Applied Economics and Finance; Vol. 13, No. 2; 2026
ISSN 2332-7294 E-ISSN 2332-7308
Published by Redfame Publishing

From Prediction to Implementation: The Role of Machine Learning in Modern Portfolio Optimization

Ying Zhang[1,2]

[1] California State Polytechnic University Pomona, Pomona, CA, USA

[2] Continuum Research & Capital LLC, San Marino, CA, USA

Correspondence: Ying Zhang, California State Polytechnic University Pomona, Pomona, CA, USA.

Received: September 11, 2026 Revised: October 28, 2026 Accepted: November 7, 2026

Available online: November 24, 2026 URL: https://doi.org/10.11114/aef.v13i2.9548

Abstract

Machine learning can improve return forecasts without improving the portfolios built from them. This narrative review examines 60 selected references to ask when predictive gains survive portfolio construction and implementation. The studies differ in objectives, allocation rules, costs, and realized fund results. Three findings emerge. First, evidence for improved forecasts and backtests is stronger than evidence for persistent net gains at scale. Second, performance reflects the full investment design, so studies rarely isolate the contribution of machine learning itself. Third, costs, capacity limits, and mixed fund results weaken broad claims of superiority. The review maps what forecast, pricing, portfolio-backtest, fund-selection, and realized-fund studies can establish, then identifies tests to clarify attribution. The evidence supports value in some applications, but does not show that machine learning itself produces general, persistent, scalable outperformance.

Keywords: machine learning, portfolio optimization, transaction costs, asset management, model uncertainty

JEL classification: C45, C58, G11, G12, G17

1. Introduction

A stronger return forecast does not guarantee a better portfolio. Investors still decide how much to hold, how much risk to accept, and whether each trade is worth its cost. Machine learning adds richer predictors and financial text to this process. Its value depends on how those inputs affect the final portfolio.

The problem begins with portfolio theory. Markowitz (1952) shows how expected returns and covariances determine portfolio choice. Both inputs require estimation. An optimizer can place excessive weight on favorable estimation errors (Michaud, 1989). Better prediction may reduce this problem. It may also create unstable rankings that require frequent trading. The relevant effect depends on how the forecast enters the allocation rule.

The studies reviewed here approach this connection from different starting points. Gu et al. (2020) examine flexible return prediction. In Chen et al. (2024), financial restrictions help determine what the model learns. DeMiguel et al. (2020) evaluate characteristics through their joint value after trading costs. Reading these papers together raises a question that individual performance figures leave open: where in the investment process does the improvement arise?

Earlier reviews survey financial machine-learning methods, practical pitfalls, and fund performance (Kelly and Xiu, 2023; Buczynski et al., 2021; Blitz et al., 2023). Giglio et al. (2022) survey machine learning in factor models and asset pricing. This review asks what each type of evidence can establish about investment performance. It groups studies by what they test, from forecasts and pricing models to portfolio backtests, fund selection, and realized fund returns. It uses gaps in this evidence to propose tests that separate a learning method's contribution from the effects of portfolio design. This map organizes the evidence; it is not a validated score.

2. Review Scope and Evidence Selection

This focused narrative review brings together 60 selected references dated from 1952 to 2026. Publication details and selected sources were checked through October 28, 2026. The cutoff identifies the scope of verification, not exhaustive coverage of publications.

This review focuses on the link between forecasts and portfolio choices. It includes foundational work on estimation risk, representative financial machine-learning studies, and research on trading costs. It also draws on three reviews of financial machine learning and quantitative asset management (Kelly and Xiu, 2023; Buczynski et al., 2021; Blitz et al.,

2023), a financial-NLP survey (Du et al., 2025), research on forecast uncertainty, and one supervisory guidance document. The reviews provide context, not independent replications. A source was included if it supplied a benchmark, a mechanism, or a limitation relevant to an investment stage. Selection did not depend on whether a study found gains from machine learning.

Titles and author names were checked against journal and publisher pages, NBER records, and author or university repositories. The supervisory document was checked against its official agency source. Targeted searches covered decision-focused learning, deep and end-to-end portfolio policies, reinforcement learning, realized mutual- and hedge-fund performance, trading costs, short-selling constraints, fund capacity, covariance estimation, tail risk, and backtest selection. These searches were purposive, not systematic. The review does not claim exhaustive coverage or a pooled estimate of investment gains.

Evidence is compared by asset setting, objective, portfolio rule, trading costs, and evaluation outcome. Theoretical results, empirical findings, and proposed tests are distinguished throughout. Each correction notice is interpreted with its original study.

Purposeful selection may underrepresent unsuccessful models and less prominent markets. Public securities receive more attention than private assets. Reinforcement learning and alternative data are covered selectively. Where only an abstract or publication record was accessible, the discussion does not infer unreported assumptions or robustness tests.

Figure 1 organizes the comparison. The information stage asks what was observable at the decision date. Prediction asks what the model learns beyond a benchmark. Portfolio construction asks how uncertainty and constraints affect positions. Implementation asks what remains after trading frictions and scale. Section 9 turns these distinctions into practical evidence requirements.

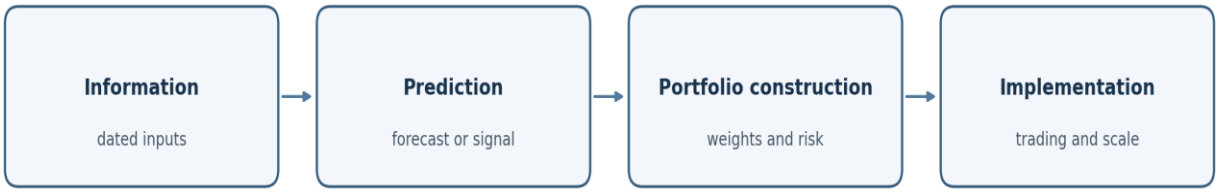

Figure 1. Evidence stages from dated information to implemented portfolios

2.1 How the Evidence Is Interpreted

The unit of comparison is the claim made by a study. A paper may contain several claims with different evidentiary strength. For example, improved prediction within one sample is narrower than successful transfer to another market. A theoretical allocation rule also has a different status from a realized trading result. Keeping these units separate reduces the risk of treating every favorable finding as evidence of implementability.

A statistically significant result may be too small to cover operating costs. A large estimated gain can instead depend on assumptions that are difficult to implement. Replication adds another consideration: whether the result persists when the sample or setting changes. These forms of evidence answer different questions. A study can make a useful contribution while supplying only some of them.

Comparisons within a common design receive greater weight. Cross-paper comparisons identify possible mechanisms, not isolated causal effects. Different samples and implementation assumptions prevent a pooled performance ranking. Bibliographic verification also cannot remove publication or selection bias. These limitations apply to the comparisons that follow.

3. Estimation Risk and the Role of Constraints

Classical portfolio research offers a necessary benchmark for machine learning. DeMiguel et al. (2009) compare 14 allocation models across seven datasets. None consistently outperforms equal weighting on their out-of-sample criteria. This is evidence about estimated allocation rules. It is not a test of every modern forecasting method. The result nevertheless raises the standard for any claim that richer inputs improve portfolio decisions.

Several remedies address different sources of fragility. Ledoit and Wolf (2004) shrink sample covariance estimates toward a structured target. Goldfarb and Iyengar (2003) formulate allocation problems around uncertain parameters. Jagannathan and Ma (2003) explain why nonnegative weights can reduce estimated portfolio risk. Their later correction addresses a proof error in the original paper (Jagannathan et al., 2019). These approaches share a concern with sensitivity, but they do not impose the same assumptions or solve identical problems.

These remedies affect different parts of the allocation process. Shrinkage changes risk inputs, while constraints limit the response to them. Robust formulations represent uncertainty explicitly. Their effects therefore need to be separated from

improvements in the return forecast itself.

Feature selection raises a related issue. Harvey et al. (2016) show why a large search space requires stronger statistical evidence. Feng et al. (2020) test whether a proposed factor adds information beyond existing factors. Kozak et al. (2020) study shrinkage in a high-dimensional stochastic discount factor. Together, they challenge the idea that more candidate predictors automatically mean more reliable information. They do not establish that the same small feature set is optimal for all investment objectives.

This distinction guides the following comparisons. Section 3 examines sensitivity in the allocation layer. Section 4 then considers what changes when the predictor or its training objective changes.

3.1 How Estimation Errors Enter Portfolio Weights

Rankings and return magnitudes enter portfolio decisions differently. A sorting strategy can use the ranking alone. An optimizer also responds to the estimated size of the advantage. Two models may therefore produce similar sorted portfolios but sharply different optimized weights. Looking only at their prediction scores would miss the source of that divergence.

Risk estimates introduce another channel. When assets appear highly correlated, the optimizer may use offsetting positions to exploit a small predicted difference. A slight change in estimated correlation can alter those positions substantially. This is why the risk model deserves its own sensitivity analysis. Treating it as a fixed technical component beneath the forecasting model can obscure how it changes the resulting positions.

