The Power of the Common Task Framework

Oliver Hellum, Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen*

Preliminary – Not for Distribution

This version: March 14, 2025

Abstract

The "Common Task Framework" (CTF) is a collaborative and competitive process in which researchers solve a task using shared data, a predefined success metric, and a leaderboard. Using an economic model, we show that the CTF incentivizes effort, increases innovation, and curbs misrepresentation by reducing research costs and improving comparability. Historical examples from computer science underscore its effectiveness. To demonstrate its broader applicability, we propose a CTF for financial economics: a platform open to all researchers designed to identify the pricing kernel and systematically evaluate asset pricing models, from traditional factor-based approaches to modern machine learning techniques.

Keywords: innovation, asset pricing, common task, competition

JEL Codes: B4, C1, G12, O3

*Hellum is at Copenhagen Business School. Jensen is at Yale School of Management; www.tijensen.com. Kelly is at Yale School of Management, AQR Capital Management, and NBER; www.bryankellyacademic.org. Pedersen is at AQR Capital Management, Copenhagen Business School, and CEPR; www.lhpedersen.com. We thank Faheem Almas for outstanding research assistance and Hellum, Jensen, and Pedersen gratefully acknowledge support from the Center for Big Data in Finance (grant no. DNRF167).

1 Introduction: A Secret Sauce in Research?

We provide a theory of how the common task framework (CTF) enhances innovation. As a case in point, show how the CTF can be applied to accelerate finance research.

To understand the potential usefulness of a CTF, consider Feynman's famous attack on "Cargo Cult Science" (Caltech commencement address, 1974).[1] Feynman argues that certain areas of research fail to progress because their results are not repeatable, not comparable, or mispresented. We show how the CTF can alleviate these scientific problems, resulting in more innovation.

One area of research that initially struggled to progress was automatic speech recognition, which therefore saw its funding cut by the Defense Advanced Research Projects Agency (DARPA) in the 1970s. DARPA only reinstated funding in the 1980s when it had developed a method of accountability, essentially inventing the CTF in the process.[2] Subsequently, speech recognition research made landmark improvements, and the CTF is also credited with accelerating the research in computer science and artificial intelligence more broadly (Liberman and Wayne, 2020). Public machine-learning competitions inspired by the CTF include the Netflix Prize, KDD Cup (Bennett and Lanning, 2007), and Kaggle Competitions (Bojer and Meldgaard, 2021). Donoho (2017) calls the CTF "the 'secret sauce' of machine learning."

What is a CTF? No official definition of the CTF exists, so, to fix ideas, we provide a definition that we hope captures the conventional wisdom among computer scientists involved with CTFs.[3] A CTF is a collaborative and competitive research process with the following elements:

(1) Shared training data.

(2) A common task with a predefined quantitative metric to judge success.

[1] A transcribed version of the speech is at calteches.library.caltech.edu/51/2/CargoCult.htm (2/13/2025).

[2] In a speech on "Reproducible Computational Experiments," Mark Liberman credits DARPA manager Charles Wayne with the development. The speech is available (as of Feb. 13, 2025) at www.simonsfoundation.org/event/reproducible-research-and-the-common-task-method/.

[3] Liberman and Wayne (2020) define the CTF as multiple parties sharing resources, competing, and collaborating to achieve a stated objective. Donoho (2017) defines a CTF as having a publicly training dataset, a set of enrolled competitors whose common task is to infer a class prediction rule from the training data, and a scoring referee with a testing dataset, which is sequestered behind a Chinese wall.

(3) A leaderboard that controls for over-fitting, e.g., via a secret test set.

For example, in our finance application, we propose using (1) shared data on returns and return-predictors; (2) a common task of maximizing the Sharpe ratio (SR) of the pricing kernel; and (3) a leaderboard based on out-of-sample performance. While a secret test set may not always exist, other approaches include inviting the winners to present at a conference, such that obvious over-fitting becomes apparent (with the additional benefit that future collaboration is encouraged).

The power of the CTF. It is useful to uncover the secret of the secret sauce. In other words, why would a CTF enhance research? If the reason is not specific to computer science, then can a CTF be used to accelerate research in other fields, including in social science?

We address these questions in an economic model of research innovation. In the model, researchers must exert costly effort, and their resulting innovative output has an element of randomness and is observed with noise. We show that the CTF can promote equilibrium effort and innovation while curbing misrepresentation of findings. These benefits of the CTF arise from two effects.

First, a CTF can promote innovation by lowering the cost of research, for example by proving easily accessible data, sharing code online, and promoting a culture of collaboration (as seen in part 1 of our definition of the CTF). By sharing code and experience, the CTF facilitates that researcher's can more easily build on the work of others, thus making research more cumulative over time.

Second, it can promote innovation by making different research contributions more directly comparable. Comparability is helpful because it leads to more high-powered incentives in equilibrium, thus increasing effort and innovation. Comparability is promoted by having all participants in the CTF use the same data and the same success criterion (as seen in part 2 of our definition of the CTF).

Menkveld et al. (2024) provide evidence of comparability problems in finance. Emphasizing the importance of comparability, Feynman tells of a time when he recommend to a psychology researcher to *"first to repeat in her laboratory the experiment of the other person—to do it under condition X to see if she could also get result A—and then change to Y and see if A changed. Then she would know that the real difference was the thing she thought*

she had under control."

Feynman continues that problems with comparability even affect *"the famous field of physics. I was shocked to hear of an experiment done at the big accelerator at the National Accelerator Laboratory... In order to compare his heavy hydrogen results to what might happen to light hydrogen he had to use data from someone else's experiment on light hydrogen, which was done on different apparatus. When asked he said it was because he couldn't get time on the program (because there's so little time and it's such expensive apparatus) to do the experiment with light hydrogen on this apparatus because there wouldn't be any new result. And so the men in charge of programs at NAL are so anxious for new results, in order to get more money to keep the thing going for public relations purposes, they are destroying—possibly—the value of the experiments themselves, which is the whole purpose of the thing."*

CTF immediately makes different findings apples-to-apples. In other words, with a CTF researchers need not repeat the experiments of others since comparability is instead built directly into the CTF. Therefore, a CTF simultaneously increases comparability and lowers the cost of research.

Another benefit of the CTF is that it curbs the incentive to misrepresent research the findings, for example overselling results or selectively reporting findings. As also emphasized by Feynman, researchers face many temptations, but warned that *"you must not fool yourself ... [or] other scientists."* In the context of early research on automatic speech recognition, John Pierce, who chaired the "Automatic Language Processing Advisory Committee," noted that *"To sell suckers, one uses deceit and offers glamor."*[4]

A CTF makes such distortions more difficult by having a pre-defined success criterion and an objective leaderboard that seeks to control over-fitting as well as presentations that can reveal distortions (as seen in part 3 of our definition of the CTF). We show that, when comparability is increased and the cost of research is lowered, researchers spend more effort on true innovation and less on distortions.

In summary, we show that a CTF can promote innovation and reduce distortions by reducing research costs and promoting comparability. Since these mechanisms are general,

[4]From Pierce's personal letter to JASA in 1969, as cited by Liberman's speech from Footnote 2.

not specific to computer science, the results should apply to any data-driven field of research in which the CTF can lower research costs and promote comparability. As an out-of-sample test of this idea, we propose using the CTF to accelerate research in finance and social science more broadly.

Applying the CTF to finance. Finance researchers have long sought to determine the pricing kernel that prices all assets. While the literature contains many proposed pricing kernels, no consensus exists on which is the leading one (i.e., not even an informal leaderboard exists, as in part 3 of the CTF). This incomparability arises partly because different papers use different time periods, test assets, samples, and screening methods (share classes, industries, size screens), so the data is not shared as in part 1 of the CTF. Further, the metric for success differs across papers (e.g, some papers focus on SR, others on the ability to explain certain asset returns, yet others on the test by Gibbons et al. (1989), some compute performance in-sample and others out-of-sample), counter to part 2 of the CTF.

To address this issue with a CTF, we proceed in two directions. To get the CTF off the ground, we first run the CTF process on a collection of proposed pricing kernels from the literature, just like one of the first CTF-like applications in speech recognition was to test existing methods (Doddington and Schalk, 1981).[5] Second, we propose a CTF open to all researchers based on a new web-based platform with a common data set of equity returns and equity features (signals that may predict returns); an invitation for researchers to participate in the task of finding the pricing kernel via this data, laying out the success criterion as the method that yields the highest out-of-sample SR; and a leaderboard with a pre-specified set of rules.

To run the CTF process on the pricing kernels from the literature, we identify a number of candidate methods and re-estimate each of them using a common dataset. This replication of the literature in a unified setting is a significant undertaking since several of these different methods are complex and computationally intensive. We first compare our implementation of these models with results reported in the original papers, finding largely consistent results, but with generally lower SRs than those reported in the literature, as seen in Figure 1(a). These performance differences likely arise from the fact that our sample and methodology is

[5]See minutes 11–15 in the Liberman speech mentioned in Footnote 2.

consistent across approaches rather then trying to exactly mimic each paper.[6]

We find that the newer machine-learning based methods generally outperform the classic factor models. Interestingly, the performance tends to be better for methods that are published later as seen in Figure 1(b), but performance does not improve monotonically, perhaps partly because the literature has not kept track of this evolution.

We also examine how robust our results are to the choice of features that predict returns, the country of interest, and the time periods. We find that the relative ranking is relatively robust across specifications, but the magnitude of the highest Sharpe ratio differs somewhat across specifications, notably showing a decline over time.

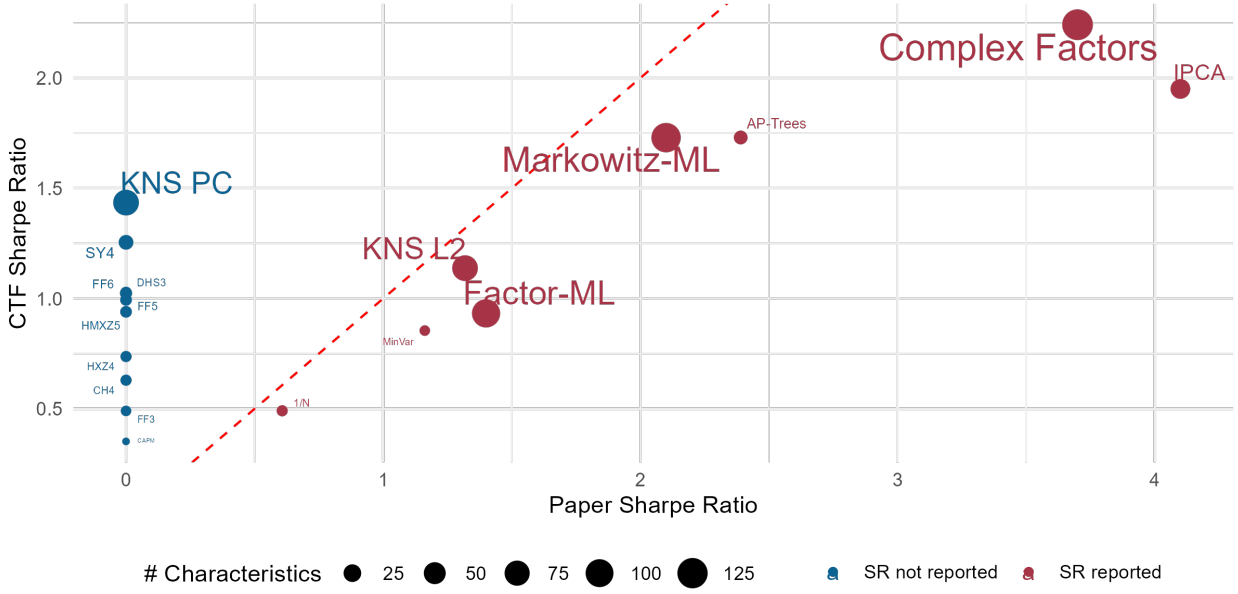
An additional benefit of the CTF is that we can combine the various models from the literature given that we have all models implemented jointly based on the same data. Combining models to get better predictions is called "ensemble methods" in machine learning. We find that ensemble models perform better than any individual models, as seen in the final point in Figure 1(b). This strong out-of-sample performance of the ensemble model shows both the power of the CTF and interesting fundamental differences across the candidate models.

Literature. Our paper is related to several literatures. First, CTFs celebrated among machine learning insiders, but to our knowledge little is written about why CTF work. Our model is related to the so-called informativeness principle (Holmström, 1979; Grossman and Hart, 1983) and the literature on competitions (Nalebuff and Stiglitz, 1983). We complement the literature by applying these principles to study CTFs and adapting the model to study the impact of correlations and potential distortions.

Second, our paper is related to the literature on the replication crises (or credibility crises) facing certain fields, including medicine (Ioannidis, 2005), psychology (Nosek et al., 2012), management (Bettis, 2012), experimental economics (Camerer et al., 2016; Maniadis et al., 2017), and some parts of finance such as corporate bonds (Dick-Nielsen et al., 2023) and green finance (Eskildsen et al., 2024), but not the literature on individual equity factors (Chen and

[6]The approach of using a consistent method is a form of "scientific replication" (as opposed to "pure replication" or "reproduction," which seeks to which to check the work of others with the same method and data) using the terminology of Hamermesh (2007). Scientific replication is also used in other finance applications in order to create a common dataset, e.g., Jensen et al. (2023); Dick-Nielsen et al. (2023); Eskildsen et al. (2024).

(a) Reported Sharpe Ratios vs. Our Replicated CTF Sharpe Ratios



(b) Replicated CTF Sharpe Ratios vs. Publication Year

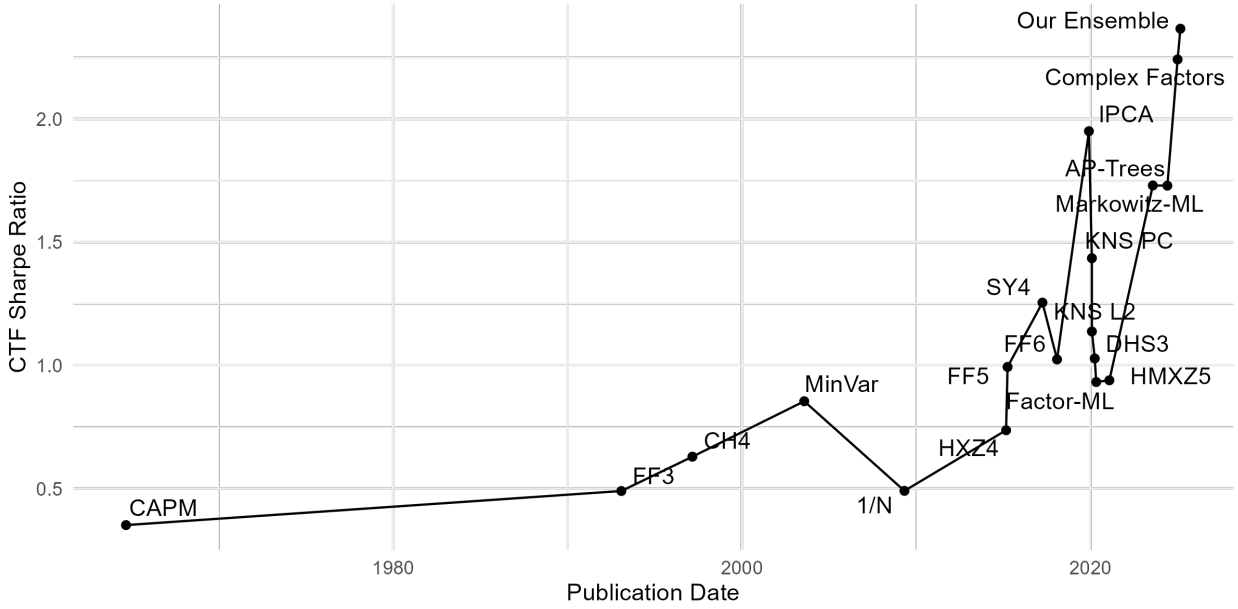


Figure 1: **Sharpe Ratios of Different Candidate Pricing Kernels.** Panel (a) shows the annualized SRs computed using our common methodology versus the SRs from the original papers (if reported). The dot sizes indicate the number of characteristics used in each study. The dashed line is the 45-degree line. Panel (b) shows the replicated SRs for each candidate pricing kernel as a function of the publication date of the corresponding paper. All results are out of sample with a test sample period of 1990–2023 using a 10-year rolling estimation window.

