imposed on all data elements to eliminate look-ahead bias. Each SEC filing is available only when its filing date and the period-end date plus 45 days are on or before the valuation date. The 45-day floor is based on the SEC's requirement to file annual reports by a certain date. For backtests, which are trial runs of the strategy on past data, adjusted prices are corrected for stock splits that occurred after the valuation date. This was done to restore the original historical prices on each evaluation date. Before combining signals, each is standardized cross-sectionally, against all the other stocks on the same date, at every evaluation date using z-scoring, which rescales each signal to a common mean of zero and standard deviation of one: $Z_{i,t}(X) = (X_{i,t} - \mu_t) / \sigma_t$, where μ_t and σ_t are the cross-sectional mean and standard deviation across all stocks at time t . This makes signals with different units and scales comparable. The composite score, a single combined score built from the standardized signals, is $SCORE_{i,t} = \sum_k w_k \times Z_{i,t}(Signal_k)$. Equal weights ($w_k = 1/3$) serve as the baseline in Options 1 and 3; Option 4 uses empirically calibrated weights of 50% for factor alpha, 35% for F-Score, and 15% for margin of safety, reflecting observed sub-group predictive power. The target variable is 36-month total return relative to SPY, computed as the difference in absolute percentage returns.

Table 1. Notation Used Across the Three Signal Methods

Symbol	Definition
$Z_{i,t}(X)$	Cross-sectional z-score of raw signal X for stock i at evaluation date t
μ_t, σ_t	Cross-sectional mean and standard deviation of a signal across all stocks at date t
w_k	Weight applied to signal k in the composite score
$SCORE_{i,t}$	Weighted composite score for stock i at date t
FCF	Free cash flow = operating cash flow – capital expenditures
CFO, Capex, NI	Operating cash flow, capital expenditures, and net income
r	Discount rate (required return on equity)
R_f	Risk-free rate (10-year U.S. Treasury yield)
β	Market beta (60-month OLS versus SPY) used in the CAPM discount rate
ERP	Equity risk premium
g_1, g_T	Stage-1 free cash flow growth rate (5%) and terminal growth rate (2.5%)
H	Stage-1 forecast horizon (3 years)
PV_k, TV, EV	Present value of period- k cash flow, terminal value, and enterprise value
IV	Intrinsic value per share
MOS	Margin of safety = $(IV - Price) / Price$
$R_{i,t}(\text{excess})$	Excess return of stock $i = R_{i,t} - R_{f,t}$
$\beta_{i,k}$	Sensitivity (factor beta) of stock i to factor k
λ_k	Factor premium = rolling 36-month mean of factor k
$E[R_i]$	Model-predicted annualized expected return
α_{implied}	Implied alpha versus SPY = $E[R_i] - E[R_{SPY}]$

Method 1: Piotroski F-Score

The Piotroski F-Score is a nine-signal binary scoring system evaluated using audited financial statements. Table 2 shows that each signal equals one if its condition holds and zero otherwise; the total score ranges from 0 to 9. Signals are organized into three groups: Profitability (F1~F4), Leverage and Liquidity (F5~F6), and Operating Efficiency (F7~F9). Higher scores indicate improvements across all three dimensions. Financial-sector (SIC 6000~6999) and utility companies (SIC 4900~4999) are excluded, as leverage and cash flow metrics have fundamentally different interpretations in those industries. Trailing twelve-month values for flow items (operating cash flow, capital expenditures, net income) are computed using the formula: $TTM = FY_last + YTD_recent_quarter - YTD_same_quarter_prior_year$, with sanity checks that reject quarterly YTD figures deviating from annual totals by more than a factor of 2.5.