A tighter position limit can restrain estimation noise, but it can also suppress a useful signal. Turnover limits create a similar tension. They save trading costs while delaying the response to new information. Shrinkage acts earlier, by changing the estimated inputs that drive the weights. Evaluating these choices requires tracing both the protection they provide and the opportunities they remove.

One diagnostic would freeze the predictor and vary covariance estimates and predeclared position limits. Changes in concentration, leverage, and realized risk would show the importance of the allocation layer. The experiment in Section 9 extends this logic to predictor–allocation interactions.

The investor's mandate also changes the benchmark. Equal weighting is informative when estimating unrestricted weights is difficult. It may be less informative for an investor with sector limits or a benchmark-relative objective. A fair benchmark satisfies the same material constraints as the proposed strategy. Otherwise, the comparison mixes the value of prediction with the value of a different investment opportunity set.

3.2 Learning Risk and Choosing the Risk Objective

Return prediction is only one route to better allocation. Risk estimation is another. Ledoit and Wolf (2017) develop nonlinear covariance shrinkage for portfolio selection. This provides a stronger risk-estimation benchmark than the raw sample covariance matrix. The relevant comparison pairs a new learning method with established shrinkage under common constraints. Lower realized volatility would support a risk-management claim. It would not establish better expected-return forecasts.

Variance also leaves some investment objectives unresolved. Rockafellar and Uryasev (2000) develop optimization with conditional value-at-risk. This directs attention to losses in the distribution's tail. Drawdown asks a different question because it depends on the path of wealth. A strategy can have an attractive average Sharpe ratio and still violate a downside-risk mandate. A useful comparison specifies the loss measure and evaluates the corresponding outcomes.

Distributional uncertainty extends parameter uncertainty. Mohajerin Esfahani and Kuhn (2018) consider distributions near the empirical distribution under a Wasserstein metric. Their framework includes mean-risk portfolio applications and conditional performance guarantees. These guarantees depend on stated assumptions. They do not ensure protection against every market break. Selecting the uncertainty radius and testing its sensitivity remain consequential choices. These methods complement machine learning by changing how uncertain inputs become decisions.

3.3 Why Prediction Gains Do Not Overturn the Equal-Weighting Evidence

DeMiguel et al. (2009) and Gu et al. (2020) can appear to offer conflicting messages. The former finds no consistently superior estimated allocation rule among its alternatives. The latter finds useful gains from flexible prediction. Yet their comparison units differ. DeMiguel et al. evaluate portfolio rules against equal weighting. Gu et al. compare forecasting methods in a large panel of individual equities. Better performance on the second task does not settle the first.

The two findings can coexist because portfolio weights may remain unstable even after average prediction accuracy improves. A few extreme forecasts can still drive large positions. A more conservative allocation rule may help by limiting that response, without improving the forecasts themselves. This reading connects the studies through estimation sensitivity. Neither paper supplies a causal decomposition of the difference between them.

Section 9.3 proposes a controlled predictor-by-allocation comparison. It holds the universe, dates, and material constraints constant. The test then traces prediction gains into net portfolio outcomes.

4. What the Empirical Evidence Establishes

Gu et al. (2020) provide a useful quantitative benchmark. In their 30-year U.S. equity test, monthly out-of-sample R^2 is 0.71% for the generalized linear model and up to 1.80% for a neural network. NN3 raises the S&P 500 timing Sharpe ratio from 0.51 for buy-and-hold to 0.77. A value-weighted long-short neural-network strategy has an annualized Sharpe ratio of 1.35. These results show predictive and portfolio gains under the paper's design. They do not establish a common net-of-cost advantage: the market-timing exercise caps leverage at 50%, turnover is high, and no model beats static UMD in the UMD timing exercise.

Freyberger et al. (2020) use U.S. stocks from July 1963 to June 2015. Their adaptive group LASSO finds that many characteristics add no incremental information, while nonlinear effects matter. The authors report out-of-sample explanatory power above linear panel regressions and Sharpe ratios about 50% higher. The result is evidence for structured nonlinear prediction and selection. It does not quantify net returns after institutional trading costs.

Chen, Pelger, and Zhu (2024) test U.S. stock returns from 1967 to 2016, with 1992–2016 held out. Their no-arbitrage GAN model has an annualized out-of-sample SDF Sharpe ratio of about 2.6, compared with about 1.5 for the feed-forward neural forecasting model. It explains 8% of individual-return variation and reports a cross-sectional R^2 of 23%. This is a strong result for a conditional pricing model. It combines network flexibility with economic restrictions and adversarial test-asset selection, so it is not a clean estimate of architecture alone or of net tradable performance.

The bond evidence shows why information timing belongs in the comparison. Bianchi, Büchner, and Tamoni (2021) forecast Treasury excess returns over a sample beginning in August 1971, with recursive out-of-sample forecasts from 1990 and the historical mean as benchmark. After correcting information availability, a comparative table reports monthly OOS R^2 of 6.10% for a one-layer neural network, 6.20% for extremely randomized trees, and 5.10% for LASSO for 13–24-month bonds; neural-network predictability remains statistically and economically significant (Leow & Lindenthal, 2025, Table 1). These are forecast statistics, not net portfolio returns.

Cakici et al. (2023) extend the test to 46 markets from 1991 to 2020, with out-of-sample evaluation from 2001. Their pooled forecast-combination portfolio earns 1.51% per month at an annualized Sharpe ratio of 1.49. A one-month data skip lowers these to 1.30% and 0.89. The strategy's average one-sided turnover is 110.5% per month, and its global break-even cost is 1.52%. This provides broad cross-market evidence and a direct implementation sensitivity. It also shows that a reported gross return cannot be read as a net return.

Table 1 brings together forecast tests, pricing-model tests, portfolio backtests, fund-selection studies, realized fund outcomes, and capacity evidence. The rows differ in sample, target, portfolio rule, risk exposure, and cost treatment, so the table is not a performance ranking. Forecast R^2, pricing-kernel Sharpe, backtested returns, and realized fund alpha answer different questions. Cost-adjusted backtests remain model-based evidence; observed fund returns are closer to realized performance but usually do not isolate a specific learning algorithm. The table shows what each design supports and what it leaves unresolved.

Table 1. Selected empirical findings and their inference limits

Study and sample	Evidence type, method, and benchmark	Main reported result	Costs, robustness, and inference boundary
DeMiguel et al. (2009). Seven empirical datasets; periods vary by dataset.	Allocation-rule test. 14 estimation-based portfolio rules versus 1/N equal weighting; evaluated by out-of-sample Sharpe ratio, certainty-equivalent return, and turnover.	No rule consistently beats 1/N across the three criteria. Under U.S.-calibrated simulations, mean-variance estimation needs about 3,000 months for 25 assets and 6,000 months for 50 assets to outperform 1/N.	Turnover is reported as a performance criterion; this is not a uniform modern transaction-cost test. It establishes a demanding allocation benchmark, not a test of today's return-prediction models.
Gu et al. (2020). U.S. equities, 1957–2016; 30-year OOS period (1987–2016).	Forecast and timing test. Linear and regularized forecasts, trees, and neural networks; OOS predictive R^2 benchmarked to a zero-return forecast. Market timing compared with S&P 500 buy-and-hold; long-short deciles compared across model classes.	Monthly OOS R^2: 0.71% for GLM and up to 1.80% for a neural network. S&P 500 timing Sharpe: 0.77 for NN3 versus 0.51 buy-and-hold. Value-weighted long-short NN spread Sharpe: 1.35.	Market-timing positions cap leverage at 50%; long-only strategies exclude short sales. Reported Sharpe ratios are not a common net-of-cost ranking. Turnover is high, and no method beats static UMD in the UMD timing test.
Freyberger et al. (2020). U.S. common stocks, July 1963–June 2015.	Characteristic-sort test. Adaptive group LASSO estimates nonparametric characteristic effects; compared with linear panel regressions and alternative selection methods.	Equal-weighted OOS Sharpe: 3.42 for the nonparametric model versus 2.26 for the linear model; value-weighted Sharpe: 1.24 versus 1.01. The nonparametric model selects eight characteristics, compared with 21 for the linear model.	The OOS Sharpe ratios are not net of a common trading-cost model. Value weighting lowers them; excluding the smallest stocks reduces the Sharpe to 1.33 for the nonparametric model and 0.93 for the linear model. This remains evidence about characteristic sorts, not institutional net returns.
Chen, Pelger, and Zhu (2024). U.S. equities, January 1967–December 2016; 25-year OOS test (1992–2016).	Pricing-model test. No-arbitrage GAN/SDF versus regularized linear SDF and a feed-forward neural forecasting model; evaluates Sharpe ratio, explained variation, and cross-sectional R^2.	Annualized OOS SDF Sharpe is about 2.6 for GAN versus about 1.5 for the feed-forward forecasting model; GAN explains 8% of return variation and has cross-sectional R^2 of 23%.	The comparison tests a full pricing design, including adversarial test assets and no-arbitrage loss. It is not a net trading strategy; trading costs are not deducted.
Bianchi, Büchner, and Tamoni (2021), original. U.S. Treasury excess returns, 1971:08–2018:12; recursive OOS forecasts from 1990.	Original forecast test. One-layer neural network with seven nodes and lagged forward-rate inputs versus the historical-mean (expectations-hypothesis) benchmark; covers several maturities.	The original Table 1 reports OOS R^2 of 6.4% for the 2-year bond and 20.6% for the equally weighted bond portfolio ($p = .028$ and .002, respectively). These are the original, uncorrected estimates.	The study also evaluates forward rates with macroeconomic variables. Its later corrigendum corrects information timing and lowers OOS R^2; the corrected results are shown separately. These are forecast statistics, not net portfolio returns.

