Financial Economics Research; Vol.3, No.4; 2026
E-ISSN: 3006-3663 P-ISSN: 3006-3655
DOI: https://doi.org/10.70267/fer.v3n4.143161
Published by: Zeus Press

The Evolution of Predictive Analytics: Machine Learning and the Predictive Step in Empirical Asset Pricing—From Factor Zoo to Factor Refinery

Kaiyu Wu*

School of Business, Nanjing University of Science and Technology Zijin College, Nanjing 210023, China

Abstract

In the last fifty years, empirical asset pricing has depended on a consistent knowledge-production pipeline: economic theory first suggests hypotheses, hypotheses are then operationalised through candidate factors, and the validity of those hypotheses is finally determined by in-sample statistical significance. As the number of candidate factors grows into the hundreds, this pipeline is no longer a reliable way of identifying genuine factors among data-mined false positives. Organised around four steps — testing, construction, discovery, and integration — the paper argues that machine learning has established, alongside the conventional pricing tests, an evaluation framework based on out-of-sample predictive performance. This does not mean, however, that the out-of-sample evaluation framework is a substitute for alpha, GRS, or risk-price tests. The main thesis is that machine learning has made genuine instrumental advances in prediction-based tasks, but cannot be regarded as a substitute for causal inference or mechanistic explanation; its benefits are real but restricted, and its drawbacks (loss of interpretability, conflation of correlation with causation, and overfitting) are significant. A more precise statement is thus not that machine learning has changed empirical asset pricing, but that it has changed the mode of execution of prediction-related tasks within the field without changing the basic logic of understanding-related tasks.

Keywords

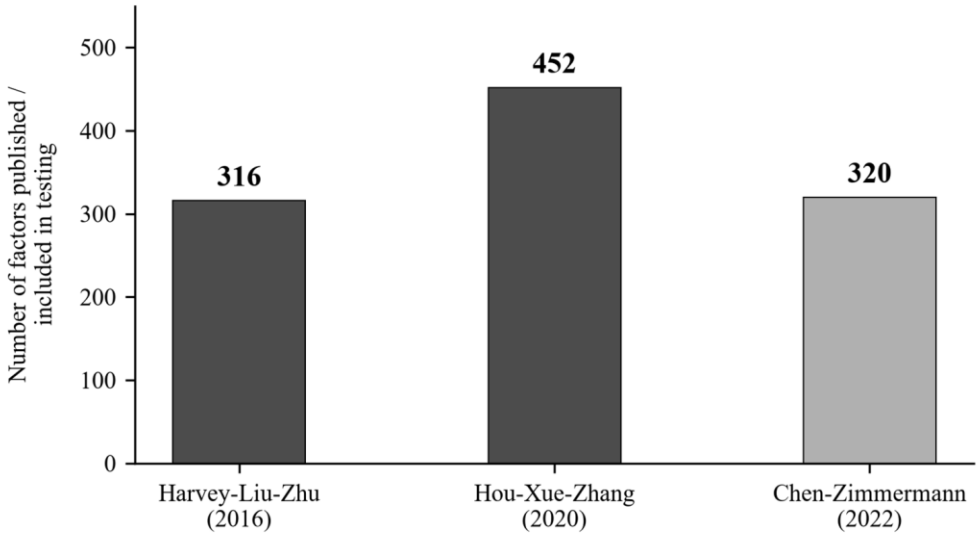
machine learning, asset pricing, factor zoo, out-of-sample prediction, evaluation framework

1. Introduction

The factor zoo is the honest self-assessment of the situation in empirical asset pricing. The term was introduced by Cochrane [1] in his discussion of discount rates; it implies that there are now so many potential factors that economic theory can no longer justify each of them individually. Harvey, Liu, and Zhu [2] reported 316 published cross-sectional return predictors; in the replication analysis of Hou, Xue, and Zhang [3], 452 anomalies were tested together; and the open-source factor data set of Chen and Zimmermann [4] lists more than 320 reproducible factors. The three counts differ in scale, but they all point in the same direction: the supply of factors passed the hundred mark long ago, far beyond what economic theory was previously expected to provide (see Figure 1, constructed and illustrated by the author).

© 2026 The Author(s). Authors retain copyright in their articles.

Figure 1: The number of factors documented in three major studies: an expansion to several hundred



Compiled and drawn by the author from Harvey–Liu–Zhu 2016, Hou–Xue–Zhang 2020, and Chen–Zimmermann 2022. Note: The three counts are derived from different statistical bases, but they point in the same direction — the supply of factors far exceeds the a priori guidance suggested by economic theory.

Empirical asset pricing should reconsider the limits of its current testing structures now that "significance" is no longer believable and the replication requirements are not homogeneous. Machine learning presents exactly such an opportunity for re-examination, but its real value has long been misconstrued as the value of a more potent prediction tool. The paper proposes that the role of machine learning is not to replace the old pricing tests but to offer another benchmark of evaluation based on out-of-sample predictive performance, thereby compelling empirical asset pricing to reconsider what the validity of a factor means.

The argument proceeds in four steps — testing (Direction II, Section 3), construction (Direction III, Section 4), discovery (Direction IV, Section 5), and integration (Direction V, Section 6) — beginning with the issues raised by the conventional paradigm (Direction I, Section 2) and ending with its limits and the criticisms they invite (Direction VI, Section 7).

2. Challenges Facing the Traditional Paradigm

Conventional factor discovery has a well-defined structure: it starts with economic theory offering a hypothesis, continues with the construction of a candidate factor, and ends by testing the importance of that factor with data. Knowledge is generated along this path by theory alone; data play the role of arbiter. One of the main assertions of this section is that once the number of candidate factors rises into the hundreds, the conventional strategy of testing factors one by one and declaring them valid whenever they are significant becomes highly susceptible to scale pressure and can no longer separate genuine factors from false positives produced by data mining. This is not a failure of one particular study but a systemic problem of the paradigm under scale pressure.

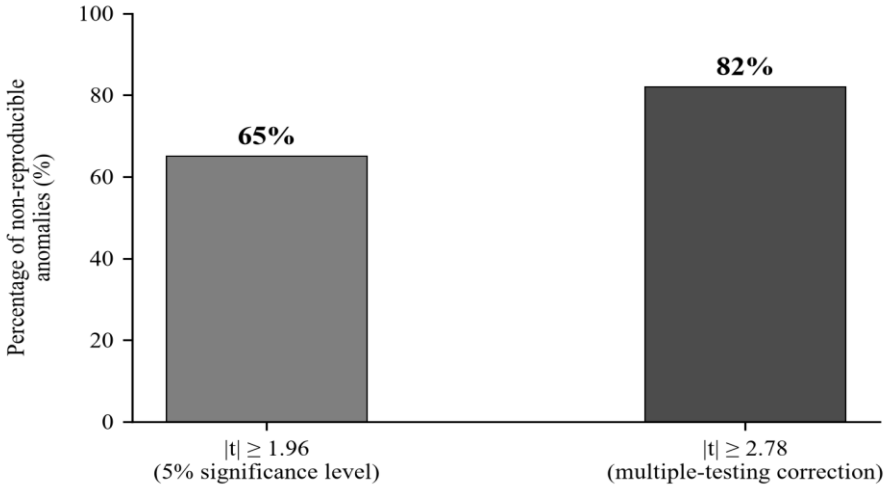
2.1 The Replication Crisis: Survival Rates under Unified Replication

The fundamental premise of the traditional paradigm is that when any element is considered important in an individual test, it is thus real. The assumption was directly tested using a so-called unified replication protocol by reexamination of 452 published anomalies with identical procedures, sample period, and control variables, as suggested by Hou, Xue, and Zhang [3].

The findings are clear: 65 percent of the anomalies were not reproduced at the significance level of $|t| \geq 1.96$; at the multiple-testing-adjusted level of $|t| \geq 2.78$, the failure rate rises to 82 percent (see Figure 2, redrawn by the author from Hou–Xue–Zhang 2020 data).

© 2026 The Author(s). Authors retain copyright in their articles.

Figure 2: Anomaly failure rates under the unified replication protocol



Redrawn by the author from Hou–Xue–Zhang 2020 data. Note: $|t| \geq 1.96$ is the 5 percent significance level and $|t| \geq 2.78$ is the threshold for multiple-testing correction; the higher the threshold, the higher the failure rate, because much of the observed significance stems from multiple testing and data mining.

The two figures suggest that much of the earlier reported significance might have been due to multiple testing and data mining as opposed to actual economic effects.