Table 2. Piotroski F-Score: Nine Binary Signals

Signal	Group	Condition	Economic Rationale
F1	Profitability	$ROA(t) > 0$	Positive return on assets indicates profitable use of capital.
F2	Profitability	$CFO(t) > 0$	Positive operating cash flow shows the core business is genuinely cash generative.
F3	Profitability	$\Delta ROA > 0$	Improving ROA signals increasing management effectiveness over time.
F4	Profitability	$CFO(t) > NI(t)$	CFO exceeding net income indicates high earnings quality; accruals are low (Sloan, 1996).
F5	Leverage	$\Delta LEV < 0$	Decreasing long-term debt-to-assets ratio reduces financial distress risk.
F6	Liquidity	$\Delta CR > 0$	Improving current ratio reduces the risk of a near-term liquidity shortfall.
F7	Financing	$Shares(t) \leq Shares(t-1)$	No share issuance; buybacks signal excess cash and return of shareholder value.
F8	Efficiency	$\Delta GM > 0$	Improving gross margin reflects pricing power or cost reduction.
F9	Efficiency	$\Delta ATO > 0$	Improving asset turnover indicates increasing operational productivity.

Piotroski (2000) recommended classifying scores of 0~2 as weak candidates, 3~6 as neutral, and 7~9 as strong long candidates. In this research, the F-Score serves as one input into the broader composite score in Options 1 and 3, a binary gate requiring F-Score ≥ 6 in Option 2, and a hard gate requiring F-Score ≥ 8 in Option 4. The distinction between thresholds is empirically motivated: in the backtests conducted here, stocks scoring exactly 7 achieved a 23% hit rate versus 60% for stocks scoring 8 or above. As a point-in-time validation, ACI Worldwide (ACIW) received an F-Score of 8/9 for fiscal year 2024 and rose approximately 57% during that year. Intel (INTC), which received a poor score driven by negative free cash flow in fiscal year 2022, declined approximately 55% over the 36 months following the January 2023 evaluation date.

Method 2: Discounted Cash Flow Engine

Table 3 shows that the DCF engine calculates the intrinsic value of each firm by discounting projected free cash flows to the present. When intrinsic value exceeds the current market price, the stock is considered undervalued and the difference is the Margin of Safety (MOS). This valuation approach is rooted in the theory of Williams (1938) and formalized by Damodaran (2000). Free cash flow is defined as operating cash flow with less capital expenditure.

Table 3. DCF Formula Chain - From Financial Data to Margin of Safety

Formula	Name	Economic Role
$FCF = CFO - Capex$	Free Cash Flow	Residual cash after sustaining capital expenditures; harder to manipulate than accrual-based earnings (Bernard & Stober, 1989).
$r = Rf + \beta \times ERP$	Discount Rate (CAPM)	Required return on equity. Rf = risk-free rate; β = 60-month OLS market beta vs. SPY; ERP = equity risk premium (Sharpe, 1964).
$FCF(t+k) = FCF \times (1+g)^k$	Stage-1 FCF Forecast	FCF projected forward $k = 1$ to $H = 3$ periods at near-term growth rate $g_1 = 5\%$.
$PV_k = FCF(t+k) / (1+r)^k$	Stage-1 Present Values	Each projected cash flow discounted to present value at discount rate r .
$TV = FCF_H \times (1+gT) / (r - gT)$	Terminal Value	Gordon Growth Model (Gordon, 1959) capturing all value beyond the 3-year forecast horizon.
$EV = \sum PV_k + PV(TV)$	Enterprise Value	Sum of all Stage-1 discounted cash flows plus the discounted terminal value.
$Equity = EV + Cash - Long-term Debt$	Equity Bridge	Converts enterprise value to equity value by adding excess cash and subtracting long-term debt.
$IV = Equity / Shares Outstanding$	Intrinsic Value per Share	Value estimate derived from discounted FCF.
$MOS = (IV - Price) / Price$	Margin of Safety	Primary ranking signal. Positive MOS = undervalued; negative MOS = overvalued.

Bernard and Stober (1989) established that cash flows are the better predictor of underlying fundamentals compared to accrual-based earnings, providing the theoretical basis for using FCF rather than net income as the valuation base. Financial-sector companies use a net income proxy model with capital expenditures set to zero, because their operating cash flow figures are distorted by loan origination activity. The discount rate follows the CAPM specification of Sharpe (1964): $r = Rf + \beta \times ERP$, where Rf is the 10-year U.S. Treasury yield, β is estimated via rolling 60-month OLS regression of the stock's monthly excess returns against SPY, and ERP is the equity risk premium. A stock with $\beta > 1$ is riskier than the market and receives a higher discount rate, producing a lower intrinsic value. The valuation formula chain proceeds from free

cash flow to intrinsic value in eight steps, as summarized in Table 3. Stage 1 projects FCF for $H = 3$ years at a uniform 5% growth rate; Stage 2 applies a Gordon Growth terminal value at $g_T = 2.5\%$. The terminal value typically constitutes 70~85% of total enterprise value, making it the most sensitive component of the valuation and the largest source of estimation error.