Study and sample	Evidence type, method, and benchmark	Main reported result	Costs, robustness, and inference boundary
Bianchi et al. (2021), corrigendum. Treasury excess returns, 1971:08–2018:12; recursive OOS forecasts from 1990.	Corrected forecast test. Neural networks, extreme trees, and penalized regressions versus the historical-mean benchmark; the comparative table reports 13–24-month bonds using forward rates and macro variables.	The corrected comparison reports monthly OOS R^2 of 6.10% for a one-layer neural network, 6.20% for extremely randomized trees, and 5.10% for LASSO; neural-network predictability remains statistically and economically significant (Leow & Lindenthal, 2025, Table 1).	The correction directly tests information timing and lowers the original OOS R^2. These are forecast results, not net portfolio returns; the 6.10%–6.20% figures apply to the stated maturity/input specification.
Cakici et al. (2023). 46 markets; January 1991–December 2020, OOS from January 2001; up to 148 characteristics.	Portfolio backtest. 11 models, including OLS, penalized regressions, trees, neural networks, and a forecast combination; value-weighted monthly long-short quintile portfolios.	The pooled global forecast-combination strategy earns 1.51% per month with annualized Sharpe 1.49. With a one-month data skip, its average return is 1.30% per month and Sharpe 0.89.	Combination-strategy one-sided turnover averages 110.5% per month; global break-even cost is 1.52%. Returns vary by market and size. The skip-period test addresses reliance on very recent information; break-even cost is not an observed net return.
DeMiguel et al. (2020). U.S. equities, 51 characteristics; OOS implementation sample May 1988–December 2014.	Cost-aware portfolio test. Characteristic-based portfolios with and without transaction-cost-aware joint selection; evaluates portfolio utility and net performance relative to conventional characteristic strategies.	The number of economically significant characteristics rises from 6 without costs to 15 with costs, as trades generated by different characteristics can offset.	The finding is within-study and depends on the portfolio and cost specification. It does not imply that trading costs generally improve returns or that the same 6-to-15 change applies across studies.
Avramov, Cheng, and Metzker (2023). U.S. stocks, OOS 1987–2017; additional credit-rating subsample starts in 1999.	Portfolio backtest with cost estimates. Neural-network and other learned signals; long-short decile portfolios compared with OLS and factor benchmarks, including FF6-adjusted returns.	For the GKX NN signal, the full-sample value-weighted long-short strategy has annualized Sharpe 0.944, FF6-adjusted return 0.92% per month, and one-sided turnover 97.6%. Excluding microcaps lowers Sharpe to 0.644.	Estimated average trading cost is 0.67%; break-even cost for this strategy is 0.94%, while most model break-even estimates are at or below estimated costs. Results weaken under restrictions and reasonable costs, but are not uniformly unprofitable.
DeMiguel et al. (2023). 8,767 U.S. equity fund share classes, January 1980–December 2020; OOS portfolio returns, 1990–2020.	Fund-selection backtest. Long-only portfolios select top-decile funds using fund characteristics; alpha benchmarked to Carhart four-factor, FF5, FF5+momentum, and FF5+momentum+liquidity models.	Gradient boosting and random forests earn 2.36% and 2.69% annual net alpha versus FF5+momentum; linear OLS and elastic net estimates are not statistically significant.	Fund expenses and transaction costs are included in net returns; alternative factor benchmarks give similar findings. The result is fund selection in this sample, not proof that stock-level ML strategies scale similarly.
Chen and Ren (2022). U.S. mutual-fund returns, January 2009–December 2019.	Realized AI-labelled mutual-fund returns. Realized AI-powered funds compared with the market and human-managed peer funds.	AI funds do not significantly outperform the market on a risk-adjusted basis, but earn 5.8% higher annual net returns than human-managed peers in the reported comparison.	The authors attribute the peer gap partly to lower turnover/costs and stock selection. The sample is short and the AI label does not isolate the causal effect of a specific model.

Study and sample	Evidence type, method, and benchmark	Main reported result	Costs, robustness, and inference boundary
Chen, Sialm, and Xu (2026). AI adviser disclosures from 2012–2024; hedge-fund performance from 2005–2024.	Realized AI-driven hedge-fund returns. Observed AI-driven hedge funds compared with non-AI peers, including within-adviser/sibling-fund analyses.	AI funds outperform peers by about 50 basis points per month before 2017; the advantage is statistically indistinguishable from zero in later years.	The study adds realized fund evidence, but AI classification and observational performance comparisons do not isolate a causal algorithm effect. Later-period findings are an important persistence check.
Jensen et al. (2026). U.S. equities, 1981–2020; investor wealth reaches $10 billion by 2020; relative risk aversion 10.	Cost-aware portfolio backtest at modeled scale. Cost-aware dynamic Portfolio-ML versus Static-ML; reports gross and net outcomes, volatility, turnover, leverage, and trading costs.	Table 2: gross/net Sharpe 1.38/1.33 versus 0.95/0.79; annualized volatility 11% versus 28%; monthly turnover 25% versus 80%; average leverage 3.00 versus 12.78.	The design incorporates trading costs and investor scale. However, the reported Sharpe gap is not a matched-risk test of allocation quality: leverage differs sharply, and the net trading-cost measure does not by itself resolve financing spreads, margin calls, or forced liquidation. The comparison therefore supports a result about complete portfolio designs under stated assumptions, not an isolated predictor contribution.
Pástor, Stambaugh, and Taylor (2015). U.S. active equity funds, 1979–2011.	Capacity context, not an ML test. Active-fund returns compared with passive benchmark portfolios; evaluates scale, skill, gross and net performance.	Across the sample, mean benchmark-adjusted monthly return is about +0.05% gross and −0.05% net of fees and expenses.	This is capacity context, not an ML test. It illustrates that fund-scale and fees can absorb gross skill; it cannot identify the capacity of a particular learned strategy.
Allena (2026). U.S. stock-return forecasts; OOS evaluation, January 1987–February 2020. Numerical values below are from the publicly available 2023 author manuscript.	Confidence-filtered portfolio backtest. NN-3 Confident-HL versus conventional high-low (HL), with equal- and value-weighted portfolios; the published article's abstract confirms the direction across models.	In the NN-3 results, monthly OOS return and annualized Sharpe are 3.84% and 1.78 for equal weights (versus 2.21% and 1.36 for HL), and 2.70% and 1.26 for value weights (versus 1.29% and 0.78). The 2026 published abstract confirms superior OOS returns and Sharpe ratios across models but does not report point estimates.	The author manuscript reports turnover-based cost adjustments and downside-risk checks. These results test forecast precision as a portfolio input; the stated cost approximation does not establish institutional capacity.
Kaniel et al. (2023). Actively managed U.S. equity mutual funds, 1980–2019; 407,158 fund-month observations.	Fund-selection backtest. Feed-forward neural network predicts locally estimated four-factor fund alpha using fund and fund-family characteristics, holdings-weighted stock characteristics, and market sentiment; ranks funds monthly into predicted-performance deciles.	The top predicted-performance decile earns about 15 basis points of monthly abnormal return and the bottom decile loses about 25 basis points; the roughly 40-basis-point monthly spread is similar after fees and persists for more than three years.	The article reports out-of-sample results and similar after-fee spreads because top- and bottom-decile funds have similar fees. This is evidence on fund selection, not a direct estimate of capacity or market impact for a scalable long-short strategy.

Note. "OOS" means out of sample. The method column identifies forecast tests, pricing-model tests, portfolio backtests, fund-selection backtests, observed AI-labelled fund outcomes, and capacity context. Cost adjustments in a backtest do not make it a realized fund return. Results retain the units and benchmarks used in each study and are not ranked across rows. Gross return, net return, forecast R^2, pricing-kernel Sharpe, and fund alpha are distinct metrics. Break-even costs are thresholds, not realized net returns.

4.1 Matching the Evaluation Metric to the Decision

Forecast evaluation is most informative when it reflects the decision the model supports. Squared-error loss penalizes large numerical mistakes. A ranking measure asks whether higher forecasts correspond to higher subsequent returns. A portfolio metric additionally depends on position sizes and risk exposures. These measures can move in different directions without any calculation being wrong. They assess different properties of the forecast and its use.

