Diverging roads: Theory-based vs. machine learning-implied stock risk premia

Joachim Grammig[1], Constantin Hanenberg[2], Christian Schlag[3], and Jantje Sönksen[4 †]

October 21, 2022

Abstract

We assess financial theory-based and machine learning methods to quantify stock risk premia and investigate the potential of hybrid strategies. The results indicate that at the one-month investment horizon, a theory-based approach using option prices is preferable, especially if risk premium estimates get updated at high frequencies. At the one-year horizon, a random forest with sufficiently long training delivers a better performance than option-based models. The integration of machine learning procedures to address the approximation errors of a theory-based approach is identified as a novel and promising hybrid strategy.

Key words: stock risk premia, option prices, machine learning
JEL: C53, C58, G12, G17

[1] University of Tübingen, Department of Statistics, Econometrics and Empirical Economics and Centre for Financial Research (CFR), Cologne. `joachim.grammig@uni-tuebingen.de`

[2] University of Tübingen, Department of Statistics, Econometrics and Empirical Economics. `constantin.hanenberg@uni-tuebingen.de`

[3] Goethe University Frankfurt and Leibniz Institute for Financial Research SAFE. `schlag@finance.uni-frankfurt.de`

[4] University of Tübingen, Department of Statistics, Econometrics and Empirical Economics. `jantje.soenksen@uni-tuebingen.de`

[†] Earlier versions of this paper were presented at the 48th Annual Meeting of the European Finance Association (EFA), the 12th Econometric Society World Congress, the 13th Annual Conference of the Society for Financial Econometrics, and several other conferences and seminars. We thank the participants, and in particular, Michael Bauer, Svetlana Bryzgalova (EFA discussant), Emanuele Guidotti, Christoph Hank, Jens Jackwerth, Alexander Kempf, Michael Kirchler, Christian Koziol, Michael Lechner, Marcel Müller, Elisabeth Nevins, Yarema Okhrin, Olaf Posch, Éric Renault, Olivier Scaillet, Julie Schnaitmann, Grigory Vilkov, Florian Weigert, and Dacheng Xiu for many helpful comments. We acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grants GR 2288/7-1, SCHL 558/7-1, and INST 35/1134-1 FUGG. Christian Schlag acknowledges general research support by SAFE.

1 Introduction

When it comes to measuring stock risk premia, two roads diverge in the finance world – or at least, so it may seem to a student of recent literature on empirical asset pricing. Two prominent studies exemplify this impression: Martin and Wagner (2019) quantify the conditional expected return of a stock by exploiting the information contained in current option prices, as implied by financial economic theory.[1] Gu et al. (2020) pursue the same end but along a completely different path, leveraging the surge of machine learning applications in economics and finance, together with advances in computer technology.[2] Approaches similar to the one adopted by Martin and Wagner (2019) derive results from asset pricing paradigms and have no need of historical data to quantify stock risk premia; Gu et al. (2020) and related papers instead do not refer substantially to financial economic theory and prefer to "let the data speak for themselves."

These radically different ways to address the same issue motivate us to conduct a fair, comprehensive performance comparison of theory-based and machine learning approaches to measuring stock risk premia and to explore the potential of hybrid strategies. The comparison is based on the fact that the risk premium is the conditional expected value of an excess return and that, in the present context, the machine learning objective is to minimize the mean squared forecast error (MSE). Because the conditional expectation is the best predictor in terms of MSE, it seems natural to compare the opposing philosophies by gauging the quality of

[1] Their strategy to quantify the risk premia of financial assets draws on Martin's (2017) derivation of a lower bound for the conditional expected return of the market, which in turn is based on concepts outlined by Martin (2011). Kadan and Tang (2020) take up Martin's (2017) idea and argue that it can be applied to quantify risk premia for a certain type of stocks. Bakshi et al. (2020) propose an exact formula for the expected return of the market that relies on all risk-neutral moments of returns. In a similar vein, Chabi-Yo et al. (2021) consider bounds for expected excess stock returns that take into account higher risk-neutral moments using calibrated preference parameters.

[2] Recent studies in a similar vein include those by Light et al. (2017), Martin and Nagel (2021), and Freyberger et al. (2020).

their excess return forecasts: A superior forecast indicates a better approximation of the risk premium. Such a comparative analysis can reveal whether the use of the information theoretically embedded in current option prices is preferable to sophisticated statistical analyses of historical data, or vice versa.

Beyond this direct comparison, we also investigate the potential of hybrid strategies that combine the theory-based and machine learning paradigms. In particular, we rely on machine learning to address the approximation errors of the theory-based approach. These residuals are functions of moments conditional on time t information, and machine learning is employed to approximate the conditional moments using time t stock- and macro-level variables. We refer to this strategy as *theory assisted by machine learning*. We also consider a machine learning approach that includes theory-implied risk premium measures computed from current option data, along with historical stock- and macro-level feature data. To ensure a fair comparison we deliberately adhere to the model specifications used in the base papers, for example regarding the features considered and the training and validation strategy adopted for machine learning.

To level the playing field, we need data for which both theory-based and machine learning approaches are applicable. For our large-scale empirical study, we use data on the S&P 500 constituents from 1964 to 2018, including firm- and macro-level variables, as well as return and option data. The analysis centers on theory-based and machine-learning-implied estimates of stock risk premia, computed at one-month and one-year investment horizons. We focus on the machine learning methods that Gu et al. (2020) identify as most promising, namely, an ensemble of artificial neural networks (ANN), gradient boosted regression trees (GBRT), and random forests (RF). We also include the elastic net (ENet), as a computationally less demanding benchmark. We consider two training and validation strategies, starting in 1974 (*long* *training*) and 1996 (*short training*), respectively. Using the short training scheme is necessary for all hybrid approaches, because the option data are not available earlier.

The main results are as follows: Of the two theory-based approaches that we consider, the one proposed by Martin and Wagner (2019) (henceforth, MW) is preferable to Kadan and Tang's (2020) approach (henceforth, KT). At the one-month horizon, MW is also superior to three of the four machine learning methods. Only MW and the ANN deliver a positive predictive R^2 of comparable size, according to the analyses that use forecasts issued at the end of each month. When using risk premium estimates at a daily frequency, the predictive R^2 by MW increases from 0.2% to 0.9%. Adapting the machine learning models to deliver daily risk premium estimates improves their performance, but it does not match that of MW; the best machine learning result is achieved by the ANN, with a predictive R^2 of 0.5%. We note that among all the machine learning approaches and stock universes considered by Gu et al. (2020), the highest reported predictive R^2 is 0.7%; the one-month horizon is a low signal-to-noise environment. Constructing prediction-sorted portfolios, we find that the alignment of predicted and realized mean excess returns works better and the cross-sectional variation of mean realized returns across prediction-sorted portfolios is highest when using MW. The advantage of the theory-based paradigm at the one-month horizon is confirmed by a complementary analysis in which we apply Chabi-Yo et al.'s (2021) option-based method to approximate stock risk premia.

The signal-to-noise ratio increases at the one-year horizon. ANN and GBRT achieve predictive R^2 around 9%, very similar to MW. While ENet and KT are less successful, the RF delivers the highest annual predictive R^2 of about 19%. The analysis of the alignment and cross-sectional variation of prediction-sorted portfolios also provides corroborative evidence. To achieve this performance, the RF relies on the long training scheme. Generally, the performance of machine learning approaches is attenuated when using a short training scheme, but hybrid strategies can compensate for this drawback. A theory assisted by machine learning strategy that takes MW as a basis and trains an RF or an ANN to deal with the approximation errors implied by the theory-based formula is particularly successful. The assistance by the RF increases the predictive R^2 delivered by MW from 9% to 16%. The analysis of prediction-sorted portfolios further establishes the expediency of this hybrid approach: It produces the best alignment and highest variation of the mean realized excess returns across the prediction-sorted portfolios. The MW+RF and MW+ANN combinations answer critiques of machine learning as measurement without theory, because they reflect financial economic paradigms and employ statistical assistance only for the components that remain unaccounted for by theory.

When risk premia need to be estimated at a daily frequency, the theory-based methods offer a natural advantage. The required option data are available at a daily frequency, whereas many stock- and all macro-level features are updated monthly at best. However, we find that a modified hybrid strategy that uses daily updated theory-based features for an RF, trained using end-of-month data, does a good job providing daily risk premium estimates. The annual predictive R^2 of the RF without theory-based features and evaluated at a daily frequency is 9%. Including theory-based features doubles this value.

Further analysis reveals that the importance of firm- and macro-level features does not differ markedly across the two applications of the RF, that is, its pure usage or when assisting the theory-based approach. At the one-year horizon, the familiar firm-level return predictive signals are most important in both applications: the book-to-market ratio, liquidity-related indicators, and momentum variables (in that order). The dominance of the short-run price reversal at the one-month horizon vanishes at the one-year horizon. The importance of the Treasury bill rate (a macro-level predictor) in both applications supports the use of short-term interest rates as state variables in variants of the intertemporal capital asset pricing model. The benefits of theory assistance by machine learning are also corroborated by a disaggregated analysis, for which we create portfolios by sorting stocks according to valuation ratios, liquidity variables, momentum indicators, and industry affiliation.

Overall, these results indicate the usefulness of hybrid strategies that combine theory-based and machine learning methods for quantifying stock risk premia. In this respect, the present study complements recent literature that links machine learning with theory-based empirical asset pricing and for which Giglio et al. (2022) provide a comprehensive survey and guide. For example, Gu et al. (2021) note that a focus of machine learning on prediction aspects does not constitute a genuine asset pricing framework, so they propose using a machine learning method (autoencoder) that takes account of the risk-return trade-off directly. Chen et al. (2021) use the results reported by Gu et al. (2020) as a benchmark and find that the inclusion of no-arbitrage considerations improves the empirical performance. In another combination of theory and data science methods, Wang (2018) employs partial least squares to account for higher risk-neutral cumulants when modeling stock risk premia. Kelly et al. (2019) use an instrumented principle components analysis to construct a five-factor model that spans the cross-section of average returns, and Kozak et al. (2020) use penalized regressions to shrink the coefficients on risk factors in the pricing kernel. Bryzgalova et al. (2021) generalize this idea and use decision trees to construct a set of base assets that span the efficient frontier. In their attempt to address the plethora of factors described in recent asset pricing literature, Feng et al. (2020) combine two-pass regression with regularization methods. In what might be considered a broad reality check, Avramov et al. (2021) take a practitioner's perspective and assess the advantages and limitations of the aforementioned approaches. Although our study is related to this strand of literature in the general sense of combining financial economic theory with machine learning, our focus is on using this framework for approximating conditional stock risk premia. We do not aim at providing hybrid approaches for the purpose of recovering the stochastic discount factor explicitly and then predict stock excess returns. Rather, our strategy to use machine learning to deal with the approximation errors inherent to the theory-based approach could be viewed as an exercise of predicting risk-adjusted returns or being related to the notion of boosting.

The remainder of the paper is structured as follows: Section 2 contrasts theory-based and machine learning methodologies for measuring stock risk premia, then outlines ideas to combine them. Section 3 explains the construction of the database and the implementation of the respective strategies. Section 4 contains a performance comparison between theory-based and machine learning methods at varying horizons and the assessment of the potential of hybrid strategies. Section 5 concludes. An appendix and online appendix provide details on methodologies, data, and implementation.

2 Methodological considerations

2.1 Two diverging roads

This section outlines the concepts and key equations associated with the theory-based and machine learning approaches that are the focus of our study. We explain how, from a common starting point, the methodologies to measure stock risk premia diverge. For conciseness, the details of the respective approaches are presented in the Appendix.

The theory-based approach (explicitly) and the machine learning approach (implicitly) take as a point of reference the basic asset pricing equation applied to a gross return of asset i from time t to T ($R^i_{t,T}$) in excess of the gross risk-free rate ($R^f_{t,T}$),

$$\mathbb{E}_t(R^{ei}_{t,T}) = \mathbb{E}_t(R^i_{t,T}) - R^f_{t,T} = -R^f_{t,T} \cdot \mathrm{cov}_t(m_{t,T}, R^i_{t,T}), \tag{2.1}$$

where expected values are conditional on time t information. In preference-based asset pricing, the stochastic discount factor (SDF) $m_{t,T}$ represents the marginal rate of substitution between consumption in t and T. In the absence of arbitrage, a positive SDF exists, such that $R^f_{t,T} = \mathbb{E}_t(m_{t,T})^{-1} > 0$. The sign and size of the risk premium, reflected in the conditional expected excess return on asset i, are determined by the conditional covariance on the right-hand side of Equation (2.1).

Theory-/option-based approach

We first take a look down the theory-based route. Using Equation (2.1) as a starting point, we delineate in Appendix A.1 how Martin and Wagner (2019) derive the following reformulation:

$$\mathbb{E}_t(R^{ei}_{t,T}) = R^f_{t,T} \cdot \left\{ \mathrm{var}^*_t\left(\frac{R^m_{t,T}}{R^f_{t,T}}\right) + \frac{1}{2} \cdot \left[\mathrm{var}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) - \sum_j w^j_t \cdot \mathrm{var}^*_t\left(\frac{R^j_{t,T}}{R^f_{t,T}}\right)\right]\right\} + a^i_{t,T}, \tag{2.2}$$

where R^m denotes the return of a market index proxy, w^j_t is the time-varying value weight of index constituent j, var^*_t denotes a conditional variance under the risk-neutral measure, and $a^i_{t,T}$ is a time-varying, asset-specific component that, as shown in Appendix A.1, is a function of conditional moments either under the risk-neutral or the physical measure. In a similar vein, Kadan and Tang (2020) advocate an even more succinct formula:

$$\mathbb{E}_t(R^{ei}_{t,T}) = \frac{1}{R^f_{t,T}} \cdot \mathrm{var}^*_t(R^i_{t,T}) - \xi^i_{t,T}, \tag{2.3}$$

where $\xi^i_{t,T} = \text{cov}_t(m_{t,T} \cdot R^i_{t,T}, R^i_{t,T})$. In Appendix A.1, we show how Kadan and Tang (2020) draw on Martin's (2017) derivation of a lower bound for the market equity premium. They argue that, depending on the acceptable level of risk aversion, $\xi^i_{t,T} < 0$ holds for a large fraction of stocks, such that $1/R^f_{t,T} \cdot \text{var}^*_t(R^i_{t,T})$ represents a lower bound for the risk premium.

According to Martin (2017), the risk-neutral variances in Equations (2.2) and (2.3) can be obtained as follows (suppressing the asset index i for notational brevity):

$$\text{var}^*_t\left(\frac{R_{t,T}}{R^f_{t,T}}\right) = \frac{\int_0^{F_{t,T}} \text{put}_{t,T}(K)dK + \int_{F_{t,T}}^{\infty} \text{call}_{t,T}(K)dK}{0.5 \cdot S^2_t \cdot R^f_{t,T}}, \tag{2.4}$$

where $\text{call}_{t,T}(K)$ and $\text{put}_{t,T}(K)$ denote the time t prices of European call and put options, respectively, with strike price K and time to maturity T. Furthermore, S_t is the spot price, and $F_{t,T}$ is the forward price of the underlying asset. The components of the right-hand sides of Equations (2.2) and (2.3), except for the residuals $a^i_{t,T}$ and $\xi^i_{t,T}$, can be approximated using current option prices for a sufficient number of strikes. For Equation (2.3), these data are only required for asset i. Equation (2.2) is more demanding, in that the option data must be provided for both the market index proxy and its constituents, along with the time-varying index weights. Martin and Wagner (2019) argue that the consequences of setting $a^i_{t,T} = 0$ should be benign, such that stock risk premia can be quantified without the need to estimate any unknown parameters, by using:

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx R^f_{t,T}\left\{\text{var}^*_t\left(\frac{R^m_{t,T}}{R^f_{t,T}}\right) + \frac{1}{2} \cdot \left[\text{var}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) - \sum_j w^j_t \cdot \text{var}^*_t\left(\frac{R^j_{t,T}}{R^f_{t,T}}\right)\right]\right\}. \tag{2.5}$$

Similarly, assuming that the negative correlation condition holds and that the lower bound in Equation (2.3) is binding, Kadan and Tang's (2020) approximative formula for the risk premium on stock i is given by:

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx \frac{1}{R^f_{t,T}} \cdot \text{var}^*_t(R^i_{t,T}). \tag{2.6}$$

Machine learning approach

Recalling that the conditional expectation is the best predictor in terms of MSE, Equation (2.1) states that the MSE-optimal forecast of $R^{ei}_{t,T}$ is given by $-R^f_{t,T} \cdot \text{cov}_t(m_{t,T}, R^i_{t,T})$. Because the functional form of the conditional covariance is not known, one can treat $-R^f_{t,T} \cdot \text{cov}_t(m_{t,T}, R^i_{t,T})$ as a function that depends on state variables $z^i_t \in \mathcal{F}_t$, such that

$$\mathbb{E}_t(R^{ei}_{t,T}) = g^0_T(z^i_t), \tag{2.7}$$

where the subindex T indicates dependence on the horizon of interest. The machine learning approach then proceeds to approximate $g^0_T(z^i_t)$ by $g_T(z^i_t, \theta_T)$, a parametric function implied by some statistical model with a parameter vector θ_T to be estimated. The estimation of θ_T using machine learning procedures (MLPs) instead of standard econometric methods may be advocated for the following reasons.

First, there are a lot of candidates for the state variables z^i_t. A myriad of stock- and macro-level return predictive signals (*features* in machine learning terms) appear in empirical finance literature, and dimension reduction and feature selection are the very domain of MLPs. Second, the suite of statistical models employed for MLPs trade analytical tractability and rigorous statistical inference for flexible functional forms and predictive performance. The prediction implications of the basic asset pricing equation (2.1) naturally establish a learning objective, that is, minimization of the forecast MSE. However, the combination of these two issues – many features and a desire for flexibility – creates a vast risk of overfitting. To deal with this concern, MLPs divide the data into a training, a validation, and a test sample and introduce regularization in the estimation process. Regularization is controlled by the tuning of hyperparameters, which might take the form of a penalty applied to the learning objective, early stopping rules applied to its optimization, or, more generally, coefficients that determine the complexity of the statistical model (e.g., number of layers in an ANN). Using a given combination of hyperparameters, the parameter vector θ_T is estimated on the training sample, and the model performance gets evaluated, in terms of forecast MSE, on the validation sample. A search across hyperparameter combinations ultimately points to the specification that delivers the best performance. Using the hyperparameter combination thus selected, θ_T is re-estimated on the merged training/validation sample. The result is the final estimated model, $g_T(z_t^i, \hat{\theta}_T)$, which is used as a machine learning-implied approximative risk premium,

$$\mathbb{E}_t(R_{t,T}^{ei}) \approx g_T(z_t^i, \hat{\theta}_T). \tag{2.8}$$

Machine learning encompasses a variety of statistical models that offer flexible approximations of $g_T^0(z_t^i)$. In this study, we consider an ENet, GBRT, RF, and ANN. We discuss the associated hyperparameter configurations in Section 3.2.

2.2 Pros and cons

As far as the empirical implementation is concerned, the theory-based and data science approaches have their own unique pros and cons.

Parameter estimation and approximation errors

Using the theory-based formulas in Equation (2.5) or (2.6) and working under the risk-neutral measure, one can dispense with the estimation of unknown model parameters altogether. However, this parsimony of the theory-based approach comes at the cost of approximation errors, the practical consequences of which are not quite clear. In contrast, the machine learning approach deals with a huge number of parameters, which must be estimated without the risk of overfitting.

Time-varying parameters

A conspicuous feature of the theory-based approach is that it can deal naturally with changing conditional distributions and even non-stationary data. The machine learning approach, like any statistical/econometric method, struggles more with ensuing problems like an incidental parameter problem that would occur if the parameters in θ_T were time-varying. This caveat can be accounted for by employing a dynamic procedure, in which the training sample is gradually extended and the validation and test sample are shifted forward in time. (Hyper-)parameter estimation is performed for each of these "sample splits." Compared with Equation (2.8), it is thus notationally more precise, albeit more cluttered, to write

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx g_{s,T}(z^i_t, \hat{\theta}_{s,T}), \tag{2.9}$$

indicating the dependence of the functional form and estimates on the sample split s and investment horizon T.