Zimmermann, 2022; Jensen et al., 2023). We present a novel and unified replication of the literature on the pricing kernel in the context of the CTF.

Third, our application to finance is obviously related to the asset pricing literature seeking to uncover the pricing kernel. We discuss all the relevant papers in detail in our methodology section.

Fourth, related papers from computer science include Donoho (2017), Liberman and Wayne (2020), and Hofman et al. (2021).

In summary, we provide an economic model of the power of the CTF, present a CTF for the existing finance research on the pricing kernel, yielding new insights on asset pricing models, and propose a CTF for new asset pricing models open to all researchers.

2 A Theory of How a CTF Can Enhance Discovery

Model. We consider a set of researchers, indexed by $i = 1, ..., I$. Each researcher can innovate by exerting an effort of $e_i \in [0, \infty)$. Depending on the effort, the researcher creates an innovation of value

$$\tilde{v}_i = e_i + \xi_i \,, \tag{1}$$

where ξ_i is a random variable with mean zero, which captures that success in research has an element of randomness. Further, $\text{Var}(\xi_i) = \sigma_\xi^2$ for all i and $\text{Cor}(\xi_i, \xi_j) = \rho_\xi \in [0, 1]$ for $i \neq j$. The effort level, e_i, is not observable to others, and neither is the true value, $\tilde{v}_i$, of each research finding. What is observed, however, is a noisy measurement of the research innovation:

$$v_i = \tilde{v}_i + \eta_i \,, \tag{2}$$

Here, η_i is a random variable with mean zero, which captures the noise in the observed research contribution. Its variance is $\text{Var}(\eta_i) = \sigma_\eta^2$ for all i, and $\text{Cor}(\eta_i, \eta_j) = \rho_\eta \in [0, 1]$ for $i \neq j$.

Each researcher receives a wage w_i, which depends on the researcher's measured inno-

vation, v_i, and the measured innovations of other researchers. The wage can be seen more broadly as a summary of the rewards from journal acceptances, tenure decisions, promotions, and other opportunities arising from academic success. Researchers enjoy a utility of income, u, and maximize their expected utility net of their disutility of effort, e_i. Researchers can also choose to leave this area of research and pursue other areas, which provide an outside option of u^*. Hence, a researcher chooses effort (e_i) and whether or not to leave this field of research (l_i, where $l_i = 1$ means leaving) as follows, taking the wage scheme as given:

$$\max_{e_i \in [0,\infty), l_i \in \{0,1\}} \left(E[u(w_i)] - \frac{\kappa}{2} e_i^2 \right) 1_{\{l_i=0\}} + u^* 1_{\{l_i=1\}} \tag{3}$$

We assume for simplicity that researchers have mean-variance utility, $E(u(w_i)) = E(w_i) - \frac{\gamma}{2} Var(w_i)$, with a risk aversion of γ, but our results hold more generally.

As a simple model of the academic reward system, a principal sets wages as

$$w_i = c + \lambda v_i - \theta v_{-i}, \tag{4}$$

consisting of a baseline wage of c, an incentive pay based on measured output with steepness λ, and considers output relative to that of other researchers' average innovation, $v_{-i} = \frac{1}{I-1} \sum_{j \neq i} v_j$, based on the parameter θ. The principal's utility is the expected total innovation net of wages, $E(\sum_i (v_i - w_i))$, as all innovations are of separate value.

An equilibrium is a set of parameters $c, \lambda, \theta \in \mathbb{R}$ such that the principal maximizes utility, taking into account that researchers' choose their optimal effort e_i and participation l_i given this wage scheme:

$$\max_{c,\lambda,\theta \in \mathbb{R}} E[\sum_i (v_i - w_i) 1_{\{l_i=0\}}] \tag{5}$$

$$\text{s.t. } e_i, l_i \text{ solve the researcher's problem given the wage } w_i \text{ in (4) with } c, \lambda, \theta$$

Due to the participation constraint, maximizing the principal's utility is the same as maximizing the social welfare, defined as the sum of the utilities of all researchers and the

principal.[7]

To solve the model, it is helpful to combine equations (1) and (2) such that the observed research innovation is written as

$$v_i = e_i + \sqrt{\rho}\,\epsilon + \sqrt{1-\rho}\,\varepsilon_i\,, \tag{6}$$

where ϵ and ε_i are independent random variables with variance $\sigma^2 = \sigma_\xi^2 + \sigma_\eta^2$, and $\rho = \rho_\xi \frac{\sigma_\xi^2}{\sigma_\xi^2+\sigma_\eta^2} + \rho_\eta \frac{\sigma_\eta^2}{\sigma_\xi^2+\sigma_\eta^2}$ is the correlation of the shocks across researchers.[8]

Definition of CTF. A CTF has two key properties as explained in the introduction: it is collaborative and competitive. The collaborative nature of a CTF means that it lowers the cost, κ, of doing research. This cost lowering is facilitated via sharing of data, code, and experience as well as conferences and discussions about the common task. The competitive nature of the CTF makes innovations more comparable, which means that ρ_η (and, hence, ρ) increases. For example, if researchers each propose a pricing kernel in finance, evaluating their relative merits becomes easier when they are tested on a common dataset with a common success metric, as common noise interferes less with comparisons. In summary, we define a CTF as an intervention that decreases κ or increases ρ or both.

Results. We first establish the equilibrium and then turn to implications of the CTF. The equilibrium is straightforward to derive. For each wage scheme (4), the researcher derives the optimal effort, e_i, under participation using (3). Since the expected wage increases linearly in e_i and the wage variance depends on e_i squared, the optimal effort, e_i^*, is the maximum of a quadratic. If the corresponding optimal utility is lower than the outside option, u^*, the researcher decides to leave the field ($l_i^* = 0$).

The principal sets the wage scheme (4) as follows. For each λ, θ, the principal chooses $c^*(\lambda, \theta)$ such that researchers' participation constraint is satisfied with equality. The principal then computes its expected utility with all researchers staying in the field, yielding

[7] In equilibrium each researcher receives the reservation utility, u^*, so the social welfare is $E[\sum_i(\tilde{v}_i - w_i)] + \sum_i u^*$, which differs from the principal's utility by the constant $\sum_i u^* = Iu^*$.

[8] The proof of Proposition 1 in appendix contains the derivation of (6). It is easily seen that $\rho = \mathrm{Cor}(\sqrt{\rho}\,\epsilon + \sqrt{1-\rho}\,\varepsilon_i, \sqrt{\rho}\,\epsilon + \sqrt{1-\rho}\,\varepsilon_j)$ for $i \neq j$.

a quadratic equation in λ, θ with a unique solution. Finally, the principal offers this wage scheme $(c^*, \lambda^*, \theta^*)$ if the resulting principal utility is positive, otherwise offering a zero wage. The following result provides an explicit solution.

Proposition 1. *A unique equilibrium exists in which the principal chooses a wage scheme* $(\lambda^*, \theta^*, c^*)$ *and all researchers choose their effort* e_i^* *and participation* l_i^**. The equilibrium wage scheme is*

$$\lambda^* = \frac{1}{1 + \kappa\gamma\sigma^2(1-\rho) + \kappa\gamma\sigma^2\rho\left(1 - \frac{\rho}{\rho + \frac{1-\rho}{I-1}}\right)} \tag{7}$$

$$\theta^* = \frac{1}{1 + \frac{1-\rho}{\rho(I-1)}}\lambda^* \tag{8}$$

$$c^* = u^* + \frac{\lambda^*}{\kappa}\left(\frac{1}{2} - (\lambda^* - \theta^*)\right) \tag{9}$$

and researchers stay in the field ($l_i^* = 0$) *with equilibrium effort* $e_i^* = \lambda^*/\kappa$ *provided that* $\frac{\lambda^*}{2\kappa} \geq u^*$*. Otherwise, the equilibrium wage scheme is* $\lambda^* = \theta^* = c^* = 0$ *and all researchers leave the dormant field* ($l_i^* = 1$)*.*

The proposition shows that there are two types of equilibria. In the first type, the research area is funded with incentives given by (7)–(9) and researchers are actively exerting an effort to contribute. In the other type of equilibrium, the area is unfunded and researchers use their talents elsewhere.

Having determined the equilibrium, we turn to our central research question, namely how the research process in equilibrium is affected by the CTF.

Proposition 2 (CTF increases innovation)**.** *Introducing a CTF in an active research area leads to stronger research incentives (higher* λ*), more competition (higher* θ*), higher expected innovation* $E(\tilde{v}_i)$*, and higher social welfare in equilibrium.*

The CTF enhances innovation for several intuitive reasons. First, given a wage scheme, the CTF makes research effort more attractive simply by lowering the cost of research. Second, by making research more comparable, the equilibrium wage scheme becomes provides stronger incentives, leading to an additional boost to effort and innovation.

Interestingly, the optimal wage scheme has the property that the wage of researcher i decreases in the innovation of other researchers (as seen from equation (4) since $\theta^* > 0$). This property of the wage scheme has a flavor of a competition, and a CTF makes the equilibrium more competitive as seen in Proposition 2.

To understand this effect, note that the principal wants to reward each researcher for their effort, but cannot observe effort directly. The principal can observe the average observed innovation, $\bar{v}$, where we use the notation that a bar means taking the cross-sectional average, $\bar{v} = \frac{1}{I}\sum_i v_i$. This average innovation is $\bar{v} = \bar{e} + \sqrt{\rho}\,\epsilon + \sqrt{1-\rho}\,\bar{\varepsilon}$ using (6), and, with a large number of researchers I, the average idiosyncratic noise is negligible, $\bar{\varepsilon} \cong 0$, based on the law of the large numbers, and the average effort, $\bar{e}$, is essentially known in equilibrium. Therefore, the principal can essentially learn the common shock ϵ from the average innovation. Hence, each researcher's innovation minus the average innovation essentially removes the common shock, ϵ, such that the principal can make a more precise estimate of the researcher's true effort and innovation.[9]

Further, a CTF that increases ρ means that the principal can better infer each researcher's effort using (6). Indeed, a higher ρ means that idiosyncratic noise ($\sqrt{1-\rho}\varepsilon_i$) is less important and, since the common shock ($\sqrt{\rho}\epsilon$) can essentially be inferred from other researchers' average innovations, the effort (e_i) can estimated more precisely, leading to stronger incentives in equilibrium (i.e., higher λ), thus more effort and innovation.

As discussed in the introduction, funding for research in automatic speech recognition was cut to nearly zero for almost 20 years, consistent with the second type of equilibrium in Proposition 1. However, funding by DARPA restarted when the CTF was developed, which can be understood in light of the model.

Proposition 3 (CTF opens a research area)**.** *Introducing a CTF in an otherwise dormant research area starts research in the area if the CTF sufficiently decreases κ or increases ρ_η.*

The CTF opens the research area for several reasons. First, by lowering the cost of research, researchers require less funding and simultaneously become more productive. Second, by making research findings easier to evaluate, the incentives become stronger, leading

[9] With a large number of researchers ($I \to \infty$), the wage depends on a researcher's outperformance ($\theta^* \nearrow \lambda^*$) and, as idiosyncratic risk is completely diversified away, incentives become stronger (λ^* increases).

to more innovation, which makes funding more worthwhile.

Extended model with research distortion. In addition to making innovations more comparable and reducing the cost of research, a CTF can also affect researchers' ability to oversell their results or otherwise distort their findings. Specifically, suppose that a researcher can affect the idiosyncratic noise in their measured innovation, that is, ε_i in equation (6). Rather than having a zero mean, the researcher can selectively report results such that $E(\varepsilon_i) = d_i \geq 0$, thus distorting the perceived innovation to be higher than the actual innovation. The researcher incurs a cost of distortion of $\frac{z}{2}d_i^2$, where $z > 0$ captures the disutility of distorting results, the risk of embarrassment, and the required extra work. Hence, the researcher's objective (3) is replaced by

$$\max_{e_i, d_i \in [0,\infty), l_i \in \{0,1\}} \left(E[u(w_i)] - \frac{\kappa}{2}e_i^2 - \frac{z}{2}d_i^2 \right) 1_{\{l_i=0\}} + u^* 1_{\{l_i=1\}} \tag{10}$$

where w_i still depends on the observed innovation $\tilde{v}_i$ using (4), and the wage scheme is chosen to the maximize the the principal's problem (5).

In this generalized model, the CTF continues to provide the benefits presented in Proposition 2 and, in addition, the CTF lowers distortions, as shown next.

Proposition 4 (CTF reduces research distortion)**.** *A CTF leads to more expected innovation, higher welfare, and less research distortion as a fraction of effort (lower d_i/e_i) in equilibrium.*

In fact, having a CTF is all the more important in the presence of potential distortions, as we show next.

Proposition 5. *The elasticity of innovation with respect to comparability, $\frac{\partial E(\tilde{v}_i)}{\partial \rho_\eta}\frac{\rho_\eta}{E(\tilde{v}_i)}$, is higher in the presence of distortion than in the model of Proposition 1 without distortion. Further, $\frac{\partial E(\tilde{v}_i)}{\partial \rho_\eta}\frac{\rho_\eta}{E(\tilde{v}_i)}$ is decreasing in z. The same is true for the elasticity with respect to effort cost, $-\frac{\partial E(\tilde{v}_i)}{\partial \kappa}\frac{\kappa}{E(\tilde{v}_i)}$.*

Propositions 4 and 5 show that a CTF mitigates potential distortions. Further, a CTF might even directly raise the cost, z, of distorting research, for example by making it more

difficult to hide such distortions. This potential effect of the CTF would further increase true innovation and welfare.

Proposition 6 (Innovation increases in the cost of distortion)**.** *Equilibrium innovation and social welfare increase in the cost, z, of distorting research. As $z \to \infty$, innovation and welfare converge to those of Proposition 1 without distortion.*

In summary, the CTF promotes innovation, increases welfare, and curbs distortions by making research more comparable and lowering the cost of research. In the context of our finance application, these properties of the CTF correspond to making the different SDFs more comparable and based on easily available high-quality data, as we discuss next.

3 A CTF for Existing Finance Models: Preliminaries

3.1 Data

Characteristics and Returns: The Jensen et al. (2023) Data

We use the comprehensive set of global equity characteristics and returns from Jensen et al. (2023), available via WRDS.[10] The paper focuses on 153 equity characteristics chosen because they had been shown to predict stock returns in the literature or contained in other replication studies. We refer to these 153 characteristics as the "original JKP characteristics." One issue with using these characteristics is look-ahead bias—they were chosen because they work well for predicting returns. To alleviate this issue, also consider the "full" set 402 JKP characteristics, which contains the 159 original characteristics as well as 249 "expanded-ex-original" JKP characteristics. The latter set of characteristics are generated by JKP as reasonable alternatives to those already in the literature in terms of their price or accounting content, but as yet not tested in terms of their predictive power.

[10]The data can be downloaded directly from WRDS at wrds-www.wharton.upenn.edu/pages/get-data/contributed-data-forms/global-factor-data/, the code to build the data set is available from GitHub at github.com/bkelly-lab/ReplicationCrisis/tree/master/GlobalFactors, and extensive documentation of the full JKP characteristics can be found at jkpfactors.s3.amazonaws.com/documents/Documentation.pdf.

Sample

The data set is restricted to large stocks listed in the U.S. from 1952/01 to 2023/12. Specifically, we only include mega, large, and small cap stocks, as defined by Jensen et al. (2023), which means that we exclude stocks with a market capitalization below the 20th percentile of NYSE stocks (sometimes called microcaps). We exclude these stocks to ensure that our test sample focuses on large and liquid stocks, whose prices are not artificially stale and are not overly affected by microstructure issues, such as the bid-ask bounce. Furthermore, we only include ordinary common stocks (i.e., stocks with a CRSP share code of 10, 11, or 12) and stocks listed on either NYSE, Nasdaq, or AMEX (CRSP exchange code of 1, 2, or 3).[11] Finally, we require that stocks have non-missing values for at least 75 of the original JKP characteristics. The test period is 1990/01 to 2023/12.

3.2 Empirical Methodology

Standardization and Imputation

We standardize each characteristic within a month by first computing the cross-sectional rank among stocks with non-missing values of the characteristic, and then transforming the ranks to lie between -0.5 and 0.5. For stock with non-missing values, we impute the median value of zero. The standardization and imputation methodology has not been optimized. It is possible that using a more sophisticated imputation methodology, such as those of Beckmeyer and Wiedemann (2022), Bryzgalova et al. (2022), or Freyberger et al. (2024), could improve the results, or perhaps simple imputation methods are close to optimal, as suggested by Chen and McCoy (2024).