For example, a model may improve predictions for the middle of the return distribution. That improvement can lower average prediction error. It may have little effect on a strategy that trades only the highest and lowest ranks. Another model may improve the extreme ranks while producing poorly calibrated return magnitudes. That model could help a sorting strategy but remain unsuitable for an unconstrained optimizer. These are analytical possibilities, not empirical results established by this review.

A portfolio can also earn a high Sharpe ratio with modest security-level forecasting accuracy. Its result may depend on a rewarded factor exposure or a stable risk-management rule. Exposure analysis helps distinguish these explanations. Reporting the forecasts alongside the resulting positions makes it easier to see what the model contributed and what the allocation rule supplied.

4.2 Separating Model Development from Evaluation

A test period can influence model development even when every training observation precedes it. Repeatedly inspecting its results may change the chosen features, transformations, or tuning ranges. The final specification then reflects that period indirectly. Reserving a final evaluation period protects against this feedback only if later redesigns are recorded and treated as new research decisions.

Preprocessing belongs inside that separation. Parameters for scaling, imputation, and feature selection should be estimated using information available to the training procedure. A full-sample transformation can transmit future information without an explicit future-return variable. Similarly, a historical dataset can contain revisions unavailable at the time. A credible study describes the availability convention, rather than relying on an out-of-sample label alone.

Overlapping forecast horizons require additional care. If a target covers several months, nearby training and test observations may share return realizations. A chronological boundary alone may not eliminate that overlap. The target horizon therefore belongs in sample separation and uncertainty estimation. This issue is distinct from choosing a particular algorithm. It concerns how observations enter the experiment.

The model-selection process deserves the same attention as the final specification. Extensive tuning gives a flexible model more chances to fit validation noise. A minimally tuned benchmark may have had far fewer opportunities. Recording the search space, selection rule, and number of configurations makes the comparison interpretable. Equal computation is not always appropriate, but undisclosed differences in search effort leave an important part of the result unexplained.

These requirements extend the multiple-testing concern raised by Harvey et al. (2016). A model can overfit through repeated research decisions as well as through its parameters. Stronger evidence would therefore include the final specification, its development history, and a genuinely subsequent evaluation. An unusually favorable test result should invite examination of that history rather than immediate extrapolation.

4.3 What Changes When the Model or Objective Changes

Gu et al. (2020) and Chen et al. (2024) illustrate an important distinction within neural-network research. Gu et al. organize estimation around return-prediction error. Chen et al. use no-arbitrage conditions and an adversarial construction of informative test assets. A pricing error concerns whether an estimated pricing model satisfies economic restrictions. A prediction error concerns the difference between a forecast and a realized return. These losses need not rank models identically.

The loss function is central to this comparison. With a different objective, the same network can learn different relationships. Better results from an economically structured model cannot therefore be assigned entirely to network depth. Keeping inputs and network capacity comparable while changing the loss would clarify this issue. Predictive accuracy and pricing performance would both be informative, followed by a portfolio comparison under a fixed allocation rule.

Freyberger et al. (2020) provide a second contrast with Gu et al. Their characteristic-selection approach uses structured nonlinear effects. Gu et al. compare broader model classes, including methods that capture interactions. Nonlinear responses and interactions are different sources of flexibility. A curved relationship for one characteristic does not establish that its effect changes with another characteristic. This distinction helps explain why variable-selection results need not match model-importance rankings.

The meaning of an important variable changes with the exercise. Conditional contribution under a specified model differs

from reliance by a fitted prediction system. Correlated inputs can make the distinction consequential. A mismatch between selected characteristics and importance rankings is therefore worth investigating, rather than treating as an immediate inconsistency. Adding interactions to a structured baseline, using common inputs and dates, would help locate the source of the difference.

The Bianchi correction discussed above adds information timing to this comparison. A model's representational capacity, training objective, and observable inputs are separate sources of variation. Keeping them separate helps explain differences in findings without assigning every improvement to algorithm choice.

4.4 Evidence on Economic Implementation and Realized Fund Performance

Evidence on trading frictions is mixed across markets and portfolio designs. In China, Leippold, Wang, and Zhou (2022) find that liquidity is an important return predictor and report economically significant out-of-sample results after transaction costs. Avramov, Cheng, and Metzker (2023) provide a detailed U.S. benchmark. Their neural-network long-short portfolio has an annualized Sharpe ratio of 0.944 and a monthly FF6-adjusted return of 0.92%. Monthly turnover is 97.6%. Excluding microcaps lowers Sharpe to 0.644. Their estimated average trading cost is 0.67%, compared with a 0.94% break-even cost for this strategy; across methods, most break-even costs are at or below the estimated cost. The implication is attenuation and uneven tradability, not universal failure.

Evidence from technical signals points to a further boundary. Brogaard and Zareei (2023) find that machine-learning technical trading rules using past prices can earn out-of-sample profits, but those profits decline over time. Murray, Xia, and Xiao (2024) show that machine-learning forecasts from historical price charts predict returns among large U.S. stocks and remain useful out of sample. These papers support predictive value in their settings; they do not, by themselves, establish that every resulting strategy earns durable net returns after realistic implementation costs.

Fund evidence separates model-based selection from funds that themselves advertise or use AI. DeMiguel et al. (2023) study 8,767 U.S. fund share classes from 1980 to 2020 and evaluate long-only portfolios from 1990 to 2020. Gradient boosting and random forests deliver 2.36% and 2.69% annual net alpha relative to FF5 plus momentum; linear OLS and elastic net are not statistically significant. Kaniel et al. (2023) use a neural network to predict abnormal returns for actively managed U.S. equity funds from 1980 to 2019. The predicted top decile earns about 15 basis points of monthly abnormal return, while the bottom decile loses about 25 basis points. The top-bottom spread is roughly 40 basis points per month, is similar after fees, and persists for more than three years. Fund flows and fund momentum are the strongest predictors. These are out-of-sample fund-selection results, not proof of strategy capacity after portfolio-level market impact. Chen and Ren (2022) find no significant risk-adjusted outperformance by AI-labelled funds against the market. In their 2009–2019 sample, those funds earn 5.8% higher annual net returns than human-managed peers. Lower turnover and costs contribute to that peer gap. These results are favorable in particular comparisons, but they do not isolate one common causal effect of machine learning.

Evidence on funds using AI in actual management is beginning to extend to hedge funds. Chen, Sialm, and Xu (2026) examine hedge-fund returns from 2005 to 2024 and adviser disclosures from 2012 to 2024. AI-driven hedge funds outperform non-AI peers by about 50 basis points per month before 2017, but the advantage is statistically indistinguishable from zero in later years. Their within-adviser comparisons strengthen the peer design, yet the classification and returns remain observational. The later-period result also cautions against assuming that an early performance gap persists.

5. Financial Text and Incremental Information

Ke et al. (2019) use supervised learning to extract return-related information from news. Du et al. (2025) survey a wider range of financial NLP applications. These sources serve different purposes. The first supplies a specific predictive method. The second maps applications and challenges. The survey maps applications and challenges; its breadth does not provide independent evidence that every text application improves investment performance.

For portfolio use, the key comparison is incremental. The informative comparison evaluates a text model alongside price and accounting predictors. Otherwise, it may simply recover information already present in conventional inputs. Its contribution may also depend on timing. Publication dates, document revisions, and processing delays determine what an investor could actually observe.

Large language models may help organize long documents. Their economic contribution still requires a separate test. This review proposes comparing a text-augmented strategy with the same strategy without text. The two strategies can hold the trading universe, allocation rule, and risk limits constant. Data acquisition and processing costs should be included when material. Output checks should address accuracy and reproducibility before any investment benefit is inferred.

This separates two claims that are often combined. A tool may make research easier without creating a new return signal. It may also create a signal that is too delayed or costly to trade. Reporting workflow efficiency and portfolio outcomes

separately would make the evidence more informative.

5.1 Measuring the Value of a New Information Source

A new dataset can affect three outcomes. It can improve a forecast, accelerate a workflow, or reduce the effort needed to maintain an existing process. These benefits have different baselines. Forecasting value requires a comparison with existing predictors. Workflow value requires a comparison with the time and resources used previously. Maintenance value requires evidence across repeated updates. A single backtest cannot establish all three.

For text, document availability is only the first timing question. The document must also be retrieved, processed, and converted into a signal. A strategy that trades immediately after publication assumes that these steps are sufficiently fast. A slower process may face a different price by the time an order is placed. Reporting both the source timestamp and assumed trade timestamp makes the information delay visible.

Language-model applications also need a stable evaluation procedure. Model versions, prompts, and retrieval sources can change the output. The relevant object is the full processing pipeline, rather than the model name alone. Reproducibility depends on retaining the inputs and outputs needed to reconstruct signals. If a model's historical training information is unclear, possible contamination also warrants explicit discussion. A past-dated financial document does not by itself establish a historically available model.

