Predicting Alpha: A Pedagogical Survey of the State of the Art in Return-Forecasting Signals*

Agustín Shehadi Candela
QUAFI Research
research@quafi.io

March 2026

Abstract

An *alpha signal* is any piece of information, available today, that helps predict the part of an asset's future return that is not simply compensation for bearing known risks. Hunting for such signals is the central activity of quantitative investing — and one of the most treacherous, because the same machinery that discovers a true signal will, if used carelessly, manufacture hundreds of false ones. This paper is a didactic survey, written for a reader with a quantitative background but no specialist training in empirical asset pricing, of what the academic literature has learned about *predicting* alpha. We proceed in four movements. First, we build the definition of an alpha signal from first principles, separating a *predictor* from the *risk-adjusted economic edge* it may or may not contain, and locating the whole enterprise inside the efficient-markets debate (Fama, 1970; Lo, 2004). Second, we map where signals have historically come from: the cross-section of firm characteristics (Jegadeesh and Titman, 1993; Carhart, 1997; Ang et al., 2006; Amihud, 2002), the time-series predictability of the market premium (Welch and Goyal, 2008; Campbell and Thompson, 2008), the microstructure of trading (Lou et al., 2019; Llorente et al., 2002), and the forecastable dynamics of volatility itself (Corsi, 2009; Andersen et al., 2003). Third, and most importantly, we explain *how to measure* whether a candidate signal is real: out-of-sample accuracy and proper scoring (Campbell and Thompson, 2008; Gneiting and Raftery, 2007), economic significance net of costs (Bailey and López de Prado, 2014; Frazzini et al., 2018), the multiple-testing and data-snooping corrections that separate discovery from noise (White, 2000; Hansen, 2005; Harvey et al., 2016), and the distribution-free uncertainty quantification that honest forecasting now demands (Vovk et al., 2005; Lei et al., 2018). Fourth, we survey the open frontier: non-stationarity, post-publication alpha decay (McLean and Pontiff, 2016; Hou et al., 2020), the high-frequency signal-to-noise ceiling, and the integration of machine learning (Gu et al.,

*QUAFI Working Paper Series, No. 2026-01. This is a survey and educational working paper of QUAFI Research. It is preliminary, circulated to inform and to invite comment; the views are the author's own. *Nothing herein is investment advice.* I thank the QUAFI Research initiative for support. All references are to published, peer-reviewed sources; any errors are mine.

2020) with rigorous, multiplicity-aware evaluation. Throughout, the pedagogy and the rigor are meant to reinforce, not trade off against, one another.

Keywords: alpha signal; return predictability; cross-section of expected returns; market efficiency; forecast combination; machine learning in finance; backtest overfitting; data snooping; conformal prediction; proper scoring rules.
JEL: C52, C53, C58, G11, G12, G14, G17.

Contents

1 Introduction: why this is hard, and why it matters

Suppose you could improve your guess of tomorrow's return on a stock by even a hair. Repeated over thousands of stocks and thousands of days, a hair of genuine predictive edge compounds into a fortune. This is the promise that animates the entire quantitative-investing industry, and the object of that pursuit is what practitioners call an *alpha signal*: a quantity, computable from information available now, that carries information about future risk-adjusted returns.

The promise comes wrapped in a paradox. If a signal truly predicted returns and were easy to find, rational investors would trade on it until the prediction disappeared — this is the logic of market efficiency (Fama, 1970). So any signal that survives must be either genuinely subtle, compensation for some risk we have not named, or a product of our own statistical self-deception. The history of the field is, to a large degree, the history of learning to tell these three apart.

This paper is a guided tour of that history and of the methodological state of the art, written to be read by a numerate non-specialist. Our aims are deliberately didactic. By the end, a reader should be able to (i) say precisely what an alpha signal is and is not (§2); (ii) recognize the major families of signals the literature has documented and the economic stories behind them (§3); (iii) understand the spectrum of models used to turn raw data into a forecast, from linear regression to modern machine learning (§4); (iv) *measure* a candidate signal honestly — the single most important and most frequently botched step (§5); and (v) appreciate what remains genuinely unsolved (§6).

A word on the spirit of the survey. The literature's hardest-won lesson is not a signal but a discipline: that the probability a published predictive finding is false is alarmingly high (Ioannidis, 2005; Harvey et al., 2016), and that the daily, single-name horizon — where the signal-to-noise ratio is smallest and the temptation to over-search is greatest — is where this danger peaks. We therefore give evaluation as much room as discovery. The reader who absorbs only §5 will already be ahead of much of the practice.