Method 3: Fama-French Five-Factor Expected Returns

Table 4 shows how the factor model estimates each stock's annualized expected return using a rolling 36-month OLS regression, a standard statistical fit re-estimated over a moving window, of monthly excess returns, meaning each stock's return above the risk-free rate, on five Fama-French factors. The five factors characterize distinct dimensions of systematic risk. The Market factor (Mkt-RF) measures aggregate market risk. The Size factor (SMB: Small Minus Big) captures the historically observed small-cap premium. The Value factor (HML: High Minus Low) captures the premium associated with high book-to-market stocks. The Profitability factor (RMW: Robust Minus Weak) measures the return spread between high and low operating-profitability firms. The Investment factor (CMA: Conservative Minus Aggressive) measures the return spread between firms with low versus high asset growth. Factor premia are estimated as the rolling 36-month mean of each factor's realized return over the same window.

Table 4. Factor Model: Key Formulas

Symbol	Name	Formula	Economic Meaning
$R_{i,te}$	Excess Return	$R_{i,te} = R_{i,t} - R_{f,t}$	Stocks return above the risk-free rate; isolates compensation for systematic risk.
$\beta_{i,k}$	Factor Beta	OLS regression coefficient over 36-month window	Stock i 's sensitivity to factor k ; estimated by regressing excess returns on five factor returns.
λ_k	Factor Premium	mean($f_{k,t}$) over rolling 36-month window	Historical average compensation per unit of exposure to factor k .
$E[R_i]$	Expected Return	$E[R_i] = R_f + \sum_k \beta_{i,k} \times \lambda_k$, annualized as $(1 + \text{monthly})^{12-1}$	Model-predicted annual return; drives portfolio selection.
α_{implied}	Implied Alpha vs. SPY	$E[R_i] - E[R_{SPY}]$	Model-implied outperformance versus the benchmark; stocks ranked by this value.

As an illustration, Apple Inc. (AAPL) as of February 28, 2025, had estimated betas of 1.20 (market), -0.30 (size), -0.20 (value), $+0.40$ (profitability), and -0.10 (investment). Combined with a 4.5% risk-free rate and historical factor premia, this yielded an expected annualized return of 15.35% and an implied alpha of $+1.85\%$ versus SPY's 13.50%, placing AAPL at the 24th percentile, just outside the top-decile portfolio cutoff, the top tenth of the ranking. A standalone 25-month backtest from February 2023 to February 2025 over approximately 3,935 tickers yielded an annualized return of $+11.2\%$ against SPY's $+22.4\%$, with a Sharpe ratio of 0.54 versus 1.87 and a portfolio beta of 1.31. While this standalone result underperformed SPY, the factor model's primary contribution in this framework is as one of three composite signals where its combination with F-Score and MOS substantially improves hit rates.

Four Architectural Options

All four options use the same three signals, F-Score, DCF MOS, and factor model that implied alpha, but differ fundamentally in how those signals are combined, gated, and weighted. This section describes each architecture and the design reasoning behind it.