However, a tension has to be recognised. Jensen, Kelly, and Pedersen [5], applying a Bayesian factor-correlation test to a global population of factors, report that most highly correlated factors are replicable, and therefore challenge the very name of a replication crisis.

These two studies are not contradictory; they differ in methodology. For Hou–Xue–Zhang, replication must be an exact re-run, that is, a test of whether the factor remains statistically significant under the same procedures as in the original study; for Jensen–Kelly–Pedersen, the methodology may vary and the question is whether the economic information behind the factor is still present. The first measures replicability, the second robustness. Since the interpretations of replication are not the same, the conclusions will also differ.

The value of this tension is that it reveals the underlying issue: the standard of replication in the traditional paradigm is not unique. What counts as valid depends on which replication standard is selected — and the traditional paradigm offers no unchallenged meta-standard for making that choice. This is why an evaluation model based on out-of-sample predictive performance is needed.

2.2 Three Supplementary Lines of Evidence: Multiple Testing, Marginal Contribution, and Publication Bias

Beyond the replication crisis, three independent chains of evidence point to the same conclusion.

The first is the multiple-testing problem. The fundamental thesis of Harvey, Liu, and Zhu [2] is a statistical truism that the classical paradigm long ignored: when each of hundreds of factors is tested separately, pure chance alone will produce a substantial number of so-called significant findings at the 5 percent level even if every factor is false. With more than 300 candidates, the conventional $t \geq 1.96$ limit is far too lax and needs to be tightened to $t > 3.0$. The argument is that the traditional pass criterion has already failed under an inflated number of factors, since it does not take into account how many factors were examined.

The second is the marginal contribution test. The more stringent question asked by Feng, Giglio, and Xiu [6] in “Taming the Factor Zoo” is whether a new factor adds any incremental information once the existing factors are controlled for. They conclude that most new factors are no longer incremental. The methodological importance of the work lies in its raising of the testing standard: the question is not simply whether a factor is significant, but whether it is incremental relative to current knowledge, which is much more demanding than the previous criterion.

© 2026 The Author(s). Authors retain copyright in their articles.

The third is the evidence on intertemporal publication bias. McLean and Pontiff [7] investigate the intertemporal performance of 97 variables and find that out-of-sample returns are 26 percent lower and post-publication returns 58 percent lower than in-sample returns. The decay is not random but systematic after publication, which can be read as evidence that the publication process has filtered out the best-performing statistical artefacts in sample.

2.3 Section Summary

The thesis of Direction I can be summarised in three statements. First, hundreds of factors have been identified, far more than economic theory could previously justify. Second, unified replication shows that 65–82 percent of anomalies do not replicate under a rigorous protocol, and the notion of replication itself is not unique in the conventional paradigm. Third, the joint indication of multiple testing, marginal contribution, and publication bias suggests that the traditional pass criterion is seriously threatened by scale pressure.

These three points establish a premise: the traditional paradigm faces systemic challenges under scale pressure. The premise provides the problem context for the expansion of the testing-evaluation framework presented in Section 3, but it should be noted that problems caused by scale pressure do not by themselves imply that a machine-learning framework can replace traditional tests.

3. Expanding the Testing-Evaluation Framework: From Significance Tests to Out-of-Sample Evaluation

Testing is the first step of empirical asset pricing and the one most directly affected by machine learning. The main argument of this part is that the obstacles facing conventional tests lie less in the severity of the significance levels than in structural flaws of the procedures themselves; machine learning has supplied an independent measure of factor assessment by introducing out-of-sample predictive performance. It should be noted that the machine-learning framework does not replace alpha, GRS, or risk-price tests but introduces, alongside the traditional tests, a benchmark evaluation based on out-of-sample predictive performance.

3.1 Inherent Limitations of the Traditional Testing Framework

Conventional tests rest on two classical tools whose logics differ and which should be considered separately.

The cross-sectional regression of Fama and MacBeth [8] first estimates the factor loadings β and then, in each cross section, uses those β to estimate the risk prices λ. It has a defect that has long been discussed: because β is itself estimated, it must contain measurement error, so when an investigator estimates λ from error-prone β, the classical errors-in-variables (EIV) problem emerges and distorts the estimate of λ. The flaw is partly hidden in portfolios but is greatly magnified in individual stocks.

The GRS test of Gibbons, Ross, and Shanken [9] is an alternative approach; in a time-series regression setting, it examines whether the alphas of all test assets are together zero (an F-test). It is not subject to the EIV critique described above because it does not require the step of estimating λ with β.

The criticism of the structure of test assets by Lewellen, Nagel, and Shanken [10] focuses on the fact that as long as test assets move in the same direction (which is almost unavoidable), nearly any factor mixture can appear to be reasonably priced. Put differently, the very criterion that cross-sectional tests must meet may be too lenient: a model need not have passed the test on its merits if the test itself was too simple to fail.

Jegadeesh et al. [11] respond by replacing single stocks with portfolios and by correcting the EIV bias of the Fama–MacBeth model with instrumental variables. The correction nevertheless remains within the bounds of the linear testing framework and does not touch the basis of the testing standard.

Note that the deficiencies mentioned above concern factor pricing only; they do not mean that the Fama–MacBeth and GRS techniques cannot be applied successfully in other contexts.

3.2 The Out-of-Sample Evaluation Benchmark of Machine Learning

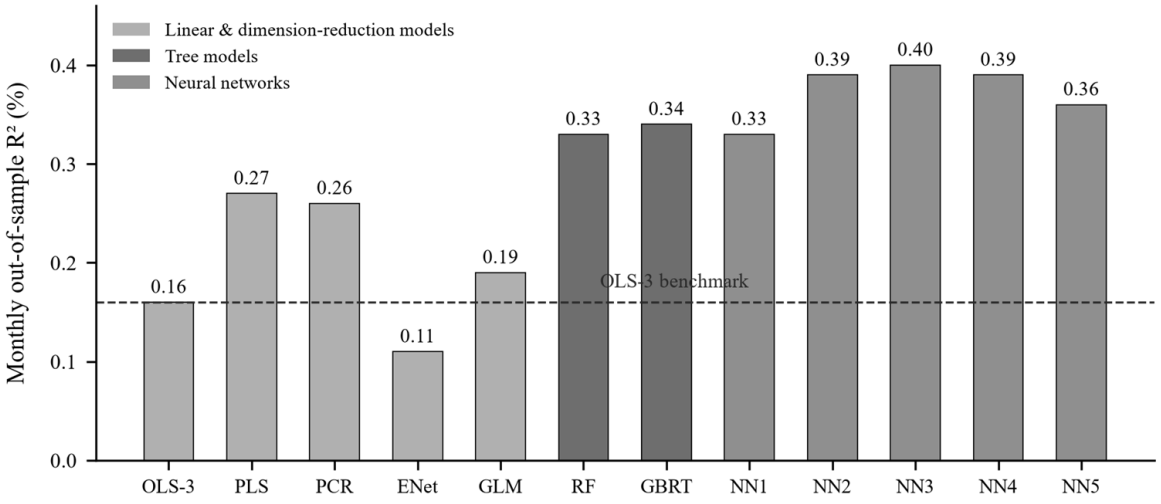
The main change brought about by machine learning is that factor validity is no longer identified with statistical significance in classical pricing tests but with strong predictive performance out of sample. This

© 2026 The Author(s). Authors retain copyright in their articles.

criterion adds a dimension to the definition of validity. It was systematically developed by Gu, Kelly, and Xiu [12].

Gu, Kelly, and Xiu [12] compare 13 prediction models, namely full-variable OLS, three-variable OLS, partial least squares (PLS), principal-component regression (PCR), elastic net (ENet), generalized linear models (GLM), random forests (RF), gradient-boosted regression trees (GBRT), and five neural networks (NN1–NN5). Their main contribution is the finding that nonparametric techniques, including tree models and neural networks, achieve the highest out-of-sample performance, with the best model reaching an out-of-sample R^2 of about 0.4 percent (see Figure 3, redrawn by the author from Gu–Kelly–Xiu 2020, Table 1). The monthly figures reported in the original paper are the following: PLS ≈ 0.27, PCR ≈ 0.26, ENet ≈ 0.11, RF ≈ 0.33, GBRT ≈ 0.34, NN3 ≈ 0.40.

Figure 3: Comparison of out-of-sample R^2 across the 13 predictive models (OLS+H omitted)



Redrawn by the author from Gu–Kelly–Xiu 2020, Table 1. Note: The R^2 of the all-variable OLS model (OLS+H) is −3.46 percent; it is not plotted because of its excessively large negative value. The remaining 12 models are plotted in full.