2 What is an alpha signal? Definitions from first principles

2.1 Returns, expected returns, and the information set

Let P_t be the price of an asset at the close of day t (adjusted for dividends and splits). Its simple return over the next day is

$$r_{t+1} \;=\; \frac{P_{t+1}}{P_t} - 1. \tag{1}$$

An investor standing at time t does not know r_{t+1}; she knows only the *information set* $\mathcal{F}_t$ — everything observable up to and including today (past prices, volumes, fundamentals, news, and so on). The best she can do is form a conditional expectation,

$$\mu_t \;=\; \mathbb{E}[\, r_{t+1} \mid \mathcal{F}_t \,], \tag{2}$$

the average return she should expect given what she knows. The entire science of return prediction is the science of estimating the function that maps $\mathcal{F}_t$ into μ_t. A *predictive signal* is any feature $s_t = s(\mathcal{F}_t)$ that carries information about r_{t+1} — formally, any s_t for which $\mathbb{E}[r_{t+1} \mid s_t]$ is not constant in s_t.

2.2 From signal to *alpha*: the risk-adjustment step

Not every predictor of returns is an alpha signal. Much of the cross-sectional and time-series variation in expected returns is simply *compensation for risk*: assets that pay off poorly in bad times must offer higher average returns to induce investors to hold them. To isolate the part of expected return that is *not* such compensation, we pass returns through a risk model. In the canonical linear factor form,

$$r_{t+1} - r^f_{t+1} \;=\; \alpha \;+\; \boldsymbol{\beta}^\top \boldsymbol{f}_{t+1} \;+\; \varepsilon_{t+1}, \tag{3}$$

where r^f is the risk-free rate, $\boldsymbol{f}_{t+1}$ is a vector of common risk factors (e.g. the market, size, value, and momentum factors of Carhart, 1997), $\boldsymbol{\beta}$ are the asset's exposures, and ε is idiosyncratic noise. The intercept α is the asset's average return that the risk factors *cannot* explain.

Definition 1 (Alpha signal)**.** *An* alpha signal *is a predictor $s_t = s(\mathcal{F}_t)$ such that conditioning on s_t shifts the* risk-adjusted *expected return — the α of Eq. (3) — rather than merely loading on known factor exposures $\boldsymbol{\beta}$. Equivalently, s_t forecasts returns in excess of what a correctly specified risk model already implies.*

This distinction is the crux of the field and the source of its deepest controversy. Whether a documented return predictor is "alpha" (a genuine, exploitable mispricing or a missing-factor premium) or merely a proxy for an unmodeled risk is rarely settled cleanly — a point Fama (1970) formalized as the *joint-hypothesis problem*: any test of return predictability is simultaneously a test of the risk model used to adjust for risk. A signal that looks like alpha under one factor model can become a risk premium under a richer one.

2.3 Efficiency, anomalies, and adaptive markets

The efficient-markets hypothesis (Fama, 1970) holds that prices already reflect available information, so that μ_t should be explainable by risk alone and stable, exploitable alpha should be competed away. Decades of evidence have produced a long list of *anomalies* — robust, repeatable patterns that a frictionless efficient market should not display (Schwert, 2003). Their very persistence is a puzzle. Lo (2004) proposes a reconciliation, the *adaptive-markets hypothesis*: efficiency is not a fixed state but an evolutionary process in which signals work until enough capital learns them, then decay, while new ones emerge as the environment changes. Under this view alpha is real but *perishable*, and prediction is a moving target — a theme we return to in §6.

3 Where do alpha signals come from? A map of the literature

The empirical literature has documented predictors in four broad territories. We sketch each with its economic intuition; the goal is a mental map, not an exhaustive catalogue (the catalogues themselves now run to hundreds of factors, Harvey et al., 2016).

3.1 The cross-section: firm characteristics

The largest body of work asks *which stocks* outperform others, sorting the cross-section on observable characteristics.

- **Momentum.** Stocks that have risen over the past 3–12 months tend to keep outperforming those that have fallen (Jegadeesh and Titman, 1993), a pattern so robust that Carhart (1997) elevated it to a fourth factor alongside market, size, and value. Earlier, Jegadeesh (1990) documented short-horizon *reversal* (last month's losers bounce back), showing that the sign of the autocorrelation flips with horizon. George and Hwang (2004) refine the signal: much of momentum's predictive power is captured by how close a stock trades to its 52-week high, a finding often read as evidence of anchoring by investors.
- **Low-risk anomalies.** Counterintuitively, stocks with *higher* idiosyncratic volatility have earned *lower* subsequent returns (Ang et al., 2006) — the opposite of a simple risk-reward trade-off, and a durable challenge to standard models.
- **Liquidity.** Amihud (2002) shows that less liquid stocks command higher expected returns as compensation for the cost and difficulty of trading them; his illiquidity measure (the average ratio of absolute return to dollar volume) remains a workhorse predictor and a reminder that *tradeability* is itself priced.