Hyperparameter Tuning

Several of the models we consider require the selection of hyperparameters, and we use the same hyperparameter tuning strategy for all of them. Each month, we use the previous ten years as training data, and we use a standard version of five-fold cross-validation to

[11] Using the variable names from our data set, the restrictions amounts to `exch_main=1`, `common=1` and `source_crsp=1`

pick hyperparameters. Specifically, we split the training data into five folds that maintain the temporal order of the data, such that the first fold contains the first two years and the last fold contains the last two years. The folds then take turns being assigned as the validation data, while the remaining four folds are assigned as training data. For example, when the first fold is the validation data, the method is estimated on the remaining four folds for each set of hyperparameters, and the estimated model is then evaluated (using the method-specific performance metric) on the first fold. We repeat this procedure five times, average the performance for each set of hyperparameters, and choose the set with the best performance. We then re-estimate the model with the chosen hyperparameters on the full training data. We repeat this process every month. Our methodology for hyperparameter tuning has not been optimized and could most likely be improved.

Performance Metric

We search for the portfolio with the highest ex-ante Sharpe ratio, which is often called the tangency portfolio. The tangency portfolio has a number of important properties for asset pricing: (i) is can be used together with the risk-free asset to create the mean-variance efficient frontier (Markowitz, 1952); (ii) is can be used to create a stochastic discount factor (SDF) of the form $m = a + bR_{\text{tpf}}$, were a and b are constants and R_{tpf} is the return of the tangency portfolio (Cochrane, 2005); and (iii) it is the SDF with the lowest possible variance (Hansen and Jagannathan, 1991).

3.3 Candidate Methods

Table 1 shows an overview of the 18 candidate methods we consider in our empirical tests. The 18 methods range from traditional low-dimensional factor models, such as the Fama and French (1993) three-factor model, to modern methods based on machine learning, such as the AP-Trees method from Bryzgalova et al. (2023). The table also shows how the original paper filtered their data set, how they selected hyperparameters, what sample they used for the test set, etc. The table highlights that the methodology in different papers reflects a wide range of small choices, which makes it difficult to compare results across papers. In contrast,

we use the same data filters, hyperparameter tuning scheme, and test set for all methods. Except for these choices, we try to stay as close as possible to the portfolio construction from the original paper. We describe the exact construction of each candidate method in Appendix B.

4 A CTF for Existing Finance Models: Results

4.1 Out-of-Sample Performance

Figure 2 shows the Sharpe ratio for each of the candidate methods on the test set. This figure is our paper's version of a leaderboard for the finance CTF (the competition described in Section 6 will be based on a similar figure that also includes the performance of new methods submitted by competition participants). The complex factor models is the best performing method overall, followed by IPCA, while AP-trees and Markowitz-ML are effectively tied for the third place. The worst performing method is the CAPM—that is, a passive investment in the market portfolio—followed by FF3 and 1/N.[12]

One key take-away from the figure is that traditional low-dimensional factor models perform worse than more modern machine learning approaches. For example, SY4, the best performing low-dimensional factor model, has a Sharpe ratio that is only around half of the complex factor models. The results suggest that the most promising way to find the highest Sharpe ratio portfolio is by using a large number of characteristics and modern machine learning techniques.[13]

But not all machine learning methods perform well. For example, Factor-ML, which uses machine learning to predict returns and then creates a simple long-short portfolio sort based on these predictions, is one of the worst performing methods. The bad performance likely comes from the naive allocation of stocks into long and short portfolios that is designed

[12]Appendix C.1 contains statistical tests related to the results in Figure 2. Table C.1 shows that only Complex Factors, IPCA, AP-Trees and Markowitz-ML receive a statistically significant weight in the in-sample mean-variance optimal portfolio. Table C.2 shows that this conclusion holds when we impose a no-shorting constraint. Finally, Figure C.1 shows pairwise tests of whether one method has alpha relative to another.

[13]Another perspective is that most of the performance can be captured by using a small number of characteristics. For example, Complex Factor use information from 153 characteristics, while SY4 only use information from 13.

Table 1: Candidate methods

Name	Paper	Share codes	Exch.	Excl. Ind.	Size screens	#Chars	Full sample	Test sample	Test SR	Weights	HP tuning	Re-fitting
CAPM	Sharpe (1964)					1						
FF3	Fama and French (1993)	10, 11	3	Financials	None	3	1962-1991	None		FF	No hyperparameters	None
CH4	Carhart (1997)					4	1963-1993	None			No hyperparameters	None
MinVar	Jaganathan and Ma (2003)	10, 11	2	None	Megacaps	3	1963-1998	1968-1998	1.2	MinVar	No hyperparameters	Yearly
1/N	Demiguel et al (2009)	10,11	3	None	None	4	1963-2004	1973-2004	0.6	EW	No hyperparameters	Monthly
HXZ4	Hou et al (2015)	?	3	Financials	None	4	1972-2012	None		FF*	No hyperparameters	None
FF5	Fama and French (2015)	10, 11	3	None	None	5	1963-2013	None		FF	No hyperparameters	None
SY4	Stambaugh and Yuan (2017)	10, 11	3	None	Price>$5	13	1967-2013	None		FF*	No hyperparameters	None
FF6	Fama and French (2018)	10,11	3	None	None	6	1963-2016	None		FF	No hyperparameters	None
IPCA	Kelly, Pruitt, and Su (2019)	10, 11	3	None	None	36	1962-2014	1972-2014	4.1	MeanVar	Expanding training	Monthly
KNS PC	Kozak et al (2020)	10, 11	All	None	Varying size cutoff	80	1964-2017	2005-2017		MeanVar*	3-fold cross-validation	Once
KNS L2	Kozak et al (2020)	10, 11	All	None	Varying size cutoff	80	1964-2017	2005-2017	1.3	MeanVar*	3-fold cross-validation	Once
DHS3	Daniel et al (2020)	10, 11	3	Financials	None	4	1972-2014	None		FF	No hyperparameters	None
Factor-ML	Gu et al (2020)	All	3	None	None	103	1957-2016	1987-2016	1.4	VW	Expanding training	Yearly
HMXZ5	Hou et al (2021)	10,11	?	Financials	None	5	1967-2018	None		FF*	No hyperparameters	None
AP-Trees	Bryzgalova et al (2023)	?	?	?	?	10	1964-2016	1994-2016	2.4	MeanVar*	Fixed train/validation	Once
Markowitz-ML	Jensen et al (2024)	10, 11, 12	1	None	Mega and largecaps	115	1952-2020	1981-2020	2.1	MeanVar	Expanding training	Yearly
Complex Factors	Didisheim et al (2025)	10, 11, 12	3	None	No nanocaps	130	1963-2023	1993-2023	3.7	MeanVar*	Rolling train/val	Monthly

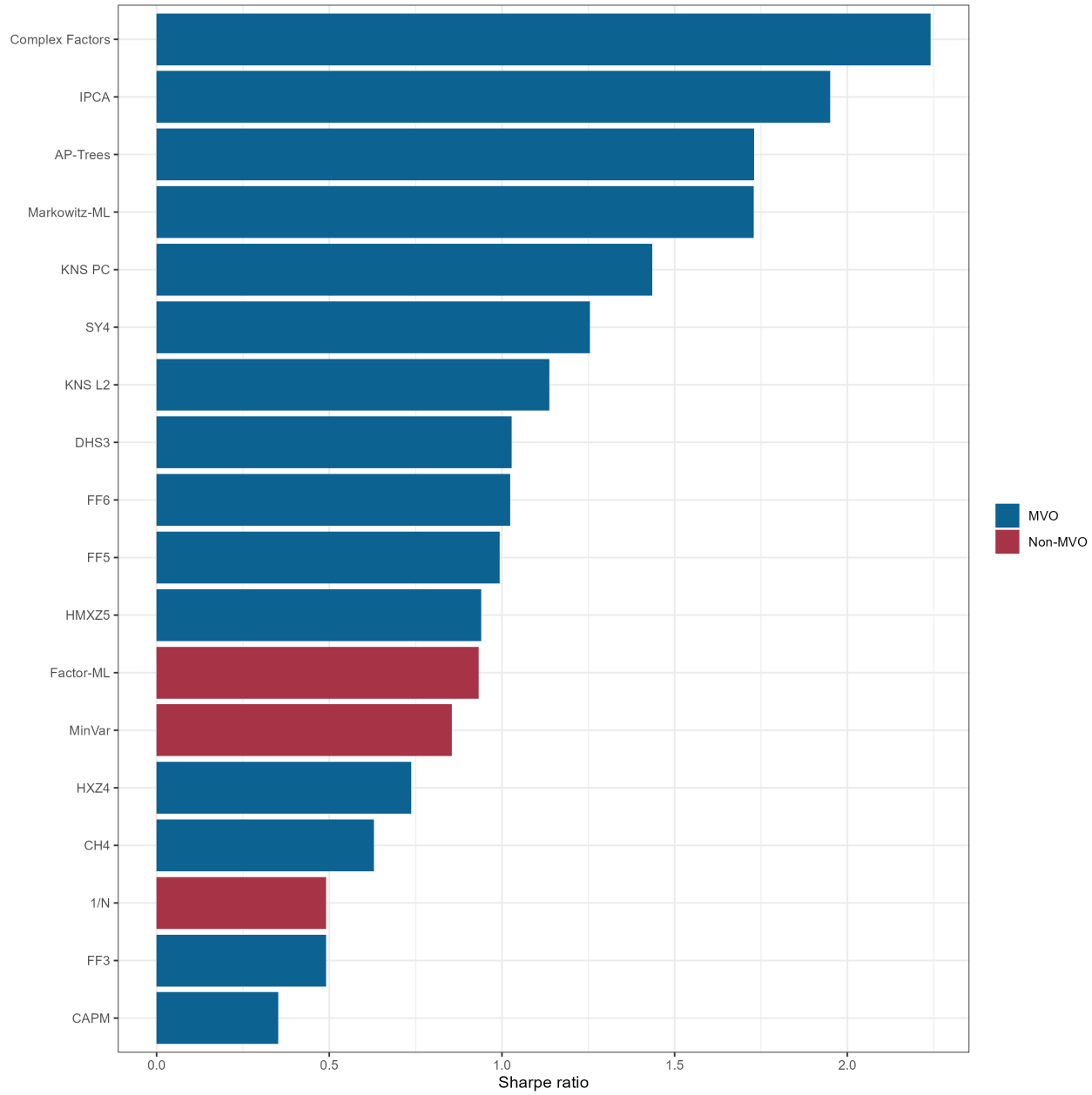


Figure 2: **Sharpe Ratio of Mean-Variance Optimal Portfolio Methods.** Portfolio methods are implemented on JKP data and their annualized Sharpe ratios are reported. Results are for mega-, large-, and small-cap US stocks. The test data period is 1990-2023. All methods use a 10-year rolling window for estimation.

to maximize the portfolio mean return but does not attempt to minimize the portfolio variance. More generally, we find that mean-variance optimized (MVO) methods generally outperform non-MVO methods. This result may seem obvious, but there are many papers arguing that non-MVO methods outperform MVO methods out-of-sample (e.g, Michaud, 1989 and DeMiguel et al., 2009).

4.2 Robustness: Other Characteristics, Countries, and Dates

A central component of the common task framework is to evaluate different methods on a secret test data. Keeping the test data secret is crucial to avoid methods that overfit the data. In our application of the CTF to finance, we have used CRSP and Compustat data from 1990 to 2023 as our test set, but this data is hardly secret. In fact, all the methods we have considered have already been evaluated on similar data, which has plausibly led to some overfitting. In this section, we consider various changes to the test set, in order to see how robust our results are to changes in the training and test data.

Figure 3 shows how well the models perform when we train the models on different stock characteristics. Specifically, the first, second, and third panel show the Sharpe ratio on the test set for models trained with, respectively, the 153 original, 249 expanded-ex-original, and 402 expanded JKP characteristics.[14] . The ranking of the models is quite similar across the three sets of characteristics. For example, the complex factor model consistently has the best performance and the Factor-ML and minimum-variance methods consistently has the worst. Looking at the magnitudes of the Sharpe ratios, we find that most models perform best with the original, second best with the expanded, and worst with the expanded-ex-original characteristics. For example, the complex factor model has a Sharpe ratio of 2.2, 1.9, and 1.5, respectively. The fact that most models perform better on the 153 original characteristics relative to the 402 expandend characteristics (of which all the original characteristics are included) is noteworthy. It is consistent with the view that training a model on the original characteristics—which were selected because they have shown to predict returns—leads to

[14]We exclude models from the non-original figures if they use a prespecified set of characteristics. For example, the Bryzgalova et al. (2023) model prespecifies 10 characteristics, which are all part of the original set of characteristics, so we do not include it in the panel where models are trained using the expanded or the expanded-ex-original characteristics.

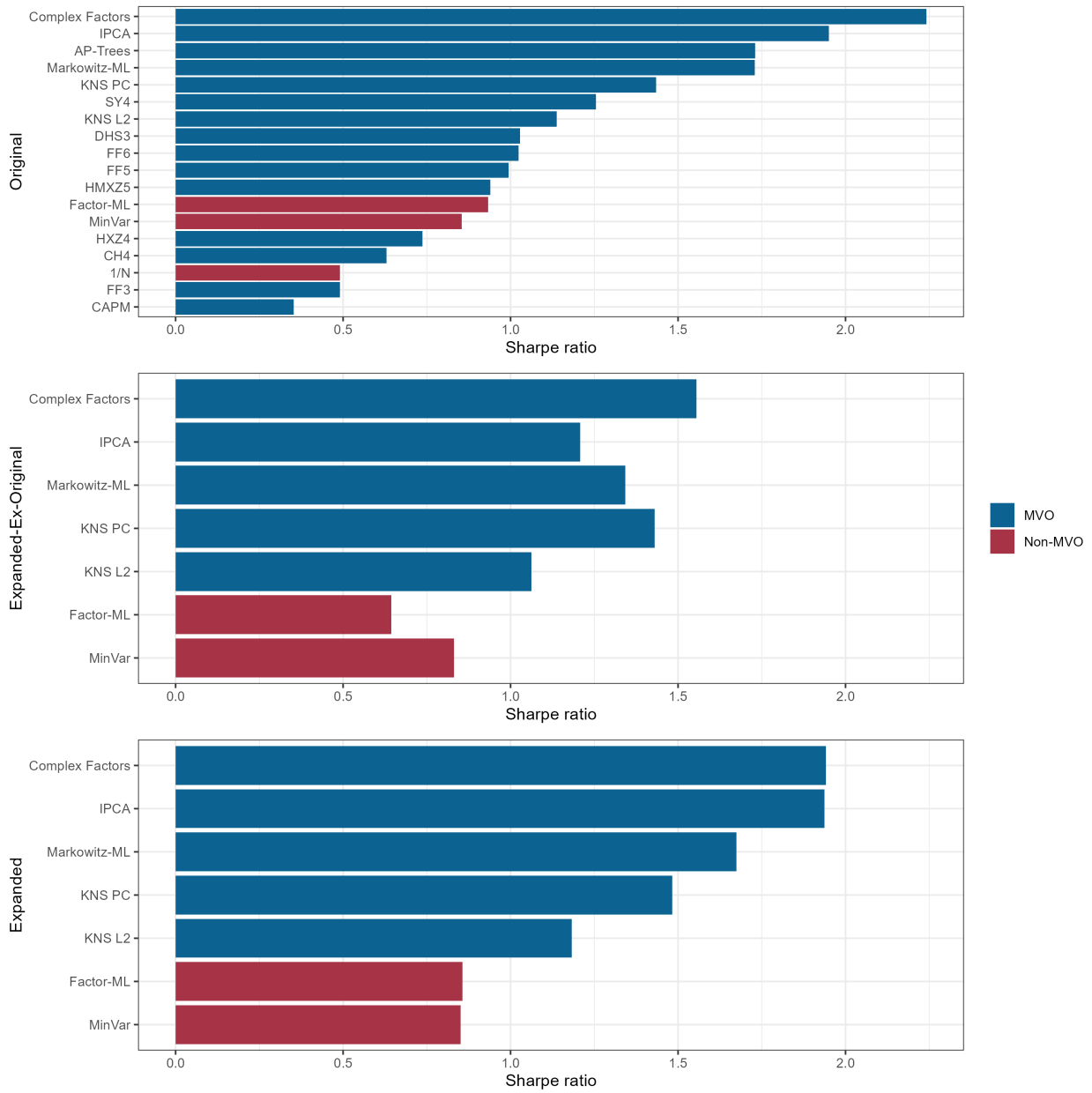


Figure 3: **Sharpe Ratio of Portfolio Methods.** Portfolio methods are implemented on JKP data and their annualized Sharpe ratios are reported. Results are for mega-, large-, and small-cap US stocks. The test data period is 1990-2023, and the training data period is 1946-1989. "Expanded" denotes the expanded feature set of 402 characteristics, and "Expanded-Ex-Original" is the expanded feature set less the original 153 JKP characteristics.

upward biased results.