Although the figure looks small, it should be read against the right benchmark: it is a predictive gain that is still substantial at the individual-stock level and on a monthly basis, and it can be converted into portfolios with fairly high out-of-sample Sharpe ratios; net returns should be evaluated separately from transaction costs.

A significant operational benefit of this format is that it bypasses the EIV problem of the Fama–MacBeth structure. Machine learning does not need a first-stage estimate of β and a second-stage estimate of λ; it predicts directly from features out of sample and therefore avoids the circularity of estimating one estimator from another. This differs from the EIV correction of Jegadeesh et al. [11] inside the classical framework: the latter improves the quality of the estimate within the original framework, whereas the former changes the evaluative question from whether pricing errors equal zero to whether out-of-sample returns can be predicted systematically.

The applicability of the out-of-sample evaluation framework has also been established in cross-asset settings. Chinco, Clark-Joseph, and Ye [13] use LASSO to identify sparse signals in the cross-section of individual stocks, and Bianchi, Büchner, and Tamoni [14] use machine learning to predict bond risk premia. In both cases, out-of-sample predictive power serves as a measure of validity outside equities.

The metric should be defined explicitly: the out-of-sample R^2 of about 0.4 percent is computed from monthly U.S. equity returns over 1957–2016 and compared across 13 predictive models. It is a measure of predictive performance and cannot be compared directly with pricing-error measures such as alpha or GRS; Sharpe ratios belong to a different class of portfolio-performance measures and should not be conflated with it. This confusion of metrics is common in the literature.

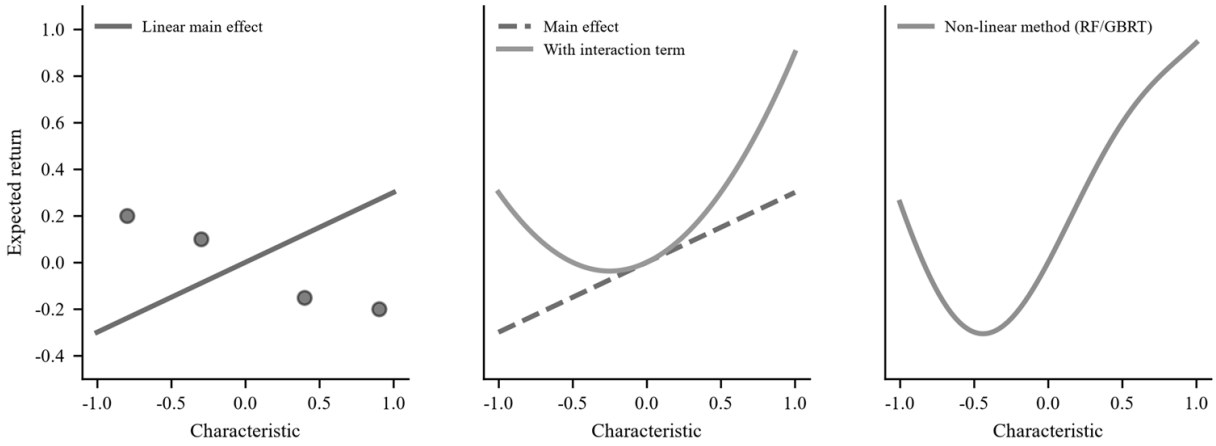
© 2026 The Author(s). Authors retain copyright in their articles.

3.3 Core Verification: Where Do the Gains of Nonlinear Methods Come From?

A deeper question needs to be answered: in which specific area do the benefits of machine learning come into play? The answer is not a matter of more complicated models but of two origins. Compared with linear parametric models (OLS, LASSO), nonparametric approaches can automatically learn the interactions between features and the nonlinearities in the data; compared with sorted portfolios, they reduce the loss of information caused by discretisation and by the impossibility of exhaustively enumerating high-dimensional interactions.

One important point is that if true returns are generated by interactive relationships among firm characteristics, linear parametric models (OLS, LASSO) represent main effects only and therefore carry no interaction information, whereas nonparametric techniques (random forests, gradient boosting) can generate interaction terms and nonlinear transformations automatically and thus deliver conditional gains (see Figure 4, a conceptual schematic).

Figure 4: Illustrative comparison between linear main effects and automatic interaction capture



Note: This is a conceptual diagram, not empirical data; it illustrates the interaction information that is lost when only main effects are modelled.

This logic matters because it turns the vague question of whether machine learning is more effective than linearity into a testable hypothesis: when features interact or the return structure is nonlinear, linear parametric methods can fail to capture information whereas nonparametric methods can achieve conditional gains. It is not a question of the superiority of algorithms but of what the evaluation model is designed to capture.

The limits should also be defined: the small improvement of nonlinear approaches holds only when the return structure really does contain interactions or nonlinearities. When the true data-generating process is linear, the benefit of nonlinear methods vanishes and can even turn negative because of overfitting. This limit leads directly to the criticism of overfitting in Direction VI and shows that the statement that "nonlinear is better than linear" is an empirical assessment, not a final verdict.

3.4 Section Summary

The Direction II argument may be summarised in three points. First, the challenges facing the traditional tests are structural: the EIV problem of Fama–MacBeth and the correlated nature of the test assets make the pass criterion excessively easy, a flaw that cannot be repaired merely by raising the t-value thresholds. Second, the machine-learning framework of out-of-sample evaluation provides a different standard for assessing factors; it avoids the EIV problem of the Fama–MacBeth framework, but out-of-sample R^2 reflects predictive ability and cannot be compared with pricing-error indicators such as alpha or GRS. Third, the extra benefit of nonlinear techniques is situational: it arises from the information loss of linear parametric techniques and from the operational constraints of sorted portfolios, and it applies only when interactions or nonlinearities are present in the return structure.

These three points support the thesis that the real value of machine learning is not confined to the release of a stronger prediction engine; machine learning offers an additional, practical standard for the notion of factor

© 2026 The Author(s). Authors retain copyright in their articles.

validity. Because in-sample significance is extremely vulnerable to scale pressure, introducing an out-of-sample evaluation framework is methodologically important; it remains, however, an addition to the existing concept rather than a replacement of it.

4. Changes in the Construction Paradigm: From Sorting to Automatic Interaction Discovery

The testing phase answers the question of whether a factor is valid; the construction phase answers the question of how it is produced. The main argument of this section is that the manual sorting procedures underlying conventional factor construction run into practical constraints in high-dimensional settings — discretisation loss, arbitrary cut points, the curse of dimensionality, and the difficulty of enumerating interactions exhaustively; machine learning addresses them by searching automatically for interactions between features and for nonlinear structures, but it lacks interpretability.

4.1 Limitations of Traditional Sorting Methods

Single- and double-sorted portfolios are the standard operations by which traditional factors are constructed: stocks are sorted on a characteristic, grouped, and the long–short return of the extreme groups is used as the factor return.

A popular misconception should be clarified beforehand: sorting is a nonparametric grouping technique and does not imply that the relationship between characteristic and return must be linear or monotonic. Decile sorts may be U-shaped, inverted U-shaped, or otherwise non-monotonic; double sorts consider return differences conditional on two dimensions and do not necessarily assume the absence of interaction between characteristics.

The genuine limitations of sorting methods are the following four.

First, discretisation causes information loss. Placing continuous characteristics into bins removes the within-bin variance and also lowers the precision of the factor returns.

Second, grouping breakpoints are subjective. The quintiles, deciles or any other divisions and their breakpoints are not standardised; inferences can vary depending on the grouping scheme.

Third is the curse of dimensionality. The number of cells in a multi-way sort grows exponentially with the number of characteristics, the sample size in each cell falls quickly, the precision of the estimates declines, and multi-characteristic interactions cannot be studied efficiently.

Fourth, it is hard to enumerate all possible multi-characteristic interactions. Manual sorting can cross two or three characteristics at best; interactions of a higher order are not feasible in practice.

The interaction effects among predictors are a significant contributor to the gains of tree models and neural networks, as Gu, Kelly, and Xiu [12] state. This is not equivalent to saying that traditional portfolio sorts assume linearity; it points rather to the operational limitation of sorting, namely its inability to represent higher-order interactions.

4.2 From Sorting to Dimension Reduction: Evolution within the Linear Path

One point that is easily misunderstood needs to be clarified. The operational constraints of sorting were discussed in Section 4.1 (information loss due to discretisation, arbitrary breakpoints, the curse of dimensionality, and the difficulty of enumerating multi-characteristic interactions).