3.2 The time series: predicting the market premium

A parallel literature asks *when* the aggregate market will pay more, using predictors such as the dividend-price ratio, valuation ratios, and interest rates. The sobering benchmark is Welch and Goyal (2008), who show that most such predictors, evaluated honestly out of sample, fail to beat the simple historical average and are unstable over time. Campbell and Thompson (2008) provide the constructive counterpoint: once forecasts are constrained to be economically sensible (e.g. non-negative equity premia), even a tiny out-of-sample R^2 — on the order of half a percent monthly — can be economically meaningful for a mean-variance investor. The pair is the canonical lesson that *small, fragile, but real* is the honest description of aggregate predictability.

3.3 Microstructure: information in the act of trading

Higher-frequency signals exploit the mechanics of trading itself. Lou et al. (2019) document a striking decomposition: the entire historical equity premium and momentum effect accrue

overnight (close-to-open), while the intraday (open-to-close) component behaves oppositely — a "tug of war" between clienteles that any daily predictor must respect. Llorente et al. (2002) show that the relationship between volume and the next period's return depends on whether trading is driven by information or by liquidity needs: high-volume moves reverse when trading is speculative and persist when it is informed. These results teach that *how* a price moved (on what volume, in which session) can matter as much as *how much* it moved.

3.4 Volatility: the predictable companion of returns

While the *level* of returns is barely forecastable, their *variance* is strongly and persistently forecastable — a fact essential to signal construction, because most signals are scaled by risk before they are traded. French et al. (1987) first linked expected returns to predictable volatility. The modern toolkit measures volatility from high-frequency data as *realized variance* (Andersen et al., 2003) and models its long-memory dynamics parsimoniously with the Heterogeneous Autoregressive (HAR) model of Corsi (2009). Refinements separate the smooth ("good"/"bad") and jump components of variance: Barndorff-Nielsen and Shephard (2006) develop the econometrics of detecting jumps via bipower variation, and Patton and Sheppard (2015) show that *signed* jumps carry distinct information for future volatility. Because a forecast of risk is a prerequisite for sizing any return forecast, volatility prediction is the quiet backbone of alpha-signal engineering.

4 From data to forecast: the modeling spectrum

Given a set of candidate signals, how does one combine them into a single forecast $\hat{\mu}_t$ of Eq. (2)? The answer has evolved from parsimonious linear models to flexible machine learning, with an important detour through *combination*.

4.1 Linear predictive regression and its hazards

The natural starting point regresses next-period returns on current signals, $r_{t+1} = a + \boldsymbol{b}^\top \boldsymbol{s}_t + \varepsilon_{t+1}$. Simple and interpretable, it is also subject to well-known traps when predictors are persistent and correlated with returns' own innovations, which inflates in-sample fit and corrupts naive significance tests — part of why Welch and Goyal (2008) find that in-sample success so often fails to survive out of sample. The discipline of reporting *out-of-sample* R^2 against a naive benchmark (Campbell and Thompson, 2008) is the first line of defense.

4.2 Forecast combination: averaging beats choosing

A robust empirical regularity, far older than finance, is that *combining* forecasts from several models usually beats trying to pick the single best one. Bates and Granger (1969) introduced the idea; Timmermann (2006) surveys the theory (diversification of model error, much as asset

diversification reduces portfolio risk); and Stock and Watson (2004) document its strength across many series. Paradoxically, simple equal weights often beat "optimal" estimated weights — the *forecast-combination puzzle* explained by Smith and Wallis (2009) as the result of estimation error in the weights themselves. The practical lesson for alpha: prefer humble, well-conditioned combinations of weak signals over a single confidently-tuned model.