Option 1: Parallel Composite Scoring

Option 1 runs the F-Score, DCF (using a CAPM-derived discount rate), and factor model independently in parallel. Each signal is winsorized, so its extreme values are capped at fixed bounds, to prevent outliers from dominating the composite: MOS is capped at $\pm 200\%$ and alpha at $\pm 50\%$. Any stock missing a signal has it imputed at the 25th percentile of the cross-sectional distribution, a deliberately low stand-in for the missing value, preserving universe breadth. Each signal is then z-scored cross-sectionally and combined with equal weights of one-third each. The top 10 stocks by composite score are selected with equal weights. The core strength of Option 1 is inclusiveness: imputation maximizes universe coverage and no single signal can alone exclude a stock. A company with a low DCF score can still rank highly if it has an exceptional F-Score and strong factor alpha. The structural weakness is internal inconsistency: the DCF uses a CAPM discount rate (one risk factor), while the alpha signal uses a five-factor model. Two stocks with identical market beta but different profitability profiles receive the same DCF discount rate yet different factor alphas, creating a systematic asymmetry that cannot be corrected by winsorization. A practical example of failure under Option 1 is CVNA from the April 2021 cohort: its missing F-Score was imputed at the 25th percentile, but its strong factor alpha propelled it into the top ten. CVNA subsequently declined approximately 70% over the next 36 months, and the absence of a valid F-Score was an early warning sign masked by imputation.

Option 2: Factor-Adjusted DCF with F-Score Gate

Option 2 resolves the inconsistency of Option 1 by replacing the CAPM discount rate in the DCF engine with the five-factor expected return for each stock. This means both the discount rate and the ranking signal are derived from the same theoretical model. The F-Score acts as a binary gate, a pass-or-fail filter a stock must clear to stay in, requiring a score of 6 or above; stocks below the threshold are excluded before ranking. Surviving stocks are ranked solely by their factor-adjusted MOS, and the top N are selected. No imputation is performed; stocks with missing data are excluded entirely. While Option 2 achieves internal theoretical consistency, it achieves the lowest average hit rate of all four options (27%). The reason is structural: ranking by a single signal provides no safeguard against estimation errors that are specific to the DCF component. A one-time FCF boost from an asset sale, or a non-recurring tax benefit can inflate the margin of safety to the top of the ranking, with no counterbalancing signal to detect the anomaly. Furthermore, because both the discount rate and the ranking score are derived from the same 36-month regression window, any estimation error in the factor model propagates simultaneously into both components, eliminating the diversifying correction that multiple independent signals would provide.

Option 3: Internally Consistent Hybrid Composite

Option 3 combines the structural strength of Option 1 (blending multiple signals) with the theoretical improvement of Option 2 (using the five-factor expected return as the DCF discount rate). The factor model output serves two purposes simultaneously: as the discount rate in the DCF calculation and as the third composite signal via the implied alpha. This creates a structural negative correlation of approximately -0.29

between the factor-adjusted MOS and the implied alpha signals. Stocks with high implied alpha have high discount rates and therefore lower (more negative) MOS; stocks with high MOS have lower required returns and therefore lower expected alpha. This complementarity means the two signals partially offset each other, providing natural tension that reduces the influence of extreme outliers in either direction. Option 3 delivered the highest individual cohort return of all four options: in the 2021~2024 cohort, its portfolio reached \$23.76 on a \$10 investment versus SPY's \$16.39, driven by URI (+144%), WAB, BLDR, and GEN. Similarly, the 2022~2025 cohort portfolio reached \$22.16 versus SPY's \$15.20, driven by TSN, BKNG, JBL, QCOM, and ANET. However, Option 3 underperformed SPY in 2020~2023 and 2023~2026 cohorts. The high variance stems from the absence of gates preventing value traps, which are stocks that look cheap but keep sinking as the business worsens, sector concentration, and weak-signal persistence.