Data quality and computational resource demands

The demands for data quality and quantity in both the theory-based and machine learning strategies are considerable, distinct, and complementary. The machine learning approach needs historical data on stock-level predictors for every asset of interest. A critical aspect is that these data suffer from a missing value problem that is most severe in the more distant past. As pointed out by Freyberger et al. (2021), the imputation of those observations is not innocuous and may hamper the application of data-intensive machine learning methods. This issue is mitigated using theory-based approaches. However, both MW and KT require high quality option data. In particular, for the option prices, the times-to-maturity must match the horizons of interest, and only a sufficiently large number of strike prices K can provide a good approximation of the integrals in Equation (2.4). Moreover, Equation (2.5) reveals that these data are required for not only the stocks of interest but also every member of the market index, as well as the index itself.

An advantage of the option-based approaches is that the computational resources needed to provide quantifications of stock risk premia are moderate. Machine learning approaches instead mandate ready access to considerable computing power. Training and hyperparameter tuning are required for each statistical model, for each horizon of interest, and for every new test sample.

2.3 Hybrid approaches

Because of the diversity of their respective pros and cons, it is intriguing to combine the theory-based and machine learning philosophies. Our primary hybrid approach is based on MW; it starts from Equation (2.2) and the approximative formula in Equation (2.5) and then employs machine learning to account for the approximation residuals $a^i_{t,T}$.[3] Let us use $\widetilde{\mathbb{E}}_t(R^{ei}_{t,T})$ to denote the right-hand side of Equation (2.5). Then $\widetilde{R}^{ei}_{t,T} = R^{ei}_{t,T} - \widetilde{\mathbb{E}}_t(R^{ei}_{t,T})$ gives the component of the excess return left unexplained by MW. Provided that the aforementioned data requirements are met, $\widetilde{R}^{ei}_{t,T}$ can be computed for every i, t, and T. Emphasizing the prediction aspect of the basic asset pricing equation, we consider the following decomposition:

$$\widetilde{R}^{ei}_{t,T} = a^i_{t,T} + \varepsilon^i_{t,T}, \tag{2.10}$$

[3] Alternatively, we could also use KT as a starting point, but MW is arguably more appropriate for a larger number of stocks.

where $\varepsilon^i_{t,T} = R^{ei}_{t,T} - \mathbb{E}_t(R^{ei}_{t,T})$ can be conceived of as the irreducible idiosyncratic forecast error. We can now apply the MLPs not to $R^{ei}_{t,T}$ and $\mathbb{E}_t(R^{ei}_{t,T})$ but rather to $\widetilde{R}^{ei}_{t,T}$ and $a^i_{t,T}$. This is a sensible approach because the approximation residual $a^i_{t,T}$ is a function of time t conditional moments, as is shown in Appendix A.1. Similar to the treatment of $g^0_T(z^i_t)$ in Equation (2.7), we can represent $a^i_{t,T}$ as a function of the time t state variables z^i_t, such that $a^i_{t,T} = h^0_T(z^i_t)$, and use a statistical model with parameters ϑ_T to approximate $h^0_T(z^i_t) \approx h_T(z^i_t, \vartheta_T)$.

The machine learning-style estimation of the parameters ϑ_T entails minimizing the MSE associated with the forecast error $\widetilde{R}^{ei}_{t,T} - h_T(z^i_t, \vartheta_T)$ instead of $R^{ei}_{t,T} - g_T(z^i_t, \theta_T)$. The hybrid risk premium quantification is then given by:

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx \widetilde{\mathbb{E}}_t(R^{ei}_{t,T}) + h_T(z^i_t, \hat{\vartheta}_T), \tag{2.11}$$

which yields the familiar decomposition:

$$R^{ei}_{t,T} - \underbrace{(\widetilde{\mathbb{E}}_t(R^{ei}_{t,T}) + h_T(z^i_t, \hat{\vartheta}_T))}_{\text{hybrid forecast}} = \underbrace{(a^i_{t,T} - h_T(z^i_t, \vartheta_T))}_{\text{approximation error}} + \underbrace{(h_T(z^i_t, \vartheta_T) - h_T(z^i_t, \hat{\vartheta}_T))}_{\text{estimation error}} + \varepsilon^i_{t,T}. \tag{2.12}$$

To account for time-varying model parameters, the dynamic hyperparameter tuning described in Section 2.3 can be applied in the same way, which yields the following hybrid approximative formula for the stock risk premium:

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx \widetilde{\mathbb{E}}_t(R^{ei}_{t,T}) + h_{s,T}(z^i_t, \hat{\vartheta}_{s,T}). \tag{2.13}$$

Neither the theory-based (“Econ”) nor the machine learning (“Metrics”) approach would be described as econometrics, the discipline founded to connect economic theory and statistics. Yet, the formula in Equation (2.13) may be seen as a novel way to combine Econ and Metrics in the modern age of data science. We refer to this hybrid strategy as *theory assisted by machine learning.*

An obvious alternative hybrid strategy is motivated by the observation that though GKX include a plethora of stock-level and macro features, they do not use the information provided by the theory-based risk premium measures, or any other conditional time t moment computed under the risk-neutral measure. By augmenting the set of features accordingly, we can assess whether the theory-based measurements enhance the explanatory power of the data science approach. We refer to this hybrid approach as *machine learning with theory features.*

A central tenet of financial economics, derived from Equation (2.1), states that marginal utility-weighted prices follow martingales. This tenet implies that return predictability should be a longer-horizon phenomenon. High frequency price processes are expected to behave like martingales, such that the MSE-optimal return prediction at very short horizons should be close to the zero forecast (cf. Cochrane (2005), Section 2.4). The signal-to-noise ratio – $\mathbb{E}_t(R^{ei}_{t,T})$ to $\varepsilon^i_{t,T}$ – is expected to increase at longer forecast horizons. So, the empirical question that we seek to address refers to which of the approaches – theory-based, machine learning or hybrid – delivers a better approximation of $\mathbb{E}_t(R^{ei}_{t,T})$, i.e. a superior out-of-sample performance, at given horizons. To answer this question we need a comprehensive database.

3 Data, implementation, and performance assessments

3.1 Assembling the database

Selection of stocks and linking databases

The universe of stocks for which we compare the alternative risk premium measures is defined by a firm's membership in the S&P 500 index.[4] One reason to choose this criterion is that if we want to compute theory/option-based risk premia according to Equation (2.5), we have to provide information about the constituents of the market index proxy. Because the S&P 500 is used for that purpose, index membership is the obvious criterion to select the cross-section of stocks considered for our analysis. For the identification of historical S&P 500 constituents (HSPC) across databases, we start by extracting information about a firm's S&P 500 membership status from Compustat. We thereby obtain, for every month from March 1964 to December 2018, a list of HSPC. In total, we find 1,675 firms that have been in the S&P 500 for at least one month. For the HSPC identified in Compustat, we retrieve price and return data from CRSP. Compustat and CRSP also supply the data used for the machine learning approaches. The option data, which are required to compute the theory-based measures, come from OptionMetrics. Section O.1 of the Online Appendix explains in detail how we link the three databases. Appendix A.2 documents the quality of the matching procedure.

Stock-level and macro features

Following GKX, we retrieve from Compustat and CRSP 93 firm-level variables that have been identified as predictors for stock returns in previous literature. We also construct 72 binary variables that identify a firm's industry (see Table 9 in Appendix A.3).[5] A cross-sectional median-based imputation is applied to deal with missing

[4] Each company in the S&P 500 may be associated with multiple securities. An S&P 500 constituent is a specific company-security combination, but we refer to them, as is common in the literature, interchangeably as "securities," "stocks" or "firms."

[5] For that purpose, we adapt the SAS program from Jeremiah Green's website, `https://sites.google.com/site/jeremiahrgreenacctg/home`, accessed January 20, 2020. The industry indicators are based on the first two digits of the standard industrial classification (SIC) code.

observations.[6]

We consider two types of transformation for firm-level fetures: standard mean-variance and median-interquartile range scaling, the latter being more robust in the presence of outliers. The choice of the scaling procedure (standard or robust) is treated as a hyperparameter.[7] In either case, we make sure that no information from the future enters the validation or tests sets in order to prevent a look-ahead bias. The stock-level features are augmented by macro-level variables, obtained from Amit Goyal's website.[8] These variables are the market-wide dividend-price ratio, earnings-price ratio, book-to-market ratio, net equity expansion, stock variance, the Treasury bill rate, term spread, and default spread. Their detailed definitions can be found in Welch and Goyal (2008).

The variables retrieved have a mixed frequency: monthly (20 stock-level + 8 macro-level variables), quarterly (13 stock-level variables), or annual (60 stock-level variables). Using the date of the last trading day of each month as a point of reference, they are aligned according to Green et al.'s (2017) assumptions about delayed availability to avoid any forward-looking bias. Features at the monthly frequency are delayed at most one month, quarterly variables by at least a four-

[6] Median-based imputation is frequently applied in related literature. However, Bryzgalova et al. (2022) point out that firm characteristics are typically not missing at random, rendering median-based imputation problematic. They propose an alternative approach that exploits cross-sectional and time series dependencies between characteristics to impute missing values. For their empirical analysis Bryzgalova et al. (2022) use a sample that comprises more than 22,000 stocks (including penny stocks) and starts in 1967. Missing data occur particularly often at the beginning of the sample and for small firms. Being aware of the missing value issue, we do not follow GKX, who use data from the late 1950s, but instead commence the training process in 1974. Focusing on HSPC, which are large firms by constructions, further mitigates the problem of missing values.

[7] Here we deviate from GKX, who achieve outlier robustness by applying a cross-sectional rank transformation and re-scaling the stock-level features to the interval -1 to 1. Various studies (e.g., Da et al., 2022 and Kelly et al., 2019) report that their results do not critically depend on the choice of scaling. To assess whether this conclusion also holds true in our setting, Section A.5 of the Appendix reports the results of a robustness check, in which the empirical analysis is conducted with rank-transformed features.

[8] See http://www.hec.unil.ch/agoyal, accessed January 20, 2020.

month lag, and annual variables by at least a six-month lag. Moreover, we match CRSP returns at horizons of one month (30 calendar days) and one year (365 calendar days), such that they are forward-looking from the vantage point of the end-of-month alignment day.

A considerable number of missing values for stock-level features arise, if we go further back in time than the mid-1970s. To mitigate the aforementioned negative consequences associated with massively imputing missing values, we start using the data in October 1974, when the problem is alleviated. Moreover, two of the originally 93 stock-level features retrieved are excluded, because they contain an excessive amount of missing values. Figure 1 shows a heatmap that illustrates how the share of missing values of stock-level features changes over time.

[Insert Figure 1 about here]

The out-of-sample analysis is performed for the period from January 1996, the starting date of OptionMetrics, until December 2018. Proceeding as described, we obtain an unbalanced panel data set at a monthly frequency that ranges from October 1974 until December 2018. The number of HSPC during that period is 1,145, with a varying number of observations per stock. In total, there are 362,306 stock/month observations.

Option data

The data to implement the option-based risk premium formulas in Equations (2.5) and (2.6) are retrieved from OptionMetrics. Two issues must be resolved in the process. First, options on S&P 500 stocks are American options, yet the computation of risk-neutral variances according to Equation (2.4) relies on European options. Second, a continuum of strike prices is not available, so the integrals in Equation (2.4) must be approximated, using a grid of discrete strikes. As pointed out by Martin (2017), a lack of a sufficient number of strikes may severely downward bias the computation of risk-neutral variances. Martin and Wagner (2019) advocate for the use of the OptionMetrics volatility surface to address these issues and compute risk-neutral variances according to Equation (2.4). Although European options are traded on the S&P 500 index, and their prices are available in OptionMetrics, we also rely on the volatility surface to compute risk-neutral index variances. Using the OptionMetrics volatility surface, we compute the theory-based risk premium measures for the selected stocks and the two horizons of interest. These data are matched, by their security identifier and end-of-month date, with the aforementioned unbalanced panel. A detailed explanation of our use of the volatility surface is provided in Section O.2 of the Online Appendix.

Risk-free rate proxies

To compute excess returns and all of the option-based measures, we need a risk-free rate proxy that matches the investment horizon. It can be computed for different horizons at a daily frequency using the zero curve provided by OptionMetrics. However, like any data supplied by OptionMetrics, the zero curve is not available before January 1996. We therefore employ the Treasury bill rate as a risk-free rate proxy for earlier periods.

3.2 Empirical implementations

In the following we provide information about the hyperparameter configurations of the statistical models, the construction of the vector of state variables z_t^i, and the long and short training schemes.

As mentioned previously, our machine learning approaches employ four popular statistical models: the ANN, RF, GBRT, and ENet. The first three were identified by GKX as the most appropriate for the task at hand. The ENet is included as an instance of penalized regression because of the less demanding hyperparameter tuning.[9] The hyperparameter configurations for these models are listed in Table 1.

[Insert Table 1 about here]

The selection of features collected in the vector z_t^i follows GKX, such that we use the 91 stock-level variables (included in the vector c_t^i) and their interactions with the eight macro predictors (included in the vector x_t). Formally, z_t^i is comprised of the vector $(1, x_t')' \otimes c_t^i$, augmented with industry dummies, such that altogether we have $91 \times 9 + 72 = 891$ features.[10]

The implementation of the sequential validation procedure mentioned in Section 2.1 is illustrated in Figure 2 (long training scheme). It shows that the length of the training period increases from 10 years initially to 31 years; the 12-year validation period shifts forward by one year with every new test sample. There are S=22 out-of-sample years with the final one-year predictions made in December 2017 for December 2018. For every sample and statistical model, hyperparameter tuning is performed at the one-month and one-year forecast horizon. When considering the one-month horizon, the number of test samples increases to S=23, because monthly forecasts are possible during the year 2018. Details on the hyperparameter tuning are provided in Appendix A.4.[11]

[Insert Figure 2 about here]

The basic setup remains the same when considering the hybrid approaches. However, the training and validation procedure changes because of the delayed availability of the OptionMetrics data beginning January 1996. We therefore consider

[9] We assume that the reader has some familiarity with these approaches, which are covered by Hastie et al. (2017).

[10] In principle, it would also be possible to explicitly consider the time series of macroeconomic variables, as proposed by Chen et al. (2021). In line with GKX, we choose to focus on the last observation of these series instead.

[11] While our implementation of the machine learning approaches draws on GKX, it deviates in some respects. Section O.3 of the Online Appendix provides a detailed juxtaposition.

the alternative, short training scheme illustrated in Figure 3; it is used for the *theory assisted by ML* and *ML with theory features* strategies.

[Insert Figure 3 about here]

The short training scheme reduces the initial training period to one year and the validation period comprises 1 year instead of 12. With this configuration, we can retain a sufficiently large number of out-of-sample years, comparable to the long training scheme.

To establish a benchmark for the performance of the hybrid approaches, we also train the models using the original feature set and the short training scheme. A comparison with the long training results is interesting for another reason too: It allows us to study how important the length of the training period is and to assess the effect of the length of the validation period.

3.3 Performance assessments

We compare the alternative approaches to measure stock risk premia by assessing their out-of-sample forecast performance. This represents a useful criterion, because the different methodologies provide approximations of the conditional expected excess return, which is the MSE-optimal prediction. The smaller the MSE, the better the approximation of the stock risk premium. We consider forecasts with horizons of one month (30 calendar days) and one year (365 calendar days), issued at an end-of-month and daily frequency, respectively.

Following Welch and Goyal (2008), we rely on a performance measure that relates the MSE of a model's out-of-sample forecast to that of a benchmark. We use the zero forecast for that purpose, which has the appeal of providing a parameter-free alternative and comparability across studies. More specifically, the performance criterion is the pooled predictive R^2 given by:

$$R^2_{oos} = 1 - \frac{\sum_t \sum_i \left(R^{ei}_{t,T} - \hat{R}^{ei}_{t,T}\right)^2}{\sum_t \sum_i \left(R^{ei}_{t,T}\right)^2}, \tag{3.1}$$

where $\hat{R}^{ei}_{t,T}$ denotes the respective forecast/risk premium estimate. The calculation is based solely on observations included in the S test sample years that were not used for training or validation.

To study performance over time, we also compute the predictive R^2 for each of the test samples separately:

$$R^2_{oos,s} = 1 - \frac{\sum_i \sum_t \left(R^{ei}_{t,T} - \hat{R}^{ei}_{t,T}\right)^2 \cdot \mathbb{1}[t \in \mathcal{S}(s)]}{\sum_i \sum_t \left(R^{ei}_{t,T}\right)^2 \cdot \mathbb{1}[t \in \mathcal{S}(s)]} \qquad s = 1, 2, \ldots, S, \tag{3.2}$$

where $\mathcal{S}(s)$ denotes the set of time indices of forecast sample s, such that $\mathbb{1}[t \in \mathcal{S}(s)]$ is equal to 1 if the observation at period t belongs to the sample year s, and 0 otherwise. For the assessment of statistical significance, we report the p-values associated with a test whether a model has no explanatory power over the zero forecast; formally, the null hypothesis that $\mathbb{E}(R^2_{oos,s}) \leq 0$. To construct a convenient test statistic, we take the mean of the $R^2_{oos,s}$ across the test samples, $\overline{R^2_{oos}} = \frac{1}{S}\sum_{s=1}^{S} R^2_{oos,s}$, and compute its standard error $\hat{\sigma}(\overline{R^2_{oos}})$, using a Newey-West correction to account for serial correlation. Provided that a central limit theorem applies, and assuming that $\mathbb{E}(R^2_{oos,s}) = 0$, the t-statistic $\overline{R^2_{oos}}/\hat{\sigma}(R^2_{oos})$ is approximately standard normally distributed, such that a one-sided p-value can be provided.[12]

As an alternative to the R^2_{oos} in Equation (3.1), we also consider the time-series R^2 used by Chen et al. (2021), which accounts for the fact that the number of stocks

[12] The Diebold-Mariano test employed by GKX to gauge differences in forecast performances is constructed in a similar vein. We provide p-values associated with this test in Section O.4 of the Online Appendix.

in period t (N_t) can change over time:

$$EV_{oos} = 1 - \frac{\sum_t \frac{1}{N_t} \sum_{i=1}^{N_t} \left(R^{ei}_{t,T} - \hat{R}^{ei}_{t,T}\right)^2}{\sum_t \frac{1}{N_t} \sum_{i=1}^{N_t} \left(R^{ei}_{t,T}\right)^2}. \tag{3.3}$$

As this study is ultimately concerned with approximating stock risk premia, both the level and cross-sectional properties of the excess return predictions should be taken into account for performance assessment. However, the R^2_{oos} can be dominated by the forecast error in levels, potentially masking the cross-sectional explanatory power of a model. To explicitly account for this dimension of return predictability, we use the following measures: First, we compute a cross-sectional out-of-sample R^2 similar to those advocated by Maio and Santa-Clara (2012) and Bryzgalova et al. (2021):

$$XS_{oos} = 1 - \frac{\text{Var}_N(\overline{\hat{\varepsilon}^i_T})}{\text{Var}_N(\overline{R^{ei}_T})}, \tag{3.4}$$

where $\text{Var}_N(\cdot)$ stands for the cross-sectional variance across the N sample stocks; $\overline{\hat{\varepsilon}^i_T}$ and $\overline{R^{ei}_T}$ are the stock-specific time-series averages of $R^{ei}_{t,T} - \hat{R}^{ei}_{t,T}$ and $R^{ei}_{t,T}$, respectively. Second, we assess cross-sectional performance by forming decile portfolios based on the respective model's excess return predictions and comparing predicted and realized mean excess returns across approaches. If an approach delivers sensible risk premium estimates then a) the mean predicted excess returns and mean realized excess returns of the prediction-sorted portfolios should align, and b) there should be sizable variation in the mean realized excess returns across these portfolios. Besides graphical assessments and rank correlations, we also compare the annualized Sharpe ratios of zero-investment portfolios long in the decile portfolio of stocks with the highest excess return prediction and short in that with the lowest. The Sharpe ratio accounts for the desideratum that the cross-sectional differentiation of the mean realized excess returns should be achieved by a small variation over the years of the test sample.