4.3 Machine learning in asset pricing

The contemporary frontier applies flexible, regularized learners — penalized regression, gradient-boosted trees (Ke et al., 2017), random forests, and neural networks — to large signal sets. The landmark study is Gu et al. (2020), who show in a comprehensive horse race that machine-learning methods substantially improve out-of-sample prediction of the cross-section of returns relative to linear baselines, with trees and neural networks performing best and largely agreeing on *which* signals matter (variants of momentum, liquidity, and volatility). Crucially, their gains come from disciplined regularization and honest out-of-sample protocol, not from raw flexibility. López de Prado (2018) provides the canonical practitioner's account of how to deploy machine learning in finance *without* fooling oneself — the subject of the next section. Standard, well-documented implementations (Pedregosa et al., 2011; Ke et al., 2017) have made these methods ubiquitous, which only sharpens the evaluation problem: when anyone can fit ten thousand models by lunch, the scarce resource is not the model but the honest test.

5 How to measure an alpha signal: the discipline of honest evaluation

This is the heart of the survey. A candidate signal must clear three rising bars: *statistical* predictive accuracy, *economic* significance net of costs, and — the bar most often skipped — *multiplicity-aware* significance that accounts for how many signals were tried. We then add the modern requirement of *calibrated uncertainty*.

5.1 Bar 1 — statistical accuracy out of sample

The first question is whether the forecast beats a trivial benchmark on data not used to build it.

- **Out-of-sample R^2.** Following Campbell and Thompson (2008), compare the forecast's squared error against that of the prevailing historical mean; a positive out-of-sample R^2, however small, is the minimal evidence of genuine predictability. In-sample fit is almost worthless here.
- **Directional accuracy** (the hit rate of predicting the sign) is intuitive but weak at daily horizons, where a hit rate a touch above 50% can still be economically large or pure noise.

- **Comparing two forecasts.** The Diebold and Mariano (1995) test asks whether one model's out-of-sample loss is significantly lower than another's, properly accounting for the serial correlation of forecast errors — the right tool for a head-to-head, rather than eyeballing average errors.

5.2 Bar 2 — economic significance net of frictions

A statistically detectable edge is worthless if trading it costs more than it earns. Two corrections are mandatory.

- **Transaction costs.** Realistic costs — spreads, market impact, borrowing fees — must be charged against every trade. Frazzini et al. (2018) provide large-sample estimates of real-world trading costs and show that paper alphas can shrink dramatically once implementation is priced, especially for the fast-turnover signals that look best on paper.

- **The Sharpe ratio, deflated.** The Sharpe ratio (mean excess return per unit of volatility) is the standard scorecard, but a backtested Sharpe ratio is upward-biased precisely because it was *selected* as the best of many. Bailey and López de Prado (2014) derive the *Deflated Sharpe Ratio*, which corrects the observed Sharpe for the number of trials, the length of the sample, and the non-normality of returns — converting a flattering backtest number into an honest one.

5.3 Bar 3 — multiplicity: the data-snooping problem

Here lies the field's central methodological insight. If you test enough signals, some will look brilliant *by chance alone*. Evaluating the winner as if it were the only candidate is the cardinal sin of quantitative finance, and the literature has built a precise apparatus against it.

- **Reality Check and Superior Predictive Ability.** White (2000) introduced the "Reality Check," a bootstrap test of whether the *best* of many strategies genuinely beats a benchmark once the search is accounted for. Hansen (2005) sharpens it into the SPA test, more powerful and robust to poor strategies padding the set. Sullivan et al. (1999) applied the idea to technical trading rules and found much apparent profitability was data-snooping artifact. The required dependence-robust resampling is supplied by the *stationary bootstrap* of Politis and Romano (1994).

- **Which signals, not just whether any.** Romano and Wolf (2005) extend the logic to *stepwise* multiple testing, identifying the full set of genuinely significant strategies while controlling the family-wise error rate.

- **Raising the bar for "discovery."** Surveying hundreds of published factors, Harvey et al. (2016) argue that the conventional $t > 2$ threshold is far too lax given the multiplicity of the

search; they recommend a hurdle closer to $t > 3$, and Harvey and Liu (2015) translate the argument into concrete backtesting practice. Arnott et al. (2019) distill it into a practical research protocol for the machine-learning era.

- **Backtest overfitting, quantified.** Bailey et al. (2017) define the *Probability of Backtest Overfitting* (PBO): the probability that the configuration chosen as best in-sample underperforms the median out-of-sample. The cross-validation scheme that makes this estimable without leakage — combinatorial purged cross-validation — is developed by López de Prado (2018).