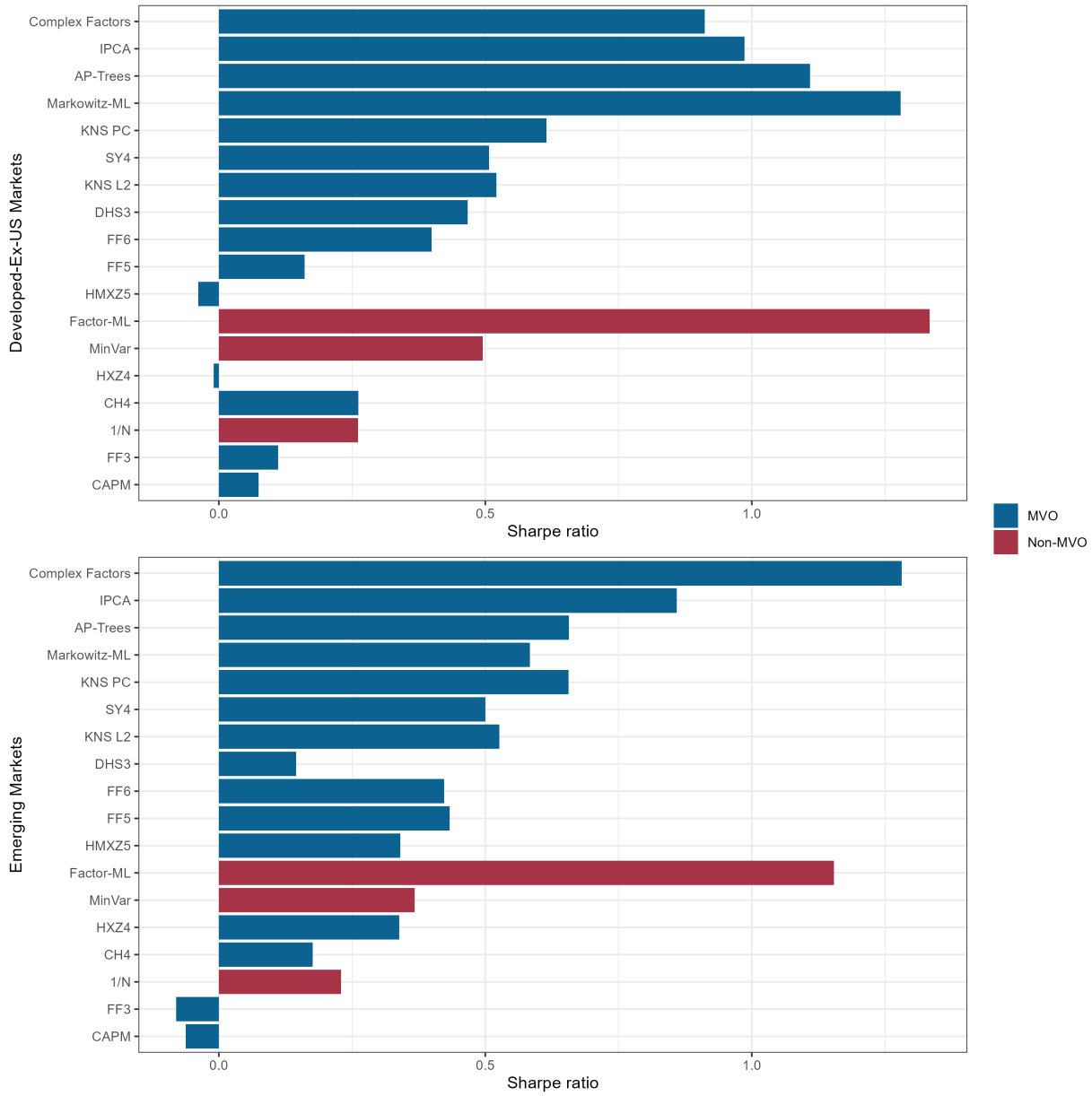


Figure 4: **Sharpe Ratio of Portfolio Methods.** Portfolio methods are implemented on JKP data, and their annualized Sharpe ratios are reported. Results are for mega-, large-, and small-cap stocks in developed ex-US and emerging markets. The test data period is 1985-2023 and 1988-2023 for developed ex-US and emerging markets respectively. Models are trained on mega-, large-, and small-cap US stocks in the period 1946-1989. Portfolio methods are ordered by their out-of-sample Sharpe ratio on US stocks.

Figure 4 shows how well the models perform when we evaluated them on stocks from

different regions. Specifically, we take the model trained on U.S. data with the 153 original JKP characteristics and apply them to stocks outside of the U.S. We apply the methods separately for stocks listed on exchanges in countries classified as developed or emerging market by MSCI. The ranking of the models is similar across the three regions, except that the Factor-ML model has a much better ranking outside of the U.S. The magnitudes of the Sharpe ratios are lower outside the U.S., but similar in developed and emerging markets. These lower magnitudes could simply reflect that the models were trained on U.S. data, and thus should be expected to perform better on a test set that is more similar to the training set. Alternatively, it could be because most papers first test their models on US data and therefore inadvertently end up overfitting US data.

Figure 5 shows how well the models perform over time. Specifically, the figure shows the realized Sharpe ratio over a 10-year rolling window for a selected set of methods. For example, the figure starts in 2000/01 with the realized Sharpe ratio from 1990/01-1999/12. Model rankings fluctuate quite a bit over time, showing the dependence of the ranking on the test sample. Most noticeable, however, is the decline in the Sharpe ratio of all methods over time. For example, the complex factor model starts out as the best performing model with a Sharpe ratio above 4 and ends with a Sharpe ratio around 1. Another less pronounced example is AP-trees, which starts out with a Sharpe ratio around 3 and ends as the best performing models with a Sharpe ratio of around 1.3. The decline over time is one reason why it is problematic to evaluate models on historical data if we want to know the future pricing kernel.

5 Combining Models: Ensemble Methods

The CTF has led to progress in many fields, and the hope is that it will lead to progress in finance. As a first step, we show that any of the individual methods we have considered are dominated by an ensemble method that combines the individual methods. We note that it is only possible to create an ensemble for methods implemented on a common data set, so the fact that we can even create a tradable ensemble is yet another benefit of the CTF framework.

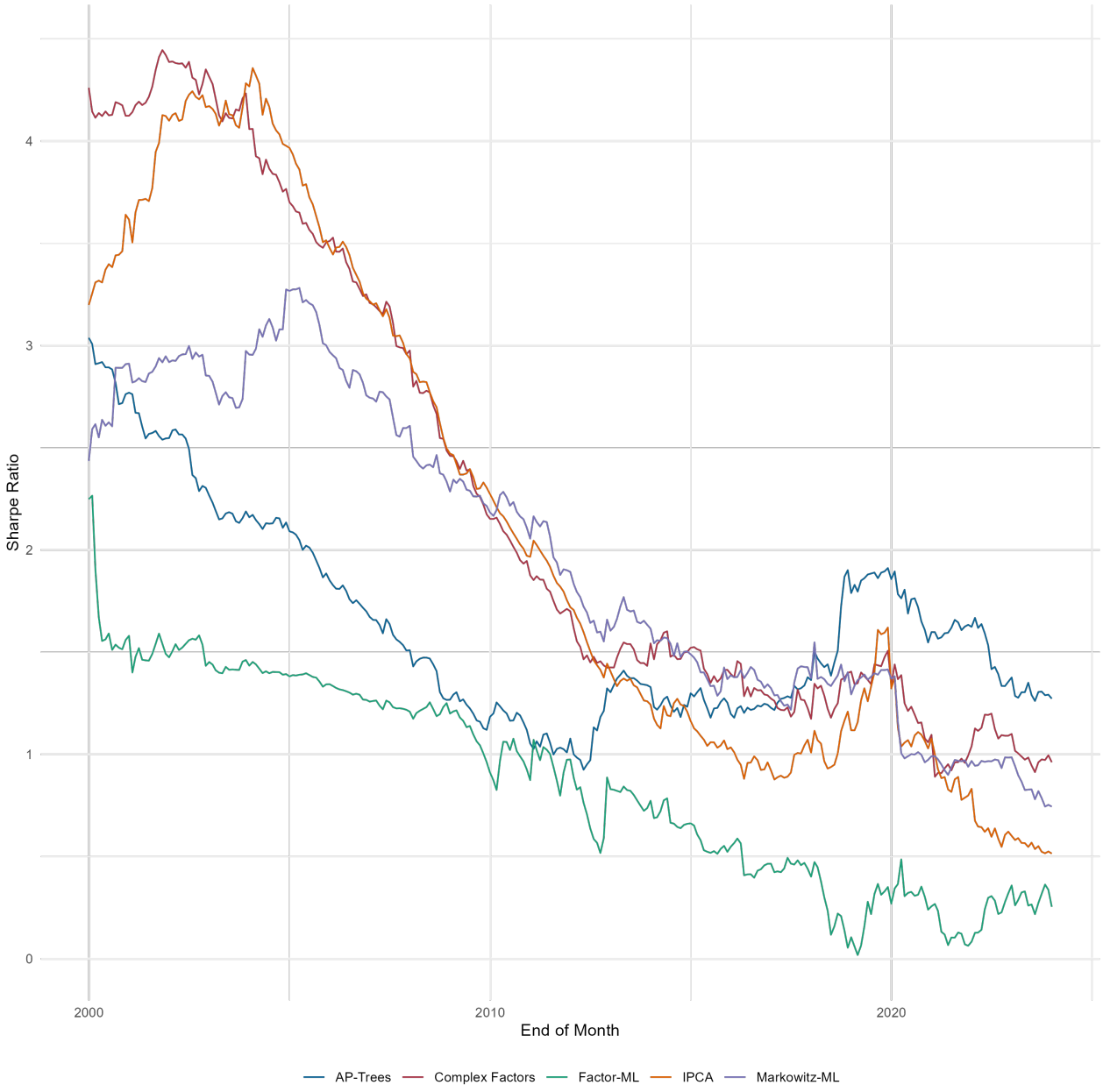


Figure 5: **10-Year Rolling Sharpe Ratio of Portfolio Methods.** Portfolio methods are implemented on JKP data and their annualized 10-year rolling Sharpe ratios are reported. Results are for mega-, large-, and small-cap US stocks. The test data period is 1990-2023, and the training data period is 1946-1989.

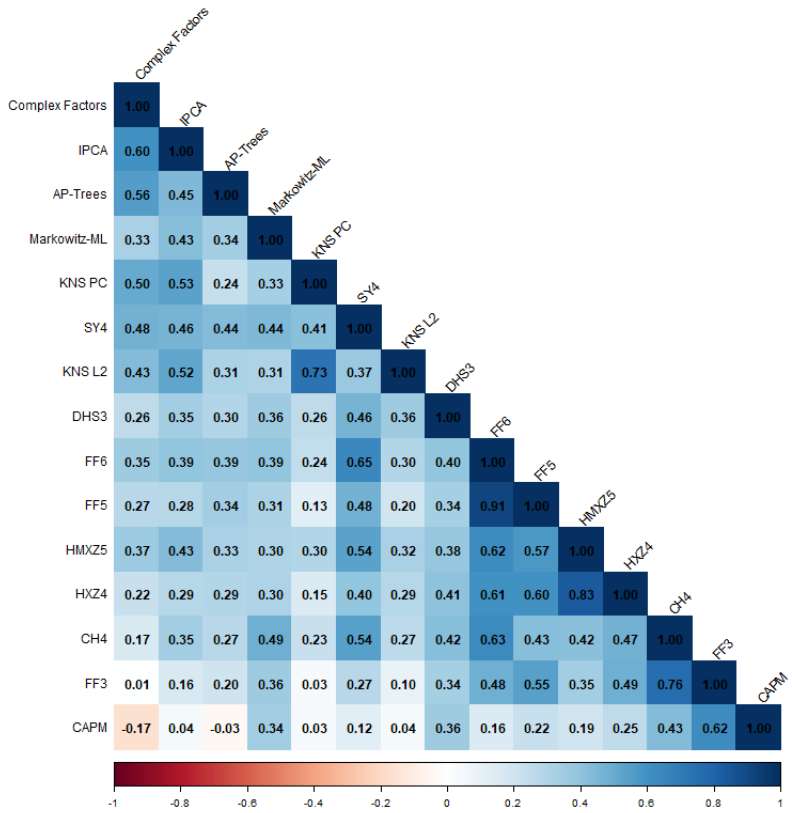


Figure 6: **Correlation between Returns of Portfolio Methods.** Portfolio methods are implemented on JKP data and their correlation matrix is reported. Results are for mega-, large-, and small-cap US stocks. The test data period is 1990-2023, and the training data period is 1946-1989. Methods are fitted using the original 153 JKP characteristics. Portfolios are ordered by their out-of-sample Sharpe ratio.

Figure 6 show the methods' return correlation matrix. Although the MVO models attempt to estimate the same underlying SDF, their estimates have surprisingly low correlations. The low correlations invite the opportunity to combine the different methods into one. In statistics and machine learning, this approach is known as "ensemble learning". Intuitively, ensemble learning aims to combine several "weak learners" to obtain a predictive performance unattainable by the individual learners. For supervised learning, low correlations between the predictions of weak learners lead to better ensemble performance. Inspired

by this, we investigate four different ensemble methods.[15]

Equal-Weighted Ensemble (EW Ensemble). A common ensemble learning method in supervised learning is to use the average prediction from the weak learners as the ensemble prediction. We can analogously take the average SDF estimate as a new SDF estimate

$$\hat{w}_{EW} = \frac{1}{M}\sum_{m=1}^{M}\hat{w}_m, \tag{11}$$

where $\hat{w}_m$ is the portfolio of model m, and M is the number of models.

Signal-Weighted Ensemble (SW Ensemble). Instead of equal-weighting every strategy, the SW Ensemble weighs them by their mean return

$$\hat{w}_{SW} = \sum_{m=1}^{M}\hat{\mu}_m\hat{w}_m, \tag{12}$$

where $\hat{\mu}_m$ is the 10-year rolling average return of model m.

Mean-Variance Optimal Ensemble (MV Ensemble). As the name suggests, the MV Ensemble is the mean-variance optimal combination of the models

$$\hat{w}_{MV} = \hat{\Sigma}_{\mathcal{M}}^{-1}\hat{\mu}_{\mathcal{M}}, \tag{13}$$

where $\mathcal{M}$ is the set of models, $\hat{\Sigma}_{\mathcal{M}}$ and $\hat{\mu}_{\mathcal{M}}$ are the 10-year rolling sample covariance matrix and average return vector of the models.

Mean-Variance Optimal Ensemble with No-Shorting Constraint. A no-shorting constaint often improves the out-of-sample performance of an MVO portfolio. The ensemble portfolio solves the optimization problem

$$\max_{w} \quad w'\hat{\mu}_{\mathcal{M}} - \frac{1}{2}w'\hat{\Sigma}_{\mathcal{M}}w \tag{14}$$

$$\text{s.t.} \quad w \geq 0. \tag{15}$$

[15]Before building an ensemble, we first standardize each of the candidate portfolios to have the same ex-ante volatility. We do so by, each month, scaling the portfolio return by its realized volatility over the last ten years.

The last three methods draw inspiration from an ensemble method known as "stacking", which entails training a model to combine the weak learners' predictions instead of using a simple average.