The machine learning models are trained on data at a monthly frequency. Accordingly, the respective excess return forecasts are updated once at the end of each month. Forecasts at these same dates are also available using the option-based approaches, which additionally can provide risk premium estimates at higher frequencies, up to daily. To facilitate comparisons at a daily frequency, we retain the most recent ML-based risk premium estimate until an update becomes available by the end of the next month. For example, the estimate of an annual horizon stock risk premium in mid-April 2015 corresponds to the last available estimate calculated at the end of March 2015. For the *ML with theory features* strategy, the hybrid model's daily estimate employs the statistical model (trained on monthly data) endowed with the prevailing end-of-month firm- and macro-level features and daily updated theory-based measures. Similarly, the adaption of the *theory assisted by ML* approach combines the theory-based daily risk premium estimate with the prevailing end-of-month ML-based residual approximation.

4 Empirical Results

4.1 Comparison at monthly and annual horizons

One-month horizon

Table 2 contains the results for the one-month horizon; in Panel A, the forecasts are issued at a daily frequency, whereas in Panel B, they are issued monthly (end-of-month). Among the machine learning approaches in Panel B, only the ANN achieves a positive predictive R^2 (0.2%); the same R^2_{oos} is delivered by the theory-based

MW.[13] Evaluating the daily MW forecasts, we find that the predictive R^2 increases to 0.9%, which represents the only instance in which we can reject the hypothesis that $\mathbb{E}(R^2_{oos,s}) \leq 0$ at significance levels below 5%. For a daily forecast frequency, the ANN achieves an R^2_{oos} of 0.5%, the highest among the machine learning approaches.[14]

[Insert Table 2 about here]

The comparatively good performance of the theory-based approach is corroborated by a complementary analysis based on the data that Chabi-Yo et al. (2021) used to introduce their alternative option-based risk premium estimate,[15] and which contain their estimates at the one-month and one-year horizons. Although the universe of stocks is different, there is an overlap with our study. When we conduct an analysis at the intersection of firms and dates, it yields a monthly R^2_{oos} of 1% implied by Chabi-Yo et al.'s (2021) method (daily forecast frequency). For this merged sample, the predictive R^2 produced by MW remains unchanged (0.9%); the R^2_{oos} of the machine learning approaches do not improve.

[Insert Figure 4 about here]

The relative advantages of the theory-based paradigm are also evident in Figure 4. Panel A (monthly forecast frequency) and conspicuously Panel B (daily) both show that MW yields a better alignment of the prediction-sorted portfolios. The rank correlation between mean predicted and mean realized excess returns is 0.96,

[13] To avoid a cluttered exposition, we focus in the main text on reporting and interpreting the R^2_{oos} results. Section O.4 of the Online Appendix includes extended tables that also report XS_{oos} and EV_{oos}. It can be seen that R^2_{oos} and EV_{oos} take on very similar values, and while the level of XS_{oos} is somewhat smaller, its pattern across approaches corresponds to that of R^2_{oos}. Accordingly, the conclusions obtained by using the alternative performance measures remain the same.

[14] A monthly predictive R^2 of about 1% may appear small, but it is actually higher than any reported by GKX. Their ANNs yield monthly predictive R^2 between 0.3% and 0.7%, depending on the universe of stocks and ANN architecture.

[15] We are grateful to Grigory Vilkov for providing access to these data.

whereas that implied by the ANN is 0.56 (monthly forecast frequency). Figure 4 also shows that the variation of the mean realized excess returns across prediction-sorted portfolios is favorably wider using MW than the variation implied by the ANN. This result is reflected in the Sharpe ratios of the zero investment portfolios (cf. Table 2), which are 0.30 (monthly forecast frequency) and 0.37 (daily) for MW, compared with 0.28 (monthly) and 0.26 (daily) for the ANN.[16] Overall, these findings indicate that at the one-month horizon, care is needed when investing in machine learning-based methods; their superiority over the theory-based paradigm is by no means a given.

An alternative conclusion might refer to the sample period and universe of stocks, for which the task at hand might be more difficult for machine learning. Compared with GKX, we consider fewer stocks for training and validation, and the training begins in a later year, both of which are factors that could prevent the machine learning approaches from reaching their full potential.

One-year horizon

Most of these concerns can be alleviated by a review of Table 3, which shows the results for the one-year horizon. Contrasting Panels A and B, we observe that it matters little whether we use daily or monthly forecasts, so we simply focus on the latter in the following discussion.

[Insert Table 3 about here]

Compared with the one-month horizon results, the annual predictive R^2 increase by an order of magnitude; the R^2_{oos} delivered by MW is about 9%. The results in Table 3 mitigate any concerns that the present selection of stocks constitutes a more difficult environment for machine learning approaches or that their training is

[16] Tables 2 and 3 also show that, in terms of predictive R^2, KT is less successful. Yet, regarding prediction-sorted portfolios, KT and MW are equivalent. Both achieve cross-sectional differentiation through risk-neutral variances $\text{var}^*_t(R^i_{t,T})$. Thus, the prediction-sorted portfolios include the same stocks and yield the same mean realized excess returns and Sharpe ratios.

flawed. For example, the ANN achieves an annual R^2_{oos} notably higher than those reported by GKX.[17] Furthermore, MW, GBRT, and the ANN perform comparably well, with R^2_{oos} ranging between 8.8% and 10.6% and p-values for the hypothesis that $\mathbb{E}(R^2_{oos,s}) \leq 0$ ranging from 3.5% to 5.1%.[18] Notably smaller predictive R^2 and higher p-values are implied by the ENet and KT; that is, not all option-based and machine learning approaches perform equally well.

[Insert Figure 5 about here]

In terms of predictive R^2, the RF stands out, delivering an annual R^2_{oos} of 19.5% with a p-value of 0.2%. The good RF results are confirmed by the favorable alignment and cross-sectional variation in realized mean excess returns of the prediction decile portfolios (cf. Panel C of Figure 5), and the highest Sharpe ratio of the long-short portfolio among the approaches considered. We thus conclude that at the one-year horizon, there exists a machine learning method that offers a comparative advantage over the theory-based approach.[19]

Time-series variation

The time-series variation of the predictive R^2 is illustrated in Figure 6. In Panel A, we present a comparison of MW with the random forest, the best-performing machine learning method; the other approaches are in Panel B. The $R^2_{oos,s}$ values depicted in Figure 6 refer to the year the forecast was issued. For example, the annual predictive R^2 associated with the year 2008 is based on forecasts issued from January to December 2007.

[17] Depending on the selection of stocks, they report annual predictive R^2 for ANNs that range from 3.4% to 5.2%.

[18] A complementary analysis using data provided by G. Vilkov yields very similar annual predictive R^2 values for MW and Chabi-Yo et al.'s (2021) alternative approach.

[19] As mentioned in Section 3.3, the R^2_{oos} can be dominated by the forecast error in levels, whereas the Sharpe ratio captures purely cross-sectional aspects. Hence, it is not necessary for R^2_{oos} and the Sharpe ratio to point into the same direction in terms of favored approaches.

[Insert Figure 6 about here]

The volatility of the $R^2_{oos,s}$ values indicated by Figure 6 is not surprising; the years 1996-2018 represent a period rife with crises and crashes. These events have a notable effect on the standard deviations of the predictive R^2 in Tables 2 and 3. We observe that at the one-year horizon, the impact of the build-up and burst of the so-called dot-com bubble is more pronounced than that of the 2008 financial crisis. Both theory-based and machine learning approaches yield large negative annual $R^2_{oos,s}$ values associated with forecasts issued during 2000 and 2001. Panel A in Figure 6 also illustrates how the RF achieves its improvement over MW at the one-year horizon.

4.2 Hybrid approaches and short training

Next, we assess the potential of hybrid strategies that combine the theory-based and machine learning paradigms. Table 4 indicates the promise of this idea: Although theory-based and machine learning forecasts covary positively, the correlations are not strong, so the two approaches seem to account for different components of the stock risk premium.

[Insert Table 4 about here]

Short-training effects and ML with theory features

Any hybrid methodology must accommodate the late availability of the OptionMetrics data. As discussed previously, we deal with this issue by applying the short-training scheme in Figure 3. Tables 5 (one-month horizon) and 6 (one-year horizon) present two sets of machine learning results obtained by short training. The first uses the same 891 features as selected for long training. The second, referred to as *ML with theory features*, results from adding the two option-based stock risk premium measures (according to MW and KT) and Martin's (2017) lower bound of the expected market return. The following discussion contains an assessment of the incremental effects of applying the short-training scheme and including the theory-based features.[20]

[Insert Table 5 about here]

We have already seen that at the one-month horizon, most of the machine learning approaches do not perform well. Table 5 shows that the results worsen when applying the short-training scheme. All MLPs, including the ANN, now yield a negative predictive R^2. Their standard deviations increase, and the Sharpe ratios of the long-short portfolios decline. The segments labeled *ML with theory features* in Table 5 reveal that this deterioration is not mitigated by the inclusion of theory-based features. Using MW to obtain risk premium estimates remains the preferred strategy at the one-month horizon.

[Insert Table 6 about here]

Table 6 shows that the short-training effects are more ambiguous with regard to end-of-month issued forecasts with a one-year horizon. While the ENet now performs poorly, the ANN benefits from short training: Its R^2_{oos} increases from 9% (long training) to 14%, with a p-value of 0.4%. In contrast, short training reduces the RF's predictive R^2 from 19.1% (long training) to 12.4%, accompanied by increases of the standard deviation and p-value. However, Panel A of Figure 7, which depicts the time-series variation of the predictive R^2, shows that the adverse effects of short training on the RF are mitigated as the training sample grows. At the start of the sequential validation procedure, there are only a few years of observations available

[20] Comparing Table 5 with Table 2, we note that the theory-based results only change because the out-of-sample evaluation period is shorter. The years 1996 and 1997 are excluded to ensure comparability with the short-trained MLPs.

for training. When the dot-com crisis confronts such an RF, it results in a sharp decline of the $R^2_{oos,s}$ associated with the one-year forecasts issued in the year 2000. This drop causes the increase of the time-series standard deviation and p-value compared with the long-trained RF.[21] As the training sample grows, the performance of the short-trained RF improves and reaches, near the end of the sample period, the level of its long-trained counterpart.

[Insert Figure 7 about here]

Table 6 also shows that the *machine learning with theory features* strategy yields a positive effect only when using the RF. Though the improvement is moderate for end-of-month-issued forecasts – the R^2_{oos} increases from 12.4% to 14.6%, and the Sharpe ratio increases from 0.59 to 0.62 – we note that the augmentation with theory features helps the short-trained RF improve the 2008 crisis year forecasts (cf. Figure 7).

[Insert Table 7 about here]

Table 7 suggests that the *ML with theory features* strategy is more rewarding for forecasts at a daily frequency, and in particular when using the RF. Augmented with daily theory-based features, the RF's predictive R^2 increases from 9.0% to 18.6%, while also reducing the time-series variation across test samples. Considering that the pure theory-based (MW) R^2_{oos} amounts to 9.5%, this hybrid approach makes particularly good use of the additional data. The highest Sharpe ratio of the long-short portfolio in the field of competitors corroborates this conclusion.

Theory assisted by machine learning

[21] Figure 7 shows that this drop is much less pronounced for the short-trained ANN, which explains the smaller standard deviation and p-value in Table 6.

For our implementation of the *theory assisted by machine learning* strategy we rely on Martin and Wagner's (2019) approach to measuring stock risk premia (MW for short), which explicitly starts from the basic asset pricing equation, the keystone of financial economics. MW is empirically not unsuccessful, and we propose building on it, as a basis, to model only that which theory cannot account for – the approximation errors – by applying machine learning techniques.

[Insert Figure 8 about here]

The segment labeled *theory assisted by ML* in Table 6 contains the results obtained from applying this idea.[22] We observe that not all machine learning assistance improves the performance of the theory-based approach; the ENet even drives the R^2_{oos} into a negative domain. GBRT yield a moderate improvement, whereas the ANN and RF are more successful. Their support increases the baseline MW R^2_{oos} by 5.1 percentage points (MW+ANN) and 7 percentage points (MW+RF), respectively. The standard deviations of the predictive R^2 grow, but Figure 8 shows that this increase is mainly due to the short-training effect, which in turn is reflected in the harsh drop of the $R^2_{oos,s}$ associated with the year 2000 forecasts, which we also identified for the short-trained RF. By zooming in on more recent forecast samples, we observe that with an increasing training sample size, the performance of the MW+RF hybrid matches that of the long-trained RF.

The prediction decile plots in Figure 9 show that the alignment of mean predicted and realized excess returns of the prediction-sorted portfolios is particularly good for the MW+RF approach and that the variation of the mean realized excess returns across the prediction-sorted portfolios is favorably high. Consistently, RF assistance increases the Sharpe ratio for the long-short portfolio from 0.37 (pure MW) to 0.65,

[22] Short-trained MLPs do not perform well at the one-month horizon, and when using them to account for the approximation errors of MW, we find no improvement. We therefore discuss in detail only the one-year horizon results.

as reported in Table 6. For daily forecasts rather than forecasts issued at the end of the month, these conclusions remain the same (cf. Table 7).

[Insert Figure 9 about here]

These results lead to the conclusion that at the one-year horizon, the MW+RF approach qualifies as a promising alternative for the task of quantifying stock risk premia. This hybrid strategy also has the appeal of effectively combining theory with measurement.

4.3 Feature importance and a disaggregated analysis

We also investigate how the importance of features with respect to stock risk premia might differ between pure machine learning and theory assisted by machine learning. We consider both pure RF and the MW+RF hybrid and focus on the one-year horizon with end-of-month issued forecasts. To gauge a feature's importance by the reduction of the predictive R^2 induced, we use a disruption of the temporal and cross-sectional alignment of the feature with the prediction target. This disruption is implemented by replacing the feature's observed values by 0 when computing the predictive R^2. We compute the importance measure on the test samples, and report the size of the induced R^2_{oos} reduction.[23] Figures 10 (RF) and 11 (MW+RF) illustrate the results.

[Insert Figures 10 and 11 here]

[23] Alternatively, it is possible to compute the importance measure on the training samples and provide a relative measure of feature importance, as done by GKX. Moreover, feature importance could be assessed by randomly drawing a feature from the empirical distribution instead of replacing it by 0. We prefer the present approach for its straightforward interpretability. Another approach to assess the importance of features is based on the absolute gradient of the loss function with respect to each feature respectively, which is very convenient in the context of neural networks (cf. Chen et al., 2021), but not suitable for all machine learning techniques. Shapley additive explanations (cf. Lundberg and Lee, 2017) would be well suited to account for dependencies between features, but are computationally infeasible given our number of characteristics.

A comparison of Figures 10 and 11 reveals that the conclusions regarding the relative importance of features remain the same, regardless of whether the RF serves to assist the theory-based approach or is applied for its original use. The pattern is similar in both applications. With respect to stock-level variables, the established return predictive signals (RPS) are most important: The book-to-market ratio ranks first (along with other valuation ratios), followed by variables associated with liquidity (dollar trading volume, Amihud illiquidity), and then momentum indicators (industry momentum and 12-month momentum). None of the other more than 80 stock level features is among the top four. The revival of the classic RPS, and in particular the conspicuous role of the book-to-market ratio, is noteworthy. In GKX's study, the short-term price reversal dominated the feature importance at the one-month horizon, whereas the book-to-market ratio remained nondescript. The consistent feature importance in both applications – RF and MW+RF – may seem surprising, because MW already accounts for a considerable part of the excess return variation. We might have expected that modeling the approximation error of the theory-based approach would reveal other important features. But it is the familiar triad – valuation ratio, liquidity, and momentum – that dominates in both applications.

A corresponding conclusion arises from an analysis of the importance of the market-wide variables (Panels B in Figures 10 and 11). In both uses of the RF, the Treasury bill rate is the most important variable. Its conspicuous role highlights the relevance of asset pricing approaches that adopt Merton's (1973) suggestion to use short-term interest rates as state variables in variants of the intertemporal CAPM (e.g., Brennan et al. (2004), Petkova (2006), Maio and Santa-Clara (2017)), as well as preference-based asset pricing models that motivate a short-term interest rate-related risk factor, as in Lioui and Maio (2014).[24]

The feature importance results provide the foundation for a disaggregated analysis, for which we form portfolios by sorting stocks into quintiles according to key characteristics associated with valuation ratios, liquidity, and momentum. As suggested by the previous results, we choose book-to-market and earnings-to-price as valuation ratios; for liquidity, we use dollar trading volume and Amihud's illiquidity measure. Momentum portfolios are based on 12-month and industry momentum.[25] The sorting of stocks into quintile portfolios on the basis of the respective characteristic gets renewed each month. We also form 10 industry portfolios based on one-digit SIC codes. For each quintile and industry portfolio and each approach of interest – MW, pure machine learning (ANN and RF), and theory assisted by machine learning (MW+RF and MW+ANN) – we compute the annual R^2_{oos} according to Eq. (3.2).

[Insert Table 8 about here]

The results in Table 8 generally corroborate the conclusions of the aggregated analysis and also reveal the following detailed insights: For all portfolios based valuation ratios, we observe an improvement of the theory-based method by machine learning assistance. Moreover, the hybrid approaches are preferred across all quintile portfolios. MW+RF is particularly successful in quintiles 2 to 5, and MW+ANN is optimal in quintile 1. For all momentum portfolios, machine learning assistance improves the performance of the theory-based approach. For momentum quintiles 1 to 4, MW+RF is the preferred strategy. For momentum quintile 1, pure ANN and MW+ANN perform better. Regarding the liquidity-sorted portfolios, machine

[24] We also check whether feature importance differs when we measure the effect of an exclusion of a feature on the cross-sectional performance, measured by the Sharpe ratio of the long-short portfolio. The conclusions remain qualitatively the same as when we use the predictive R^2. Details of this analysis are available in Section O.4 of the Online Appendix.

[25] We report the results for quintile portfolios based on other characteristics in Section O.4 of the Online Appendix.

learning assistance again improves the theory-based results, but we note that MW+RF does not perform well on the high liquidity portfolios. The explanation is that the short training effect that we discussed previously has the strongest effect on the performance of both RF and MW+RF in the high liquidity portfolios.[26] The pure ANN, less affected by short training, delivers more consistent performance across liquidity portfolios. Nevertheless, a hybrid strategy is preferred over pure machine learning for four (dollar trading volume), respectively three (Amihud illiquidity) quintile portfolios.

Panel B of Table 8 shows that for all industry portfolios, RF assistance improves the performance of MW; the ANN assistance does so in seven of ten cases. With the exception of one of the sector portfolios for which the pure ANN is preferred, the hybrid strategies yield the highest predictive R^2. In addition, MW+RF is preferred in seven of ten sector portfolios, and MW+ANN is preferred in two. The complementary advantage of the two hybrid approaches is thus a recurring result.

5 Conclusion

In this study, we took two diverging paths to measure stock risk premia in an attempt to assess and reconcile the opposing philosophies that underlie them. The comparison, at one-month and one-year investment horizons, reveals that the theory/option-based method offers an advantage at the shorter horizon, especially if stock risk premium estimates are to be delivered at higher frequencies. At the one-year horizon, the picture is more complex. Of the four machine learning methods considered in this study, one delivers weaker performance than the theory-based strategy (elastic net), two are comparable (gradient boosted regression trees and artificial neural networks),

[26] For more details, refer to Section O.4 of the Online Appendix, which contains time series plots of the predictive R^2 corresponding to Figure 6. They illustrate the short training effect broken down by quintile portfolios based on Amihud illiquidity.

and one (random forest) offers the best results. To achieve this performance, a sufficiently long training period is required though.

Noting the concerns regarding the use of agnostic machine learning procedures in a theoretically well-developed discipline like finance, we put forth a methodology that takes Martin and Wagner's (2019) theory-based approximate formula for the stock risk premium as its basis and then applies machine learning to account for the approximation error. Although a pure theory-based method remains the preferred choice at the one-month horizon, the empirical performance of this *theory assisted by machine learning* approach at the one-year horizon is encouraging. Using a random forest, the theory-based component provides 57% of the hybrid model's explanatory power in terms of the predictive R^2; 43% is attributable to machine learning assistance. The conclusion that such a supportive use of machine learning captures fundamental components of stock risk premia is supported by the conspicuous role of valuation ratios and liquidity indicators in an analysis of feature importance. A disaggregated analysis based on stock portfolios sorted according to these characteristics corroborates the expediency of the proposed hybrid approach. We view it as a promising alternative for bringing together the diverging paths in finance.