These considerations do not rule out language models. They define the evidence required for a particular claim. A model may be useful for present-day research without supporting a clean historical trading experiment. A carefully designed forward evaluation could still measure its contribution. The distinction helps prevent a useful operational tool from being judged by an unsuitable backtest, or a contaminated backtest from being mistaken for investment evidence.

6. Why Portfolio and Trading Choices Change the Result

Trading costs change more than the level of returns. They can alter which signals should be combined and how quickly positions should move. G âleanu and Pedersen (2013) derive partial adjustment toward an aim portfolio. That portfolio reflects current and expected future frictionless targets. Persistent signals receive greater weight. This establishes a mechanism linking forecast decay to trading speed. Applying the mechanism to a new strategy still requires estimates of costs and persistence.

Jensen et al. (2026, Table 2) compare monthly U.S. equity strategies from 1981 to 2020. Their investor’s wealth reaches $10 billion by 2020, and relative risk aversion is 10. Portfolio-ML reports gross and net Sharpe ratios of 1.38 and 1.33; Static-ML reports 0.95 and 0.79. Portfolio-ML has 25% monthly turnover versus 80% for Static-ML, and average leverage of 3.00 versus 12.78. The framework explicitly models scale and trading costs, which makes this unusually direct evidence on implementation. Still, the methods differ in leverage, turnover, holdings, and portfolio construction. Their Sharpe gap is not a matched-risk estimate of the causal contribution of machine learning. A fair attribution test would align the risk budget and constraints, and disclose financing spreads, margin requirements, and forced-liquidation rules.

Capacity evidence adds another qualification. Korajczyk and Sadka (2004) examine momentum under trading costs and price impact. The feasible scale depends on the strategy and the securities traded. Novy-Marx and Velikov (2016) show that implementation rules can preserve some anomaly profits while high-turnover strategies remain difficult to exploit. Thus, a negative net result may reflect an unsuitable trading rule. A positive result may depend on a specific cost-mitigation choice.

Costs can change which signals are useful. DeMiguel et al. (2020) find that the number of economically significant characteristics rises from 6 to 15 when portfolio selection accounts for transaction costs. Trades generated by different characteristics can offset one another. Avramov et al. (2023) show the opposite pressure on some strategies: high turnover and estimated costs erode a large share of the return spread. Together, these results show that costs can change both the value assigned to a signal and the return left after trading. The findings remain specific to their samples and cost models.

6.1 How Trading Choices Explain Different Findings

Freyberger et al. (2020) and DeMiguel et al. (2020) attach value to a characteristic at different points in the process. Predictive selection asks what information it adds conditional on other inputs. Selection with trading costs also depends on the orders it creates alongside existing signals. A characteristic can change the route from current holdings to the next portfolio. Its economic contribution consequently depends on what the investor already holds.

Consider an illustrative case. One signal calls for selling a security while another calls for buying it. Combining the demands can reduce the order sent to the market. The second signal could then have economic value even if it adds little forecast information. This mechanism explains why predictive redundancy does not imply trading redundancy. The example illustrates offsetting demands; it is not an additional empirical result.

The increase from six to 15 significant characteristics is a within-study finding of DeMiguel et al. (2020). It is not a

numerical comparison with the selection count in Freyberger et al. (2020). The papers use different designs, so their counts cannot measure the isolated effect of costs across studies. A direct comparison would hold the characteristic universe fixed. It would then evaluate selections under predictive loss and net investment utility. Reporting order netting and turnover would reveal whether an added characteristic helps through information or implementation.

Jensen et al. (2026, Table 2) report gross and net Sharpe ratios of 1.38 and 1.33 for Portfolio-ML, versus 0.95 and 0.79 for Static-ML. The comparison also reports annualized volatility of 11% versus 28%, monthly turnover of 25% versus 80%, and average leverage of 3.00 versus 12.78. These results are not a matched-risk test of allocation quality. The reported net Sharpe ratios deduct modeled trading costs, but they do not by themselves show whether financing spreads, margin calls, or forced liquidation would change the comparison. The lower-leverage portfolio's stronger Sharpe is evidence about the complete portfolio designs under the study's assumptions, not an isolated machine-learning effect. A sharper attribution test would hold risk budgets and leverage constant, then report net outcomes under explicit financing, margin, and liquidation assumptions.

6.2 A Transparent Illustration of a Ranking Reversal

Consider two hypothetical portfolios with the same 10% volatility and gross excess returns of 10.5% and 8.5%. Their gross Sharpe ratios are 1.05 and 0.85; after annual costs of 3.0% and 0.8%, they fall to 0.75 and 0.77, reversing the ranking. This example illustrates a possible ranking reversal, not an empirical estimate of any strategy's net return.

Actual costs are uncertain, vary with market conditions, and can alter volatility. Empirical comparisons therefore require full portfolio evaluation under stated execution assumptions.

6.3 Persistence and the Decision to Trade

The relevant cost is often the cost of moving from the existing portfolio. A target may look attractive in isolation yet offer little improvement over current holdings. An investor must compare the expected benefit of changing positions with the cost of the change. This makes yesterday's holdings economically relevant to today's decision. It also explains why repeatedly optimizing a frictionless portfolio can generate avoidable trading.

Novy-Marx and Velikov (2016) examine rules that distinguish entering a position from continuing to hold it. Such rules can reduce turnover around a selection boundary. For a machine-learning strategy, the analogous question is whether small changes in predicted ranks justify a trade. A prediction can change without its economic value changing enough to cover the transaction. This motivates comparing sophisticated joint optimization with simpler holding rules.

A smooth signal may look easier to trade. Yet smoothing can delay useful information as well as reduce turnover. The comparison accounts for the return lost through delay as well as the costs avoided. The balance can change with the forecast horizon. A monthly strategy cannot simply inherit an implementation result obtained at a much shorter interval.

6.4 Capacity and Adverse Trading Conditions

Capacity depends on more than the average spread. Trade size relative to available liquidity can change price impact. Several positions may also need to be reduced simultaneously during stress. An average-cost assumption may understate the cost of that joint exit. Sensitivity analysis is more informative when it includes concentrated rebalancing episodes as well as increases in a constant cost coefficient.

An advantage at one capital level may disappear at another. Reporting that threshold would make capacity claims more useful, provided the execution assumptions are stated. The threshold belongs to the strategy and cost model together. It is not a fixed attribute of the learning algorithm.

Trading constraints can also interact. A short position may face both limited availability and a borrowing charge. A liquidity restriction can remove the securities with the largest forecasted advantage. Financing and leverage limits can alter the risk budget. Joint tests can reveal interactions that separate average penalties may miss, particularly when restrictions change which trades are feasible.

Capacity research gives a broader basis for interpreting scale. Pástor, Stambaugh, and Taylor (2015) model how scale relates to skill and performance in active management. Ciccotello et al. (2011) find that capacity effects vary with fund liquidity and investment style. Using corporate-bond strategies, Ivashchenko and Kosowski (2024) find that high-turnover strategies reach capacity quickly as fund size increases, while lower-turnover credit strategies can support greater scale. Frazzini, Israel, and Moskowitz (2012) estimate break-even capacity using nearly one trillion dollars of institutional trading data. They show that cost-aware portfolio design can change those estimates. Their later live-trading study finds that execution costs vary with trade size, stock characteristics, and time (Frazzini et al., 2018). These studies are not all machine-learning tests. They show why capacity depends on the strategy, execution process, asset liquidity, and investor size rather than on the algorithm alone.

Short-sale costs add another constraint on net performance. Chu, Hirshleifer, and Ma (2020) use a natural experiment that

relaxed short-sale restrictions for selected stocks. The combined returns of anomaly long-short portfolios fell by 72 basis points per month, driven by the short legs. Across 162 anomaly portfolios, Muravyev, Pearson, and Pollet (2025) report an average long-short return of 0.14% per month before stock-borrow fees and approximately zero after the fees are included. These findings concern anomaly portfolios rather than machine-learning strategies specifically; they are relevant because many ML tests also form long-short positions and may inherit similar borrowing and liquidity frictions.

Capacity analysis also requires an explicit method. O'Neill and Warren (2019) review how capacity is defined and identify ten drivers. They outline methods for estimating how trading and portfolio construction change as assets under management grow. In a direct ML application, Jensen et al. (2026) evaluate net returns across investor wealth levels and show that the implementable frontier contracts as scale and market impact increase. Together, these studies support reporting strategy-specific capacity thresholds under transparent execution assumptions; neither a generic asset-size figure nor an algorithm label alone establishes scalable performance.

Figure 2 summarizes the dependence of net value on complexity and implementation. It is a conceptual illustration, not an estimated curve or an identified optimal complexity level.