Figure 7 shows the Sharpe ratio for each ensemble, based on candidate methods fitted using either the original, full-ex-original, or full JKP characteristics. For all three sets of characteristics, at least one of the ensemble methods outperform the best performing individual method. This result is not obvious ex-ante as the way the ensembles weigh the individual methods is determined in an out-of-sample way. The ensembles generally do better on the non-original characteristic sets, which is consistent with the individual models being to some extent tailored to the original characteristics. When looking at the full-ex-original characteristics, all four ensemble methods outperform the best individual method.

The ranking of the four ensemble methods is similar across the three sets of characteristics. The mean-variance ensemble with the no shorting constraint is always best, the unconstrained mean-variance ensemble is the second best for two characteristic sets, the signal-weighted ensemble is the second-best for one characteristic set, and the equal-weighted ensemble is always worst. The results suggest that ensemble methods can outperform individual methods, but only by going beyond naive alternatives, such as equal-weighting, into more sophisticated alternatives, such as mean-variance weighting with a no short constraint.

6 A CTF for New Finance Models: A Competition

We now describe a competition that is designed to be an instance of the common task framework applied to finance. The competition will be hosted on the website https://jkpfactors.com/common-task-framework, and, like any instance of the common task framework, consists of three elements:

1. A shared training data set
2. A common task with a predefined quantitative metric to judge success
3. A leaderboard that controls for over-fitting, e.g., via a secret test set

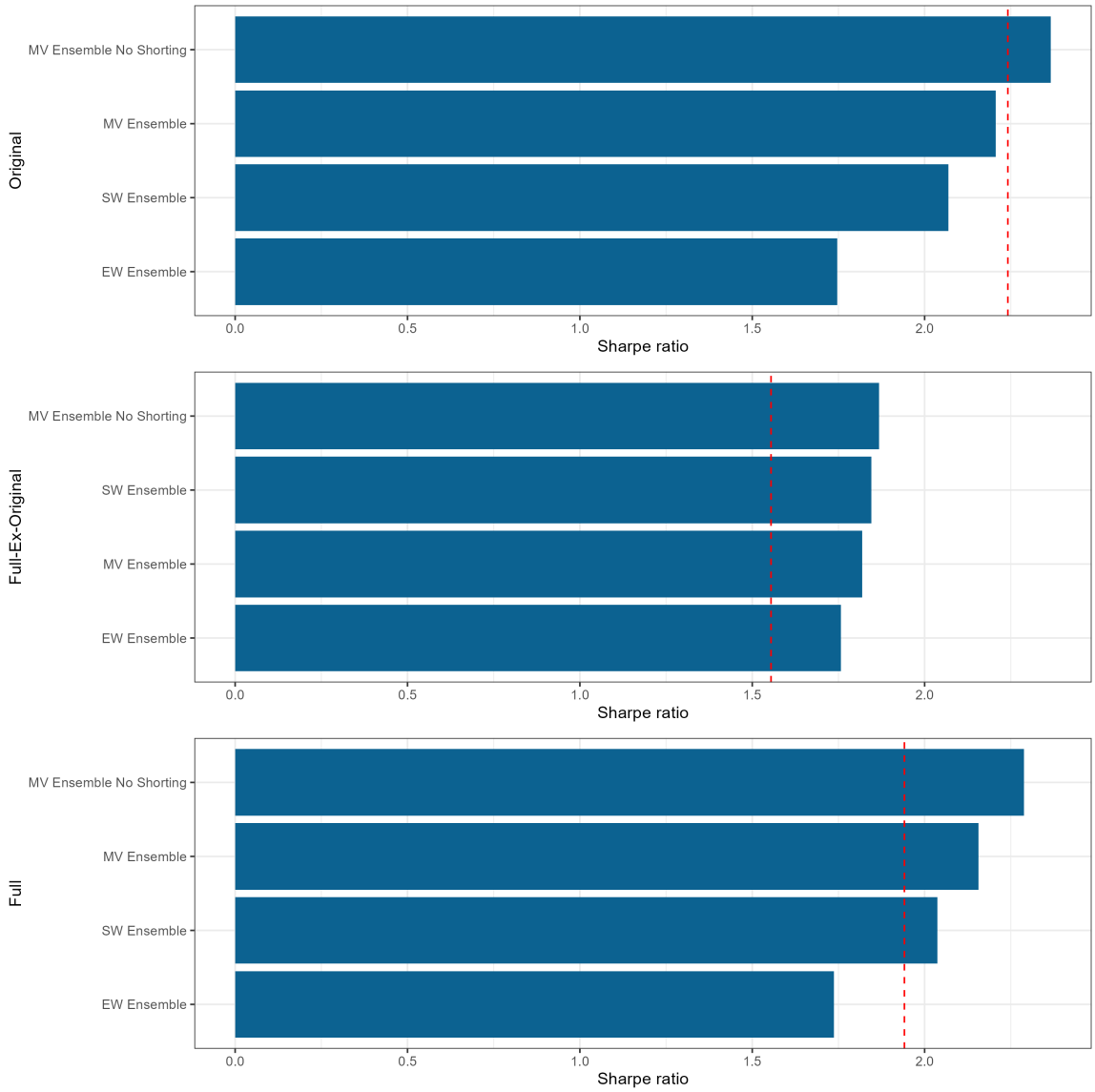


Figure 7: **Sharpe Ratio of Ensemble Methods.** Portfolio methods are combined in "ensemble" portfolios by the 10-year rolling mean-variance optimal combination (with and without no-shorting constraint), equal-weighted, or weighted by their 10-year rolling mean. The underlying portfolio methods are implemented on JKP data and their annualized Sharpe ratios are reported. Results are for mega-, large-, and small-cap US stocks. The test data period is 1990-2023, and the training data period is 1946-1989. "Original", "Expanded", and "Expanded-Ex-Original" denote whether the underlying models were fitted using the original 153 JKP characteristics, the expanded 402 characteristics, or the difference between the two, respectively. The dashed line indicates the highest Sharpe ratio across the underlying portfolio methods.

The shared training data set. The data set that competition participants should use to fit their models is the one we have used in the paper (described in Section 3.1). Researchers can follow the guide on the competition website to apply for access.

The common task. Participants are judged by the Sharpe ratio realized by their portfolio on the test data set. Section 3.2 explains how this common task is equivalent to finding the tangency portfolio that can be used to create the minimum variance SDF.

A leaderboard based on a secret test set. Creating a secret test data set in finance is both easy and difficult. Easy because the passing of time will create new unseen test data, but difficult because returns are so noisy that it takes a long time to separate the signal from the noise. In the context of our model, equation (2), we observe the Sharpe ratio on the test set (ν_i), but it reflects both the true expected Sharpe ratio ($\tilde{\nu}$) that we care about and the noise in returns (η_i). The noise should cancel out over time, but it may be unsatisfactory for participants to wait, say, ten years before knowing their final score.

To balance these considerations, we create three separate competitions based on three different test data sets. The first competition is a "hindcast" competition, where the test data covers the same data and period as in our paper, 1990/01-2023/12. The benefit of the hindcast competition is that it allows us to provide a competition score immediately and that it is based on a long sample, but the downside is that the test set is not secret. However, the hindcast competition is similar to the way researchers usually evaluate their models. The second competition is a "forecast" competition, where test data is created using the same filters as in the paper but start in 2026/01 and end in 2026/12. To enter this competition participants must submit their model by December 31, 2025, which ensures that the test set is truly secret. The third competition is a "secret" competition, where the test data is secret at the beginning of the competition and will be revealed at a later point.[16]

How to enter the competitions. To enter into the three competitions, researchers only need to make one submission, which contains three components: (1) a self-contained script in R or Python that takes the shared training data and assigns a portfolio weight for each stock-

[16]We view these as "pilot" competitions, and the competition structure may change over time as we learn about what works and what does not. In this spirit, we welcome any feedback on how to improve the competition structure.

month combination in a generic test set,[17] (2) a CSV file with the portfolio weights assigned by the script for all stock-month combinations in the test set from the hindcast competition, and (3) a PDF file containing a description of their portfolio construction methodology, which can be anything from a full research paper to a short step-by-step document. The CSV file will be used to compete in the hindcast competition. The R or Python script will be used by us to re-estimate a selected set of models on the new data for the forecast and secret competition. The documentation will be used to select which models we will enter into the forecast and secret competition.[18]

7 Conclusion

We present a theory of how a collaborative and competitive CTF can promote research by lowering the cost of research, making findings easily comparable, and reducing distortions. Taking the theory to practice, we put forward a CTF to discover the pricing kernel used in finance.

More generally, we propose that the CTF be applied throughout economics and other sciences. For example, the CTF can be applied to predict inflation, GDP growth, unemployment (overall or by industry or region), real estate price changes, default rates, market risk, carbon prices, global temperatures, schoolchildren's reading abilities, or mortality rates.

[17] The script must assign portfolio weights in an out-of-sample way, meaning that portfolio weights assigned at time t must only be based on information that was known at t.

[18] We will do our best to re-estimate as many models as possible, but time and computational constraints may prevent us from re-estimating them all. We will select models based on the quality of their documentation (favoring full research papers over step-by-step documents) and the computational resources necessary to estimate them. For example, a model with a full research paper that can be estimated in less than one hour will almost surely be re-estimated.

References

Asness, C. and A. Frazzini (2013). The devil in hml's details. *Journal of Portfolio Management 39*(4), 49–68.

Beckmeyer, H. and T. Wiedemann (2022). Recovering missing firm characteristics with attention-based machine learning.

Bennett, J. and S. Lanning (2007). The netflix prize. *Proceedings of KDD Cup and Workshop 2007*, 3–6.

Bettis, R. A. (2012). The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal 33*(1), 108–113.

Bojer, C. S. and J. P. Meldgaard (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting 37*(2), 587–603.

Britten-Jones, M. (1999). The sampling error in estimates of mean-variance efficient portfolio weights. *The Journal of Finance 54*(2), 655–671.

Bryzgalova, S., S. Lerner, M. Lettau, and M. Pelger (2022). Missing financial data. *Available at SSRN 4106794*.

Bryzgalova, S., M. Pelger, and J. Zhu (2023). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science 351*(6280), 1433–1436.

Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. *the Journal of Finance LII*(1), 57–82.

Chen, A. Y. and J. McCoy (2024). Missing values handling for machine learning portfolios. *Journal of Financial Economics 155*, 103815.

Chen, A. Y. and T. Zimmermann (2022). Open source cross-sectional asset pricing. *Critical Finance Review 11*(2), 207–264.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Cochrane, J. H. (2005). *Asset Pricing: Revised Edition*. Princeton, NJ: Princeton University Press.

Daniel, K., D. Hirshleifer, and L. Sun (2020). Short-and long-horizon behavioral factors. *The Review of Financial Studies 33*(4), 1673–1736.

DeMiguel, V., L. Garlappi, and R. Uppal (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies 22*(5), 1915–1953.

Dick-Nielsen, J., P. Feldhütter, L. H. Pedersen, and C. Stolborg (2023). Corporate bond factors: Replication failures and a new framework. *Available at SSRN 4586652*.

Didisheim, A., S. B. Ke, B. T. Kelly, and S. Malamud (2023). Complexity in factor pricing models. Technical report, National Bureau of Economic Research.

Doddington, G. R. and T. B. Schalk (1981). Computers: Speech recognition: Turning theory to practice: New ics have brought the requisite computer power to speech technology; an evaluation of equipment shows where it stands today. *IEEE spectrum 18*(9), 26–32.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics 26*(4), 745–766.

Eskildsen, M., M. Ibert, T. I. Jensen, and L. H. Pedersen (2024). In search of the true greenium. *Available at SSRN*.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Freyberger, J., B. Höppner, A. Neuhierl, and M. Weber (2024). Missing data in asset pricing panels. *The Review of Financial Studies*, hhae003.

Gibbons, M. R., S. A. Ross, and J. Shanken (1989). A test of the efficiency of a given portfolio. *Econometrica: Journal of the Econometric Society*, 1121–1152.

Grossman, S. J. and O. D. Hart (1983). An analysis of the principal-agent problem. *Econometrica 51*, 7–45.

Hamermesh, D. S. (2007). Replication in economics. *Canadian Journal of Economics/Revue canadienne d'économique 40*(3), 715–733.

Hansen, L. P. and R. Jagannathan (1991). Implications of security market data for models of dynamic economies. *Journal of political economy 99*(2), 225–262.

Hofman, J. M., D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, et al. (2021). Integrating explanation and prediction in computational social science. *Nature 595*(7866), 181–188.

Holmström, B. (1979). Moral hazard and observability. *The Bell journal of economics*, 74–91.

Hou, K., C. Xue, and L. Zhang (2015). Digesting anomalies: An investment approach. *Review of Financial Studies 28*(3), 650–705.

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

Jensen, T. I., B. T. Kelly, S. Malamud, and L. H. Pedersen (2024). Machine learning and the implementable efficient frontier. *Review of Financial Studies (forthcoming)*.

Jensen, T. I., B. T. Kelly, and L. H. Pedersen (2023). Is there a replication crisis in finance? *The Journal of Finance 78*(5), 2465–2518.

Kelly, B. T., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

Liberman, M. and C. Wayne (2020). Human language technology. *AI Magazine 41*(2), 22–35.

Maniadis, Z., F. Tufano, and J. A. List (2017). To replicate or not to replicate? exploring reproducibility in economics through the lens of a model and a pilot study. *The Economic Journal 127*, F209–F235.

Markowitz, H. M. (1952). Portfolio selection. *The Journal of Finance 7*(1), 77–91.

Menchero, J., D. Orr, and J. Wang (2011). The barra us equity model (use4), methodology notes. *MSCI Barra*.

Menkveld, A. J., A. Dreber, F. Holzmeister, J. Huber, M. Johannesson, M. Kirchler, S. Neusüss, M. Razen, U. Weitzel, D. Abad-Díaz, et al. (2024). Nonstandard errors. *The Journal of Finance 79*(3), 2339–2390.

Michaud, R. O. (1989). The markowitz optimization enigma: Is 'optimized'optimal? *Financial analysts journal 45*(1), 31–42.

Nalebuff, B. J. and J. E. Stiglitz (1983). Prizes and incentives: towards a general theory of compensation and competition. *The Bell Journal of Economics 14*, 21–43.

Nosek, B. A., J. R. Spies, and M. Motyl (2012). Scientific utopia: ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science 7*(6), 615–631.

Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance 19*(3), 425–442.

Stambaugh, R. F. and Y. Yuan (2017). Mispricing factors. *The Review of Financial Studies 30*(4), 1270–1315.

APPENDIX

Appendix A contains proofs of our propositions, Appendix B explains each of the candidate asset pricing methods that we examine, and Appedix C.1 contains additional empirical results.