A Appendix

A.1 Theory-based stock risk premium formulas

This section provides details for the stock risk premium formulas in Equations (2.2) and (2.3) and the nature of the approximation residuals $a^i_{t,T}$ and $\xi^i_{t,T}$. We delineate the assumptions and rationales behind their omission, which provide the theory-based approximation formulas in Equations (2.5) and (2.6).

Martin and Wagner's (2019) derivations originate from the basic asset pricing equation, with a focus on the gross return of a portfolio with maximal expected log return ($R^g_{t,T}$). This growth-optimal return has the unique property among gross returns that its reciprocal is an SDF, such that $m_{t,T} = 1/R^g_{t,T}$. Using this SDF to price the payoff $X^i_{t,T} = R^i_{t,T} \cdot R^g_{t,T}$ gives:

$$\mathbb{E}_t\Big(m_{t,T} \cdot X^i_{t,T}\Big) = \mathbb{E}_t\Big(R^i_{t,T}\Big) = \frac{1}{R^f_{t,T}} \mathbb{E}^*_t\left(R^i_{t,T} \cdot R^g_{t,T}\right), \tag{A-1}$$

where the $*$ notation indicates that the expected value is computed with respect to the risk-neutral measure. Division by $R^f_{t,T}$ and subtracting $\mathbb{E}^*_t\Big(R^i_{t,T}/R^f_{t,T}\Big) \times \mathbb{E}^*_t\Big(R^g_{t,T}/R^f_{t,T}\Big) = 1$ (the price of any gross return is 1) yields:

$$\mathbb{E}_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) = 1 + \operatorname{cov}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}, \frac{R^g_{t,T}}{R^f_{t,T}}\right). \tag{A-2}$$

An orthogonal projection under the risk-neutral measure of $R^i_{t,T}/R^f_{t,T}$ on $R^g_{t,T}/R^f_{t,T}$ and a constant gives:

$$\frac{R^i_{t,T}}{R^f_{t,T}} = \alpha^i_{t,T} + \beta^i_{t,T} \cdot \frac{R^g_{t,T}}{R^f_{t,T}} + u^i_{t,T}, \tag{A-3}$$

where the moment conditions $\mathbb{E}_t^*(u_{t,T}^i) = 0$ and $\mathbb{E}_t^*(u_{t,T}^i \cdot R_{t,T}^g) = 0$ define the projection coefficients

$$\beta_{t,T}^i = \frac{\text{cov}_t^*\left(\frac{R_{t,T}^i}{R_{t,T}^f}, \frac{R_{t,T}^g}{R_{t,T}^f}\right)}{\text{var}_t^*\left(\frac{R_{t,T}^g}{R_{t,T}^f}\right)},$$

and $\alpha_{t,T}^i = 1 - \beta_{t,T}^i$. Inserting these insights into Equation (A-2) produces:

$$\mathbb{E}_t\left(\frac{R_{t,T}^i}{R_{t,T}^f}\right) = 1 + \beta_{t,T}^i \cdot \text{var}_t^*\left(\frac{R_{t,T}^g}{R_{t,T}^f}\right). \tag{A-4}$$

Moreover, Equation (A-3) implies:

$$\text{var}_t^*\left(\frac{R_{t,T}^i}{R_{t,T}^f}\right) = (\beta_{t,T}^i)^2 \cdot \text{var}_t^*\left(\frac{R_{t,T}^g}{R_{t,T}^f}\right) + \text{var}_t^*(u_{t,T}^i). \tag{A-5}$$

To make these results practically usable, Martin and Wagner (2019) propose to linearize $(\beta_{t,T}^i)^2 \approx 2\beta_{t,T}^i - k$, which for $k = 1$ amounts to a first-order Taylor approximation at $\beta_{t,T}^i = 1$. Using this approximation and inserting it into Equation (A-4) (for $k = 1$) removes the dependence on $\beta_{t,T}^i$,

$$\mathbb{E}_t\left(\frac{R_{t,T}^i}{R_{t,T}^f}\right) \approx 1 + \frac{1}{2}\text{var}_t^*\left(\frac{R_{t,T}^i}{R_{t,T}^f}\right) + \frac{1}{2}\text{var}_t^*\left(\frac{R_{t,T}^g}{R_{t,T}^f}\right) - \frac{1}{2}\text{var}_t^*(u_{t,T}^i). \tag{A-6}$$

The term neglected on the right-hand side of Equation (A-6) due to the linearization is $-\text{var}_t^*(R_{t,T}^g/R_{t,T}^f)(\beta_{t,T}^i - 1)^2$. The approximation thus should be reasonable for stocks whose $\beta_{t,T}^i$ is close to 1.

Using w_t^j, the weight of stock j in a market index with gross return $R_{t,T}^m$, Martin and Wagner (2019) perform a value-weighting of Equation (A-6) to obtain:

$$\mathbb{E}_t\left(\frac{R_{t,T}^m}{R_{t,T}^f}\right) \approx 1 + \frac{1}{2}\sum_j w_t^j \text{var}_t^*\left(\frac{R_{t,T}^j}{R_{t,T}^f}\right) + \frac{1}{2}\text{var}_t^*\left(\frac{R_{t,T}^g}{R_{t,T}^f}\right) - \frac{1}{2}\sum_j w_t^j \cdot \text{var}_t^*(u_{t,T}^i). \tag{A-7}$$

Subtracting Equation (A-7) from (A-6) removes the dependence on the unobservable optimal growth portfolio, such that

$$\begin{aligned}\mathbb{E}_t\left(R^i_{t,T}\right) \approx \mathbb{E}_t\left(R^m_{t,T}\right) + \frac{R^f_{t,T}}{2}\left[\mathrm{var}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) - \sum_j w^j_t \cdot \mathrm{var}^*_t\left(\frac{R^j_{t,T}}{R^f_{t,T}}\right)\right] \\ - \frac{R^f_{t,T}}{2}\left(\mathrm{var}^*_t(u^i_{t,T}) - \sum_j w^j_t \cdot \mathrm{var}^*_t(u^j_{t,T})\right). \qquad \text{(A-8)}\end{aligned}$$

Keeping track of the approximation error due to the linearization, we note that the term that is omitted on the right-hand side of Equation (A-8) is

$$\kappa^i_{t,T} = -\frac{1}{2R^f_{t,T}}\mathrm{var}^*_t\left(R^g_{t,T}\right) \cdot \left[(\beta^i_{t,T} - 1)^2 - \sum_j w^j_t \cdot (\beta^i_{t,T} - 1)^2\right].$$

To account for the first term on the right-hand side of Equation (A-8), Martin and Wagner (2019) draw on a result by Martin (2017), who derives a lower bound for the expected return of a market index. His starting point is again the basic asset pricing Equation (2.1), which can be written in terms of the price of the payoff $(R^i_{t,T})^2$ using an add-and-subtract strategy:

$$\mathbb{E}_t(R^i_{t,T}) - R^f_{t,T} = \left(\mathbb{E}_t[m_{t,T} \cdot (R^i_{t,T})^2] - R^f_{t,T}\right) - \left(\mathbb{E}_t[m_{t,T} \cdot (R^i_{t,T})^2] - \mathbb{E}_t(R^i_{t,T})\right). \quad \text{(A-9)}$$

The first term on the right-hand side of Equation (A-9) can be related to a risk-neutral variance, and the second term to a covariance under the physical measure, such that

$$\mathbb{E}_t(R^i_{t,T}) - R^f_{t,T} = \frac{1}{R^f_{t,T}}\mathrm{var}^*_t(R^i_{t,T}) - \mathrm{cov}_t(m_{t,T} \cdot R^i_{t,T}, R^i_{t,T}). \qquad \text{(A-10)}$$

As noted in the main text, Kadan and Tang (2020) use Equation (A-10) for their quantification and approximation of stock risk premia.

Martin (2017) argues that for an asset return that qualifies as a market return proxy (denoted $R^m_{t,T}$), it should be the case that

$$\xi_{t,T} = \text{cov}_t\big(m_{t,T} \cdot R^m_{t,T}, R^m_{t,T}\big) < 0. \tag{A-11}$$

Intuitively, an investor's marginal rate of intertemporal substitution should be negatively correlated with any portfolio that qualifies as a market index. Accordingly,

$$\mathbb{E}_t(R^m_{t,T}) - R^f_{t,T} \geq \frac{1}{R^f_{t,T}} \text{var}^*_t(R^m_{t,T}). \tag{A-12}$$

Assuming that the inequality (A-12) is binding, we can use it with Equation (A-8), which yields:

$$\begin{aligned}\mathbb{E}_t\big(R^i_{t,T}\big) - R^f_{t,T} \approx R^f_{t,T} \cdot &\left[\text{var}^*_t\left(\frac{R^m_{t,T}}{R^f_{t,T}}\right) + \frac{1}{2}\left\{\text{var}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) - \sum_j w^j_t \cdot \text{var}^*_t\left(\frac{R^j_{t,T}}{R^f_{t,T}}\right)\right\}\right] \\ &- \frac{R^f_{t,T}}{2} \cdot \left[\text{var}^*_t(u^i_{t,T}) - \sum_j w^j_t \cdot \text{var}^*_t(u^j_{t,T})\right],\end{aligned} \tag{A-13}$$

where the approximative formula in Equation (A-13) omits the term $\kappa^i_{t,T} - \xi_{t,T}$ on the right-hand side. Equation (2.2) thus results from

$$a^i_{t,T} = \kappa^i_{t,T} - \xi_{t,T} - \zeta^i_{t,T}, \tag{A-14}$$

where

$$\zeta^i_{t,T} = \frac{1}{2} R^f_{t,T} \cdot \big[\text{var}^*_t(u^i_{t,T}) - \sum_j w^j_t \cdot \text{var}^*_t(u^j_{t,T})\big]. \tag{A-15}$$

Working with the abbreviated formula in Equation (2.5) thus entails three approximations: (1) the linearization of $(\beta^i_{t,T})^2$, (2) the assumption that Martin's (2017) lower bound for the expected return of the market is binding, and (3) the assumption that the residual variances $\mathrm{var}_t^*(u_{t,T}^i)$ are very similar across stocks, such that $\zeta_{t,T}^i$ is negligibly small in absolute terms.

A.2 Construction of the database (details)

Detailed information on how we identify HSPC in Compustat, CRSP, and OptionMetrics and how we retrieve information from these databases is provided in Section O.3 of the Online Appendix. Section O.6 explains how to access the Python programs that we use for this purpose.

The starting point for HSPC identification is Compustat. The number of HSPC we can trace in Compustat during the period of March 1964 to December 2018, is depicted in Panel A of Figure 12. We successfully recover many of the Compustat-identified HSPC also in CRSP, in particular after October 1974, the first month used for training the MLPs.

[Insert Figure 12 about here]

Panel A in Figure 12 shows that the matching procedure can identify a large fraction of the Compustat-identified HSPC also in OptionMetrics. The approximation formula in Equation (2.5) indicates that the higher the coverage of index stocks, the better the theory-based approach should perform, whereas a poor match adds another source of approximation error. The coverage rate that we achieve with our procedure is higher than that reported by Martin and Wagner (2019). Averaged over the respective sample periods, we succeed in recovering 483/500 HSPC; Martin and Wagner's (2019) coverage ratio is 451/500. Panel B of Figure 12 shows that the true S&P 500 market capitalization is closely tracked by that of the HSPC identified in Compustat, CRSP, and OptionMetrics.

A.3 Theory-based, stock-level, and macro-level variables

Table 9 give a description of the variables used in this study. The content from Panel B1 is obtained from Table A.6 in GKX. The stock-level features are retrieved using the SAS program kindly provided by Jeremiah Green that we update and modify for our purposes. These variables are originally used for the study by Green et al. (2017).

[Insert Table 9 about here]

A.4 Hyperparameter tuning and computational details

We adapt the search space for the hyperparameters of each machine learning model to the requirements of our restricted sample. In particular, GKX set the maximum depth of each tree in their random forest to 6. We increase this upper boundary to 30, which improves the validation results, especially at the one-year horizon. We also extend the search space for the elastic net's L1-ratio, which in GKX is fixed at 0.5, to allow for a more flexible combination of L1- and L2-penalization. For the gradient boosted regression trees, we limit the number of trees to the interval $[2, 100]$, increase the maximum tree depth to 3, and extend the interval for the learning rate to $[0.005, 0.12]$. In the case of the neural networks, we switch from the seed value-based ensemble approach advocated by GKX to dropout regularization, in combination with a structural ensemble approach, such that each neural network in the ensemble can have a different architecture. Ensemble methods have proven to be the gold standard in many machine learning applications, because they can subsume the different aspects learned by each individual model within a single prediction. However, creating ensembles can become prohibitively expensive if the number of sample observations is large and/or each individual model is highly complex.

Srivastava et al. (2014) address this issue by proposing dropout regularization, which retains the capability of neural networks to learn different aspects of the data while also being computationally more efficient than the standard ensemble approach. We also introduce a maximum weight norm for each hidden layer. By applying both dropout regularization and a structural ensemble approach with ten different neural networks per ensemble, we seek to combine the best of both worlds. Compared to GKX, we also reduce the batch size; a smaller batch size typically improves the generalization capabilities of a model that is trained with stochastic gradient descent (cf. Keskar et al., 2017). For a detailed comparison of the hyperparameter search spaces, refer to Table 1 in the main text and Table A.5 in GKX.

We implement our machine learning procedures using Python's scikit-learn ecosystem. To train neural networks, we rely on Python's deep learning library Keras with the Tensorflow backend. Although scikit-learn also supports the training of neural networks, it is less flexible than Keras and lacks some degrees of freedom in the construction of network architectures. To achieve maximum parallelization during our extensive hyperparameter search, we combine scikit-learn with the parallel computing environment Dask. Computations are performed on a high performance computing cluster.

A.5 Alternative feature transformation

Discussion

As described in the main text, we apply standard mean-variance or robust median-interquartile range scaling to the firm characteristics z_t^i, pooling across i and t. To prevent future information from leaking into the validation and test sets, the transformation of a feature within those sets is based on the mean, variance, median, and interquartile range in the associated training sets. In contrast, GKX scale firm characteristics to the interval $[-1, 1]$ period-by-period using cross-sectional ranks, as advocated by Freyberger et al. (2020). More specifically, they transform their set of firm characteristics according to

$$\tilde{c}_t^i = 2 \cdot \frac{\text{rank}(c_t^i)}{N_t + 1} - 1, \tag{A-16}$$

where N_t is the number of sample firms in period t.[27] The macroeconomic features x_t are not scaled, because for the individual time series there is no cross-section on the basis of which a rank transformation could be performed. As a consequence, the set of combined firm-level and macro features originates from

$$\tilde{z}_t^i = (1, x_t')' \otimes \tilde{c}_t^i. \tag{A-17}$$

Which feature scaling strategy is more suitable for the present application? The rank transformation in (A-16) invokes the idea of portfolio sorting, the hallmark of which is that "[one is] typically not interested in the value of a characteristic in isolation, but rather in the rank of the characteristic in the cross section" (Freyberger et al., 2020, pp. 16-17). In the same vein, Kozak et al. (2020) argue that by transforming firm characteristics according to their rank, they can focus on the "purely cross-sectional aspect of return predictability." However, the present study does not exclusively focus on the cross-section, but is also concerned with the *level* of stock risk premia. Using rank-transformed features, one cannot account for structural changes in the level of firm characteristics.[28]

Kelly et al. (2019) and Gu et al. (2021), point out that the rank transformation

[27] GKX give no indications as to their treatment of stocks that are tied in the ranking. We assume that they rank tied stocks as in Kozak et al. (2020) by assigning the average rank to each of the stocks.

[28] An obvious thing to note is that without scaling the macro features, the $\tilde{z}_t^i$ are not elements of $[-1, 1]$.

renders models less susceptible to outliers. However, Kelly et al. (2019) also report that the "results are qualitatively unchanged" compared to those obtained without rank transformation. Da et al. (2022) arrive at a similar conclusion, reporting that the rank transformation "barely changes any follow-up results." As we aim at finding the model that delivers MSE-optimal excess return predictions, the question of how to transform and scale firm characteristics is ultimately a matter of out-of-sample forecast performance (cf. Freyberger et al., 2020). Accordingly, we leave it up to the validation process whether to apply standard or robust scaling, noting that the latter mitigates the issue of outlier susceptibility.

To investigate whether our conclusions from the main analysis are affected by the chosen feature transformation strategy, we perform a supplementary analysis using rank-transformed firm-level features according to (A-16) and (A-17). We thereby acknowledge the code of conduct for research in empirical finance formulated by Arnott et al. (2019).

Results using rank-transformed firm-level features

Table 10 contains the *long training* results for both horizons. It is the counterpart of Panels B of Tables 2 and 3 from the main analysis.

[Insert Table 10 about here]

At the *one-month horizon* (Panel A of Table 10), RF and GBRT perform worse than the zero forecast, while ANN and ENet benefit from using rank transformed features. Compared to the main analysis, the predictive R^2 increase from 0.2% to 0.4% in case of the ANN and, quite conspicuously, from -0.3% to 0.5% in case of the ENet. Figure 13 depicts the results for prediction-sorted portfolios. It should be compared with Figure 4, the counterpart from the main analysis. The plots confirm the conclusion that the theory-based approach is difficult to beat at the one-month horizon, but also that the ENet is emerging as a new competitor.

[Insert Figure 13 about here]

Panel B of Table 10 shows that at the *one-year horizon* ENet, GBRT, and ANN by and large maintain their performance levels from the main analysis (cf. Panel B of Table 3). The ENet's R^2_{oos} increases from 5.5% to 6.9%, the predictive R^2 of ANN (from 9.0% to 8.1%) and GBRT (from 10.6% to 9.7%) decrease. In terms of R^2_{oos}, the RF is not as conspicuous as in the main analysis. The R^2_{oos} decreases from 19.5% to 9.6%, but with a Sharpe ratio of 0.67 (increasing from 0.58), the RF is the best approach when prediction-sorted portfolios are used for performance assessment. The ANN is ranked second with a Sharpe ratio of 0.63 (increasing from 0.50) and a favorable alignment of the prediction-sorted portfolios (see Figure 14, the counterpart of Figure 5 from the main analysis).

[Insert Figure 14 about here]

As can be seen in Table 11 – which should be compared to Table 5 from the main analysis – the *short training* effect is somewhat mitigated at the *one-month horizon*. Although still negative, the predictive R^2 delivered by the machine learning approaches no longer tend to extremes. As in the main analysis, the inclusion of theory features does not improve the one-month horizon results.

[Insert Table 11 about here]

At the *one-year horizon* with *short training*, our assessment of the model performances does not differ substantially from that of the main analysis (compare Table 12 with Table 6).

[Insert Table 12 about here]

In terms of R^2_{oos} and Sharpe ratio, the RF is the preferred model. Its R^2_{oos} increases from 12.4% to 15% and the Sharpe ratio of 0.59 remains unchanged. The ANN ranks second according to both criteria, with an R^2_{oos} of 11.5% (down from 14.1% in the main analysis) and a Sharpe ratio of 0.50 (up from 0.47). As in the main analysis, GBRT (deteriorating) and ENet (though notably improving) are no strong competitors. Table 12 further shows that the inclusion of theory features does not improve the performance of the machine learning models, at least when a monthly forecast frequency is considered.The conclusions regarding the *theory assisted by ML* strategy also hold with rank-transformed features, insofar as the predictive R^2 of 9.1% and the Sharpe ratio of 0.37 delivered by MW are notably improved by RF assistance. The MW+RF hybrid delivers an R^2_{oos} of 13.0%, a Sharpe ratio of 0.58, and a favorable alignment of the prediction-sorted portfolios (see Figure 15). Similar to the main analysis, the ANN assistance proves useful, too (the R^2_{oos} of MW+RF is 11.2%, the Sharpe ratio is 0.45), while GBRT or ENet assistance does not.

[Insert Figure 15 about here]

Overall, we find that the conclusions of the main analysis are also supported when using rank-transformed firm-level features.

A.6 Online Appendix

The contents of the Online Appendix are accessible at

https://drive.google.com/file/d/1kXeAq42zXkv5hqd5kS7AKxXXEmhxio_x/view?usp=sharing.

The Online Appendix is comprised of the following sections:

O.1 explains details regarding the construction of the database, in particular, on how to achieve the identification and selection of S&P 500 constituents, and on how the merge of Compustat, CRSP, and OptionMetrics data is performed.

O.2 explains in detail the use of the OptionMetrics volatility surface when computing risk-neutral variances.