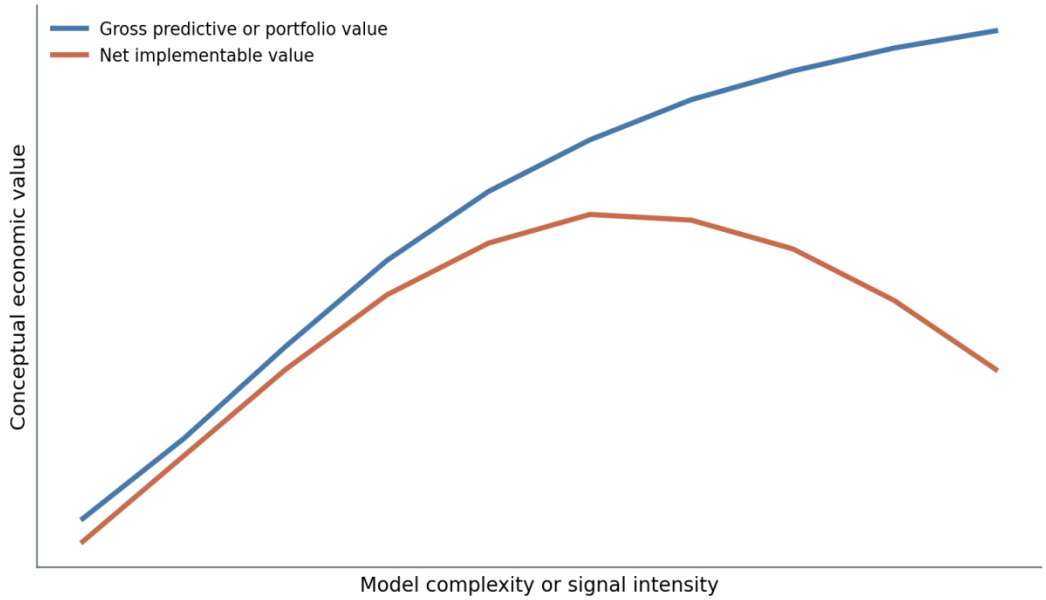

Figure 2. Conceptual relation between model complexity and net economic value. The vertical axis denotes portfolio value after implementation costs, shown schematically without a numerical unit; the curves are not estimated returns

6.5 Comparing Approaches to Direct Allocation

These implementation issues also influence how learning is organized. Butler and Kwon (2023) integrate regression prediction with an explicit mean-variance optimization problem. Their historical simulations include global futures data. Training can therefore account for the portfolio consequences of prediction errors. The optimization layer makes the specified risk-return tradeoff and constraints visible.

Jiang et al. (2017) learn portfolio actions through a reward-based framework. Their preprint reports cryptocurrency backtests at a short trading interval. Here, investment preferences enter through the reward and decision environment. The design can depend strongly on simulated execution and training history. These dependencies differ from the estimation choices in a mean-variance layer.

Online portfolio selection provides a third perspective. Li and Hoi (2014) survey methods that update allocations as observations arrive. Wealth growth and performance relative to specified benchmark strategies are common objectives. The interpretation of their guarantees depends on the benchmark and trading assumptions. Those guarantees do not directly establish net profitability under market frictions.

These branches extend the framework to models without a separate return forecast. For a direct policy, the prediction-stage question becomes whether learned information improves decisions relative to a policy benchmark. Information timing, portfolio risk, and implementation remain relevant across the designs.

Comparing Butler and Kwon (2023) with Jiang et al. (2017) does not isolate the value of reinforcement learning. Asset class, objective, and decision horizon change together. The cryptocurrency trading interval is also much shorter than the monthly setting in Jensen et al. (2026). This changes how often execution costs can affect the strategy.

A proposed comparison would align data, initial holdings, decision dates, feasible positions, and execution. A simple partial-adjustment strategy would supply a dynamic benchmark, following Gârleanu and Pedersen (2013). This would test whether learned sequencing adds value beyond delayed trading alone. Section 9 places this comparison within the broader attribution experiment.

6.6 Direct Learning of Portfolio Policies

Recent work expands the evidence on policies that learn allocations directly or place portfolio objectives inside training. Babiak and Barunık (2026) find that deep feed-forward and recurrent networks improve utility and Sharpe outcomes in a two-asset market-timing problem. The results remain in tests with modeled transaction costs, borrowing limits, and short-selling constraints. Simon, Weibels, and Zimmermann (2026) use deep neural networks to learn portfolio weights under investor utility. They report higher certainty-equivalent returns than linear portfolio policies. Gains vary by risk aversion and remain in their transaction-cost and short-selling tests. Guijarro-Ordonez, Pelger, and Zanotti (2026) combine conditional latent factors, a convolutional transformer, and a constrained trading policy for daily U.S. equity statistical arbitrage, reporting out-of-sample performance above their benchmarks.

Reinforcement-learning studies also embed risk and trading costs in allocation decisions. Jiang, Olmo, and Atwi (2024) use a deep deterministic policy-gradient method with transaction-cost-sensitive mean-variance rewards; their tests use large U.S. equity indices. The same authors study multi-period allocation with transaction costs and risk aversion in a broader asset setting (Jiang et al., 2025). Zhou et al. (2026) combine factor estimation and mean-variance allocation in an end-to-end direct reinforcement-learning design. These results show that direct policies can be tested against economic objectives. They are not directly comparable because they use different asset universes, horizons, reward functions, risk budgets, cost models, and benchmarks. Claims about general net gains therefore require the design details and robustness results to be reported alongside the headline performance.

Cong, Tang, and Wang (2026) extend this strand with AlphaPortfolio, an NBER working paper combining attention-based neural networks and deep reinforcement learning for direct, path-dependent portfolio construction. Their U.S. equity backtests report strong out-of-sample performance and robustness to exclusions such as small and illiquid stocks; the framework can incorporate transaction costs and alternative objectives. These are model-based backtests, not realized returns from an operating fund. Interpret the estimates in light of each study's universe, cost calibration, risk objective, and benchmark.

7. Forecast Uncertainty and Changing Market Conditions

Allena (2026) studies whether forecast precision can improve portfolio sorts. In the publicly available 2023 author manuscript, NN-3 Confident-HL raises the equal-weighted (value-weighted) monthly OOS return from 2.21% (1.29%) to 3.84% (2.70%) and annualized Sharpe from 1.36 (0.78) to 1.78 (1.26). The published abstract confirms that Confident-HL outperforms across models but does not give point estimates. The author manuscript also reports turnover-based cost adjustments, which do not establish capacity at institutional scale (Allena, 2023, 2026).

A model can estimate sampling uncertainty carefully and still encounter a market in which the learned relationship no longer holds. More precise estimation within the training environment does not resolve that change. Market states and alternative training windows therefore matter when interpreting the intervals. They test a different weakness from finite-sample imprecision.

Confidence-based selection can change the number of holdings, sector weights, and exposure to smaller firms. Its evaluation must therefore trace how precision affects positions. The next subsection develops the relevant calibration and portfolio checks.

These checks begin with reliable inputs. The Bianchi correction shows why information availability must be established before forecast precision is interpreted (Bianchi, Büchner, Hoogteijling, & Tamoni, 2021). Data-version controls connect this requirement to the governance practices in Section 8.

7.1 Testing Whether Confidence Deserves Capital

Confidence estimates are most informative when evaluated on the same object that guides allocation. An interval for an expected return describes uncertainty about a conditional mean. It is different from an interval for the next realized return. The latter also includes unpredictable return variation. Confusing them can make an estimate appear safer than the position actually is. Reports should state which quantity the interval concerns.

Acceptable coverage over a full sample can conceal poor coverage in one market state. That state may also be where the allocation rule takes its largest positions. Examining forecast groups and market conditions can expose this mismatch. Those checks are more persuasive when chosen before the final results are inspected.

Lower confidence does not dictate a single trading response. The investor might retain the holding, reduce leverage, or use a simpler model temporarily. Each choice changes exposure and turnover. Comparing these responses would help

establish whether the confidence estimate adds useful information. Otherwise, a general reduction in risk could explain the apparent benefit.

Monitoring can distinguish input failures, changes in forecast distributions, and unexpected realized outcomes. Each calls for a different investigation. Automatic retraining may conceal a data error or trigger unnecessary trading. This review proposes these categories as a monitoring structure, not as a tested trading rule.

7.2 Different Uncertainty Methods Protect against Different Errors

Allena (2026), Ledoit and Wolf (2017), and Mohajerin Esfahani and Kuhn (2018) address distinct objects of uncertainty. Allena uses precision around expected-return forecasts in portfolio formation. Covariance shrinkage addresses instability in estimated dependence and risk. Distributionally robust optimization evaluates decisions across a specified neighborhood of distributions. Calling all three uncertainty-aware can obscure what each actually changes.

Consider a portfolio with precise expected-return estimates but a poor covariance estimate. Selecting confident forecasts alone does not directly repair that risk input. Conversely, a stable covariance matrix does not establish accurate return magnitudes. A distributional neighborhood can represent broader misspecification, but its choice determines which departures receive protection. These are implications of the methods' designs. They do not constitute a published ranking of their empirical performance.

The methods may therefore complement one another. A proposed comparison would test each adjustment separately and then combine them. Concentration, realized risk, tail losses, turnover, and net returns would help identify which error was reduced. Section 9 includes this distinction in the proposed sensitivity tests.