5.4 The replication lens: do discoveries survive?

Multiplicity is not a hypothetical worry; it is visible in the data after the fact. McLean and Pontiff (2016) show that the average documented predictor's return falls by roughly half (about 58%) after publication — partly because some findings were statistical flukes, partly because real edges get arbitraged once known, exactly as Lo (2004) would predict. Hou et al. (2020) replicate hundreds of anomalies under a uniform, careful methodology and find that a large majority lose significance, especially once microcaps are handled properly. The broader scientific backdrop is Ioannidis (2005)'s argument that, in any field with many tested hypotheses and flexible analysis, most published positive findings are false. Finance is not exempt; if anything, its incentives make it a prime example (Harvey et al., 2016).

5.5 Bar 4 — calibrated uncertainty, not just point forecasts

A modern forecast should report not only a number but an honest sense of how uncertain it is. Two complementary technologies have matured.

- **Proper scoring rules.** To grade probabilistic forecasts one needs a score that cannot be gamed by misreporting one's true belief. Gneiting and Raftery (2007) characterize *strictly proper* scoring rules (such as the logarithmic score and the Continuous Ranked Probability Score, CRPS), which reward honest, well-calibrated predictive distributions. Quantile forecasts can be scored with the pinball loss of quantile regression (Koenker, 2005).

- **Conformal prediction.** A signal's point forecast should come with a prediction interval that actually contains the truth at the stated rate. Conformal prediction (Vovk et al., 2005; Lei et al., 2018) produces such intervals with *distribution-free, finite-sample* coverage guarantees under exchangeability. Because financial data are emphatically *not* exchangeable — they drift and shift regimes — the recent extensions matter most: Gibbs and Candès (2021) adapt the intervals online to distribution shift, and Barber et al. (2023) establish coverage guarantees *beyond* exchangeability. These tools let an alpha signal say "up, but I am this unsure," which is exactly what a risk-aware investor needs.

5.6 A practical checklist

Synthesizing the above, an honest evaluation of a candidate alpha signal should: (1) freeze a true out-of-sample period before any search; (2) report out-of-sample R^2 and a Diebold and Mariano (1995) comparison against a real benchmark; (3) charge realistic costs (Frazzini et al., 2018) and deflate the Sharpe ratio for the number of trials (Bailey and López de Prado, 2014); (4) apply a data-snooping-robust test of the *best* strategy (White, 2000; Hansen, 2005) and a PBO check (Bailey et al., 2017); (5) hold discovery to a multiplicity-appropriate threshold (Harvey et al., 2016); and (6) emit calibrated, distribution-free intervals scored by proper rules (Gneiting and Raftery, 2007; Gibbs and Candès, 2021). A signal that clears all six is rare — and that rarity is the point.

6 Open problems and directions for further research

The state of the art is better at *disproving* false signals than at *producing* durable true ones. Several frontiers remain genuinely open.

1. **Non-stationarity and regime shift.** Markets evolve, so a model estimated on the past is always slightly mis-specified for the future (Lo, 2004). Distribution-free inference under dependence and drift is advancing (Gibbs and Candès, 2021; Barber et al., 2023), but a fully satisfactory theory of *when* a signal has decayed enough to retire it does not yet exist.

2. **Alpha decay and the arms race.** If publication and capital erode edges (McLean and Pontiff, 2016; Hou et al., 2020), then the economically relevant object is not a signal's historical strength but its *remaining* strength and capacity. Modeling decay — and designing signals that are robust to being crowded — is an open, high-value problem.

3. **The high-frequency signal-to-noise ceiling.** At the daily, single-name horizon the predictable component of returns is minuscule relative to noise, and honest evaluation frequently returns a calibrated *null*. Whether this ceiling is fundamental or merely awaiting better data and features (e.g. the overnight/intraday and volume-conditioned structure of Lou et al., 2019; Llorente et al., 2002) is unresolved.

4. **Prediction versus explanation.** Machine learning excels at prediction (Gu et al., 2020) but is silent on mechanism; a signal without an economic story is more likely to be a multiplicity artifact (Harvey et al., 2016). Integrating causal reasoning with predictive performance is an active and important direction.

5. **Uncertainty-aware portfolio construction.** Most of the evaluation apparatus stops at the signal; comparatively little connects calibrated predictive *distributions* (Gneiting and Raftery, 2007; Vovk et al., 2005) to downstream *allocation* in a way that propagates

uncertainty into position sizing under realistic costs (Frazzini et al., 2018). This is precisely the seam where forecasting meets portfolio optimization, and where much practical value remains unclaimed.