O.3 provides a juxtaposition of our implementation of the machine learning procedures with that by GKX, and explains where and why we deviate.

O.4 reports the results when applying a Diebold-Mariano test for significant differences in the forecast performances, and when using alternative performance measures. There are also a feature importance analysis that focuses on the cross-sectional performance of MW and MW+RF based on Sharpe ratios instead of predictive R^2 and additional results pertaining to the disaggregated analysis.

O.5 provides access to our program code in order to enable reproduction studies: Python and SAS programs for the extraction and management of data, the computation of the theory-based measures, as well as the training of machine learning models.

References

Arnott, R., C. R. Harvey, and H. Markowitz (2019): "A Backtesting Protocol in the Era of Machine Learning," *The Journal of Financial Data Science.*

Avramov, D., S. Cheng, and L. Metzker (2021): "Machine Learning versus Economic Restrictions: Evidence from Stock Return Predictability," forthcoming: Management Science.

Bakshi, G., J. Crosby, X. Gao, and W. Zhou (2020): "A New Formula for the Expected Excess Return of the Market," Working Paper.

Brennan, M. J., A. W. Wang, and Y. Xia (2004): "Estimation and Test of a Simple Model of Intertemporal Capital Asset Pricing," *Journal of Finance*, 59(4), 1743–1775.

Bryzgalova, S., S. Lerner, M. Lettau, and M. Pelger (2022): "Missing Financial Data," Working Paper.

Bryzgalova, S., M. Pelger, and J. Zhu (2021): "Forest Through the Trees: Building Cross-Sections of Stock Returns," Working Paper.

Chabi-Yo, F., C. Dim, and G. Vilkov (2021): "Generalized Bounds on the Conditional Expected Excess Return on Individual Stocks," Working Paper.

Chen, L., M. Pelger, and J. Zhu (2021): "Deep Learning in Asset Pricing," forthcoming: Management Science.

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

Da, R., S. Nagel, and D. Xiu (2022): "The Statistical Limit of Arbitrage," Working Paper.

Feng, G., S. Giglio, and D. Xiu (2020): "Taming the Factor Zoo: A Test of New Factors," *Journal of Finance*, 75(3), 1327–1370.

Freyberger, J., B. Höppner, A. Neuhierl, and M. Weber (2021): "Missing Data in Asset Pricing Panels," Working Paper.

Freyberger, J., A. Neuhierl, and M. Weber (2020): "Dissecting Characteristics Nonparametrically," *Review of Financial Studies*, 33(5), 2326–2377.

Giglio, S., B. T. Kelly, and D. Xiu (2022): "Factor Models, Machine Learning, and Asset Pricing," forthcoming: Annual Review of Financial Economics.

Green, J., J. R. M. Hand, and X. F. Zhang (2017): "The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns," *Review of Financial Studies*, 30(12), 4389–4436.

Gu, S., B. Kelly, and D. Xiu (2020): "Empirical Asset Pricing via Machine Learning," *Review of Financial Studies*, 33(5), 2223–2273.

——— (2021): "Autoencoder Asset Pricing Models," *Journal of Econometrics*, 222(1), 429–450.

Hastie, T., R. Tibshirani, and J. Friedman (2017): *The Elements of Statistical Learning*, Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Kadan, O., and X. Tang (2020): "A Bound on Expected Stock Returns," *Review of Financial Studies*, 33(4), 1565–1617.

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.

Keskar, N., J. Nocedal, P. Tang, D. Mudigere, and M. Smelyanskiy (2017): "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima," 5th International Conference on Learning Representations, ICLR 2017; Conference date: 24-04-2017 through 26-04-2017.

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

Light, N., D. Maslov, and O. Rytchkov (2017): "Aggregation of Information About the Cross Section of Stock Returns: A Latent Variable Approach," *Review of Financial Studies*, 30(4), 1339–1381.

Lioui, A., and P. Maio (2014): "Interest Rate Risk and the Cross Section of Stock Returns," *Journal of Financial and Quantitative Analysis*, 49(2), 483–511.

Lundberg, S. M., and S.-I. Lee (2017): "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, 30.

Maio, P., and P. Santa-Clara (2012): "Multifactor Models and their Consistency with the ICAPM," *Journal of Financial Economics*, 106(3), 586–613.

——— (2017): "Short-Term Interest Rates and Stock Market Anomalies," *Journal of Financial and Quantitative Analysis*, 52(3), 927–961.

Martin, I. (2011): "Simple Variance Swaps," Working Paper.

——— (2017): "What is the Expected Return on the Market?," *Quarterly Journal of Economics*, 132(1), 367–433.

Martin, I., and S. Nagel (2021): "Market Efficiency in the Age of Big Data," forthcoming: Journal of Financial Economics.

Martin, I. W. R., and C. Wagner (2019): "What Is the Expected Return on a Stock?," *Journal of Finance*, 74(4), 1887–1929.

Merton, R. C. (1973): "An Intertemporal Capital Asset Pricing Model," *Econometrica*, 41(5), 867–887.

Petkova, R. (2006): "Do the Fama-French Factors Proxy for Innovations in Predictive Variables?," *Journal of Finance*, 61(2), 581–612.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014): "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, 15(56), 1929–1958.

Wang, K. (2018): "Risk-Neutral Cumulants, Expected Risk Premia, and Future Stock Returns," Working Paper.

Welch, I., and A. Goyal (2008): "A Comprehensive Look at The Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 21(4), 1455–1508.

Figures and Tables

Figure 1: Proportion of non-missing observations for each stock-level feature and year. This figure illustrates, for each of the stock-level features used in the machine learning approaches, the proportion of non-missing firm-date observations per year. The sample period ranges from 1964 to 2018, and the features are sorted from top to bottom in ascending order, according to their average proportion of non-missing observations. The darker the color, the more observations are available. The lighter the color, the less observations are available. All white indicates 100% missing values, the darkest blue means no missing values. The red vertical line indicates the year 1974, which is the first year that we use in the long training scheme described in Figure 2. Because of the excessive amount of missing values, we exclude the variables *real estate holdings* and *secured debt* from the empirical analysis.

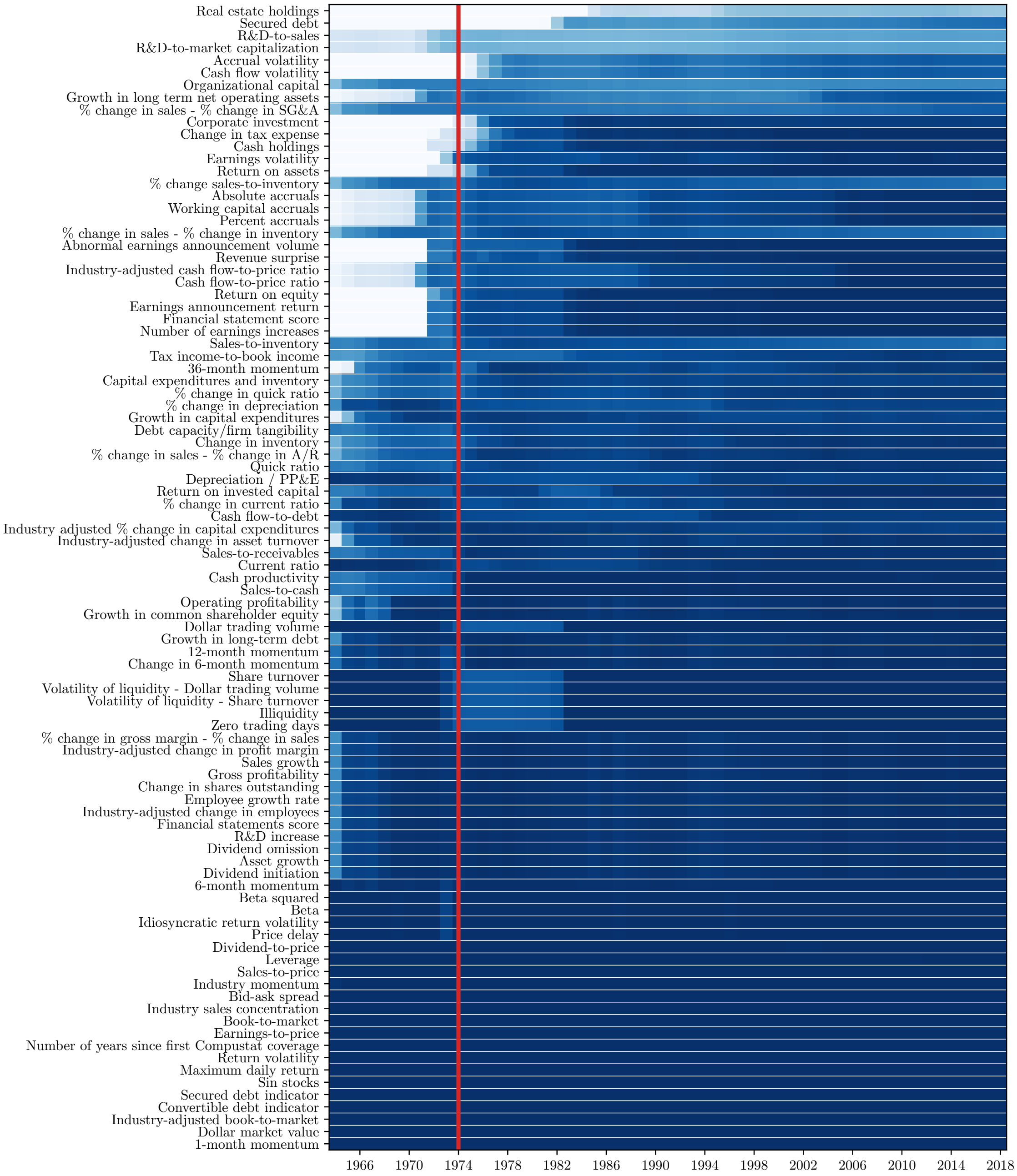

Figure 2: Long training scheme. The figure depicts the annual horizon variant of the long training scheme. The data range from October 1974 to December 2017. The training period (red/dark grey) initially spans 10 years and increases by one year after each validation step. Each of the 22 validation steps delivers a new set of parameter estimates. Each validation window (gold/light grey) covers 12 years and is rolled forward with a fixed width, followed by one year of out-of-sample testing (checkered blue).

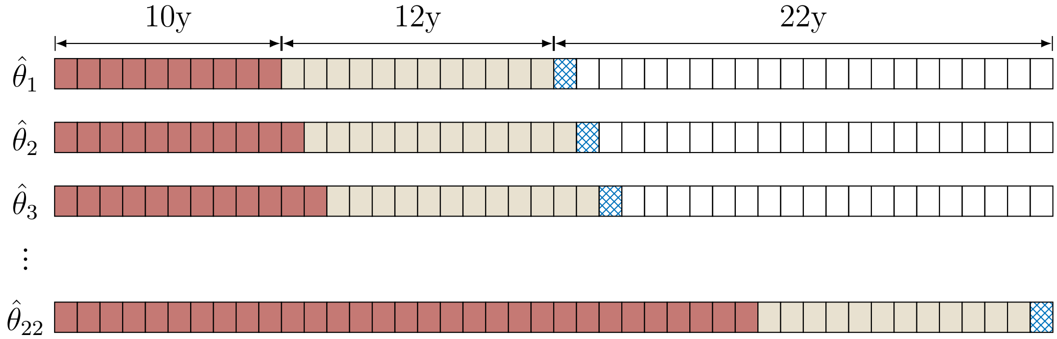

Figure 3: Short training scheme. The figure depicts the annual horizon variant of the short training scheme. The data range from January 1996 to December 2017. The training period (red/dark grey) initially spans one year and increases by one year after each validation step. Each of the 20 validation steps delivers a new set of parameter estimates. Each validation window (gold/light grey) covers one year, followed by one year of out-of-sample testing (checkered blue).

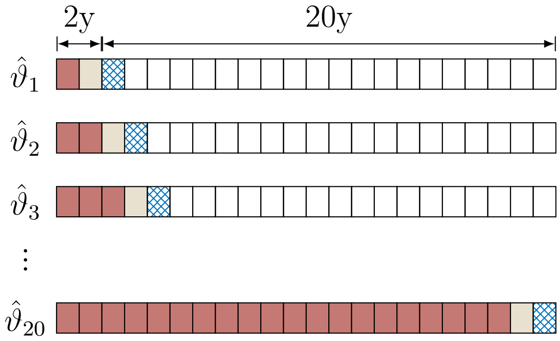

Figure 4: Prediction-sorted portfolios, one-month horizon: long training. The stocks are sorted into deciles according to the one-month horizon excess return prediction implied by the respective approach, and realized excess returns are computed for each portfolio. The prediction-sorted portfolios are formed either at the end of each month or daily. The four panels plot the predicted against realized portfolio excess returns (in %), averaged over the sample period. The numbers indicate the rank of the prediction decile. The rank correlation between predicted and realized excess returns in each panel is Kendall's τ. Approaches considered are MW (Panel A), an ANN (Panel C), and RF (Panel D). Panel B shows the MW results when the prediction-sorted portfolios are formed at a daily frequency. The out-of-sample period ranges from January 1996 to November 2018. Machine learning results are based on the long training scheme depicted in Figure 2.

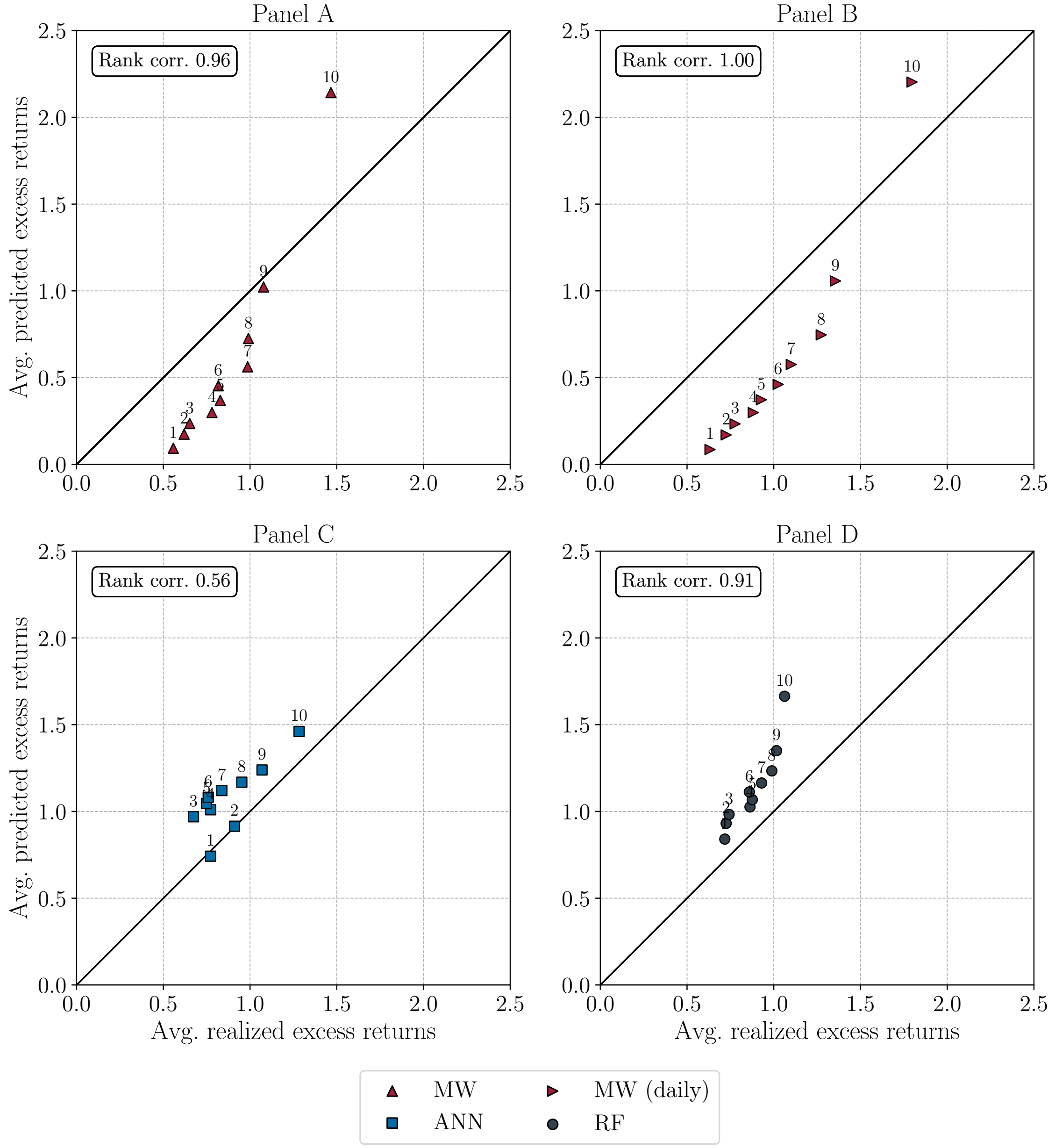

Figure 5: Prediction-sorted portfolios, one-year horizon: long training. The stocks are sorted into deciles according to the one-year horizon excess return prediction implied by the respective approach, and realized excess returns are computed for each portfolio. The prediction-sorted portfolios are formed either at the end of each month or daily. The four panels plot predicted against realized portfolio excess returns (in %), averaged over the sample period. The numbers indicate the rank of the prediction decile. The rank correlation between predicted and realized excess returns in each panel is Kendall's τ. Approaches considered are MW (Panel A), an ANN (Panel C), and RF (Panel D). Panel B shows the MW results when the prediction-sorted portfolios are formed at a daily frequency. The out-of-sample period ranges from January 1996 to December 2017. Machine learning results are based on the long training scheme depicted in Figure 2.

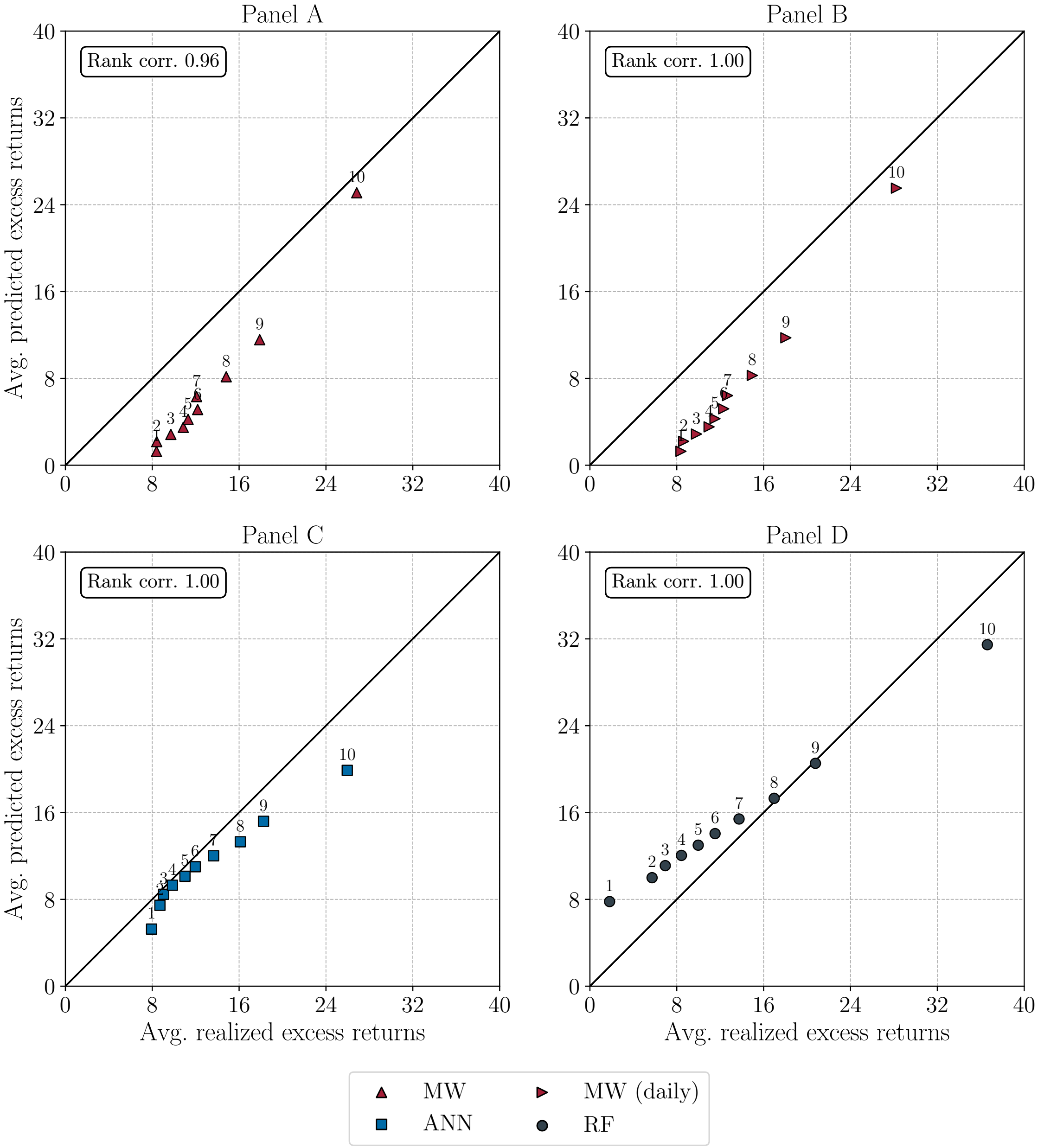

Figure 6: Time series of predictive R^2, one-year horizon: long training. The figure depicts the $R^2_{oos,s}$ time series based on annual test samples. The forecast horizon is one year; the prediction frequency is monthly (end-of-month). The out-of-sample period ranges from January 1996 to December 2017. Panel A contrasts the MW results with the RF, which in terms of R^2_{oos} is the best among the machine learning approaches. Panel B shows the $R^2_{oos,s}$ time series of the remaining approaches. The machine learning results are obtained using the long training scheme depicted in Figure 2.