Sorting is not, however, the only way of constructing factors. Econometricians had already suggested an alternative long before machine learning entered the subject, namely dimension reduction and extraction, that is, the direct derivation of latent factors from a characteristic panel without grouping. The three-pass regression filter (3PRF) of Kelly and Pruitt [15] is in fact a supervised dimension-reduction approach to forecasting with many predictors and is not, strictly speaking, an asset-pricing factor-construction technique; it is the IPCA of Kelly, Pruitt, and Su [16] that truly brought dimension reduction into factor construction.

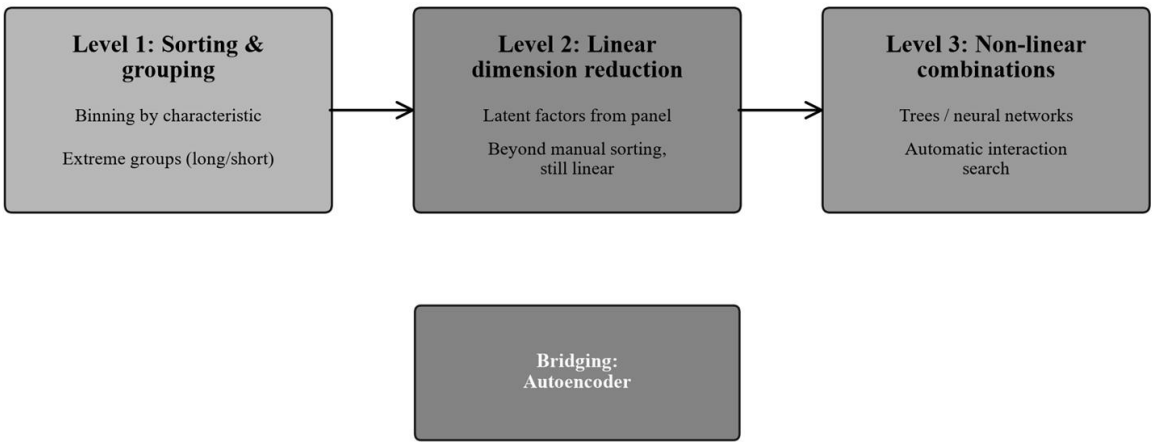
The most important difference is that dimension reduction addresses the discretisation loss of sorting but still resides in the linear-combination paradigm. It represents factors as linear functions of characteristics and

© 2026 The Author(s). Authors retain copyright in their articles.

cannot represent nonlinear relationships. Kelly, Pruitt, and Su also develop the idea that characteristics are covariances by transforming characteristics into time-varying factor loadings [16]. It should be observed, nevertheless, that the statement that characteristics are covariances is itself a hypothesis requiring further verification; it has not yet gained enough validation to be treated as universally applicable.

Between linear dimension reduction and nonlinear combination, the autoencoder asset-pricing model of Gu, Kelly, and Xiu [17] acts as a transitional bridge: the autoencoder compresses features and identifies latent factors in a nonlinear manner, thereby breaking the linearity presumption at the dimension-reduction step (see Figure 5, drawn by the author).

Figure 5: Evolution of the Factor-Construction Paradigm: From Linear Dimension Reduction to Nonlinear Combination via Autoencoders



Drawn by the author. Note: This is a conceptual diagram without data; it shows the three-tier evolution of the factor-construction paradigm.

4.3 Decision Trees and Automatic Interaction Capture

Tree-based methods such as random forests and gradient boosting are instances of automatic interaction discovery, since they do not rely on a predetermined set of interaction patterns but explore interactions between features and nonlinear partitions in a recursive manner.

The conclusion of Gu, Kelly, and Xiu [12] is that tree models and neural networks are the highest-performing of the 13 predictive models, which is exactly what this ability implies. Tree models can construct interactive features automatically, such as high turnover combined with low volatility, whereas manual sorting requires the researcher to see the interaction in advance.

Bryzgalova, Pelger, and Zhu [18], in “Forest Through the Trees”, go further and construct test assets with decision trees: instead of sorting manually on a size–B/M grid, the sorting itself is performed by decision trees. The importance of this work is that it extends the decision tree from a prediction tool to a test-asset construction method. It should be noted that the paper is a working paper that has not completed peer review; it counts as frontier exploration rather than as a settled conclusion.

This advance prompts a deeper question: since decision trees can uncover interactions that manual sorting misses, have conventional factors already lost information at the construction stage? At present it can be stated that when a return structure contains interactions and nonlinearities, traditional sorting is prone to information loss. This is an internal link between Direction III and Direction II: the problems of traditional tests (Direction II) may be rooted partly in the shortcomings of the construction approach with respect to discretisation loss and the enumeration of high-dimensional interactions (Direction III) — although such an evaluation is not obligatory.

© 2026 The Author(s). Authors retain copyright in their articles.

4.4 The Cost of Interpretability

The cost of this change in construction method concentrates on a single issue: interpretability.

Sorting and linear dimension reduction have an unassailable benefit: they can answer precisely the question of which characteristic, at what weight, drives returns. This makes the factor economically narratable, explainable by theory, intelligible to regulators, and trustworthy to practitioners.

Decision trees and neural networks are nonlinear and give up exactly this benefit. It is hard to compress the split paths of a trained random forest into a succinct economic mechanism, and the internal representations of a deep neural network cannot be converted into claims of the form "characteristic X determines return Y". This is a significant loss for a field whose central goal is mechanistic understanding.

The cost is not fixed. Feature-importance measures and SHAP (Shapley additive explanations) are interpretability methods developed in recent years that reduce the black-box problem to some degree, but they supply information about which feature is statistically important rather than why the feature influences returns in economic terms. The first is statistical explanation, the second is economic mechanism, and the two should not be confused. This distinction leads directly to the notion of economic retrospective validation developed in Direction IV.

4.5 Section Summary

The Direction III argument may be summarised in three points. First, traditional sorting does not rest on a linearity assumption, but discretisation loss, arbitrary breakpoints, the curse of dimensionality, and the difficulty of enumerating interactions all cause information loss. Second, the machine-learning construction process has two stages: the first is linear dimension reduction, which removes the discretisation loss of sorting, and the second breaks the linear-combination structure itself through nonlinear solutions. Third, the price of this transition is interpretability; nonlinear construction comes at the cost of the economic story conveyed by the characteristic–mechanism link, and statistical explanation is no substitute for economic mechanism.

These three points together formulate the question that Direction IV should answer: when factors are constructed by algorithms rather than by human hands, how does the source of knowledge about factors change? How can algorithmically extracted factors be shown to carry economic significance rather than being mere statistical artefacts?

5. The Discovery Path: Toward Data-Driven Methods and the Backward Shift of Theory's Role

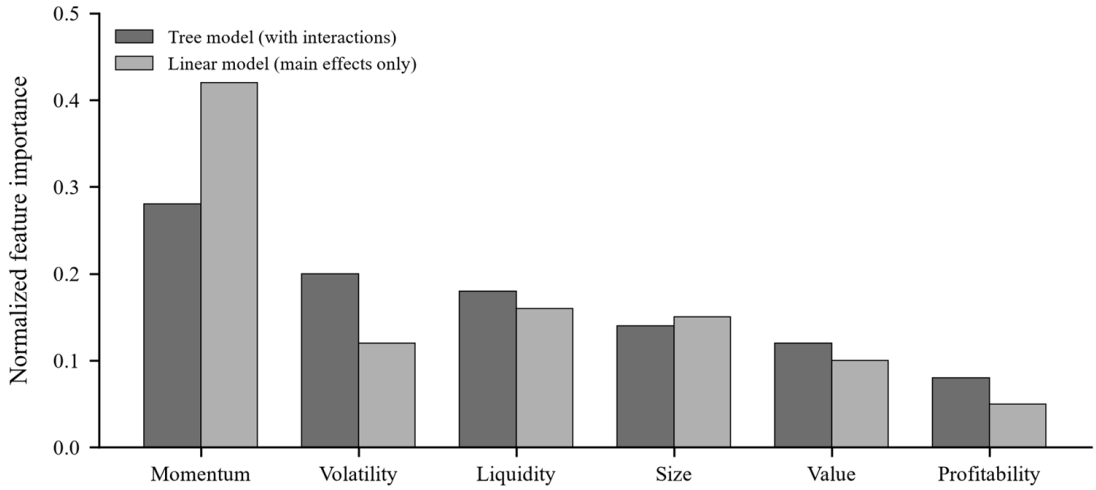
The construction phase alters the way in which factors are produced; the discovery phase alters the origin of factors. The main point of this section is that machine learning provides a data-based route of discovery, which leads to a partial shift of the source of knowledge from human prior theory to algorithmic search. This transition does not and will not remove the need for economic validation; rather, it changes the role of economics from that of a pre-existing hypothesis to that of a retroactive examination.