Option 4: Conviction Gate Pipeline

Option 4 was designed systematically from the failure modes observed in Options 1~3 across all four backtesting cohorts. It applies six sequential gates before selecting the final portfolio. Gate 1 (Anti-value-trap filter) removes stocks where $MOS > 0$ but $\alpha < 0$ simultaneously. This combination of apparent DCF undervaluation but expected factor underperformance appeared in eleven stock-cohort observations across Options 1~3 and yielded an average hit rate of only 18% and average return of -0.5% . Repeat offenders included CVS Health (three observations), AT&T (two), and Tyson Foods (two), all cases where historical FCF appeared strong, but the factor model was pricing in deteriorating business conditions not yet visible in financial statements. Gate 2 (Hard F-Score threshold) requires $F\text{-Score} \geq 8$ with no missing data and no imputation. Gate 3 (Reweighted composite) assigns weights of 50% to factor alpha, 35% to F-Score, and 15% to MOS, reflecting sub-group hit rates of 62%, 60%, and 48% respectively. Notably, MOS alone underperforms random selection at 48%, justifying its reduced weight. Gate 4 (Persistence bonus) adds $+0.3$ z-score to stocks that appeared in the prior-period top 20 with $F\text{-Score} \geq 8$. The motivating example is Wabtec Corporation (WAB), which received the bonus across five consecutive scoring periods and averaged $+92.7\%$ return versus SPY's approximately 26% over those same windows. The counterexample is CVS Health, which received the bonus twice but underperformed SPY by 40% and 29% respectively, as pharmacy benefit management reimbursement pressures had not yet appeared in backward-looking financial statements. Gate 5 (Sector cap) limits the final selection to three stocks per GICS sector, a standard industry grouping, addressing the Consumer Staples concentration observed in the April 2021 cohort under Option 1, where four Consumer Staples stocks were selected simultaneously. Gate 6 (Conviction Tier) selects the top five surviving stocks with equal weights of 20% each.

Results

To compare all four options on equal footing, each was applied to the S&P 500 universe at four scoring dates on April 1 of 2020, 2021, 2022, and 2023. Then, actual returns over the following 36 months were tracked. The primary evaluation metric is the hit rate: the fraction of top-ranked stocks that outperformed SPY over the subsequent 36 months. Table 5 reports the hit rates for each option and cohort and provides a dimension-by-dimension comparison of all four options. The April 2020 cohort produced uniformly poor hit rates across all options. SPY delivered $+60.9\%$ following the V-shaped COVID-19 recovery, a regime driven by high-beta growth stocks rather than the financially conservative, fundamentals-quality stocks that

all four options tend to select. This is a structural feature of the framework, not an estimation failure. In the 2021~2024 cohort, Option 3 delivered the highest absolute return (\$23.76 vs. SPY \$16.39), but Option 4 demonstrated the most consistent cross-cohort behavior, with \$21.00, \$16.29, and \$16.92 across the three most recent cohorts compared to SPY's \$16.39, \$15.20, and \$17.84 respectively.

These differences come from the architecture, not the signals, which are the same across options. Option 2 ranks on a single signal whose discount rate and ranking score both come from the same 36-month regression window, so one estimation error throws off both at once, with no independent signal to offset it; that is why its hit rate is low. Option 1 imputes missing signals and weights them equally, which gives it wide coverage but also lets in value traps such as CVNA. Option 3 carries a built-in negative correlation (about -0.29) between factor-adjusted MOS and implied alpha, so the two signals partly cancel, and its results swing more from one cohort to the next. Option 4's gates were built to remove the specific failures seen in Options 1 through 3, which is why it is the steadiest across cohorts even when it does not post the highest return in any single one. A full risk-adjusted evaluation, with the Sharpe ratio, maximum drawdown, and beta computed consistently across all four options, would meaningfully strengthen these comparisons, but it was beyond the time available for this study and is left for future work.

Table 5. Multi-Cohort Hit Rate Comparison Across All Four Options

Cohort	SPY 36M Return	Option 1 (top 10)	Option 2 (top 10)	Option 3 (top 10)	Option 4 (top 5)
Apr 2020 → Apr 2023	+60.9%	—	3/10 = 30%	4/10 = 40%	1/5 = 20%
Apr 2021 → Apr 2024	+31.1%	3/10 = 30%	3/10 = 30%	6/10 = 60%	3/5 = 60%
Apr 2022 → Apr 2025	+23.3%	5/10 = 50%	2/10 = 20%	6/10 = 60%	2/5 = 40%
Apr 2023 → Apr 2026	+22.9%	6/10 = 60%	3/10 = 30%	4/10 = 40%	4/5 = 80%
Average (2021~2023)	—	47%	27%	53%	56%*

* Option 1 excludes 2020 as it was not evaluated for that cohort.