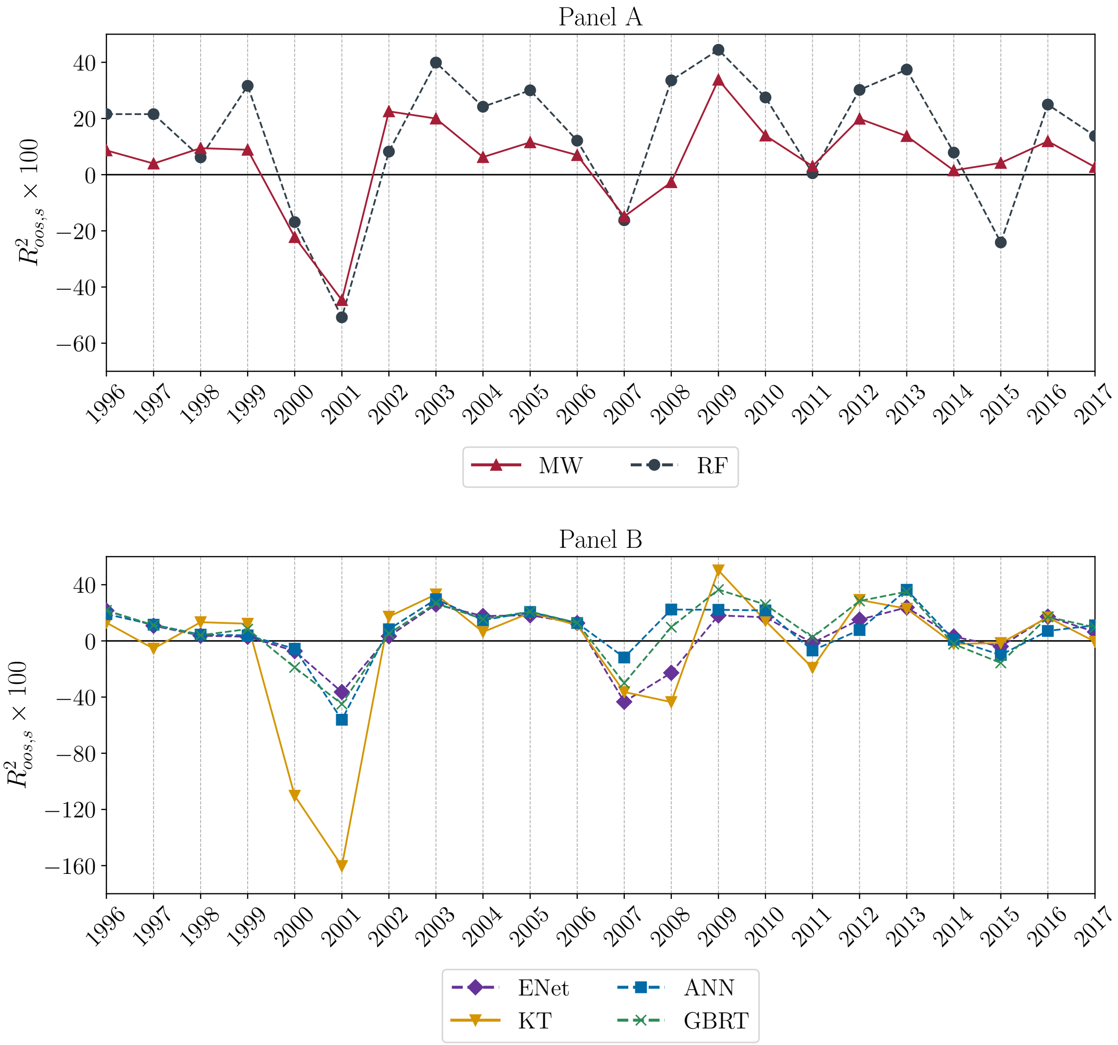

Figure 7: Time series of predictive R^2, one-year horizon: theory-based vs. machine learning with and without theory features. The figure depicts the $R^2_{oos,s}$ time series based on annual test samples. The forecast horizon is one year; the prediction frequency is monthly (end-of-month). The out-of-sample period ranges from January 1998 to December 2017. The machine learning results are obtained using the short training scheme depicted in Figure 3. For a comparison, we also display the $R^2_{oos,s}$ for MW and the long-trained RF from Panel A of Figure 6.

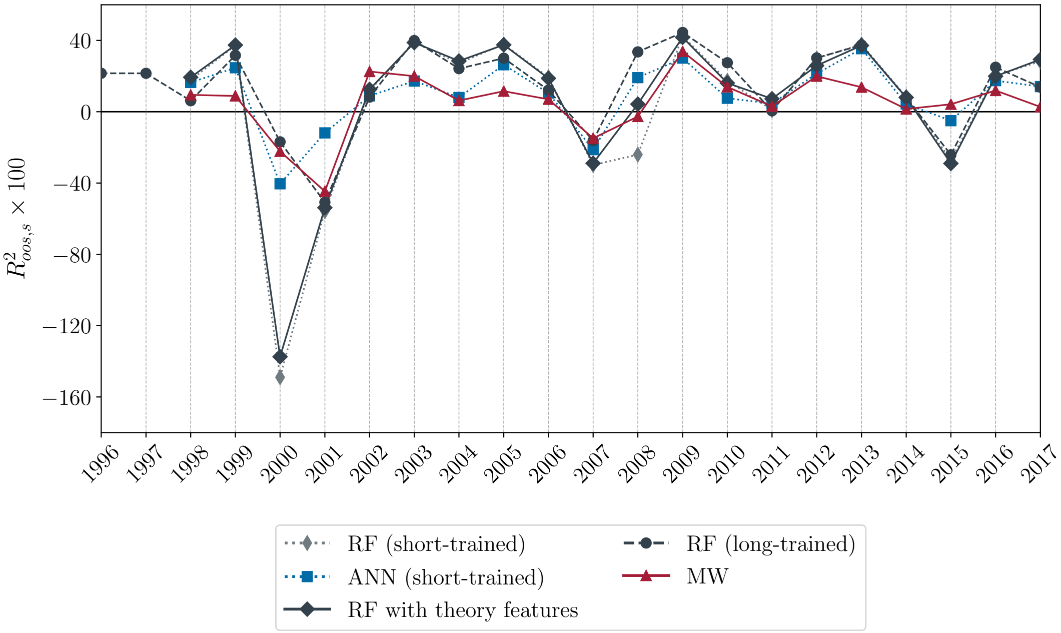

Figure 8: Time series of predictive R^2, one-year horizon: MW+RF vs. pure RF (long-training) vs. MW. The figure depicts the $R^2_{oos,s}$ time series based on annual test samples for the MW+RF hybrid (theory assisted by machine learning). The forecast horizon is one year; the prediction frequency is monthly (end-of-month). The out-of-sample period ranges from January 1998 to December 2017. The MW+RF results are based on the short training scheme depicted in Figure 3. For a comparison, we also display the $R^2_{oos,s}$ for MW and the long-trained RF from Panel A of Figure 6.

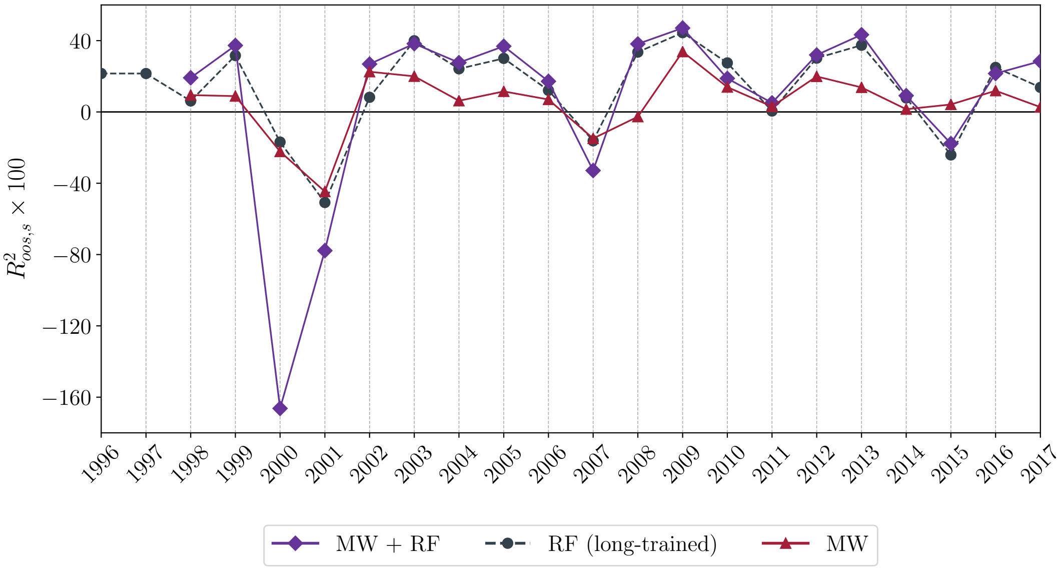

Figure 9: Prediction-sorted portfolios, one-year horizon: theory assisted by machine learning approaches. The stocks are sorted into deciles according to the one-year horizon excess return prediction implied by the respective approach, and realized excess returns are computed for each portfolio. The prediction-sorted portfolios are formed at the end of each month. The two panels plot predicted against realized portfolio excess returns (in %), averaged over the sample period. The numbers indicate the rank of the prediction decile. The rank correlation between predicted and realized excess returns in each panel is Kendall's τ. Approaches considered are MW assisted by an ANN (MW + ANN, Panel A) and MW assisted by RF (MW+RF, Panel B). The out-of-sample period ranges from January 1998 to December 2017. Results are based on the short training scheme depicted in Figure 3.

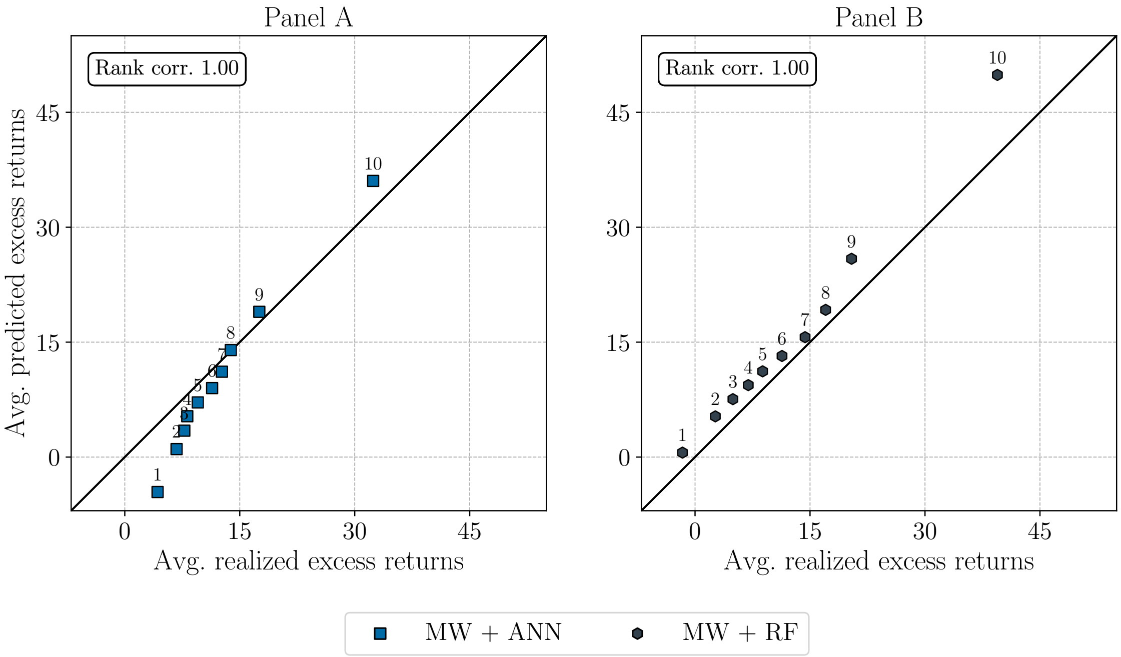

Figure 10: Feature importance, one-year horizon: random forest (short training). The figure depicts feature importance (Panel A: firm-level features, Panel B: macro-level features) for the RF. The forecast horizon is one year; the prediction frequency is end-of-month. A feature's importance is measured by the reduction of the predictive R^2 that is induced by setting the feature's values in the test samples to 0. In both panels, the features are sorted in descending order of importance. Panel A focuses on the ten most important firm-level features. The dashed vertical line, included for reference, represents the R^2_{oos} that is obtained without setting any feature's values to 0. The out-of-sample period ranges from January 1998 to December 2017. Results are based on the short training scheme depicted in Figure 3.

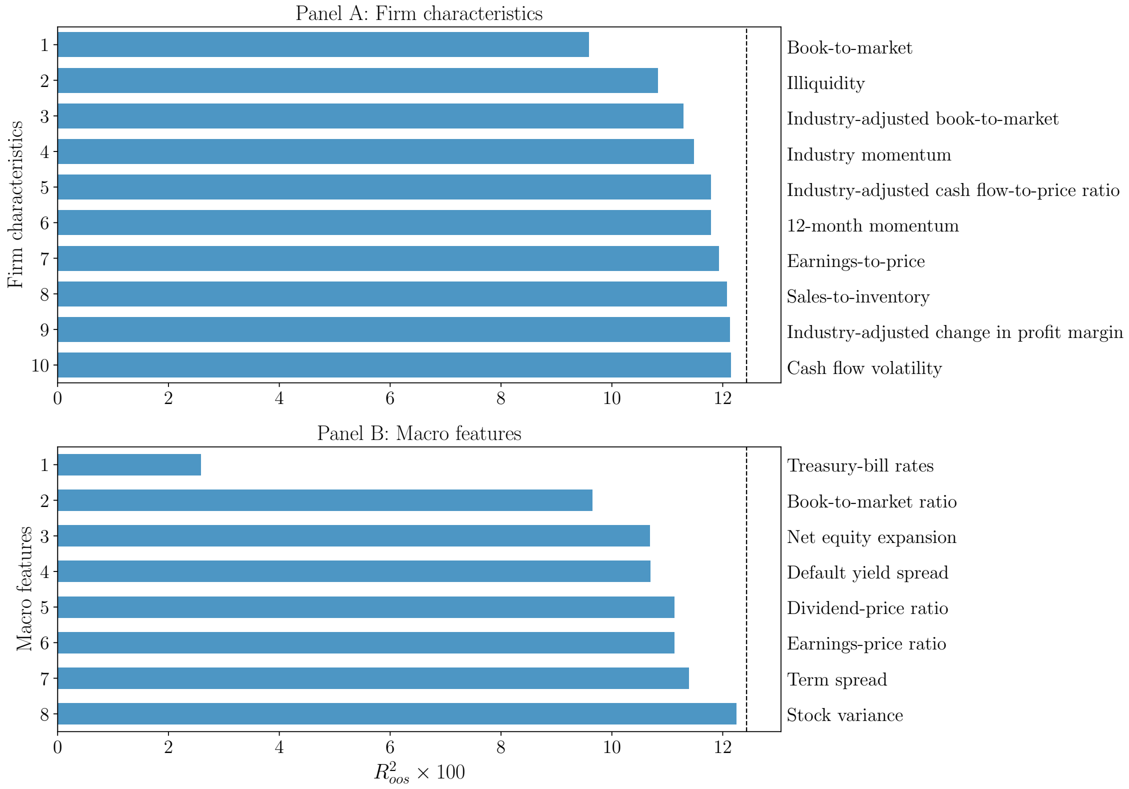

Figure 11: Feature importance, one-year horizon: MW+RF. The figure depicts feature importance (Panel A: firm-level features, Panel B: macro-level features) for the MW assisted by RF strategy. The forecast horizon is one year; the prediction frequency is end-of-month. A feature's importance is measured by the reduction in R^2 that is induced by setting the feature's values in the test samples to 0. In both panels, the features are sorted in descending order of importance. Panel A focuses on the ten most important firm-level features. The dashed vertical line, included for reference, represents the R^2_{oos} that is obtained without setting any feature's values to 0. The out-of-sample period ranges from January 1998 to December 2017. Results are based on the short training scheme depicted in Figure 3.

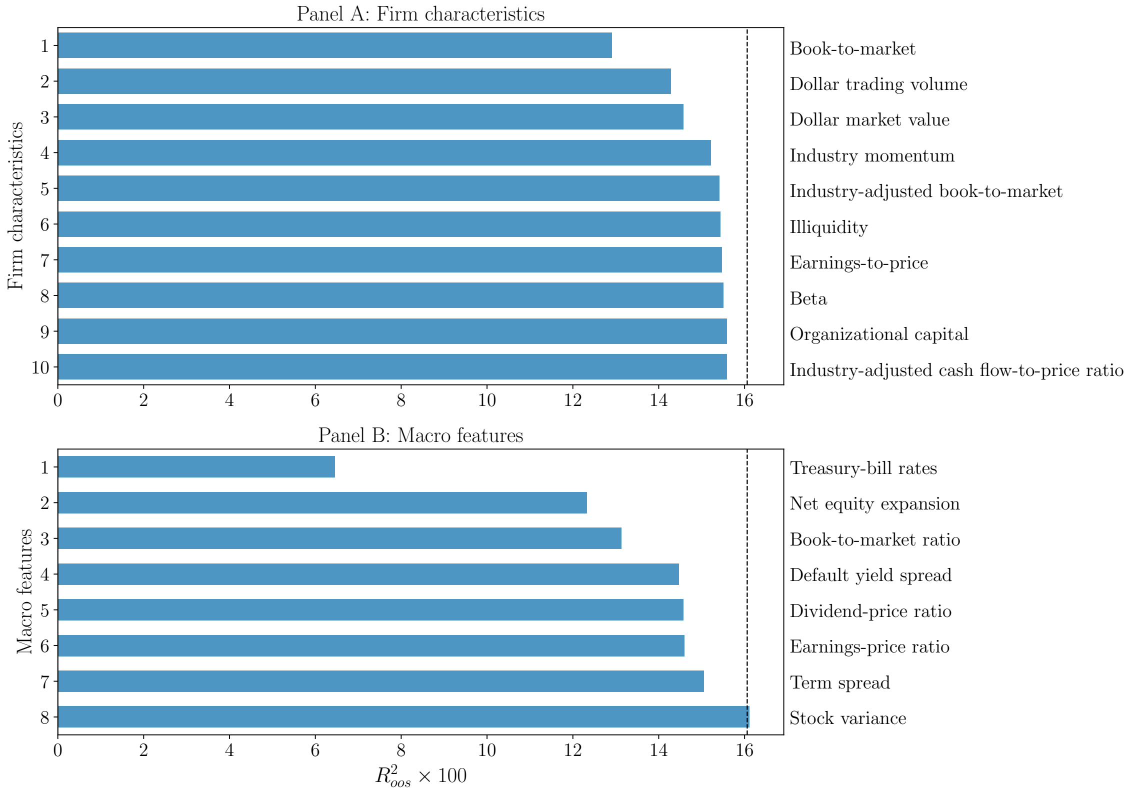

Figure 12: Identification of S&P 500 constituents. The figure illustrates the ability to detect historical S&P 500 constituents according to the implemented identification strategy. Panel A presents the coverage of HSPC achieved at different stages of the data processing. The line in light grey refers to the HSPC found in Compustat. The blue line shows for how many of these constituents it is possible to find stock price information in CRSP. The red line starting in 1996 illustrates for how many HSPC it is also possible to find information in OptionMetrics. Panel B depicts the aggregate market capitalization for each of these three groups of HSPC.