5.1 The Transfer of the Source of Knowledge

Conventional factor discovery follows one route: theory first builds the factor, which is then tested; knowledge starts from theory and data merely decide whether the theory passes. Machine learning offers a different route: signals are mined by data-driven methods in a high-dimensional feature library without the need to predict in advance which features are correct. The role of data in the production of knowledge therefore increases, and theory at certain points becomes a post hoc explanation (see Figure 6, a schematic by the author, not an empirical finding).

© 2026 The Author(s). Authors retain copyright in their articles.

Figure 6: Illustrative comparison of feature-importance rankings between tree-based and linear models



Note: Illustrative data only. The tree model attributes larger contributions to more features (such as volatility and liquidity) than the linear model does.

The change is clearly visible in the methodological spectrum.

Variable selection is the first layer. The sparse selection introduced by Tibshirani [19] for high-dimensional features through penalised regression is one of the oldest forms of data-driven discovery. Chinco et al. [13] apply it to the cross-section of individual stocks, deriving sparse predictors from a large number of lagged-return signals.

The second layer is factor extraction. The IPCA of Kelly, Pruitt, and Su [16] does not require the economic meaning of the latent factors to be specified in advance; it estimates simultaneously the latent factors and their characteristic-driven time-varying loadings (although the number K of latent factors still has to be set). The advance is to move from selecting existing features to estimating latent factors.

The third layer is semantic and alternative-data signal mining. Loughran and McDonald [20] build a specialised finance dictionary and show that general-purpose dictionaries perform poorly on financial text, which makes a domain-specific dictionary a basic instrument for quantifying textual signals. Jiang, Kelly, and Xiu [21] extract trend signals from price-chart images with convolutional neural networks, extending the carriers of signals from numeric features to raw image data.

The next level is semantic knowledge from large language models. Lopez-Lira and Tang [22] use GPT-4 to read news headlines and anticipate the stock-price responses to them; the paper has been published in the Journal of Financial Economics.

A caveat on evidence: the most recent layers of this lineage (images, large language models) are largely new work, and some are still working papers or forthcoming. They are the leading directions of the discovery path, not its final ones. Citations should therefore be explicit about distinguishing evidence that has been published and peer-reviewed from evidence that is still a working paper or forthcoming, so that frontier exploration is not confused with mature discovery.

5.2 The Core Risk of the Discovery Path: Statistical Artefacts Masquerading as Economic Factors

Data-driven discovery has a fundamental risk: when the feature space is large enough, the algorithms may be able to discover some apparently meaningful factor of in-sample historical returns, but the factor might just be an artefact of noise.

This risk is not hypothetical; it can be observed. A plausible explanation for why 65 percent of the analysed anomalies could not be replicated by Hou, Xue, and Zhang [3] is that many of them were statistical artefacts

© 2026 The Author(s). Authors retain copyright in their articles.

created by high-dimensional searching rather than real factors explicable by economic mechanisms. As the discovery process shifts its emphasis from theory-first to data-first, this risk increases systematically.

Data-driven discovery should therefore include two non-negotiable steps: out-of-sample validation and economic-mechanism retrospection. The first examines the strength of the signal on a new data set (repeating the out-of-sample R^2 criterion of Direction II); the second explains the existence of the signal in terms of economic reasoning. Neither can be neglected: out-of-sample validation without mechanistic explanation is a black box that predicts without understanding, while mechanistic explanation without out-of-sample validation produces a story that is coherent but unreliable.

5.3 The Necessity of Economic Retrospective Validation

The threat identified in the previous subsection leads directly to the argument of this one: economics has not been pushed aside by the discovery path; it has moved to a different point in it.

In the conventional paradigm, economics stands at the beginning of the discovery chain: theory first puts forward hypotheses and data subsequently test them. On the route outlined by machine learning, economics brings up the rear: algorithms first extract signals and theory explains them afterwards. This backward shift may appear to undermine the power of theory, but in reality it enhances its importance. The number of signals extracted by algorithms is far larger than the number of hypotheses presented by people; without an economic mechanism to act as a filter, researchers would be overwhelmed by unintelligible statistical signals.

As Giglio, Kelly, and Xiu [23] state clearly, machine learning is not a panacea and does not replace economic theory. The judgment reached here is that machine learning has changed the effectiveness and the scope of what is called discovery, but not the academic baseline of the field, according to which a factor must be explicable through an economic mechanism.

5.4 Section Summary

The claim of Direction IV can be summarised in three points. First, the machine-learning discovery path partially displaces the source of knowledge from prior theory towards data-driven search, and its methodological span runs from LASSO variable selection and IPCA factor extraction to image- and LLM-based semantic mining. Second, the central danger of this turn is the appearance of statistical artefacts posing as economic variables: in a highly dimensional feature space, algorithms may end up fitting false signals that are non-trivially significant in sample. Third, the requirement of economic validity has not been eliminated but has moved into the category of retrospective testing — algorithms widen the range of discovery, theory screens and explains, and the two complement each other rather than competing.

6. Cross-Asset Extension of the Machine-Learning Framework

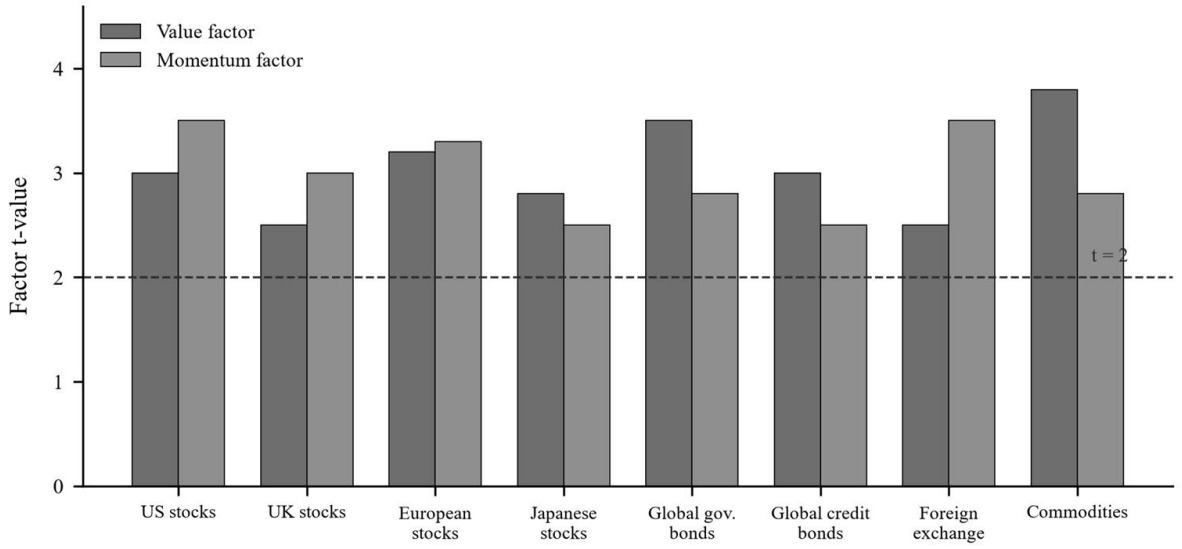
Directions II, III, and IV concern changes at the testing, construction, and discovery stages respectively, but they share one weakness of premise: these changes mostly take place in the equity asset class. Were the machine-learning paradigm to apply only to the U.S. equity cross-section, it would be of local importance. This section argues that if machine learning is to be applied to cross-asset domains, three further issues must be confronted: whether factor isomorphism holds across assets, whether the methodological challenges of joint estimation can be overcome, and how institutional differences limit the transfer of models. These points are emphasised in order to draw a sharp line between the contribution of machine learning and the existing cross-asset literature, so that the findings of Asness et al. [24] on cross-asset factors, those of Giglio and Xiu [27] on omitted factors, and the Chinese CH-3 model [28] are not attributed directly to machine learning. The discussion of the Chinese market serves as an illustration of institutional dependence and transfer boundaries, not as proof that a unified framework has been created.

6.1 Cross-Asset Factor Isomorphism: Value and Momentum Everywhere

The empirical basis of cross-asset factors is the work of Asness, Moskowitz, and Pedersen [24]. The article examines value and momentum factors in eight different markets and asset classes and shows that they are both prevalent and positively correlated across markets and asset classes (see Figure 7, redrawn by the author from Asness et al. 2013).