The tradeoff is clear: Option 4 reduces variance at the cost of capping upside in the best-performing cohorts. Table 6 illustrates how the same five stocks receive different treatment across the four options. CVS and AT&T rank highly in Options 1 and 2 due to strong DCF MOS signals driven by high historical free cash flows but are excluded entirely from Option 4 because their F-Scores fall below the hard threshold and they exhibit the value-trap pattern (positive MOS, negative alpha).

Table 6. Sample Output Comparison Across Four Options, April 2023 Scoring Date

Ticker	F-Score	Opt. 1 Rank	Opt. 2 Rank	Opt. 3 Rank	Opt. 4 Status
WAB	9/9	#1	#3	#2	Conviction Tier (persistence bonus)
BLDR	8/9	#3	#7	#1	Conviction Tier
CVS	7/9	#2	#1	#5	Excluded (F < 8; value trap flag)
PSX	8/9	#5	#4	#4	Conviction Tier
T (AT&T)	6/9	#8	#2	#9	Excluded (F < 8; value trap flag)

Figure 2 illustrates the performance of the Option 4 hybrid prediction system, specifically focusing on the conviction-tier results for a cohort initiated on April 1, 2022. This visualization serves as a crucial proof-of-concept for the model's ability to generate alpha over a long-horizon timeframe, even when facing significant macroeconomic challenges. The blue line represents the concentrated five-stock portfolio (WY, WAB, MAS, META, and URI) selected by Option 4's rigorous multi-stage filtering process. Initially, during the broad market downturn of 2022, the portfolio tracked closely with the S&P 500 (SPY) benchmark, demonstrating the inherent risk of high beta equities in a rising-rate environment. However, a distinct divergence began in early 2023. This "inflection point" highlights the efficacy of the Margin of Safety (MOS) and F-score gates, which ensured that only companies with robust fundamental recovery potential remained in the conviction tier.



Figure 2. Example of Option 4 results

Option 4's unique features, specifically its sector diversification cap and value-trap filters, prevented overexposure to declining industries, while the temporal memory (persistence bonus) allowed the model to maintain positions in high-conviction stocks, such as META, during their recovery phases.

Table 7. Comparative Summary of All Four Options Across Key Evaluation Dimensions

Dimension	Option 1	Option 2	Option 3	Option 4
Risk model	Inconsistent (CAPM + FF5)	Consistent (FF5 only)	Consistent (FF5 only)	Consistent (FF5 only)
Signals used for ranking	3, equal weight	1 (MOS_ff5 only)	3, equal weight	3, weights 50/35/15
Missing data handling	Imputed at p25	Excluded	Excluded	Excluded, hard
F-Score role	Graded signal	Binary gate (≥ 6)	Graded signal	Hard gate (≥ 8)
Value-trap filter	None	None	None	Yes (MOS+ and alpha-)
Sector diversification	None	None	None	Cap 3 per GICS sector
Temporal memory	None	None	None	Persistence bonus (+0.3Z)

Portfolio size (S&P 500)	Top 10 of ~200	Top 10 of ~40–60	Top 10 of ~30–50	Top 5 of ~15–30
Avg hit rate (2021–23)	47%	27%	53%	56%*
Computing time (S&P 500)	~6–8 min	~10–12 min	~10–12 min	~14–17 min

By the conclusion of the evaluation period in early 2025, the Option 4 portfolio achieved a terminal value of \$21.00, significantly outperforming the SPY's \$18.29. This outperformance confirms that the integration of weighted signal ranking (50/35/15) and hard F-Score gates (≥ 8) successfully mitigates the "noise" inherent in long-horizon forecasting. The results suggest that while the hybrid system remains subject to short-term market volatility, its structural emphasis on fundamental quality and valuation alignment provides a superior risk-adjusted trajectory, one judged against the risk taken rather than the raw return alone, compared to a passive index strategy. The same framework also produced clear failures: CVNA (–70%) survived an imputed F-Score under Option 1, CVS Health trailed SPY by 40% and 29% even with the persistence bonus, and Option 2 fell to a 27% hit rate overall. We include these so the picture is not built only from the cases that worked. Table 7 shows that WAB ranks first in Option 1 and receives the persistence bonus in Option 4 due to five consecutive prior-period qualifications, subsequently returning approximately +83% over 36 months.