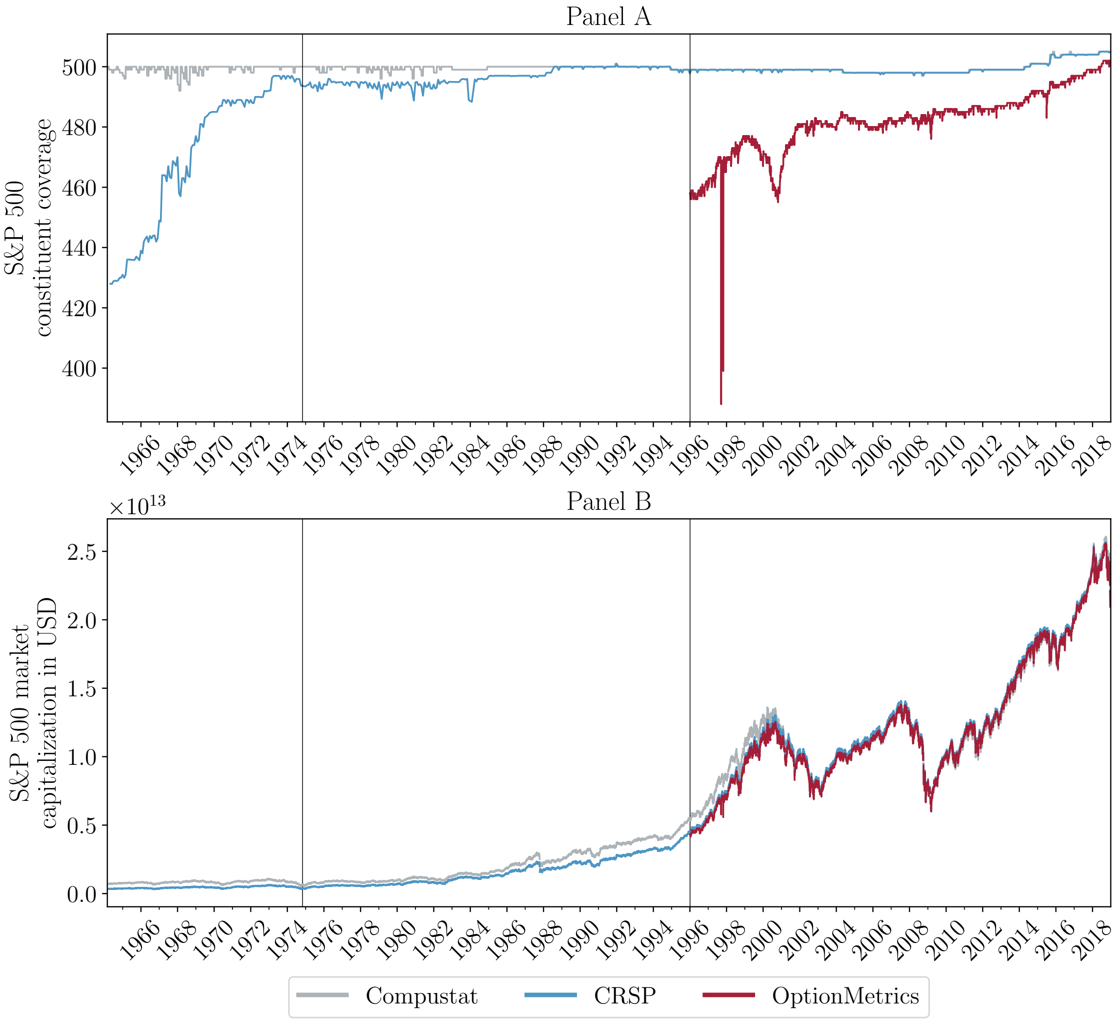

Figure 13: Prediction-sorted portfolios, one-month horizon: long training, rank transformation. The stocks are sorted into deciles according to the one-month horizon excess return prediction implied by the respective approach, and realized excess returns are computed for each portfolio. The prediction-sorted portfolios are formed either at the end of each month or daily. The four panels plot the predicted against realized portfolio excess returns (in %), averaged over the sample period. The numbers indicate the rank of the prediction decile. The rank correlation between predicted and realized excess returns in each panel is Kendall's τ. Approaches considered are MW (Panel A), ENet (Panel C), and RF (Panel D). Panel B shows the MW results when the prediction-sorted portfolios are formed at a daily frequency. The out-of-sample period ranges from January 1996 to November 2018. The features are rank-scaled as described in Appendix A.5. Machine learning results are based on the long training scheme depicted in Figure 2.

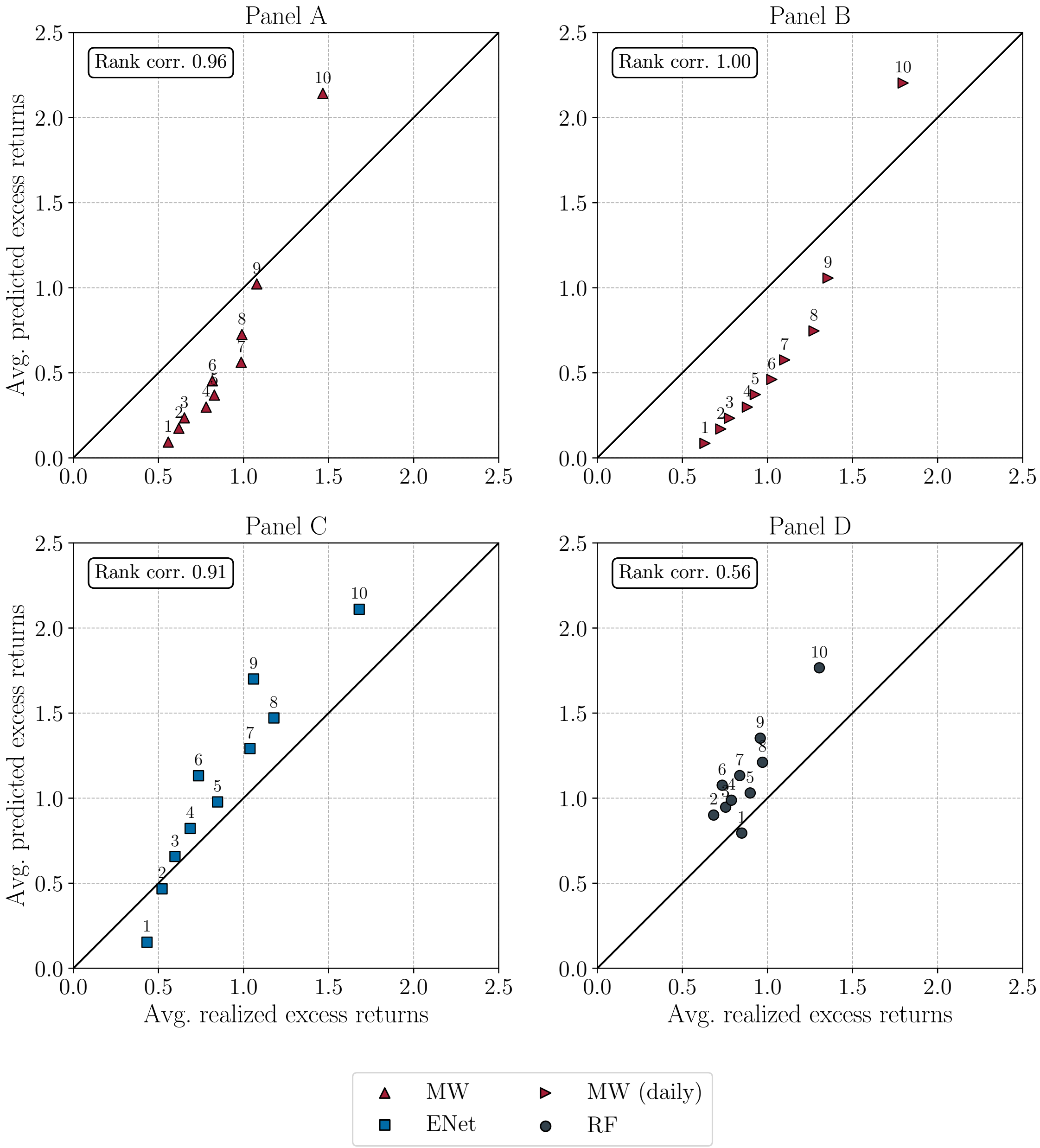

Figure 14: Prediction-sorted portfolios, one-year horizon: long training, rank transformation. The stocks are sorted into deciles according to the one-year horizon excess return prediction implied by the respective approach, and realized excess returns are computed for each portfolio. The prediction-sorted portfolios are formed either at the end of each month or daily. The four panels plot predicted against realized portfolio excess returns (in %), averaged over the sample period. The numbers indicate the rank of the prediction decile. The rank correlation between predicted and realized excess returns in each panel is Kendall's τ. Approaches considered are MW (Panel A), an ANN (Panel C), and RF (Panel D). Panel B shows the MW results when the prediction-sorted portfolios are formed at a daily frequency. The out-of-sample period ranges from January 1996 to December 2017. The features are rank-scaled as described in Appendix A.5. Machine learning results are based on the long training scheme depicted in Figure 2.

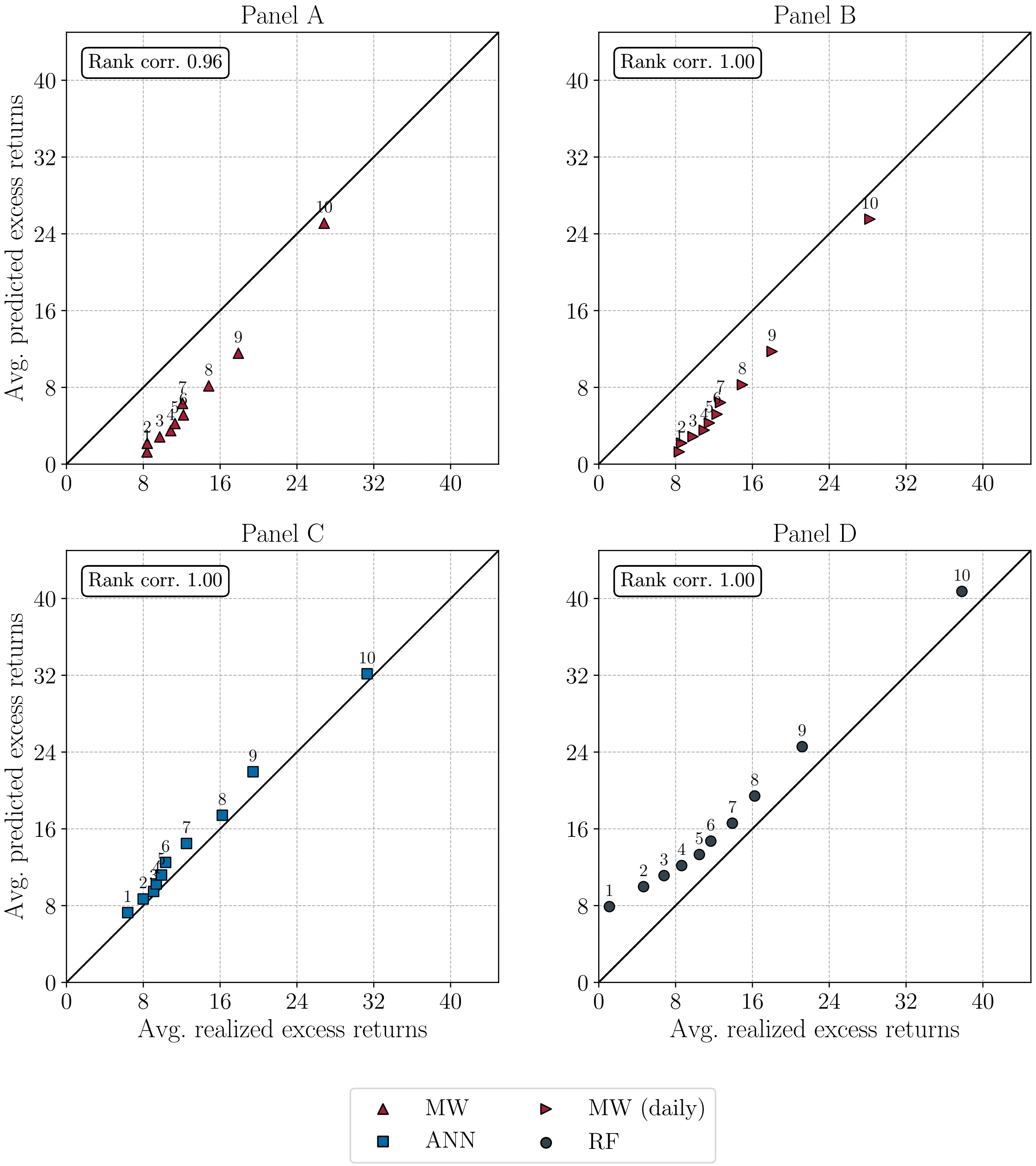

Figure 15: Prediction-sorted portfolios, one-year horizon: theory assisted by machine learning approaches (rank transformation). The stocks are sorted into deciles according to the one-year horizon excess return prediction implied by the respective approach, and realized excess returns are computed for each portfolio. The prediction-sorted portfolios are formed at the end of each month. The two panels plot predicted against realized portfolio excess returns (in %), averaged over the sample period. The numbers indicate the rank of the prediction decile. The rank correlation between predicted and realized excess returns in each panel is Kendall's τ. Approaches considered are MW assisted by an ANN (MW + ANN, Panel A) and MW assisted by RF (MW+RF, Panel B). The out-of-sample period ranges from January 1998 to December 2017. The features are rank-scaled as described in Appendix A.5. Results are based on the short training scheme depicted in Figure 3.

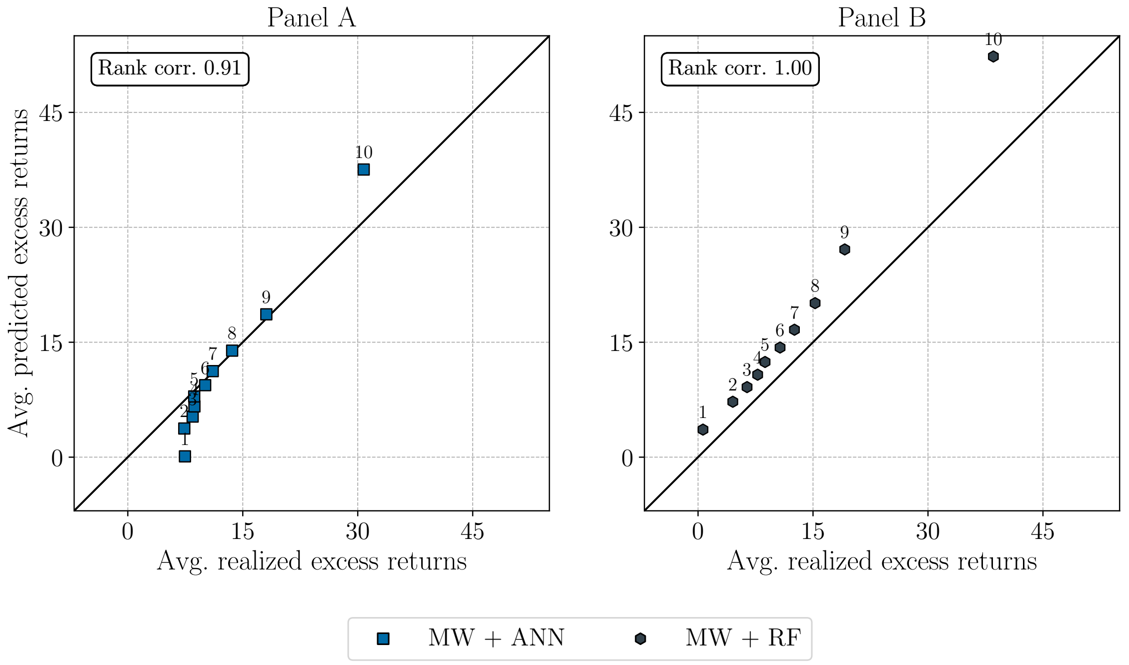

Table 1: Hyperparameter search space. This table shows the hyperparameter search space and the Python packages used for both long and short training. Parameter configurations not listed here correspond to the respective default settings.

Panel A: ENet	**Panel B: RF**
Package: Scikit-learn (SGDRegressor)	*Package:* Scikit-learn (RandomForestRegressor)
Feature transformation: Standard & robust scaling Selection by variance threshold	*Feature transformation:* Standard & robust scaling Selection by variance threshold
Model parameters: L1-L2-penalty: $\{x \in \mathbb{R} : 10^{-5} \leq x \leq 10^{-1}\}$ L1-ratio: $\{x \in \mathbb{R} : 0 \leq x \leq 1\}$	*Model parameters:* Number of trees: 300 Max. depth: $\{x \in \mathbb{N} : 2 \leq x \leq 30\}$ Max. features: $\{x \in \mathbb{N} : 2 \leq x \leq 150\}$
Optimization: Stochastic gradient descent Tolerance: 10^{-4} Max. epochs: $1,000$ Learning rate: $10^{-4}/t^{0.1}$	
Random search: Number of combinations: $1,000$	*Random search:* Number of combinations: 500

Panel C: GBRT	**Panel D: ANN**
Package: Scikit-learn (GradientBoostingRegressor)	*Package:* Tensorflow/Keras (Sequential)
Feature transformation: Standard & robust scaling Selection by variance threshold	*Feature transformation:* Standard & robust scaling Selection by variance threshold
Model parameters: Number of trees: $\{x \in \mathbb{N} : 2 \leq x \leq 100\}$ Max. depth: $\{x \in \mathbb{N} : 1 \leq x \leq 3\}$ Max. features: $\{20, 50, \text{All}\}$ Learning rate: $\{x \in \mathbb{R} : 5 \times 10^{-3} \leq x \leq 1.2 \times 10^{-1}\}$	*Model parameters:* Activation: TanH (Glorot), ReLU (He) Hidden layers: $\{1, 2, 3, 4, 5\}$ First hidden layer nodes: $\{32, 64, 128\}$ Network architecture: Pyramid Max. weight norm: 4 Dropout rate: $\{x \in \mathbb{R} : 0 \leq x \leq 0.5\}$ L1-penalty: $\{x \in \mathbb{R} : 10^{-7} \leq x \leq 10^{-2}\}$
	Optimization: Adaptive moment estimation Batch size: $\{100, 200, 500, 1,000\}$ Learning rate: $\{x \in \mathbb{R} : 10^{-4} \leq x \leq 10^{-2}\}$ Early stopping patience: 6 Max. epochs: 50 Batch normalization before activ. Number of networks in ensemble: 10
Random search: Number of combinations: 300	*Random search:* Number of combinations: $1,000$

Table 2: Performance comparison, one-month horizon: long training. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. For Panel A, the one-month horizon forecasts are issued at a daily frequency. For Panel B, the one-month horizon forecasts are issued at the end of each month. The out-of-sample testing period starts in January 1996 and ends in November 2018. The machine learning results are obtained using the long training scheme depicted in Figure 2.