8. Interpretation and Governance

An explanation of a model output is a starting point for investigation. Explainable-AI methods can identify influential inputs (Černevičienė & Kabašinskas, 2024). A plausible account of those inputs still leaves the investment result to be tested. Explanations are most useful when read alongside exposures, realized outcomes, and alternative specifications. Their diagnostic value depends on whether they help account for the model’s behavior.

Governance becomes more important when models and data change frequently. This review recommends version control, independent review, monitoring, and procedures for retiring a model. The 2026 U.S. supervisory guidance offers relevant principles for traditional quantitative models. It is mainly relevant to banking organizations with more than $30 billion in assets. It excludes generative and agentic AI and does not establish enforceable standards (Board of Governors of the Federal Reserve System et al., 2026). Applying its principles more broadly here is an analytical recommendation, not a claim about every asset manager's legal obligations.

These controls make the investment process identifiable. They connect each reported result to the data, model, and decisions that produced it. Documentation is therefore part of the evidence needed to assess deployment.

8.1 Documentation as Evidence of an Investment Process

A reviewable model needs a record of its intended use. That record identifies the predicted quantity, the allowed decisions, the main limitations, and who can approve changes. These details connect technical validation with the institution's actual investment mandate. A model that performs well for one horizon may be inappropriate for another.

Effective change control distinguishes a routine data refresh from a redesign. Adding a predictor or changing the loss function can alter the strategy's economic behavior, so earlier validation does not automatically carry over. Comparing old and new versions in parallel can reveal differences before capital is reassigned.

A change in influential variables can be informative, but its cause may be mundane. A revised normalization procedure or a data error can resemble a new market relationship. Exposure records and data versions help distinguish these possibilities. The explanation tool identifies where to investigate; the surrounding research record is needed to establish what changed.

9. Research Progress and Remaining Questions

9.1 What the Selected Literature Establishes

Across the selected studies, flexible forecasts and structured pricing models find information that standard benchmarks miss (Gu et al., 2020; Freyberger et al., 2020; Chen et al., 2024). Results are less consistent when studies test trading: costs, portfolio design, and leverage affect performance (DeMiguel et al., 2020; Avramov et al., 2023; Jensen et al., 2026). Some fund-selection studies report positive net or after-fee spreads in particular samples (DeMiguel et al., 2023; Kaniel et al., 2023). Results from observed AI-labelled funds are mixed, and early hedge-fund advantages do not persist (Chen & Ren, 2022; Chen et al., 2026). Together, these findings support value in specific tasks, not a general performance advantage from machine learning.

Table 1 groups studies by evidence type: forecasts and pricing, portfolio backtests, fund selection, observed fund results, and capacity. Table 2 lists tests for each investment stage. Together, the tables show how far the evidence reaches, from prediction toward implementation. They do not combine unlike metrics or assign an overall score.

9.2 What Remains Unresolved

Research has not yet isolated how much each component contributes. When inputs, allocation rules, and trading behavior all change, one Sharpe ratio combines their effects. Even within one sample, strategies may use different leverage or tuning effort. Paired studies suggest possible mechanisms, but do not measure each contribution separately.

Backtests also differ from live trading. A strategy may fit an inaccurate execution simulator or hide severe losses behind a strong average. Spreads, liquidity, fund size, and changing markets affect results in different ways. Existing studies do not show how these factors interact.

Research selection creates another challenge. Bailey et al. (2017) propose a way to estimate how likely a backtest is to be overfit. Similar concerns arise when researchers test many factors or strategies. A strategy may reflect choices made throughout its development, not just its final test. Reports should include attempted variants and chronological tests. These checks cannot fix missing historical data or unrealistic execution assumptions.

9.3 What Future Studies Should Test

Table 2 turns the review's findings into four practical evidence checks. It is a guide for research design, not a validated score. For a direct portfolio policy, test whether it improves decisions over an appropriate benchmark.

Table 2. Proposed evidence requirements for each investment stage

Stage	Question	Minimum evidence	Failure signal
Information	Was the input available?	Dated data and revisions; realistic publication and processing lags.	Apparent benefit disappears after timing corrections.
Prediction	Does the signal add information?	Chronological tests; common tuning budget; simple and established benchmarks.	Gain relies on one split or duplicates existing predictors.
Portfolio construction	Does allocation preserve the benefit?	Common risk budget; stable constraints; exposure and concentration comparisons.	Gain comes mainly from leverage or unstable weights.
Implementation	Can the strategy be traded at scale?	Net performance; turnover; cost and capacity sensitivity; execution assumptions.	Modest cost or fund-size changes remove the benefit.

An initial experiment could compare two predictors—a transparent model and a flexible one—under two allocation rules: fixed sorts and risk-aware optimization. Use the same dated information, investment universe, and constraints in all four combinations. This would show whether the flexible predictor helps under both rules or only one.

Directly trained and reinforcement-learning systems need similar comparisons. Test them with the same data, risk budget, and execution assumptions as modular systems. Their performance initially supports a claim about the full system. Ablation tests can then isolate the contribution of individual components. Repeated runs can show how results depend on initialization and training instability.

Compare cost-aware learning with simpler ways to reduce trading, such as slower rebalancing and turnover penalties. Vary spreads, market impact, fund size, and liquidity to see when each approach loses its advantage. Compare volatility and tail losses with the investor's mandate, especially in adverse periods.

Replication across markets can show which results generalize. Use realistic information delays. For text models, record when documents became available and the model's training-data cutoff. Preserve data versions and the model-search history. If data cannot be shared, publish exact construction rules and executable evaluation steps. Include unsuccessful periods.

A future preregistered review could apply this framework to a larger set of strategies. It could record validation methods, cost assumptions, risk-matched comparisons, and realized outcomes, then test whether these distinctions identify results that survive implementation.

10. Conclusion

The evidence supports a conditional conclusion. Some machine-learning methods improve forecasts and certain portfolio designs; some fund-selection models report positive net alpha. But costs, mixed or fading fund results, and differences in leverage and exposures prevent a broader claim: current evidence does not show that machine learning itself delivers persistent, scalable outperformance.

This review links each type of evidence to the claims it can support and proposes tests to separate model effects from portfolio-design effects. The framework helps researchers state what their results establish and what remains unknown. It does not combine unlike performance measures or treat modeled outcomes as realized fund performance.

Several questions remain. Which applications retain value after realistic trading costs, and at what scale? Do learning methods add value beyond portfolio rules, leverage, and risk exposures? Do gains persist across markets, asset classes, and out-of-sample periods? Can end-to-end policies be reproduced as data and execution conditions change? Until these questions are answered, conclusions should be limited to the data, horizon, benchmark, and implementation assumptions each study tests.

Acknowledgments

Not applicable.

Author Contributions

Ying Zhang designed the review, conducted and synthesized the literature review, drafted and revised the manuscript, and approved the final version.

Funding

Not applicable.

Competing Interests

Ying Zhang declares no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed Consent

Obtained.

Ethics Approval

The Publication Ethics Committee of the Redfame Publishing.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and Peer Review

Not commissioned; externally double-blind peer reviewed.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data Sharing Statement

No additional data are available.

Open Access

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

References

Allena, R. (2023). *Confident risk premiums and investments using machine learning uncertainties*.

Allena, R. (2026). Confident risk premiums and investments using machine learning uncertainties. *The Review of Financial Studies*, *39*(5), 1463-1505. https://doi.org/10.1093/rfs/hhaf087

Avramov, D., Cheng, S., & Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, *69*(5), 2587-2619. https://doi.org/10.1287/mnsc.2022.4449

Babiak, M., & Barun ķ, J. (2026). Deep learning, predictability, and optimal portfolio returns. *Journal of Empirical*

Finance, *87*, 101705. https://doi.org/10.1016/j.jempfin.2026.101705

Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2017). The probability of backtest overfitting. *The Journal of Computational Finance*, *20*(4), 39-69. https://doi.org/10.21314/jcf.2016.322

Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, *34*(2), 1046-1089. https://doi.org/10.1093/rfs/hhaa062

Bianchi, D., Büchner, M., Hoogteijling, T., & Tamoni, A. (2021). Corrigendum: Bond risk premiums with machine learning. *The Review of Financial Studies*, *34*(2), 1090-1103. https://doi.org/10.1093/rfs/hhaa098

Blitz, D., Hoogteijling, T., Lohre, H., & Messow, P. (2023). How can machine learning advance quantitative asset management. *The Journal of Portfolio Management*, *49*(9), 78-95. https://doi.org/10.3905/jpm.2023.1.460

Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, & Office of the Comptroller of the Currency. (2026, April 17). Supervisory guidance on model risk management (SR 26-2).