Discussion

The most striking finding is not the absolute performance of any single option, but how much performance varies based on architectural design. Options 2 and 4 use identical underlying signals: the same factor model, DCF engine, and F-Score. Yet their average hit rates differ by 29 percentage points. This variance is generated entirely by architecture: rank on one signal versus three, gate before scoring or after, uniform weights versus empirically calibrated weights, and the presence or absence of temporal memory. The signal construction itself becomes almost secondary. This observation carries an important implication for quantitative finance practice: academic literature typically evaluates signals in isolation, asking whether a given factor predicts returns. But a signal that works alone can produce sub-par results when poorly integrated with complementary signals. The combination architecture generates additional predictive power that is not adequately recognized in the existing literature. Option 2's failure is the most instructive result in the paper. On the surface, using the same risk model for both the discount rate and the ranking signal achieves clean internal consistency. The problem is that both components rely on the same 36-month regression window: an estimation error in the factor model simultaneously distorts both the discount rate and the rank score, with no counterbalancing signal available. This is a case where theoretical elegance reduces empirical robustness.

Option 4's Persistence Bonus is the most nuanced feature retrospectively. Wabtec Corporation (WAB) received the bonus for five consecutive scoring periods and outperformed significantly each time, because the underlying business quality was genuine and the F-Score accurately reflected it. CVS Health received the bonus in two successive cohorts and underperformed by 40% and 29% respectively, because pharmacy benefit management reimbursement pressures were building in ways the income statement had not yet

captured. This illustrates the fundamental limitation of any backward-looking accounting model: financial statements reflect the past, and business trends can shift before the F-Score detects them. Several limitations apply, and the results should be read as preliminary. First, Option 4's six gates were designed after the fact from failure modes seen in Options 1 through 3 over these same cohorts, so the 56% hit rate is an in-sample number, measured on the very data used to build the model, and we cannot tell here how much of it is real signal rather than fitting. A confirmatory study would need walk-forward backtesting, an expanding-window method that trains on past data, tests on the next stretch, then steps forward and repeats, an untouched holdout period, a stretch of data set aside and never touched during design, and a multiple-testing correction, an adjustment for having tried many variants, along the lines of Harvey, Liu, and Zhu (2016), none of which we claim to have passed. Second, the four cohorts use overlapping 36-month windows and so are not independent, which is why we report descriptive hit-rate comparisons and make no claim of statistical significance, that is, evidence that a gap is unlikely to be down to chance; showing significance would take an independent, non-overlapping design. Third, the sample is only four cohorts, and the 29-point gap between Options 2 and 4, though striking, cannot yet be separated cleanly from sampling noise. Fourth, the Option 4 weights (50/35/15) are heuristic, set by rule of thumb, taken from sub-group hit rates of 62%, 60%, and 48% rather than formally optimized, and we have not tested how sensitive the results are to other weightings. Fifth, limiting the universe to S&P 500 proxies probably understates the signals, since factor effects tend to be stronger in mid- and small-cap stocks with thinner analyst coverage. Sixth, the margin of safety signal on its own reached only a 48% hit rate, below chance, because terminal value (70–85% of enterprise value) is so sensitive to the long-run growth assumption; a simpler metric such as free cash flow yield may rank stocks more reliably in future work. Seventh, because the same dataset was used to design and to evaluate the architectures, live performance may fall short of what we report; the strict point-in-time policy keeps look-ahead bias out of each window but is not a substitute for a true holdout. For practitioners, the practical guidance differs by use case. For broad quarterly coverage of the S&P 500, Option 1 provides the easiest entry point: signal imputation maximizes universe coverage, equal weights prevent any single signal from dominating, and the CAPM-FF5 inconsistency is mitigated by cross-sectional winsorization. For concentrated high-conviction selection, Option 4 provided the most consistent hit rates across the cohorts evaluated here, though the persistence bonus requires at least one prior scoring period to activate fully. An ensemble approach of selecting stocks ranking highly in at least two of the four options offers the best risk-adjusted starting point under regime uncertainty. Wabtec Corporation ranked first in Option 1, second in Option 3, and received the persistence bonus in Option 4, passing all six gates, and subsequently returned approximately +83% over 36 months.