© 2026 The Author(s). Authors retain copyright in their articles.

Figure 7: The prevalence of value and momentum across eight markets and asset classes



Redrawn by the author from Asness–Moskowitz–Pedersen 2013 data. Note: The two factors are consistently significant and positively correlated across asset classes, which supports a unified framework.

The theoretical implication of this result is that if value and momentum are both present across markets and asset classes, they are not phenomena specific to a particular asset but cross-asset pricing regularities. This provides an empirically credible justification for a unified factor framework in which factors are no longer seen as unique to equities but as common pricing factors across asset classes.

The same isomorphism is confirmed in fixed-income assets. Cochrane and Piazzesi [25] show that a single return-forecasting factor constructed from forward rates predicts bond excess returns; Chordia, Goyal, Nozawa, and Subrahmanyam [26] show that several equity anomalies are also present in corporate bond returns. Taken together, these works suggest that cross-asset factor isomorphism may be a general phenomenon in asset pricing, although the magnitude of a particular factor differs across assets, and citations should be careful about the strength of the claim.

As a caveat on evidence, cross-asset isomorphism has been documented in research covering different asset types, frequencies, and sample periods. These studies agree on the direction of the finding (factors are present across asset classes), but the claim that value holds in commodities and the claim that it holds in equities cannot be treated as resting on the same weight of evidence.

6.2 Unified Estimation under Omitted Factors: The Theoretical Basis of Integration

Cross-asset isomorphism implies that a common framework can be developed, but it addresses only whether such a framework is feasible, not how it should be implemented. The conceptual barrier to a unified framework is that when a researcher tries to estimate a unified stochastic discount factor (SDF), any omitted pricing factor biases the estimate. This is exactly the issue tackled by Giglio and Xiu [27].

As shown by the three-step procedure of Giglio and Xiu [27] in “Asset Pricing with Omitted Factors”, the risk prices of observable factors can be estimated consistently even in the presence of omitted factors. The importance of the work lies in providing a methodological basis for estimating a common SDF: even when the researcher cannot take all factors into account, consistent estimates remain achievable in the presence of omitted ones.

It should be made clear that there is a connection between Directions II and III: the omitted-factor model of Giglio–Xiu, the machine-learning-based out-of-sample performance evaluation of Direction II, and the latent-factor model of IPCA in Direction III offer possible methodologies for the three questions underlying cross-asset unification — what is omitted, how to predict performance, and how to build factors. These are three interrelated steps; they do not imply that machine learning triggered cross-asset unification.

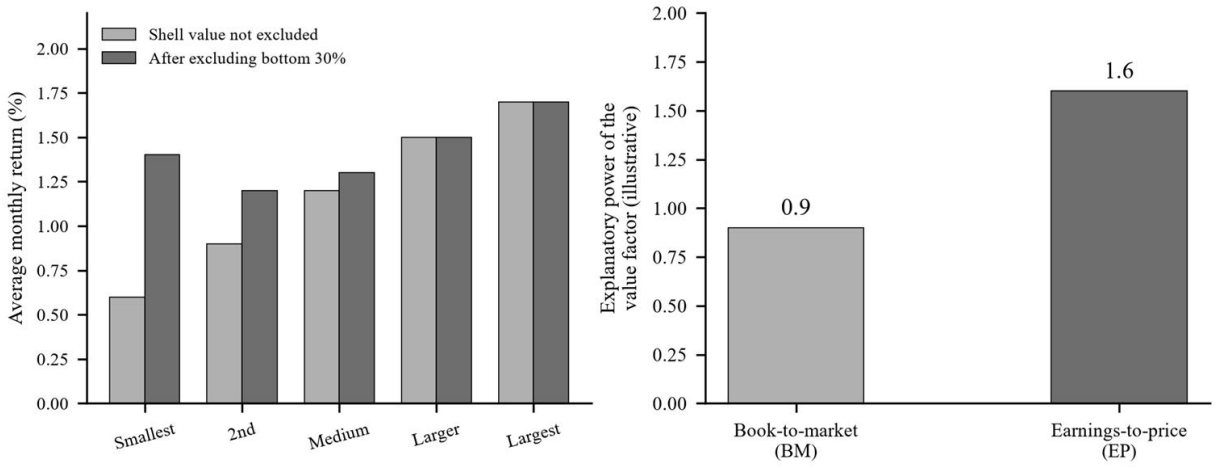
© 2026 The Author(s). Authors retain copyright in their articles.

6.3 The Chinese Market as an Illustration of Institutional Dependence and Transfer Boundaries

The final test of a unified structure is not whether it fits U.S. equities but whether it applies to other markets with very different institutional environments. From this point of view, the Chinese market provides important evidence: institutional differences have non-trivial effects on the construction and the performance of factors.

The Chinese three-factor model (CH-3) proposed by Liu, Stambaugh, and Yuan [28] illustrates the essence of this specificity: the smallest 30 percent of A-share market-capitalisation stocks are contaminated by shell value, which means that their prices are substantially affected by the reverse-merger option value rather than by operating fundamentals. Including this group would therefore create a systematic bias in the size effect. At the same time, book value under Chinese accounting rules is quite noisy, so the earnings-to-price ratio (EP) should be used instead of book-to-market (BM) as the value measure (see the schematic in Figure 8, reproduced by the author from the adjustment scheme of Liu–Stambaugh–Yuan 2019; it is not an empirical regression result).

Figure 8: Chinese-market institutional calibration: CH-3 adjustments relative to FF



Schematic drawn by the author from the adjustment scheme of Liu–Stambaugh–Yuan 2019. Note: The chart is schematic. The key insight is that excluding shell value and using EP instead of BM are essential steps of institutional calibration.

Excluding the lowest 30 percent by market capitalisation and replacing BM with EP are not merely data-manipulation techniques but essential adjustments through which institutional specificity enters the pricing model. It follows that a machine-learning framework cannot be transplanted into the Chinese market directly from the U.S. equity model; institutional calibration must come first.

Empirical work published in Chinese supports this assessment. Li et al. [29] find that the Fama–French five-factor model outperforms the three-factor model in the Chinese equity market, and that the profitability and investment factors become priced after the split-share structure reform; Zhao et al. [30] provide a systematic comparison of the five-factor and three-factor models. All of this work implies that a unified approach can be implemented in the Chinese market only through institutional calibration, and that institutional calibration itself shows cross-market transfer to be a conditional rather than an unconditional generalisation.

This judgment is a contribution of the paper: to our knowledge, English-language reviews rarely examine the Chinese-market evidence, which shows that transferring the framework first requires institutional adjustment. It is at once a test of the unified framework and a delineation of the boundary beyond which it can no longer be claimed to generalise without qualification.

6.4 Section Summary

Direction V can be summarised in three statements. First, cross-asset factor isomorphism provides empirical support for a unified framework but cannot be equated with machine learning driving cross-asset

© 2026 The Author(s). Authors retain copyright in their articles.

unification. Second, methods of unified estimation in the presence of omitted factors provide a theoretical basis for constructing a unified SDF; they relate to machine learning as candidate tools for the same problem domain rather than as extensions of machine learning results. Third, because of its institutional particularity, the Chinese market provides significant information about the limits of transferring the machine-learning framework.

7. Evaluating the Evidence: Are the Gains of Machine Learning Worth Their Costs?

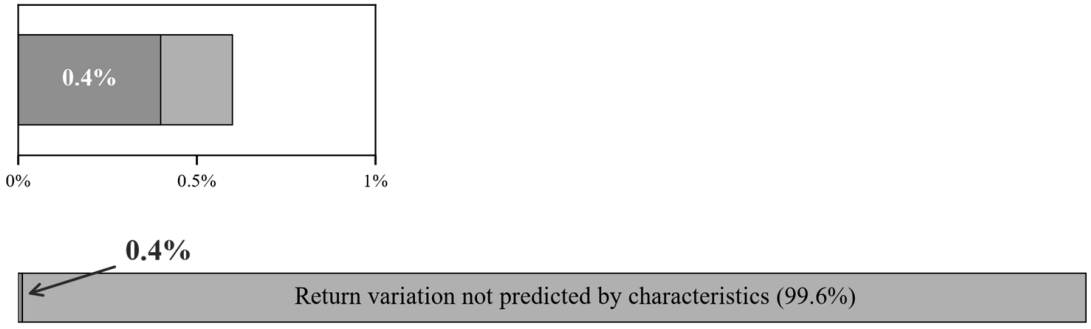
It would be easy to slip into an unexamined, optimistic conclusion from the arguments of Directions I through V: that machine learning is now comprehensively and unconditionally better than the old paradigm. The aim of this section is to stop that slide. The main assertion is that the advantages of machine learning are genuine but restricted, and that it carries losses (interpretability loss, conflation of correlation with causation, and overfitting) that are substantive rather than purely technical; machine learning therefore does not constitute an unconditional endorsement of the approach but a conditional judgment.