A Proofs

Proof of Propositions 1-2. Decompose

$$\xi_i = \sqrt{\rho_\xi}\xi + \sqrt{1-\rho_\xi}\tilde{\xi}_i \tag{A.1}$$

$$\eta_i = \sqrt{\rho_\eta}\eta + \sqrt{1-\rho_\eta}\tilde{\eta}_i\,, \tag{A.2}$$

such that ξ, $\tilde{\xi}_i$, η, $\tilde{\eta}_i$ are independent random variables with mean zero and variances $\text{Var}(\xi) = \text{Var}(\tilde{\xi}_i) = \sigma_\xi^2$ and $\text{Var}(\eta) = \text{Var}(\tilde{\eta}_i) = \sigma_\eta^2$. Then let

$$\epsilon \equiv \frac{\sqrt{\rho_\xi}\eta + \sqrt{\rho_\eta}\eta}{\sqrt{\rho}} \tag{A.3}$$

$$\varepsilon_i \equiv \frac{\sqrt{1-\rho_\xi}\tilde{\eta}_i + \sqrt{1-\rho_\eta}\tilde{\eta}_i}{\sqrt{1-\rho}}, \tag{A.4}$$

where

$$\rho \equiv \rho_\xi \frac{\sigma_\xi^2}{\sigma_\xi^2+\sigma_\eta^2} + \rho_\eta \frac{\sigma_\eta^2}{\sigma_\xi^2+\sigma_\eta^2}\,, \tag{A.5}$$

with $\frac{\partial \rho}{\partial \rho_\eta} > 0$. The measured innovation is then given in (6). The variance of the new random variables is

$$\text{Var}(\epsilon) = \frac{1}{\rho}\left(\rho_\xi\sigma_\xi^2 + \rho_\eta\sigma_\eta^2\right) = \sigma_\eta^2 + \sigma_\xi^2 \equiv \sigma^2 \tag{A.6}$$

$$\text{Var}(\varepsilon_i) = \frac{1}{1-\rho}\left((1-\rho_\xi)\sigma_\xi^2 + (1-\rho_\eta)\sigma_\eta^2\right) = \sigma^2\,. \tag{A.7}$$

Since ϵ and ε_i are linear combinations of random variables with covariance 0, the covariance between ϵ and ε_i is also 0. The correlation between the shocks of two researchers i, j is

$$\text{Cor}(\sqrt{\rho}\epsilon + \sqrt{1-\rho}\varepsilon_i, \sqrt{\rho}\epsilon + \sqrt{1-\rho}\varepsilon_i) = \frac{\rho(\sigma_\xi^2 + \sigma_\eta^2)}{\sigma_\xi^2 + \sigma_\eta^2} = \rho\,. \tag{A.8}$$

If the researcher chooses $l^* = 0$, they have utility net of effort costs

$$\mathbb{E}[w_i] - \frac{\gamma}{2}\text{Var}[w_i] - \frac{\kappa}{2}e_i^2 \tag{A.9}$$

For a given wage contract, the researcher's wage has mean

$$\mathbb{E}[w_i] = c + \lambda\left(\mathbb{E}[v_i] - \frac{1}{I-1}\sum_{j\neq i}\mathbb{E}[v_i]\right) = c + \lambda e_i - \theta e_{-i}$$

and variance

$$\begin{aligned}
\text{Var}[w_i] &= \text{Var}\left[\lambda(e_i + \sqrt{\rho}\epsilon + \sqrt{1-\rho}\varepsilon_i) - \theta\frac{1}{I-1}\sum_{j\neq i}(e_j + \sqrt{\rho}\epsilon + \sqrt{1-\rho}\varepsilon_j)\right] \\
&= \text{Var}\left[(\lambda-\theta)\sqrt{\rho}\epsilon + \lambda\sqrt{1-\rho}\varepsilon_i - \theta\frac{1}{I-1}\sum_{j\neq i}\sqrt{1-\rho}\varepsilon_j\right] \\
&= \rho\sigma^2(\lambda-\theta)^2 + (1-\rho)\sigma^2\lambda^2 + (1-\rho)\sigma^2\frac{1}{I-1}\theta^2
\end{aligned}$$

Inserting these into the objective function (A.9) yields

$$\max_{e_i} \quad c + \lambda e_i^* - \theta e_{-i}^* - \frac{\gamma}{2}\rho\sigma^2(\lambda-\theta)^2 - \frac{\gamma}{2}(1-\rho)\sigma^2\lambda^2 - \frac{\gamma}{2}(1-\rho)\sigma^2\frac{1}{I-1}\theta^2 - \frac{\kappa}{2}(e_i^*)^2,$$

with the first-order condition

$$e_i^* = \frac{\lambda}{\kappa}.$$

Since the utility net of effort costs is strictly concave, e_i^* is a maximum. Since all researchers choose this level of effort, the average effort of others is $e_{-i}^* = e_i^*$. It is optimal for the

researcher to choose effort level e_i^* and $l^* = 0$ if

$$c + \lambda e_i^* - \theta e_{-i}^* - \frac{\gamma}{2}\rho\sigma^2(\lambda-\theta)^2 - \frac{\gamma}{2}(1-\rho)\sigma^2\lambda^2 - \frac{\gamma}{2}(1-\rho)\sigma^2\frac{1}{I-1}\theta^2 - \frac{\kappa}{2}(e_i^*)^2 \geq u^*,$$

i.e. their expected utility is greater than their outside option. Otherwise, they choose $l^* = 1$ and $e_i^* = 0$.

Suppose the principal chooses c, λ and θ such that researchers are willing to participate, i.e. $l_i^* = 0$. Then, the principal solves

$$\begin{aligned}
\max_{c,\lambda,\theta} \quad & \mathbb{E}\sum_i (v_i - w_i)1_{l_i=0} \\
\text{s.t.} \quad & c + \lambda e_i^* - \theta e_{-i}^* - \frac{\gamma}{2}\rho\sigma^2(\lambda-\theta)^2 - \frac{\gamma}{2}(1-\rho)\sigma^2\lambda^2 - \frac{\gamma}{2}(1-\rho)\sigma^2\frac{1}{I-1}\theta^2 - \frac{\kappa}{2}(e_i^*)^2 \geq u^*.
\end{aligned}$$

Since every researcher chooses the same e_i^*, the journal chooses c such that the constraint binds

$$\begin{aligned}
u^* =& c + \frac{(\lambda-\theta)}{\kappa}\lambda - \frac{\gamma}{2}\rho\sigma^2(\lambda-\theta)^2 - \frac{\gamma}{2}(1-\rho)\sigma^2\lambda^2 - \frac{\gamma}{2}(1-\rho)\sigma^2\frac{1}{I-1}\theta^2 - \frac{\lambda^2}{2\kappa} \\
c =& u^* - \frac{(\lambda-\theta)}{\kappa}\lambda + \frac{\gamma}{2}\rho\sigma^2(\lambda-\theta)^2 + \frac{\gamma}{2}(1-\rho)\sigma^2\lambda^2 + \frac{\gamma}{2}(1-\rho)\sigma^2\frac{1}{I-1}\theta^2 + \frac{\lambda^2}{2\kappa}.
\end{aligned}$$

The maximization problem is thus reduced to

$$\max_{\lambda,\theta} \quad I\left(\frac{\lambda}{\kappa} - u^* - \frac{\gamma}{2}\rho\sigma^2(\lambda-\theta)^2 - \frac{\gamma}{2}(1-\rho)\sigma^2\lambda^2 - \frac{\gamma}{2}(1-\rho)\sigma^2\frac{1}{I-1}\theta^2 - \frac{\lambda^2}{2\kappa}\right)$$

where the first-order conditions yield the solution

$$\begin{aligned}
\theta^* =& \frac{\rho}{\rho + \frac{1-\rho}{I-1}}\lambda^* \\
\lambda^* =& \frac{1}{1 + \kappa\gamma(1-\rho)\sigma^2 + \kappa\gamma\rho\sigma^2\left(1 - \frac{\rho}{\rho+\frac{1-\rho}{I-1}}\right)}
\end{aligned}$$

The objective function is strictly concave, making λ^*, θ^* the maximum. Using the first-order

conditions

$$\frac{1}{\kappa} = \gamma\rho\sigma^2(\lambda^* - \theta^*) + \gamma(1-\rho)\sigma^2\lambda^* + \frac{\lambda^*}{\kappa}$$
$$\gamma(1-\rho)\sigma^2\frac{1}{I-1}\theta^* = \gamma\rho\sigma^2(\lambda^* - \theta^*),$$

the expression for c^* simplifies to

$$c^* = u^* + \frac{\lambda^*}{\kappa}\left(\frac{1}{2} - (\lambda^* - \theta^*)\right)$$

The principal's utility in equilibrium is given by

$$I\left(\frac{\lambda^*}{\kappa} - u^* - \frac{\lambda^*}{\kappa}\left(\frac{1}{2} - (\lambda^* - \theta^*)\right) - \frac{(\lambda^*)^2}{\kappa} + \frac{\lambda^*\theta^*}{\kappa}\right) = I\left(\frac{\lambda^*}{2\kappa} - u^*\right).$$

The derivative of λ^* wrt. ρ is given by

$$\begin{aligned}
\frac{\partial e_i^*}{\partial \rho_\eta} &= \frac{\partial e_i^*}{\partial \rho}\frac{\partial \rho}{\partial \rho_\eta} \\
&= -(\lambda^*)^2 \cdot \frac{\partial}{\partial \rho}\kappa\gamma\sigma^2\left(1 - \rho + \rho\frac{\frac{1-\rho}{I-1}}{\frac{1-\rho}{I-1} + \rho}\right)\frac{\sigma_\eta^2}{\sigma^2} \\
&= (\lambda^*)^2\kappa\gamma\sigma^2\rho\frac{\rho + \frac{1}{I-1}}{\left(\frac{1-\rho}{I-1} + \rho\right)^2}\frac{\sigma_\eta^2}{\sigma^2} \\
&> 0,
\end{aligned}$$

meaning λ^* increases in ρ. And since $\frac{\partial \rho}{\partial \rho_\eta} > 0$, the same holds for increases in ρ_η. Furthermore, it is easily seen that λ^* increases when κ decreases.

Thus, the journal's utility increases when ρ_η increases or κ decreases, when $l_i^* = 0$. If the utility is negative, the principal would rather not pay any wage

$$\frac{\lambda^*}{2\kappa} < u^* \Rightarrow c^* = \lambda^* = \theta^* = 0.$$

Then, the researcher's utility is 0, meaning they would rather leave the field, $l_i^* = 1$. Similarly, the journal receives a utility of 0.

Conversely, if $\frac{\lambda^*}{2\kappa} > u^*$, the journal receives a strictly positive utility in equilibrium. Thus, a sufficient increase in ρ_η or decrease in κ until $\frac{\lambda^*}{2\kappa} > u^*$ leads researchers to not leave the field $l_i^* = 0$.

Note that since the journal sets c^* such that the constraint is binding, the total welfare is

$$I\left(\frac{\lambda^*}{2\kappa} - u^*\right) + Iu^*,$$

which also increases when ρ_η increases or κ decreases. □

Proof of Proposition 4-6. The distortion affects the expected wage, but not the wage variance

$$\begin{aligned}
\mathbb{E}[w_i] &= c + \lambda(e_i + \sqrt{1-\rho}\, d_i) - \theta(e_{-i} + \sqrt{1-\rho}\, d_{-i}) \\
\mathrm{Var}[w_i] &= \rho\sigma^2(\lambda-\theta)^2 + (1-\rho)\sigma^2\lambda^2 + (1-\rho)\sigma^2\frac{1}{I-1}\theta^2.
\end{aligned}$$

Assuming the researcher doesn't leave the field, $l_i^* = 0$, the researcher solves the optimization

$$\max_{e_i, d_i} \quad \mathbb{E}[w_i] - \frac{\gamma}{2}\mathrm{Var}[w_i] - \frac{\kappa}{2}e_i^2 - \frac{z}{2}d_i^2,$$

which is maximized at

$$\begin{aligned}
e_i^* &= \frac{\lambda}{\kappa} \\
d_i^* &= \frac{\lambda}{z}\sqrt{1-\rho}.
\end{aligned}$$

If the journal chooses c, λ, and θ such that the researcher participates, then similarly to the proof of Proposition 1, the c is set to bind the constraint The journal's utility is

$$\mathbb{E}[w_i] = u^* + \frac{\gamma}{2}\rho\sigma^2(\lambda-\theta)^2 + \frac{\gamma}{2}(1-\rho)\sigma^2\lambda^2 + \frac{\gamma}{2}(1-\rho)\sigma^2\frac{1}{I-1}\theta^2 + \frac{\lambda^2}{2\kappa} + \frac{\lambda^2}{2z}(1-\rho).$$

The maximization problem is then

$$\max_{\lambda,\theta} \quad I\left(\frac{\lambda}{\kappa} - u^* - \frac{\gamma}{2}\rho\sigma^2(\lambda-\theta)^2 - \frac{\gamma}{2}(1-\rho)\sigma^2\lambda^2 - \frac{\gamma}{2}(1-\rho)\sigma^2\frac{1}{I-1}\theta^2 - \frac{\lambda^2}{2\kappa} - \frac{\lambda^2}{2z}\right),$$

and the first-order conditions imply

$$\begin{aligned}
\theta^* =& \frac{\rho}{\rho + \frac{1-\rho}{I-1}}\lambda^* \\
\lambda^* =& \frac{1}{1 + \frac{\kappa}{z}(1-\rho) + \kappa\gamma(1-\rho)\sigma^2 + \kappa\gamma\rho\sigma^2\left(1 - \frac{\rho}{\rho+\frac{1-\rho}{I-1}}\right)} \\
c^* =& u^* + \frac{\lambda^*}{\kappa}\left(\frac{1}{2} - (\lambda^* - \theta^*)\right).
\end{aligned}$$

As $z \to \infty$, we have $d_i^* = 0$ and recover the same solution in λ_i^*, θ^*, c^*, and e_i^* as when no distortion was possible.

Furthermore, the ratio of distortion to effort $d_i^*/e_i^* = \frac{\kappa}{z}\sqrt{1-\rho}$ decreases when either ρ increases or κ decreases.

An increase in ρ increases λ^* by the same argument used in the proof for Proposition 2. Since c^* has the same form as in Proposition 1, the principal's utility is similar to the one in Proposition 2

$$I\left(\frac{\lambda^*}{2\kappa} - u^*\right).$$

Thus, the utility of the principal increases in ρ_η. Furthermore, it increases when κ decreases.

For the elasticities

$$\begin{aligned}
\frac{\partial e_i^*}{\partial \rho_\eta}\frac{\rho_\eta}{e_i^*} &= \frac{\partial e_i^*}{\partial \rho}\frac{\partial \rho}{\partial \rho_\eta}\frac{\rho_\eta}{e_i^*} \\
&= \lambda^*\kappa\left(\frac{1}{z} + \gamma\sigma^2\right)\rho\frac{\rho + \frac{1}{I-1}}{\left(\frac{1-\rho}{I-1} + \rho\right)^2}\frac{\rho_\eta\sigma_\eta^2}{\sigma^2} \\
&= \frac{\rho\frac{\rho+\frac{1}{I-1}}{\left(\frac{1-\rho}{I-1}+\rho\right)^2}}{\frac{1}{\kappa\left(\frac{1}{z}+\gamma\sigma^2\right)} + (1-\rho) + \frac{\gamma\rho\sigma^2}{\kappa\left(\frac{1}{z}+\gamma\sigma^2\right)}\frac{1}{1+\frac{\rho}{1-\rho}(I-1)}}\frac{\rho_\eta\sigma_\eta^2}{\sigma^2},
\end{aligned}$$

which is decreasing in z. Similarly for κ

$$\begin{aligned}\frac{\partial e_i^*}{\partial \kappa}\frac{\kappa}{e_i^*} &= -(1-\rho)\lambda^*\kappa\left(\frac{1}{z}+\gamma\sigma^2\right)\\ &= -\frac{1-\rho}{\frac{1}{\kappa\left(\frac{1}{z}+\gamma\sigma^2\right)}+(1-\rho)+\frac{\gamma\rho\sigma^2}{\kappa\left(\frac{1}{z}+\gamma\sigma^2\right)}\frac{1}{1+\frac{\rho}{1-\rho}(I-1)}},\end{aligned}$$

which is increasing in z.

□

B Candidate Asset Pricing Methods

B.1 IPCA

We choose $K = 5$ for the number of factors in the IPCA, and we focus on the restricted model with no alphas. The characteristic-to-exposure mapping, $\Gamma_\beta \in \mathbb{R}^{N_t,K}$, and factor realizations, $f_{t+1} \in \mathbb{R}^K$, in a K-factor model solves the optimization problem

$$\min_{\Gamma_\beta,(f_{t+1})_{t=1}^{T-1}} \sum_{t=1}^{T-1} \|r_{i,t+1} - X_t\Gamma_\beta f_{t+1}\|_2^2, \tag{B.1}$$

where $\|\cdot\|_2$ is the L^2-norm and $X_t \in \mathbb{R}^{N_t\times P}$ is the matrix of stock-level characteristics for N_t stocks at time t. The parameters, Γ_β and f_{t+1}, are estimated following Kelly et al. (2019) using a 10-year rolling window. The factors in this restricted model can be thought of as managed portfolios

$$\hat{f}_{t+1} = (\hat{\Gamma}_\beta' X_t' X_t \hat{\Gamma}_\beta)^{-1}\hat{\Gamma}_\beta' X_t' r_{t+1},$$

and the rolling 10-year mean-variance optimal portfolio of these factors, $\hat{w}_f$, can then be readily computed. The final stock-level weights are

$$\hat{w}_t = X_t\hat{\Gamma}_\beta(\hat{\Gamma}_\beta' X_t' X_t \hat{\Gamma}_\beta)^{-1}\hat{w}_f.$$

We note that treating the number of factors, K, as a tunable hyperparameter could lead to further improvements in performance.

B.2 KNS

We construct a rank-weighted factor for each characteristic. We focus on factors created from single characteristics and avoid creating factors based on interactions. Including interaction factors can be problematic for the 153 JKP characteristics (i.e. $\frac{1}{2} \cdot 153 \cdot 152 + 3 \cdot 153 = 12,087$ factors). It is similarly computationally problematic for an expanded set of 402 characteristics, corresponding to 81,807 factors.