Conclusion

This research demonstrated that three well-established stock evaluation signals of the Piotroski F-Score, DCF margin of safety, and Fama-French five-factor implied alpha, can be combined into a practical stock-ranking system, but that the combination architecture determines outcomes more than any individual signal. Options 2 and 4 use identical inputs yet differ by 29 percentage points in average hit rate, driven entirely by sequencing, gating, and weighting decisions. Among the architectures tested, Option 4's Conviction Gate Pipeline was the steadiest, with a 56% average hit rate and 80% in the most recent 2023–2026 cohort. With only four overlapping cohorts behind that number, we read it as promising but preliminary, not as a validated result. Its six sequential gates target specific failure modes identified from Options 1–3: value

traps, weak F-Score qualifiers, over-concentrated sectors, and cross-period inconsistency. The primary drawback is portfolio concentration where five equally weighted stocks carry meaningful risk. The Conviction Tier should be treated as a high-quality shortlist for further research rather than a ready portfolio. All four options share one non-negotiable methodological commitment for live deployment: strict point-in-time data alignment with a 45-day SEC filing lag. This ensures backtest results are conservative and reproducible estimates of live performance, free from the look-ahead bias that inflates most published backtests.

Acknowledgements

The authors would like to acknowledge the use of generative AI tools to assist in search references and language refinement during the development of this manuscript.

References

- Anthropic. (2025). Claude (Version Claude Sonnet 4.6) [Large language model]. <https://claude.ai>
- Bernard, V. L., & Stober, T. L. (1989). The nature and amount of information in cash flows and accruals. *The Accounting Review*, 64(4), 624–652.
- Bernard, V. L., & Thomas, J. K. (1990). Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. *Journal of Accounting and Economics*, 13(4), 305–340.
- Cohen, G., Aiche, A., & Eichel, R. (2025). Artificial intelligence models for predicting stock returns using fundamental, technical, and entropy-based strategies: A semantic-augmented hybrid approach. *Entropy*, 27(6), 550. <https://doi.org/10.3390/e27060550>
- Damodaran, A. (2000). Discounted cash flow valuation. New York University Online Library.
- Fabozzi, F. J., Markowitz, H. M., & Gupta, F. (2008). Portfolio selection. *Handbook of Finance*, 2, 3–13.
- Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, 47(2), 427–465.
- Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22.
- French, K. R. (2025). Kenneth R. French data library [Data set]. Dartmouth College. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

- Gkolemis, G. M., Lee, A. R., & Roudani, A. (2024). A multi-factor market-neutral investment strategy for New York Stock Exchange equities. arXiv preprint arXiv:2412.12350. <https://arxiv.org/pdf/2412.12350>
- Gordon, M. J. (1959). Dividends, earnings, and stock prices. *The Review of Economics and Statistics*, 41(2), 99–105.
- Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1), 5–68.
- 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.
- Kapinus, M., Liashenko, K., Ljepava, N., Liashenko, L., & Danylov, V. (2024). Predicting stock market trends with Python. *Grail of Science*, (40), 109–116.
- McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1), 5–32.
- Novy-Marx, R. (2010). The other side of value: Good growth and the gross profitability premium (No. w15940). National Bureau of Economic Research.
- Phan, J., & Chang, H.-F. (2024). Leveraging fundamental analysis for stock trend prediction for profit. arXiv preprint arXiv:2410.03913. <https://arxiv.org/pdf/2410.03913>
- Piotroski, J. D. (2000). Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research*, 1–41.
- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3), 425–442.
- Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, 289–315.
- U.S. Securities and Exchange Commission. (2025). EDGAR XBRL company facts API [Data set]. <https://data.sec.gov/api/xbrl/companyfacts/>
- Williams, J. B. (1938). *The theory of investment value*. Harvard University Press.

Zarei, F., & Nourahmadi, M. (2025). Graph theory in action: Forming an investment portfolio. *Interdisciplinary Journal of Management Studies*, 19(1), 1–19.