7.1 The Limited Nature of the Gains: The Meaning of an Out-of-Sample R^2 of 0.4 Percent

The first point to acknowledge openly is that the total gains of machine learning in predicting cross-sectional returns are limited. Gu, Kelly, and Xiu [12] report an out-of-sample R^2 of about 0.4 percent per month. This number was cited in Direction II to show that an out-of-sample evaluation model can be implemented; here it must be revisited from the opposite perspective: even the most effective machine-learning approaches provide only a relatively low degree of out-of-sample predictive capability. It is important to note that the out-of-sample R^2 of Gu et al. [12] takes the sum of squared un-demeaned excess returns as the denominator and a zero forecast as the benchmark; it does not carry the same meaning as the proportion of variance explained in an ordinary regression and therefore cannot be interpreted directly as explaining 0.4 percent of return variation.

This statement must not be misread either as saying that machine learning is ineffective or as exaggerating its value. It has two interpretations. First, although an out-of-sample R^2 of 0.4 percent is a very small number, the resulting predictive ranking may still generate portfolios with fairly high out-of-sample Sharpe ratios; machine-learning strategies, however, tend to have high turnover and should therefore be evaluated on net returns after transaction costs. Second, the absolute value implies that machine learning remains far from solving the central challenge of asset pricing: most of the cross-sectional variation in returns cannot be predicted from characteristics (see Figure 9, a conceptual schematic).

Figure 9: The "limited nature" of the 0.4% out-of-sample R^2



Conceptual schematic. Note: The 0.4 percent figure is an out-of-sample R^2 calculated relative to a zero-prediction benchmark, using the sum of squared un-demeaned excess returns as the denominator. It does not correspond to the "proportion of variance explained" of standard regression analysis. The bar length in this chart is for visual illustration only and should not be read as "machine learning explains 0.4 percent of return variance". This 0.4 percent is real yet limited: it is enough to form a tradable strategy, but it falls far short of solving the core challenge of asset pricing.

This double meaning is the starting point of Direction VI: if the gain is only 0.4 percent, the claim that machine learning has changed empirical asset pricing must be restricted to the precise statement that it has altered approaches without resolving the underlying problem.

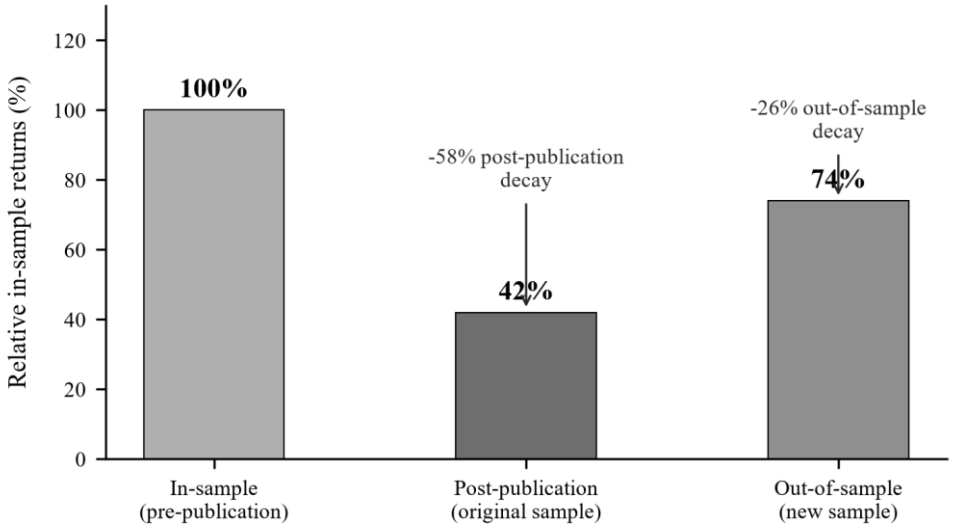
© 2026 The Author(s). Authors retain copyright in their articles.

7.2 Decay Risk: Out-of-Sample Predictive Power Is Not Naturally Immune to Publication Bias

One fact must be accepted: the out-of-sample predictive ability of machine learning is not naturally immune to publication bias and decay.

McLean and Pontiff [7] observe that post-publication returns on factors are 58 percent below in-sample returns. Although that result concerns conventional factors, the same process applies to signals discovered by machine learning: every published strategy or factor has passed through researcher selection, and such selection inevitably promotes the outcomes that worked best in sample (see Figure 10, reproduced by the author from McLean–Pontiff 2016).

Figure 10: Return decay after factor publication



Redrawn by the author from McLean–Pontiff 2016. Note: Post-publication decay is 58 percent and out-of-sample decay is 26 percent; the decay appears systematically after publication, suggesting that the publication process filters for the best statistical artefacts in the sample.

A deceptive myth surrounds machine learning on this point: out-of-sample validation is treated as a vaccine against overfitting. The truth is different. Out-of-sample validation can prevent overfitting to a fixed sample, but it does not protect against overfitting the validation step itself; when a researcher can keep changing the model, the features, and the validation windows until out-of-sample performance looks good, out-of-sample validation becomes a form of in-sample fitting.

The essence of the critique is that out-of-sample R^2 is a harsher condition than in-sample significance; it is not the end of overfitting, however, but a new front line that must be monitored continuously. The efficacy of the out-of-sample evaluation framework set out in Direction II will depend on the rigour of the validation protocol itself, not on any inherent protection of the standard.

7.3 Three Substantive Costs: Interpretability, Causality, and Overfitting

The restrictiveness of the gains and the risk of decay are quantitative concerns; machine learning also presents three qualitative costs that can affect the very principles of asset pricing as a field.

The first cost is interpretability. Section 4.4 already explained that nonlinear construction eliminates the economic story of the "characteristic–mechanism" form, which is the foundation of the discipline. Here it need only be connected with the other two costs: the loss of interpretability is substantive because it makes it harder to detect the conflation of correlation with causation, and overfitting itself.

The second cost is the confusion of correlation with causation. The out-of-sample predictive ability of machine learning answers the question of whether one can predict, not why one can predict. A black-box model with high predictive performance cannot be used to understand the causal process of pricing. The distinction

© 2026 The Author(s). Authors retain copyright in their articles.

is crucial because the ultimate objective of empirical asset pricing is not to forecast returns but to learn how risk is priced, which is a causal rather than a correlational hypothesis.

The third cost is overfitting. A sufficiently large feature space allows data-driven mining to find signals that are significant in sample yet fail to persist out-of-sample. This is not a problem that stricter validation can fix; it is a problem that accompanies the data-driven discovery path (in line with Direction IV).

Kelly, Palhares, and Pruitt [31], in "Modeling Corporate Bond Returns", present a possible response to these three costs: if machine learning is used together with economic constraints — that is, if the algorithm is embedded in an economic structure instead of being allowed to fit freely — predictability and interpretability can be achieved at the same time. The importance of this work is that it turns the contradiction between machine learning and economic theory into a complementarity, providing a methodology that leads to the conclusion of Direction VI.

7.4 Conditional Judgment: In What Sense Does the Contribution of Machine Learning Hold?

Synthesising Directions I through VI, this paper offers a conditional judgment.

Giglio, Kelly, and Xiu [23] state clearly in their survey that machine learning is not a panacea and not a replacement for economic theory. The precise interpretation of that judgment is the theoretical foundation of the conditional judgment offered here.

The conditional judgment can be stated as follows.

Machine learning has produced legitimate instrumental advances in prediction-oriented tasks: it has offered out-of-sample predictive performance as an additional standard for measuring factors, automatic interaction discovery as a way of building factors, and high-dimensional search as a realistic means of discovering factors. All three changes are real and structural.

With respect to causal inference and mechanistic understanding, however, machine learning has not been able to make a substitutive contribution: it can neither explain why risk is priced nor replace the explanatory role of economic theory, and its gains come at the cost of interpretability, causal reasoning, and robustness to overfitting.

It is therefore more accurate to say that machine learning has not transformed the empirical approach to asset pricing, but that it has changed the nature of the prediction-related processes in empirical asset pricing without altering the internal logic of the understanding-related processes. This distinction constitutes the conclusion of the paper. Table 1 summarises the correspondence between the traditional paradigm and the machine-learning paradigm on six task dimensions.