Brogaard, J., & Zareei, A. (2023). Machine learning and the stock market. *Journal of Financial and Quantitative Analysis*, *58*(4), 1431-1472. https://doi.org/10.1017/s0022109022001120

Buczynski, W., Cuzzolin, F., & Sahakian, B. (2021). A review of machine learning experiments in equity investment decision-making: Why most published research findings do not live up to their promise in real life. *International Journal of Data Science and Analytics*, *11*(3), 221-242. https://doi.org/10.1007/s41060-021-00245-5

Butler, A., & Kwon, R. H. (2023). Integrating prediction in mean-variance portfolio optimization. *Quantitative Finance*, *23*(3), 429-452. https://doi.org/10.1080/14697688.2022.2162432

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

Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review*, *57*, 216. https://doi.org/10.1007/s10462-024-10854-8

Chen, L., Pelger, M., & Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, *70*(2), 714-750. https://doi.org/10.1287/mnsc.2023.4695

Chen, R., & Ren, J. (2022). Do AI-powered mutual funds perform better? *Finance Research Letters*, *47*, 102616. https://doi.org/10.1016/j.frl.2021.102616

Chen, S., Sialm, C., & Xu, D. X. (2026). The growth and performance of artificial intelligence in asset management. NBER Working Paper No. 35273. https://doi.org/10.3386/w35273

Chu, Y., Hirshleifer, D., & Ma, L. (2020). The causal effect of limits to arbitrage on asset pricing anomalies. *The Journal of Finance*, *75*(5), 2631-2672. https://doi.org/10.1111/jofi.12947

Ciccotello, C., Greene, J., Ling, L., & Rakowski, D. (2011). Capacity and factor timing effects in active portfolio management. *Journal of Financial Markets*, *14*(2), 277-300. https://doi.org/10.1016/j.finmar.2010.10.002

Cong, L. W., Tang, K., & Wang, J. (2026). *AlphaPortfolio: Goal-oriented investment management through deep reinforcement learning*. NBER Working Paper No. 35195. https://doi.org/10.3386/w35195

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies*, *22*(5), 1915-1953. https://doi.org/10.1093/rfs/hhm075

DeMiguel, V., Gil-Bazo, J., Nogales, F. J., & Santos, A. A. P. (2023). Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics*, *150*(3), 103737. https://doi.org/10.1016/j.jfineco.2023.103737

DeMiguel, V., Martín-Utrera, A., Nogales, F. J., & Uppal, R. (2020). A transaction-cost perspective on the multitude of firm characteristics. *The Review of Financial Studies*, *33*(5), 2180-2222. https://doi.org/10.1093/rfs/hhz085

Du, K., Zhao, Y., Mao, R., Xing, F., & Cambria, E. (2025). Natural language processing in finance: A survey. *Information Fusion*, *115*, 102755. https://doi.org/10.1016/j.inffus.2024.102755

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*(3), 1327-1370. https://doi.org/10.1111/jofi.12883

Frazzini, A., Israel, R., & Moskowitz, T. J. (2012). *Trading costs of asset pricing anomalies* [Working paper]. AQR Capital Management. https://doi.org/10.2139/ssrn.2294498

Frazzini, A., Israel, R., & Moskowitz, T. J. (2018). *Trading costs* [Working paper]. SSRN.

https://doi.org/10.2139/ssrn.3229719

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, *33*(5), 2326-2377. https://doi.org/10.1093/rfs/hhz123

Gârleanu, N., & Pedersen, L. H. (2013). Dynamic trading with predictable returns and transaction costs. *The Journal of Finance*, *68*(6), 2309-2340. https://doi.org/10.1111/jofi.12080

Giglio, S., Kelly, B., & Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, *14*, 337-368. https://doi.org/10.1146/annurev-financial-101521-104735

Goldfarb, D., & Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of Operations Research*, *28*(1), 1-38. https://doi.org/10.1287/moor.28.1.1.14260

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

Guijarro-Ordonez, J., Pelger, M., & Zanotti, G. (2026). Deep learning statistical arbitrage. *Management Science*, *72*(9), 7502-7549. https://doi.org/10.1287/mnsc.2022.03132

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5-68. https://doi.org/10.1093/rfs/hhv059

Ivashchenko, A., & Kosowski, R. (2024). Transaction costs and capacity of systematic corporate bond strategies. *Financial Analysts Journal*, *80*(4), 53-80. https://doi.org/10.1080/0015198x.2024.2360390

Jagannathan, R., & Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, *58*(4), 1651-1683. https://doi.org/10.1111/1540-6261.00580

Jagannathan, R., Ma, T., & Zhang, J. (2019). A note on "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps". *The Journal of Finance*, *74*(5), 2689-2696. https://doi.org/10.1111/jofi.12824

Jensen, T. I., Kelly, B., Malamud, S., & Pedersen, L. H. (2026). Machine learning and the implementable efficient frontier. *The Review of Financial Studies*, *39*(10), 3035-3078. https://doi.org/10.1093/rfs/hhag022

Jiang, Y., Olmo, J., & Atwi, M. (2024). Deep reinforcement learning for portfolio selection. *Global Finance Journal*, *62*, 101016. https://doi.org/10.1016/j.gfj.2024.101016

Jiang, Y., Olmo, J., & Atwi, M. (2025). High-dimensional multi-period portfolio allocation using deep reinforcement learning. *International Review of Economics & Finance*, *98*, 103996. https://doi.org/10.1016/j.iref.2025.103996

Jiang, Z., Xu, D., & Liang, J. (2017). *A deep reinforcement learning framework for the financial portfolio management problem* [Preprint]. arXiv:1706.10059.

Kaniel, R., Lin, Z., Pelger, M., & Van Nieuwerburgh, S. (2023). Machine-learning the skill of mutual fund managers. *Journal of Financial Economics*, *150*(1), 94-138. https://doi.org/10.1016/j.jfineco.2023.07.004

Ke, Z. T., Kelly, B. T., & Xiu, D. (2019). *Predicting returns with text data*. NBER Working Paper No. 26186. https://doi.org/10.3386/w26186

Kelly, B. T., & Xiu, D. (2023). Financial machine learning. *Foundations and Trends in Finance*, *13*(3-4), 205-363. https://doi.org/10.1561/0500000064

Korajczyk, R. A., & Sadka, R. (2004). Are momentum profits robust to trading costs? *The Journal of Finance*, *59*(3), 1039-1082. https://doi.org/10.1111/j.1540-6261.2004.00656.x

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

Ledoit, O., & Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, *30*(4), 110-119. https://doi.org/10.3905/jpm.2004.110

Ledoit, O., & Wolf, M. (2017). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *The Review of Financial Studies*, *30(12)*, 4349-4388. https://doi.org/10.1093/rfs/hhx052

Leippold, M., Wang, Q., & Zhou, W. (2022). Machine learning in the Chinese stock market. *Journal of Financial Economics*, *145*(2), 64-82. https://doi.org/10.1016/j.jfineco.2021.08.017

Leow, K., & Lindenthal, T. (2025). Enhancing real estate investment trust return forecasts using machine learning. *Real Estate Economics*, *53(*3), 574-606. https://doi.org/10.1111/1540-6229.12527

Li, B., & Hoi, S. C. H. (2014). Online portfolio selection: A survey. *ACM Computing Surveys*, *46(3)*, Article 35, 1-36. https://doi.org/10.1145/2512962

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77-91. https://doi.org/10.2307/2975974

Michaud, R. O. (1989). The Markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, *45*(1), 31-42. https://doi.org/10.2469/faj.v45.n1.31

Mohajerin Esfahani, P., & Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, *171*(1-2), 115-166. https://doi.org/10.1007/s10107-017-1172-1

Muravyev, D., Pearson, N. D., & Pollet, J. M. (2025). Anomalies and their short-sale costs. *The Journal of Finance*, *80*(6), 3639-3694. https://doi.org/10.1111/jofi.13501

Murray, S., Xia, Y., & Xiao, H. (2024). Charting by machines. *Journal of Financial Economics*, *153*, 103791. https://doi.org/10.1016/j.jfineco.2024.103791

Novy-Marx, R., & Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *The Review of Financial Studies*, *29*(1), 104-147. https://doi.org/10.1093/rfs/hhv063

O'Neill, M. J., & Warren, G. J. (2019). Evaluating fund capacity: Issues and methods. *Accounting & Finance*, *59*(S1), 773-800. https://doi.org/10.1111/acfi.12268

Pástor, Ľ., Stambaugh, R. F., & Taylor, L. A. (2015). Scale and skill in active management. *Journal of Financial Economics*, *116*(1), 23-45. https://doi.org/10.1016/j.jfineco.2014.11.008

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *The Journal of Risk*, *2*(3), 21-41. https://doi.org/10.21314/jor.2000.038

Simon, F., Weibels, S., & Zimmermann, T. (2026). *Deep parametric portfolio policies. Management Science.* Advance online publication. https://doi.org/10.1287/mnsc.2025.00721

Zhou, K., Huang, X., Chen, X., & Gao, J. (2026). An end-to-end direct reinforcement learning approach for multi-factor based mean-variance portfolio optimization. *The European Journal of Finance*, *32*(4-6), 450-481. https://doi.org/10.1080/1351847x.2025.2558114