Rank-weighted factors are created by

$$rc_{i,t,p} = \frac{\text{rank}(x_{i,t,p})}{n_t + 1} \tag{B.2}$$

$$z_{i,t,p} = \frac{rc_{i,t,p} - rc_{t,p}}{\sum_{i=1}^{N_t} |rc_{i,t,p} - \bar{rc}_{t,p}|} \tag{B.3}$$

$$F_t = Z'_{t-1} R_t, \tag{B.4}$$

where $rc_{t,p} = \frac{1}{N_t} \sum_{i=1}^{n_t} rc_{i,t,p}$ and Z_{t-1} is a $N_t \times P$ matrix of the normalized ranks of characteristics $z_{i,t,p}$, P being the number of characteristics and N_t the number of assets at time t.

As in Kozak et al. (2020), we orthogonalize the rank-weighted factors with respect to the market by estimating their exposures β. To make the method more comparable to the others, we add the market factor to the set of rank-weighted factors.

Kozak et al. (2020) solves the penalized minimization problem

$$\min_w (\hat{\mu} - \hat{\Sigma} w)' \hat{\Sigma}^{-1} (\hat{\mu} - \hat{\Sigma} w) + \frac{\lambda_2}{2} \|w\|_2^2 + \lambda_1 \|w\|_1, \tag{B.5}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are the sample mean and covariance matrix of returns of the basis assets, and $\|\cdot\|_1$ is the L^1 norm. The basis assets are either the rank-weighted factors or the principal components of the factors.

We include only L^2 shrinkage when the rank-weighted factors are the basis assets, and

include L^1 shrinkage when working with principal components. Usually, L^1 shrinkage with multiple features precludes any closed-form solution. With principal components, however, a closed-form solution does exist since they are uncorrelated by construction

$$\tilde{w}_{PC} = \left(\hat{\Sigma}_{PC} + \lambda_2 I\right)^{-1} \hat{\mu}_{PC} \tag{B.6}$$

$$\hat{w}_{PC,p} = \text{sign}(\tilde{w}_p) \cdot (|\tilde{w}_p| - \lambda_1)^+, \tag{B.7}$$

where $\hat{w}_{PC,p}$ is the portfolio weight in the principal components p. The weights $\hat{w}_{PC}$ can be rotated back to weights in the rank-weighted factors.

We determine the hyperparameters λ_1 and λ_2 using the methodology outlined in section 3.2. Each month, we estimate the principal components for each fold to be used in the cross-validation when the principal components are the basis assets. The market exposures of each rank-weighted factor are also estimated for each fold. This ensures an unbiased estimate of the performance of the entire method pipeline using cross-validation. After picking a set of hyperparameters λ_1 and λ_2, the market exposures and principal components of the total rolling 10-year window are computed.

B.3 AP-Trees

Creating the basis assets. Bryzgalova et al. (2023) adapts the decision tree algorithm to a portfolio optimization context, and calls the resulting method "Asset Pricing Trees" (AP-Trees). AP-Trees are created from k characteristics and some depth D that controls how many layers of branching the tree has. The root node of the branch contains all stocks and the non-root nodes are created by splitting the higher nodes according to some stock characteristic. In a standard decision tree, each node is split once using a greedy algorithm that selects the characteristics and the value of that characteristic that will lead to the biggest gain in some objective function. In an AP-Tree, by contrast, each node is split k times—once for each characteristic—and the splitting point is fixed at the cross-sectional median within the node. Each node is then transformed into a portfolio by calculating the value-weighted excess return of stocks within the node, and the node-specific portfolio returns serve as the basis assets.

Concretely, suppose that we create an AP-Tree from $k = 2$ characteristics, say, market equity and book-to-market equity, with a depth of $D = 1$. The root node will contain all stocks. This node will then be split into two nodes containing stocks with market equity above and below the cross-sectional median and two nodes containing stocks with book-to-market equity above and below the cross-sectional median. By value-weighting the excess returns within each node, we get five portfolios that can serve as basis assets. The portfolio associated with the root node is the value-weighted market portfolio, while the portfolios associated with the non-root nodes capture, respectively, the returns of small (below median market equity), large (above median market equity), value (below median book-to-market equity), and growth (above median book-to-market equity) stocks.[19]

As in Bryzgalova et al. (2023), we weight each portfolio by $\frac{1}{\sqrt{2^{d_i}}}$ where d_i is the depth of the node i to account for the higher variance of deeper portfolios.

Limiting the computational complexity. An AP-Tree will produce $\sum_{d=0}^{D}(2k)^d$ basis assets. Bryzgalova et al. (2023) considers $k = 10$ and $D = 4$ in most of their analysis, so applying the aforementioned procedure would produce $168,421$ basis assets.[20] To handle the computational complexity, the authors instead create multiple AP-Trees for each combination of three characteristics. In addition, they require that market equity is one of the three characteristics. This approach produces $\frac{1}{2}(k-1)(k-2) = \frac{1}{2} \cdot 9 \cdot 8 = 36$ AP-Trees, each containing $\sum_{d=0}^{4}(2 \cdot 3)^4 = 1,555$ basis assets, for a total of 55,980 basis assets.

This method makes it feasible to handle ten characteristics, but it would require a lot of computational resources for the larger sets of characteristics that we consider. Specifically, we consider between 153 and 402 characteristics, which would require the creation of $11,628 \cdot$

[19] If we added a third characteristic ($k = 3$), say, past 12-month returns, we would add two more nodes containing stocks with a past 12-month return above and below the cross-sectional median. These portfolios would capture, respectively, the return of past winners and losers. If we instead added an additional layer ($D = 2$), then we would split each of the four non-root nodes into four additional nodes by repeating the procedure we used to split the root node. For example, the node containing stocks with a market equity below the cross-sectional median will be split into two nodes containing stocks with market equity above and below the conditional cross-sectional median, and two nodes containing stocks with book-to-market equity above and below the conditional cross-sectional median. By conditional cross-sectional median, we mean that the median is computed for the stocks in the node being split (in this example, the subset of stocks with a market equity below the unconditional cross-sectional median).

[20] The ten characteristics are (with names from the Jensen et al. (2023) data set in parentheses): size (`market_equity`), accruals (`oaccruals_at`), book-to-market equity (`be_me`), idiosyncratic volatility (`ivol_ff3_21d`), asset growth (`at_gr1`), long-term reversal (`ret_60_12`), turnover (`turnover_126d`), operating profitability (`ope_be`), momentum (`ret_12_1`), and short-term reversal (`ret_1_0`).

$1,555$ and $80,601 \cdot 1,555$ basis assets, respectively. In our implementation, we therefore limit ourselves to the same 10 characteristics analyzed in Bryzgalova et al. (2023).

Finding the SDF. For a sparse representation of the SDF, the authors "prune" nodes by solving the optimization problem

$$\min_{w} \quad \frac{1}{2}(\hat{\mu}^{robust} - \hat{\Sigma}^{robust}w)'(\hat{\Sigma}^{robust})^{-1}(\hat{\mu}^{robust} - \hat{\Sigma}^{robust}w) + \lambda_1\|w\|_1, \tag{B.8}$$

$$\text{with} \quad \hat{\mu}^{robust} = \hat{\mu} + \lambda_0\mu, \quad \hat{\Sigma}^{robust} = \hat{\Sigma} + \lambda_2 I_N, \tag{B.9}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are, respectively, the sample mean vector and the sample variance-covariance matrix of the basis assets, and μ is the average value in $\hat{\mu}$.

In the formulation from (B.8), λ_0, λ_1, and λ_2 are hyperparameters, but in Bryzgalova et al. (2023) the value of λ_1 is selected rather than optimized. Specifically, the authors chose λ_1 such that the number of non-pruned nodes taken *from each* AP-Tree is $K = 40$. We follow this approach, which means that we only do a formal hyperparameter search for λ_0 and λ_2.

We chose λ_0 and λ_2 by finding the values that lead to the highest Sharpe ratio on the validation data. We employ a technique from the literature on Bayesian hyperparameter optimization to lower time spent on hyperparameter tuning. Concretely, we use a Gaussian Process with an RBF kernel to model a density over functions mapping hyperparameters to the Sharpe ratio on the validation data. This probabilistic model is then used to determine the next set of hyperparameters by picking the hyperparameter set with highest e.g. 99% upper confidence bound (UCB) under the model.[21] Intuitively, regions of hyperparameters with high validation scores have high expected scores, while unexplored regions have high uncertainty, which can both lead to high UCBs. Thus, Bayesian optimization balances exploration of hyperparameter regions with high uncertainty and "exploiting" regions that look promising.

We follow Bryzgalova et al. (2023) and do 3-fold cross-validation using all data from before the start of the test period, 1990. We refrain from using a 10-year rolling window to

[21] The UCB algorithm needs a parameter κ denoting the number of standard deviations from the mean to evaluate the UCB at. We use a high $\kappa = 10$ to ensure adequate exploration. Several packages exist for Bayesian hyperparameter optimization, e.g. "optuna" for Python and "rBayesianOptimization" for R.

determine hyperparameters due to the high computational costs of pruning. [22]

With $K = 40$ and 36 AP-Trees, we have 1,440 non-pruned basis assets from this first stage of pruning. Some will be duplicate portfolios from different trees, e.g. the market portfolio, making the total number of basis assets somewhat lower than 1,440. We then do a second stage of pruning using the same methodology on the non-pruned basis assets from the first stage, except that we now also treat K as a hyperparameter.

We follow Bryzgalova et al. (2023) and use the hyperparameters λ_0 and λ_2 tuned from this second stage, as well as the basis assets that survived the pruning, to fit the mean-variance optimal portfolio:

$$\hat{w} = (\hat{\Sigma} + \lambda_2 I_N)^{-1}(\hat{\mu} + \lambda_0 \mu), \tag{B.10}$$

which is the solution to the minimization problem (B.8) when $\lambda_1 = 0$.

B.4 Markowitz-ML

We implement Markowitz-ML as in Jensen et al. (2024) with some modifications. The method solves for the mean-variance optimal portfolio using the classical method from Markowitz (1952):

$$\pi_t = \Sigma_t^{-1} \mu_t, \tag{B.11}$$

where π_t is the portfolio weight vector, μ_t is a the expected return vector, and Σ_t is the variance-covariance matrix.

We estimate expected returns by using the XGBoost model from Chen and Guestrin (2016) to predict realized excess returns over the next month as a function of security characteristics today. Our approach to selecting hyperparameters followed the same split into training, validation, and test data as in Section 3.2. We use a two-stage procedure for finding the hyperparameters. In the first stage, we train 20 different models on the training data based on the 20 sets of hyperparameters from Table B.1 and a fixed learning rate of 0.15.

[22] An alternative could be to only retune λ_0 and λ_2.

We select the set of hyperparameters that lead to the lowest mean squared error on the validation data. In the second stage, we train a new model on the training data using the parameters found in the first stage, except that we now train the model with a learning rate of 0.01. We then record the optimal number of iterations for the model (i.e., the number of individual decision trees to include in the ensemble before the model starts to overfit), which is the number of model iterations that resulted in the lowest mean squared error on the validation data. Finally, we train a model on the training and validation data using a learning rate of 0.01, the non-learning rate parameters from stage 1, and the optimal number of iterations from stage 2, and use this model to estimate μ_t.

To estimate the variance-covariance matrix, we use an approach similar to the one used by MSCI Barra, which is based on the assumption that returns follow a linear factor model. The idea is to treat security characteristics as observable factor loadings and infer the latent factor returns from cross-sectional regressions of excess returns on security characteristics. Specifically, each day we estimate the cross-sectional regression:

$$r_{i,t+1} = x'_{i,t}\hat{f}_{t+1} + \hat{\epsilon}_{i,t+1}, \tag{B.12}$$

where $r_{i,t+1}$ is stock's realized excess return, x is a vector of the stock's characteristics, $\hat{\epsilon}_{i,t+1}$ is the regression residual, and $\hat{f}_{t+1}$ is the estimated regression parameters, which we treat as estimated factor returns. The stock characteristics are either the 153 original, the 249 expanded ex-original, or the 402 expanded JKP features, plus 12 industry dummies (which is based on the 12 industry definition from Kenneth French's webpage).[23] The structure in (B.12) implies that the variance-covariance matrix is:

$$\hat{\Sigma}_t = X_t \text{Var}_t(\hat{f}_{t+1}) X'_t + \text{diag}(\text{Var}_t(\hat{\epsilon}_{t+1})), \tag{B.13}$$

where X_t is the matrix of stock characteristics, $\text{Var}(\hat{f}_{t+1})$ is the variance-covariance matrix

[23] Our approach differs slightly from Jensen et al. (2024), who uses a compressed version of 13 factor themes (made from the original JKP characteristics) plus the 12 industry dummies. We use the raw characteristics to enable the variance-covariance estimate to change as we change the set of characteristics. The Sharpe ratio of Markowitz-ML is similar but slightly stronger if we use the 13 factor themes instead of the 153, 249, or 402 raw characteristics.

Table B.1: XGBoost Hyperparameters

	Features	Tree depth	Sample size	Penalty	Learning rate
1	0.96	1	0.79	6.79	0.15/0.01
2	0.12	6	0.84	4.37	0.15/0.01
3	0.02	1	0.28	8.96	0.15/0.01
4	0.08	6	0.24	33.70	0.15/0.01
5	0.88	6	0.61	0.18	0.15/0.01
6	0.31	7	0.28	0.01	0.15/0.01
7	0.49	7	0.27	2.70	0.15/0.01
8	0.92	3	0.54	0.02	0.15/0.01
9	0.29	2	0.80	4.72	0.15/0.01
10	0.07	7	0.35	0.54	0.15/0.01
11	0.18	2	0.40	0.63	0.15/0.01
12	0.39	4	0.42	6.08	0.15/0.01
13	0.73	4	0.21	0.19	0.15/0.01
14	0.93	3	0.22	2.51	0.15/0.01
15	0.98	6	0.99	4.13	0.15/0.01
16	0.65	3	0.82	97.33	0.15/0.01
17	0.66	6	0.67	47.64	0.15/0.01
18	0.89	1	0.30	0.17	0.15/0.01
19	0.65	5	0.28	88.63	0.15/0.01
20	0.91	4	0.59	5.97	0.15/0.01

Note: The table shows 20 sets of hyperparameters that we search over when fitting the XGBoost model to predict next month's realized excess return. "Features" is the fraction of the features chosen randomly for each decision tree, "Tree depth" is the maximum depth of each decision tree, "Sample size" is the fraction of the observations chosen randomly for each decision tree, and "Penalty" is an L2 (ridge) penalty, and "Learning rate" is the weight each new tree gets in the ensemble. We use a two-stage tuning strategy, where the learning rate is 0.15 is the first stage, and 0.01 in the second. We get the hyperparameter sets by specifying a tolerable range for each hyperparameter and then use the `parameters` function from the dials package (https://dials.tidymodels.org/) with the type set to "max_entropy" to get 20 sets that aim to cover the associated parameter space. The ranges are features $\in [\frac{1}{\#features}, 1]$, tree depth $\in [1, 7]$, sample size $\in [0.2, 1]$, and penalty $\in [10^{-}2, 10^{2}]$. All parameters are chosen directly from their natural scales, except for penalty, which is chosen from a logarithmic (base 10) scale.

of factor returns, and $\text{diag}(\text{Var}(\epsilon_{t+1}))$ is a matrix with the idiosyncratic variances in the diagonal and zero elsewhere.

We estimate $\text{Var}(\hat{f}_{t+1})$ as the exponentially weighted sample variance-covariance matrix of the last ten years of daily returns. The exponential weighting scheme gives observations j days from t a weight of $w_{t-j} = c0.5^{j/\text{half-life}}$, where c is a constant ensuring that the weights sum to one, and it ensures that recent observations affect the estimate more than distant ones. Following the MSCI Barra's USE4S model, we use a half-life of 504 days for correlations and 84 days for variances (Menchero et al., 2011, Table 4.1). Similarly, we estimate each stock's idiosyncratic variance, $\text{Var}_t(\hat{\epsilon}_{i,t+1})$, as the exponentially weighted moving average of squared residuals, $\epsilon_{i,t+1}$ from (B.12) with a half-life of 84 days. The half-life is again chosen as the one from the MSCI Barra USE4S model (Menchero et al., 2011, Table 5.1). To estimate the idiosyncratic variance we require at least 200 non-missing observations within the last 252 trading days. For stocks without an idiosyncratic variance estimate, we estimate the function, $\ln\left(\sqrt{\hat{\text{Var}}_t(\hat{\epsilon}_{i,t+1})}\right) = \hat{f}_t(x_{i,t})$, using a ridge regression model with a small ridge penalty of 10^{-4}, and use it to etimate the missing variances.[24]

B.5 Factor-ML

The Factor-ML method uses the same expected return estimates as Markowitz-ML, and buys the 10% of stocks with the highest expected returns, while shorting the 10% of stocks with the lowest expected returns. Stocks are equal-weighted within the long and short portfolio.