Table 1: Comparison of Tasks between the Traditional Paradigm and the Machine-Learning Paradigm

Task Dimension	Traditional Paradigm	Machine-Learning Paradigm	Substitution?
Validity judgment	In-sample significance (t-stat, Alpha)	Out-of-sample predictive performance (OOS R^2)	No
Factor construction	Manual sorted portfolios	Dimension reduction + automatic interaction discovery	Partial substitution
Factor discovery	Theory-prior hypotheses	High-dimensional signal search	Partial substitution
Pricing-error tests	Fama–MacBeth / GRS	No direct counterpart	No
Mechanism explanation	Economic-theory narrative	Feature importance / SHAP	No
Cross-asset unification	Factor isomorphism + SDF estimation	Can handle high dimensions, but faces transfer boundaries	No

7.5 Section Summary

The Direction VI argument can be summarised in three points. First, the achievements of machine learning are real but constrained: an out-of-sample R^2 of 0.4 percent is at once an indication of genuine effectiveness and an indication of how far the problem remains from being solved. Second, out-of-sample predictive capacity is not inherently immune to publication bias and decay, and out-of-sample validation can itself become a form of overfitting. Third, interpretability, causality, and overfitting are three substantive costs, and Kelly, Palhares,

© 2026 The Author(s). Authors retain copyright in their articles.

and Pruitt [31] show that these costs may be reduced by combining machine learning with economic restrictions.

8. Conclusion

By organising the argument along the four steps of testing–construction–discovery–integration, the paper has shown how machine learning has systematically changed the research tasks of empirical asset pricing and has reached a conditional conclusion.

The shifts in the four stages are real: the testing stage has established an out-of-sample evaluation benchmark (Direction II), the construction stage has opened a path of automatic interaction discovery (Direction III), the discovery stage has introduced data-driven search strategies (Direction IV), and the integration stage has led to the consideration of cross-asset unification and institutional dependence (Direction V). All these changes rest on the fact that the traditional paradigm cannot cope with scale pressure (Direction I): when up to 82 percent of anomalies cannot be reproduced under a strict protocol, the standard of "significant, therefore valid" is called into question.

The change nonetheless has limits (Direction VI). Although machine learning has made genuine instrumental progress on prediction-related problems, it has made no substitutive contribution to causal inference or mechanistic insight. The out-of-sample R^2 of 0.4 percent is an indication both that the gains are real and that the original problem persists; the three costs of interpretability, causality, and overfitting are substantive rather than merely technical.

The practical consequences of this conditional judgment are clear. For researchers, machine learning is useful not as a substitute for economic theory but as a tool that gradually turns theory from a prior hypothesis into a retrospective test: the algorithm widens the set of possible factors, and theory acts as a sieve and as explanation. Practitioners may use out-of-sample predictive performance as an operating standard, but should be careful about overfitting the out-of-sample validation step itself. In the Chinese market, implementing the machine-learning framework requires institutional calibration (filtering out shell value, substituting EP for BM) and cannot be a straightforward transplant of the U.S. equity model.

The precise meaning of the phrase "from factor zoo to factor refinery" is therefore not that algorithms have replaced theory, but that algorithms have transformed how knowledge is produced while theory remains the ultimate judge of its truth or falsity. The refinery refines prediction, not comprehension; and comprehension has always been the essence of asset pricing research.

References

[1] Cochrane, J. H. (2011). Presidential address: Discount rates. The Journal of Finance, 66(4), 1047–1108. https://doi.org/10.1111/j.1540-6261.2011.01671.x.

[2] 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.

[3] Hou, K., Xue, C., & Zhang, L. (2020). Replicating anomalies. The Review of Financial Studies, 33(5), 2019–2133. https://doi.org/10.1093/rfs/hhy131.

[4] Chen, A. Y., & Zimmermann, T. (2022). Open source cross-sectional asset pricing. Critical Finance Review, 11(2), 207–264. https://doi.org/10.1561/104.00000112.

[5] Jensen, T. I., Kelly, B., & Pedersen, L. H. (2023). Is there a replication crisis in finance? The Journal of Finance, 78(5), 2465–2518. https://doi.org/10.1111/jofi.13249.

[6] 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.

[7] McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? The Journal of Finance, 71(1), 5–32. https://doi.org/10.1111/jofi.12365.

© 2026 The Author(s). Authors retain copyright in their articles.

[8] Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. Journal of Political Economy, 81(3), 607–636. https://doi.org/10.1086/260061.

[9] Gibbons, M. R., Ross, S. A., & Shanken, J. (1989). A test of the efficiency of a given portfolio. Econometrica, 57(5), 1121–1152. https://doi.org/10.2307/1913625.

[10] Lewellen, J., Nagel, S., & Shanken, J. (2010). A skeptical appraisal of asset pricing tests. Journal of Financial Economics, 96(2), 175–194. https://doi.org/10.1016/j.jfineco.2009.09.001.

[11] Jegadeesh, N., Noh, J., Pukthuanthong, K., Roll, R., & Wang, J. (2019). Empirical tests of asset pricing models with individual assets: Resolving the errors-in-variables bias in risk premium estimation. Journal of Financial Economics, 133(2), 273–298. https://doi.org/10.1016/j.jfineco.2019.02.010.

[12] 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.

[13] Chinco, A., Clark-Joseph, A. D., & Ye, M. (2019). Sparse signals in the cross-section of returns. The Journal of Finance, 74(1), 449–492. https://doi.org/10.1111/jofi.12733.

[14] 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.

[15] Kelly, B. T., & Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. Journal of Econometrics, 186(2), 294–316. https://doi.org/10.1016/j.jeconom.2015.02.011.

[16] Kelly, B. T., Pruitt, S., & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics, 134(3), 501–524. https://doi.org/10.1016/j.jfineco.2019.05.001.

[17] Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. Journal of Econometrics, 222(1), 429–450. https://doi.org/10.1016/j.jeconom.2020.07.009.

[18] Bryzgalova, S., Pelger, M., & Zhu, J. (2020). Forest through the trees: Building cross-sections of stock returns. Working paper. https://ssrn.com/abstract=3493458.

[19] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[20] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x.

[21] Jiang, J., Kelly, B., & Xiu, D. (2023). (Re-)Imag(in)ing price trends. The Journal of Finance, 78(6), 3193–3249. https://doi.org/10.1111/jofi.13268.

[22] Lopez-Lira, A., & Tang, Y. (2026). Can ChatGPT forecast stock price movements? Return predictability and large language models. Journal of Financial Economics, 184, 104335. https://doi.org/10.1016/j.jfineco.2026.104335.

[23] 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.

[24] Asness, C. S., Moskowitz, T. J., & Pedersen, L. H. (2013). Value and momentum everywhere. The Journal of Finance, 68(3), 929–985. https://doi.org/10.1111/jofi.12021.

[25] Cochrane, J. H., & Piazzesi, M. (2005). Bond risk premia. American Economic Review, 95(1), 138–160. https://doi.org/10.1257/0002828053828581.

[26] Chordia, T., Goyal, A., Nozawa, Y., & Subrahmanyam, A. (2017). Are capital market anomalies common to equity and corporate bond markets? An empirical investigation. Journal of Financial and Quantitative Analysis, 52(4), 1301–1342. https://doi.org/10.1017/S0022109017000515.

[27] Giglio, S., & Xiu, D. (2021). Asset pricing with omitted factors. Journal of Political Economy, 129(7), 1947–1990. https://doi.org/10.1086/714090.

© 2026 The Author(s). Authors retain copyright in their articles.

[28] Liu, J., Stambaugh, R. F., & Yuan, Y. (2019). Size and value in China. Journal of Financial Economics, 134(1), 48–69. https://doi.org/10.1016/j.jfineco.2019.03.008.

[29] Li, Z., Yang, G., Feng, Y., & Jing, L. (2017). An empirical test of the Fama–French five-factor model in the Chinese stock market [in Chinese]. Journal of Financial Research, 444(6), 191–206.

[30] Zhao, S., Yan, H., & Zhang, K. (2016). Does the Fama–French five-factor model outperform the three-factor model? Evidence from the Chinese A-share market [in Chinese]. Nankai Economic Studies, (2), 41–59.

[31] Kelly, B. T., Palhares, D., & Pruitt, S. (2023). Modeling corporate bond returns. The Journal of Finance, 78(4), 1967–2008. https://doi.org/10.1111/jofi.13233.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

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

© 2026 The Author(s). Authors retain copyright in their articles.