6. **Reproducibility as a first-class deliverable.** Given Ioannidis (2005) and Harvey et al. (2016), the field is moving toward pre-registration, append-only trial ledgers, and machine-checkable claims (Arnott et al., 2019; López de Prado, 2018). Making honest, multiplicity-aware evaluation *cheap and standard* — rather than heroic — may do more for real-world alpha than any single new signal.

7 Conclusion

An alpha signal is a deceptively simple object — information today about risk-adjusted return tomorrow — wrapped in two hard problems: separating it from risk (Fama, 1970), and separating it from luck (White, 2000; Harvey et al., 2016; Bailey et al., 2017). The literature has mapped where signals live (the cross-section, the time series, microstructure, and volatility), built an increasingly powerful modeling toolkit culminating in disciplined machine learning (Gu et al., 2020), and — most valuably — forged an evaluation discipline rigorous enough to tell true edges from the artifacts of a vast search. For the student and the practitioner alike, the central message is the same and worth repeating: in a world where models are free and trials are infinite, the scarce and decisive skill is honest measurement. QUAFI's own quantitative methodology is built on exactly this premise — that the credibility of any number we show a client rests on the rigor of the test behind it.

References

Yakov Amihud. Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, 5(1):31–56, 2002.

Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Paul Labys. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625, 2003.

Andrew Ang, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang. The cross-section of volatility and expected returns. *Journal of Finance*, 61(1):259–299, 2006.

Robert D. Arnott, Campbell R. Harvey, and Harry Markowitz. A backtesting protocol in the era of machine learning. *Journal of Financial Data Science*, 1(1):64–74, 2019.

David H. Bailey and Marcos López de Prado. The deflated sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5):94–107, 2014.

David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. The probability of backtest overfitting. *Journal of Computational Finance*, 20(4):39–69, 2017.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023.

Ole E. Barndorff-Nielsen and Neil Shephard. Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, 4(1):1–30, 2006.

John M. Bates and Clive W. J. Granger. The combination of forecasts. *Operational Research Quarterly*, 20(4):451–468, 1969.

John Y. Campbell and Samuel B. Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531, 2008.

Mark M. Carhart. On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82, 1997.

Fulvio Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417, 1970.

Andrea Frazzini, Ronen Israel, and Tobias J. Moskowitz. Trading costs. Technical report, AQR Capital Management, Working Paper, 2018.

Kenneth R. French, G. William Schwert, and Robert F. Stambaugh. Expected stock returns and volatility. *Journal of Financial Economics*, 19(1):3–29, 1987.

Thomas J. George and Chuan-Yang Hwang. The 52-week high and momentum investing. *Journal of Finance*, 59(5):2145–2176, 2004.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672, 2021.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273, 2020.

Peter Reinhard Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005.

Campbell R. Harvey and Yan Liu. Backtesting. *Journal of Portfolio Management*, 42(1):13–28, 2015.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016.

Kewei Hou, Chen Xue, and Lu Zhang. Replicating anomalies. *Review of Financial Studies*, 33 (5):2019–2133, 2020.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

Narasimhan Jegadeesh. Evidence of predictable behavior of security returns. *Journal of Finance*, 45(3):881–898, 1990.

Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48(1):65–91, 1993.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154, 2017.

Roger Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Guillermo Llorente, Roni Michaely, Gideon Saar, and Jiang Wang. Dynamic volume-return relation of individual stocks. *Review of Financial Studies*, 15(4):1005–1047, 2002.

Andrew W. Lo. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management*, 30(5):15–29, 2004.

Marcos López de Prado. *Advances in Financial Machine Learning*. John Wiley & Sons, Hoboken, NJ, 2018.

Dong Lou, Christopher Polk, and Spyros Skouras. A tug of war: Overnight versus intraday expected returns. *Journal of Financial Economics*, 134(1):192–213, 2019.

R. David McLean and Jeffrey Pontiff. Does academic research destroy stock return predictability? *Journal of Finance*, 71(1):5–32, 2016.

Andrew J. Patton and Kevin Sheppard. Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics*, 97(3):683–697, 2015.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994.

Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005.

G. William Schwert. Anomalies and market efficiency. *Handbook of the Economics of Finance*, 1: 939–974, 2003.

Jeremy Smith and Kenneth F. Wallis. A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355, 2009.

James H. Stock and Mark W. Watson. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430, 2004.

Ryan Sullivan, Allan Timmermann, and Halbert White. Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance*, 54(5):1647–1691, 1999.

Allan Timmermann. Forecast combinations. In Graham Elliott, Clive W. J. Granger, and Allan Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, pages 135–196. Elsevier, 2006.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Ivo Welch and Amit Goyal. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508, 2008.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.