B.6 Minimun Variance

The Minimum Variance method uses the same variance-covariance estimate as Markowitz-ML, and creates the portfolio weights as

$$\pi_t^{\text{MinVar}} = \frac{1}{1'\hat{\Sigma}_t 1}\hat{\Sigma}_t 1, \tag{B.14}$$

[24] Our approach follows MSCI Barra (Menchero et al., 2011, Eq. 5.3), except that they use an OLS regression. We use a ridge regression because some of our characteristics are highly correlated, and using a small penalty helps make the estimates robust to near multicollinearity.

where 1 is a conformable vector of ones.

B.7 Complex Factors

This method constructs many...

Didisheim et al. (2023) choose the 130 features with the fewest missing observations across the entire sample. We include the 130 features with the fewest missing observations in the training data before 1990 to avoid look-ahead bias. Didisheim et al. (2023) exclude stocks with more than 30% missing characteristics. This is an additional filter on the universe of stocks, so we exclude it.

We choose the lowest reported shrinkage $z = 10^{-5}$, refraining from any hyperparameter tuning. Instead of using a 30-year rolling window as in Didisheim et al. (2023), we use a 10-year rolling window to make the method more comparable to the other methods. In their paper, the performance appears stable from ∼36,000 factors, so we only construct 36,000 factors instead of their full 1,000,000 factors.

These choices make it easier to compare with the other methods while keeping the spirit of the original methodology. We note that the last 2-3 choices could easily be improved and further performance could be gained.

B.8 Factor models

We start by explaining how we create a single portfolio from a generic factor model with K pricing factors, and then explain how we create the specific pricing factors we consider. To create a single portfolio from a generic factor model, we draw inspiration from the SDF representation of a linear factor model. A linear factor model imply an SDF that is linear in the model's pricing factors, and, if the model correctly price all assets, then the mean-variance efficient combination of the pricing factors will lie on the efficient frontier (Cochrane, 2005). Therefore, we transform each factor model into a single portfolio by creating the mean-variance efficient combination of the model's pricing factors:

$$\pi_t = \Sigma_t^{-1} \mu_t, \tag{B.15}$$

Table B.2: Candidate factor models

Name	Paper	Pricing Factors
The CAPM	Sharpe (1964)	MKT
Fama-French-3	Fama and French (1993)	MKT, SMB^{FF3}, HML
Carhart-4	Carhart (1997)	MKT, SMB^{FF3}, HML, UMD
Fama-French-5	Fama and French (2015)	MKT, SMB^{FF5}, HML, RMW, CMA
Hou-Xue-Zhang-4	Hou et al. (2015)	MKT, SMB^{HXZ}, ROE, AG
Stambaugh-Yuen-4	Stambaugh and Yuan (2017)	MKT, SMB^{SY}, MGMT, PERF
Daniel-Hirshleifer-Sun-4	Daniel et al. (2020)	MKT, FIN, PEAD

Note: The table shows the candidate factor models we consider, the paper reference, and the pricing factors used by the model. The construction of the pricing factors is described in Section B.8.

where Σ_t is the $K \times K$ variance-covariance matrix and μ_t is the $K \times 1$ expected excess return vector of the model's pricing factors. We estimate Σ_t and μ_t each month with their sample counterparts over the past ten years.

The factor models we consider are in Table B.2, along with their pricing factors. Some of the pricing factors are present in multiple models, so we explain each of the unique pricing factors separately. The pricing factors are created solely on stocks that survive our data filters, which is generally speaking a smaller and more liquid set of stocks than what was used in the original paper underlying the factor models. Text formatted as `<text>` refers to the name of the characteristic in the data set from Jensen et al. (2023).

The market factor (MKT). The return of the market factor is the value-weighted average of all stocks included in the test data in a specific month.

The value factor (HML). We follow the construction from Fama and French (1993). Specifically, we independently assign stocks to two size groups and three book-to-market groups. The size breakpoint is the median market equity (`market_equity`) among NYSE stocks. The book-to-market breakpoints are the 30^{th} and 70^{th} percentiles of book-to-market equity (`be_me_ff`) among NYSE stocks.[25] The intersection of these groups creates six non-overlapping portfolios of stocks, and we compute the resulting portfolio return using value weights. The value factor is the average return on the two portfolios with high book-to-

[25]The Jensen et al. (2023) data set contains two book-to-market factors: `be_me_ff` and `be_me`. The `be_me` characteristic is updated every month using the most market equity and book equity, which follows the construction from Asness and Frazzini (2013). The `be_me_ff` characteristic is only updated once every year in June using the market equity from June in year t and the most recent book equity with fiscal year end in $t-1$, which more closely follows the construction from Fama and French (1993).

market ratios minus the average return on the two portfolios with low book-to-market ratios.

The momentum factor (UMD). The momentum factor is created like the HML factor, except that the non-size sorting variable is a stock's past return over the last 12 months skipping the most recent month (`ret_12_1`).

The operating profitability factor (RMW). We follow the construction from Fama and French (2015). The operating profitability factor is created like the HML factor, except that the non-size sorting variable is a stock's operating profitability (`ope_be`).

The asset growth factor (AG). We follow the construction from Fama and French (2015) and Hou et al. (2015). The asset growth factor is created like the HML factor, except that the non-size sorting variable is the negative of a stock's asset growth over the last year (`at_gr1`). Said differently, the factor is long stocks with low asset growth and short stocks with high asset growth.

The return on equity factor (ROE). We follow the construction from Hou et al. (2015). The return on equity factor is created like the HML factor, except that the non-size sorting variable is a stock's return on equity over the most recent quarter (`niq_be`).

The management mispricing factor (MGMT). We follow the construction from Stambaugh and Yuan (2017). The management mispricing factors is created like the HML factor, except that the non-size sorting variable is a composite measure of mispricing characteristics under the managements control (`mispricing_mgmt`), and that the breakpoints are the 20^{th} and 80^{th} percentile.

The performance mispricing factor (PERF). We follow the construction from Stambaugh and Yuan (2017). The performance mispricing factor is created like the HML factor, except that the non-size sorting variable is a composite measure of mispricing characteristics related to the firm's performance (`mispricing_perf`), and that the breakpoints are the 20^{th} and 80^{th} percentile.

The post-earnings announcement drift factor (PEAD). We follow the construction in Daniel et al. (2020). The post-earnings announcement drift factor is created like the HML factor, except that the second sorting variable is a stock's return in excess of the market return on the day before, the day of, and the two days after its most recent earnings announcement date (`ear`), and that the breakpoints are the 20^{th} and 80^{th} percentile.

The equity financing factor (FIN). We follow the construction in Daniel et al. (2020). The equity financing factor is created based on a separate sort of two characteristics. First, stocks are sorted into three groups based on a 5-year composite share issuance measure (`csi_60m`) with the breakpoints equal to the 20^{th} and 80^{th} percentile among NYSE stocks. We refer to these groups as the CSI groups. Second, stocks are sorted into three groups based on their 1-year net stock issuance (`chcsho_12m`). We refer to these groups as the NSI groups. The low NSI group is created by taking stocks that repurchase shares (i.e., stocks with a negative `chcsho_12m`) and selecting the ones with a repurchasing rate above the NYSE median. The high NSI group is created by taking stocks that issue shares, and selecting the once with an issuing rate above the 70^{th} percentile of NYSE stocks. The remaining stocks are in the middle NSI group. Finally, stocks are assigned to one of three financing groups. High financing stocks are those in the high groups for both NSI and CSI, or in the high group of NSI with CSI is missing, or in the high group of CSI with NSI is missing. Low financing stocks are those in the low groups for both NSI and CSI, or in the low group of NSI with CSI is missing, or in the low group of CSI with NSI is missing. We then create six portfolios based on the interaction between the financing groups and two size groups (above or below the NYSE median). The financing factor is the average of the two low-financing portfolios minus the average of the two high-financing portfolios.

The size factors (SMB^{m}). All models except Daniel-Hirshleifer-Sun-3 include a size factor, but the construction of the size factor is not consistent across the models. The SMB^{FF3} factor is built from the six portfolios constructed for the HML factor, as the average return on the three portfolios with low market equity minus the average return on the three portfolios with high market equity. The SMB^{FF5} factor is the average return on the small portfolios created for the HML, RMW, and AG factor minus the average return on the large portfolios based on the same factors. The SMB^{HXZ} factor is the average return on the small portfolios created for the ROE and AG factors minus the average return on the large portfolios based on the same factors. The SMB^{SY} factor is the average return on the small portfolios created for the MGMT and PERF factors minus the average return on the large portfolios based on the same factors.

C Additional Empirical Results

C.1 Statistical Tests Related to Figure 2

Table C.1 regresses the returns on all methods on a constant to test whether any of the methods receive a statistically significant weight in the in-sample mean-variance optimal portfolio, following the method proposed by Britten-Jones (1999). Table C.2 repeats the test after imposing a non-negativity constraint on the parameters, and the confidence interval is computed using a bootstrap estimator. Figure C.1 regresses the returns of one method on the returns of another, and reports the intercept estimate and whether it is statistically significant at conventional levels.

Table C.1: **SDF Test for Out-of-Sample Returns of Portfolio Methods.** In-sample test of the portfolio weight of each portfolio method in an SDF combining all methods. Weights are normalized to sum to 1, with t-statistics reported in parenthesis. Results are for mega-, large- and small-cap US stocks for 1990-2023. A 10-year rolling window has been used for estimation for all methods.

	SDF Weights
AP-Trees	0.147^{*}
	(1.804)
CAPM	0.122
	(1.323)
CH4	-0.032
	(-0.094)
Complex Factors	0.417^{***}
	(4.497)
DHS3	0.030
	(0.380)
FF3	-0.092
	(-0.329)
FF5	0.347
	(0.717)
FF6	-0.314
	(-0.596)
HMXZ5	-0.149
	(-1.119)
HXZ4	0.084
	(0.634)
IPCA	0.210^{**}
	(2.314)
KNS L2	-0.122
	(-1.228)
KNS PC	0.126
	(1.225)
Markowitz-ML	0.251^{***}
	(3.120)
SY4	-0.028
	(-0.271)

Table C.2: **Mean-Variance Optimal Combined Portfolio with No Shorting.** A mean-variance optimal portfolio of the portfolio methods' out-of-sample returns is constructed with a no-shorting constraint. Weights are normalized to sum to 1. Bootstrap 95%-percentile intervals with 100,000 samples are reported in parenthesis. Whether the 90%-, 95%-, or 99%-percentile intervals include 0 is indicated with 1, 2, or 3 stars, respectively. Results are for mega-, large- and small-cap US stocks for 1990-2023. A 10-year rolling window has been used for estimation for all methods.

	SDF Weights
AP-Trees	0.117*
	(0.00, 0.25)
CAPM	0.093
	(0.00, 0.19)
CH4	0.000
	(0.00, 0.00)
Complex Factors	0.419***
	(0.23, 0.57)
DHS3	0.000
	(0.00, 0.08)
FF3	0.000
	(0.00, 0.06)
FF5	0.000
	(0.00, 0.11)
FF6	0.000
	(0.00, 0.00)
HMXZ5	0.000
	(0.00, 0.00)
HXZ4	0.000
	(0.00, 0.01)
IPCA	0.140*
	(0.00, 0.27)
KNS L2	0.000
	(0.00, 0.00)
KNS PC	0.017
	(0.00, 0.13)
Markowitz-ML	0.213***
	(0.06, 0.35)
SY4	0.000
	(0.00, 0.00)

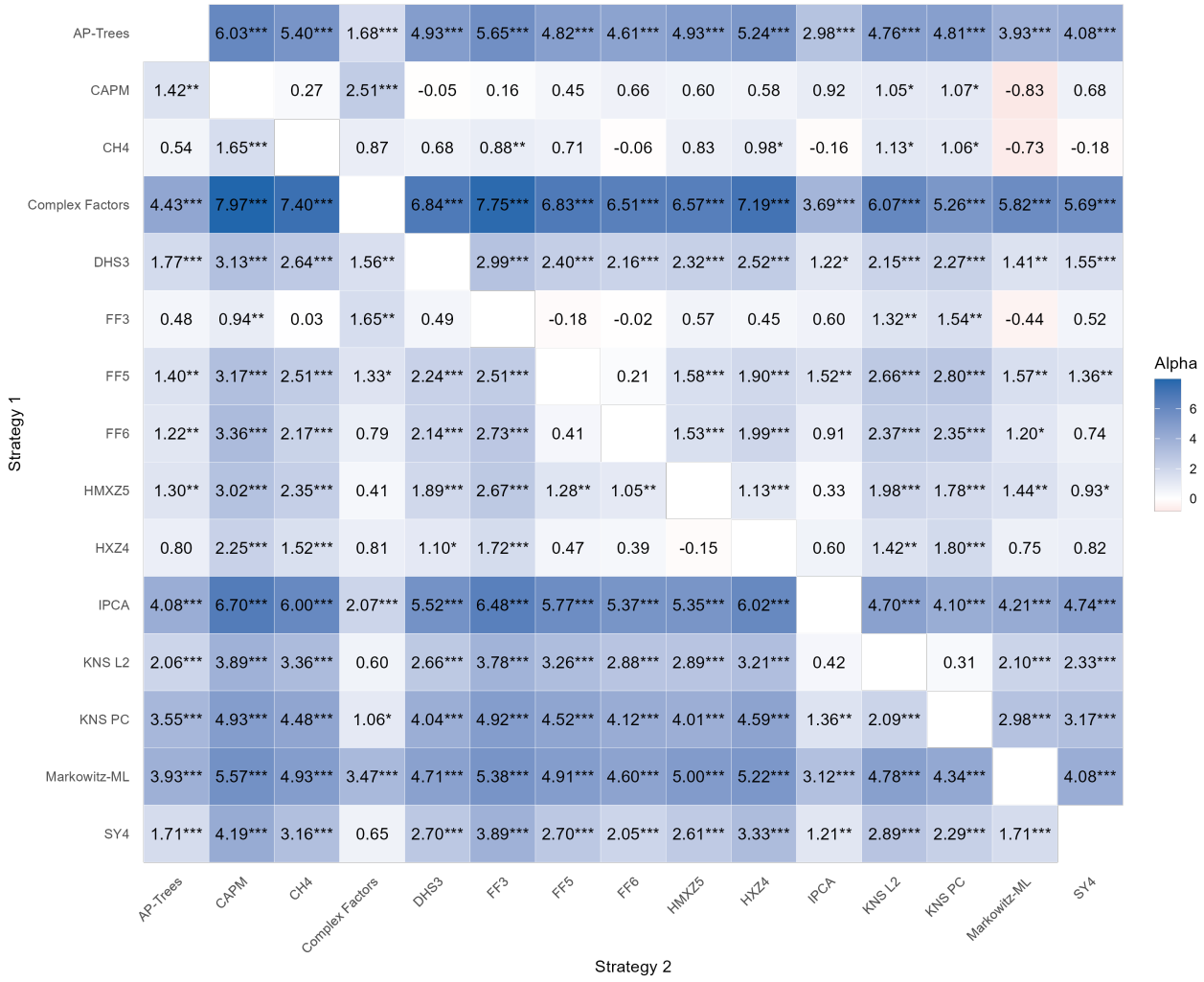


Figure C.1: **Pairwise alpha between Returns of Portfolio Methods.** Mean-variance optimal portfolio methods are implemented on JKP data. For each strategy (Strategy 1), their annualized alpha to another strategy (Strategy 2) is computed. The reported matrix consists of each pairwise alpha along with their statistical significance. Results are for mega-, large-, and small-cap US stocks. The test data period is 1990-2023, and a 10-year rolling window is used for all methods for estimation.