Panel A: daily forecast frequency

		$R^2_{oos}\times100$	Std Dev	p-val.	SR
Theory-Based	MW	0.9	2.3	0.008	0.37
	KT	−0.5	5.3	0.530	0.37
Machine Learning	ENet	0.0	2.9	0.072	0.07
	ANN	0.5	3.1	0.038	0.26
	GBRT	0.3	2.9	0.036	0.29
	RF	−0.5	3.8	0.215	0.15

Panel B: monthly forecast frequency

		$R^2_{oos}\times100$	Std Dev	p-val.	SR
Theory-Based	MW	0.2	3.2	0.154	0.30
	KT	−1.8	6.9	0.704	0.30
Machine Learning	ENet	−0.3	3.5	0.161	0.00
	ANN	0.2	3.5	0.096	0.28
	GBRT	−0.6	4.2	0.248	0.20
	RF	−1.6	5.2	0.435	0.13

Table 3: Performance comparison, one-year horizon: long training. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized SR (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. For Panel A, the one-year horizon forecasts are issued at a daily frequency. For Panel B, the one-year horizon forecasts are issued at the end of each month. The out-of-sample testing period starts in January 1996 and ends in December 2017. The machine learning results are obtained using the long training scheme depicted in Figure 2.

Panel A: daily forecast frequency

		$R^2_{oos}\times 100$	Std Dev	p-val.	SR
Theory-Based	MW	9.1	16.0	0.040	0.38
	KT	3.5	47.5	0.675	0.38
Machine Learning	ENet	4.0	19.5	0.201	0.35
	ANN	8.2	17.6	0.029	0.49
	GBRT	9.9	19.9	0.039	0.36
	RF	18.2	22.6	0.003	0.56

Panel B: monthly forecast frequency

		$R^2_{oos}\times 100$	Std Dev	p-val.	SR
Theory-Based	MW	8.8	16.3	0.051	0.37
	KT	3.1	47.6	0.694	0.37
Machine Learning	ENet	5.5	18.5	0.125	0.36
	ANN	9.0	19.0	0.028	0.50
	GBRT	10.6	20.5	0.035	0.36
	RF	19.5	23.6	0.002	0.58

Table 4: Forecast correlations. The table reports Pearson correlation coefficients for the out-of-sample forecasts of the theory-based approaches (Martin and Wagner (2019); Kadan and Tang (2020)) and the four machine learning models with the long training scheme depicted in Figure 2. Panel A refers to a forecast horizon of one month with a testing period from January 1996 to November 2018. Panel B refers to a forecast horizon of one year and a testing period from January 1996 to December 2017. All forecasts are issued at the end of each month.

Panel A: One-month horizon

	ANN	RF	GBRT	ENet	KT
MW	0.01	0.25	0.32	−0.06	0.98
KT	0.02	0.25	0.31	−0.04	
ENet	0.32	0.70	0.45		
GBRT	0.11	0.82			
RF	0.22				

Panel B: One-year horizon

	ANN	RF	GBRT	ENet	KT
MW	0.19	0.33	0.34	0.00	0.98
KT	0.20	0.32	0.35	0.02	
ENet	0.69	0.49	0.57		
GBRT	0.70	0.72			
RF	0.59				

Table 5: Performance comparison, one-month horizon: theory-based vs. machine learning approaches vs. hybrid approach. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches, the four machine learning models, and a hybrid approach in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*). The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. For Panel A, the one-month horizon forecasts are issued at a daily frequency, and for Panel B, the one-month horizon forecasts are issued at the end of each month. The out-of-sample testing period starts in January 1998 and ends in November 2018. The machine learning results are obtained using the short training scheme depicted in Figure 3.

Panel A: daily forecast frequency

		R^2_{oos}×100	Std Dev	p-val.	SR
Theory-Based	MW	0.8	2.4	0.017	0.37
	KT	−0.7	5.5	0.590	0.37
Machine Learning	ENet	−4.0	8.1	0.844	0.33
	ANN	−2.7	5.0	0.864	0.22
	GBRT	−22.6	30.7	0.884	0.12
	RF	−5.4	7.8	0.924	−0.04
ML with theory features	ENet	−3.0	6.4	0.870	0.46
	ANN	−30.7	68.7	0.853	0.20
	GBRT	−10.7	21.5	0.844	0.37
	RF	−3.0	5.8	0.868	0.17

Panel B: monthly forecast frequency

		R^2_{oos}×100	Std Dev	p-val.	SR
Theory-Based	MW	0.1	3.4	0.206	0.32
	KT	−2.0	7.2	0.739	0.32
Machine Learning	ENet	−4.0	8.6	0.840	0.21
	ANN	−3.1	5.0	0.853	0.13
	GBRT	−29.5	57.7	0.860	0.15
	RF	−8.4	15.1	0.869	0.00
ML with theory features	ENet	−3.2	7.1	0.790	0.29
	ANN	−36.0	69.5	0.859	0.07
	GBRT	−25.6	53.1	0.855	0.20
	RF	−7.6	13.3	0.871	0.01

Table 6: Performance comparison, one-year horizon, monthly forecast frequency: theory-based vs. machine learning approaches vs. hybrid approaches. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. Results of two hybrid approaches, one in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*), and another in which the machine learning models are trained to account for the approximation residuals of MW (*Theory assisted by ML*), are also reported. The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. All results refer to a one-year forecast horizon and use the out-of-sample testing period January 1998 to December 2017. All forecasts are issued monthly (end-of-month). The machine learning results are obtained using the short training scheme depicted in Figure 3.

		$R^2_{oos}\times 100$	Std Dev	p-val.	SR
Theory-Based	MW	9.1	17.1	0.072	0.37
	KT	3.1	49.9	0.706	0.37
Machine Learning	ENet	–31.6	153.6	0.873	0.36
	ANN	14.1	18.1	0.004	0.47
	GBRT	10.3	36.6	0.308	0.45
	RF	12.4	45.1	0.329	0.59
ML with theory features	ENet	–32.6	160.3	0.868	0.36
	ANN	14.1	19.7	0.013	0.57
	GBRT	9.7	39.7	0.356	0.42
	RF	14.6	42.3	0.244	0.62
Theory assisted by ML	MW+ENet	–38.2	192.9	0.885	0.45
	MW+ANN	14.2	25.8	0.073	0.51
	MW+GBRT	9.2	45.2	0.440	0.40
	MW+RF	16.1	50.6	0.259	0.65

Table 7: Performance comparison, one-year horizon, daily forecast frequency: theory-based vs. machine learning approaches vs. hybrid approaches. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. Results of two hybrid approaches, one in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*), and another in which machine learning models are trained to account for the approximation residuals of MW (*Theory assisted by ML*), are also reported. The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. All results refer to a one-year forecast horizon and use the out-of-sample testing period January 1998 to December 2017. All forecasts are issued daily. The machine learning results are obtained using the short training scheme depicted in Figure 3.

		$R^2_{oos}\times100$	Std Dev	p-val.	SR
Theory-Based	MW	9.5	16.8	0.057	0.37
	KT	3.4	49.8	0.689	0.37
Machine Learning	ENet	–35.5	140.9	0.898	0.36
	ANN	12.0	18.7	0.032	0.45
	GBRT	8.8	36.9	0.394	0.44
	RF	9.0	46.1	0.462	0.56
ML with theory features	ENet	–27.4	138.6	0.861	0.38
	ANN	16.1	20.0	0.005	0.58
	GBRT	11.6	38.5	0.308	0.44
	RF	18.6	39.9	0.126	0.67
Theory assisted by ML	MW+ENet	–41.2	176.6	0.902	0.45
	MW+ANN	12.8	26.3	0.154	0.50
	MW+GBRT	8.2	47.1	0.522	0.40
	MW+RF	14.1	51.9	0.355	0.62

Table 8: Disaggregated performance comparison, one-year horizon, monthly forecast frequency. To obtain the results in Panel A, we sort the sample stocks into quintiles, according to the size of stock-specific valuation ratios (book-to-market and earnings-to-price), liquidity (Amihud illiquidity and dollar trading volume), and momentum (industry and 12-month). The sorting is renewed each month, taking into account the availability conditions outlined in Section 3. The pooled $R^2_{oos} \times 100$ according to Equation (3.2) is reported for each quintile portfolio and the approaches of interest, namely, MW, pure ML (ANN and RF), and theory assisted by machine learning (MW+RF and MW+ANN). Panel B shows the pooled $R^2_{oos} \times 100$ for each of the 10 industry portfolios based on the one-digit SIC code. The machine learning results are obtained using the short training scheme depicted in Figure 3.

Panel A: $R^2_{oos} \times 100$ for quintile portfolios

		Book-to-market					Earnings-to-price				
		Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3	Q4	Q5
Valuation ratios	MW	8.1	7.1	8.7	9.1	12.6	8.9	7.3	8.8	10.1	11.6
	ANN	14.7	17.1	11.9	14.0	12.1	13.1	14.6	16.8	13.4	14.1
	RF	6.7	16.2	9.4	17.8	15.4	8.0	13.0	17.7	16.1	16.7
	MW+ANN	14.9	15.7	10.8	13.4	14.9	13.1	13.8	16.7	14.5	15.5
	MW+RF	8.9	19.0	13.4	21.8	21.4	10.1	17.0	22.4	20.4	22.5
		Dollar trading volume					Amihud illiquidity				
		Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3	Q4	Q5
Liquidity	MW	15.7	10.5	10.2	6.2	−0.9	−1.0	4.1	7.3	10.7	14.9
	ANN	17.2	13.1	14.5	15.8	8.0	8.2	12.4	12.8	16.0	16.5
	RF	21.8	16.0	16.8	14.0	−11.3	−8.9	4.8	12.4	19.4	20.1
	MW+ANN	19.6	13.9	15.7	16.2	2.9	4.1	10.2	12.7	17.3	18.7
	MW+RF	27.5	20.0	20.7	17.1	−11.2	−7.5	8.1	15.1	23.3	25.0
		12-month momentum					Industry momentum				
		Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3	Q4	Q5
Momentum	MW	13.9	9.4	7.5	5.9	5.8	7.7	11.1	10.3	10.3	6.4
	ANN	13.9	11.1	14.8	13.2	15.9	13.0	17.4	15.3	14.0	10.8
	RF	15.2	12.7	13.1	15.3	7.2	13.1	18.9	19.6	11.1	0.5
	MW+ANN	17.0	10.8	13.1	12.2	14.3	11.9	17.8	16.1	16.2	9.5
	MW+RF	21.4	18.1	16.3	18.4	7.4	15.6	23.4	23.6	17.5	1.8

Panel B: $R^2_{oos} \times 100$ for industry portfolios (one digit SIC code)

	0	1	2	3	4	5	6	7	8	9
MW	6.6	5.4	11.9	8.0	9.0	8.7	12.0	8.0	16.9	2.1
ANN	23.9	12.7	12.2	15.8	16.6	8.1	12.0	17.3	3.6	12.9
RF	29.3	15.6	10.8	13.2	16.5	7.7	11.9	11.4	9.5	15.2
MW+ANN	22.7	8.3	13.4	14.3	19.2	8.6	15.6	16.5	11.0	18.4
MW+RF	31.6	18.1	14.6	16.0	22.5	12.5	18.1	12.4	21.5	12.6

Table 9: Variable description. The table contains information on the variables used for the empirical analysis. Panel A covers the theory/option-based risk premium measures proposed by Martin and Wagner (2019), Kadan and Tang (2020), and Martin (2017). The information in Panels B1 and B2 is taken from Table A.6 in Gu et al. (2020). For each variable, the table reports its debut in finance literature (author(s), year, journal), from which database it can be constructed (source), and at which frequency it is reported (freq.). For the stock-level features, we also supply the name of the respective variable used in the SAS program supplied by Jeremiah Green. The updated and modified program is provided in the Online Appendix, and can be used to trace the construction of each variable. The names of the macro-level variables come from Amit Goyal's original data files.

Panel A: Theory-based variables	**Source**	**Freq.**	**Author(s)**	**Year**	**Jnl.**
MW	Compustat, CRSP, OptionMetrics	Daily	Martin & Wagner	2019	JF
KT	Compustat, CRSP, OptionMetrics	Daily	Kadan & Tang	2019	RFS
Lower bound market equity premium	Compustat, CRSP, OptionMetrics	Daily	Martin	2017	QJE

Panel B1: Stock-level variables	**Code name**	**Source**	**Freq.**	**Author(s)**	**Year**	**Jnl.**
1-month momentum	mom1m	CRSP	Monthly	Jegadeesh & Titman	1993	JF
6-month momentum	mom6m	CRSP	Monthly	Jegadeesh & Titman	1993	JF
12-month momentum	mom12m	CRSP	Monthly	Jegadeesh	1990	JF
36-month momentum	mom36m	CRSP	Monthly	Jegadeesh & Titman	1993	JF
Abnormal earnings announcement volume	aeavol	Compustat, CRSP	Quarterly	Lerman, Livnat & Mendenhall	2007	WP
Absolute accruals	absacc	Compustat	Annual	Bandyopadhyay, Huang & Wirjanto	2010	WP
Accrual volatility	stdacc	Compustat	Quarterly	Bandyopadhyay, Huang & Wirjanto	2010	WP
Asset growth	agr	Compustat	Annual	Cooper, Gulen & Schill	2008	JF
Beta	beta	CRSP	Monthly	Fama & MacBeth	1973	JPE
Beta squared	betasq	CRSP	Monthly	Fama & MacBeth	1973	JPE
Bid-ask spread	baspread	CRSP	Monthly	Amihud & Mendelson	1989	JF
Book-to-market	bm	Compustat, CRSP	Annual	Rosenberg, Reid & Lanstein	1985	JPM
Capital expenditures and inventory	invest	Compustat	Annual	Chen & Zhang	2010	JF
Cash flow-to-debt	cashdebt	Compustat	Annual	Ou & Penman	1989	JAE
Cash flow-to-price	cfp	Compustat	Annual	Desai, Rajgopal & Venkatachalam	2004	TAR
Cash flow volatility	stdcf	Compustat	Quarterly	Huang	2009	JEF
Cash holdings	cash	Compustat	Quarterly	Palazzo	2012	JFE
Cash productivity	cashpr	Compustat	Annual	Chandrashekar & Rao	2009	WP
Change in 6-month momentum	chmom	CRSP	Monthly	Gettleman & Marks	2006	WP
Change in inventory	chinv	Compustat	Annual	Thomas & Zhang	2002	RAS
Change in shares outstanding	chcsho	Compustat	Annual	Pontiff & Woodgate	2008	JF
Change in tax expense	chtx	Compustat	Quarterly	Thomas & Zhang	2011	JAR
Convertible debt indicator	convind	Compustat	Annual	Valta	2016	JFQA
Corporate investment	cinvest	Compustat	Quarterly	Titman, Wei & Xie	2004	JFQA
Current ratio	currat	Compustat	Annual	Ou & Penman	1989	JAE

Table 9 continued . . .

Table 9 continued . . .

. . .	Code name	Source	Freq.	Author(s)	Year	Jnl.
Debt capacity/firm tangibility	tang	Compustat	Annual	Almeida & Campello	2007	RFS
Depreciation/PP&E	depr	Compustat	Annual	Holthausen & Larcker	1992	JAE
Dividend initiation	divi	Compustat	Annual	Michaely, Thaler & Womack	1995	JF
Dividend omission	divo	Compustat	Annual	Michaely, Thaler & Womack	1995	JF
Dividend-to-price	dy	Compustat	Annual	Litzenberger & Ramaswamy	1982	JF
Dollar market value	mve	CRSP	Monthly	Banz	1981	JFE
Dollar trading volume	dolvol	CRSP	Monthly	Chordia, Subrahmanyam & Anshuman	2001	JFE
Earnings announcement return	ear	Compustat, CRSP	Quarterly	Kishore, Brandt, Santa-Clara & Venkatachalam	2008	WP
Earnings-to-price	ep	Compustat	Annual	Basu	1977	JF
Earnings volatility	roavol	Compustat	Quarterly	Francis, LaFond, Olsson & Schipper	2004	TAR
Employee growth rate	hire	Compustat	Annual	Bazdresch, Belo & Lin	2014	JPE
Financial statement score (q)	ms	Compustat	Quarterly	Mohanram	2005	RAS
Financial statements score (a)	ps	Compustat	Annual	Piotroski	2000	JAR
Gross profitability	gma	Compustat	Annual	Novy-Marx	2013	JFE
Growth in capital expenditures	grcapx	Compustat	Annual	Anderson & Garcia-Feijoo	2006	JF
Growth in common shareholder equity	egr	Compustat	Annual	Richardson, Sloan, Soliman & Tuna	2005	JAE
Growth in long term net operating assets	grltnoa	Compustat	Annual	Fairfield, Whisenant & Yohn	2003	TAR
Growth in long-term debt	lgr	Compustat	Annual	Richardson, Sloan, Soliman & Tuna	2005	JAE
Idiosyncratic return volatility	idiovol	CRSP	Monthly	Ali, Hwang & Trombley	2003	JFE
(Amihud) Illiquidity	ill	CRSP	Monthly	Amihud	2002	JFM
Industry momentum	indmom	CRSP	Monthly	Moskowitz & Grinblatt	1999	JF
Industry sales concentration	herf	Compustat	Annual	Hou & Robinson	2006	JF
Industry-adjusted book-to-market	bm_ia	Compustat, CRSP	Annual	Asness, Porter & Stevens	2000	WP
Industry-adjusted cash flow-to-price ratio	cfp_ia	Compustat	Annual	Asness, Porter & Stevens	2000	WP
Industry-adjusted change in asset turnover	chatoia	Compustat	Annual	Soliman	2008	TAR
Industry-adjusted change in employees	chempia	Compustat	Annual	Asness, Porter & Stevens	1994	WP
Industry-adjusted change in profit margin	chpmia	Compustat	Annual	Soliman	2008	TAR
Industry-adjusted % change in capital exp.	pchcapx_ia	Compustat	Annual	Abarbanell & Bushee	1998	TAR
Leverage	lev	Compustat	Annual	Bhandari	1988	JF
Maximum daily return	maxret	CRSP	Monthly	Bali, Cakici & Whitelaw	2011	JFE
Number of earnings increases	nincr	Compustat	Quarterly	Barth, Elliott & Finn	1999	JAR

Table 9 continued . . .

Table 9 continued . . .

. . .	Code name	Source	Freq.	Author(s)	Year	Jnl.
Number of years since first Compustat coverage	age	Compustat	Annual	Jiang, Lee & Zhang	2005	RAS
Operating profitability	operprof	Compustat	Annual	Fama & French	2015	JFE
Organizational capital	orgcap	Compustat	Annual	Eisfeldt & Papanikolaou	2013	JF
% change in current ratio	pchcurrat	Compustat	Annual	Ou & Penman	1989	JAE
% change in depreciation	pchdepr	Compustat	Annual	Holthausen & Larcker	1992	JAE
% change in gross margin - % change in sales	pchgm_pchsale	Compustat	Annual	Abarbanell & Bushee	1998	TAR
% change in quick ratio	pchquick	Compustat	Annual	Ou & Penman	1989	JAE
% change in sales - % change in A/R	pchsale_pchrect	Compustat	Annual	Abarbanell & Bushee	1998	TAR
% change in sales - % change in inventory	pchsale_pchinvt	Compustat	Annual	Abarbanell & Bushee	1998	TAR
% change in sales - % change in SG&A	pchsale_pchxsga	Compustat	Annual	Abarbanell & Bushee	1998	TAR
% change sales-to-inventory	pchsaleinv	Compustat	Annual	Ou & Penman	1989	JAE
Percent accruals	pctacc	Compustat	Annual	Hafzalla, Lundholm & Van Winkle	2011	TAR
Price delay	pricedelay	CRSP	Monthly	Hou & Moskowitz	2005	RFS
Quick ratio	quick	Compustat	Annual	Ou & Penman	1989	JAE
R&D increase	rd	Compustat	Annual	Eberhart, Maxwell & Siddique	2004	JF
R&D-to-market capitalization	rde_mve	Compustat	Annual	Guo, Lev & Shi	2006	JBFA
R&D-to-sales	rd_sale	Compustat	Annual	Guo, Lev & Shi	2006	JBFA
Real estate holdings	realestate	Compustat	Annual	Tuzel	2010	RFS
Return on assets	roaq	Compustat	Quarterly	Balakrishnan, Bartov & Faurel	2010	JAE
Return on equity	roeq	Compustat	Quarterly	Hou, Xue & Zhang	2015	RFS
Return on invested capital	roic	Compustat	Annual	Brown & Rowe	2007	WP
Return volatility	retvol	CRSP	Monthly	Ang, Hodrick, Xing & Zhang	2006	JF
Revenue surprise	rsup	Compustat	Quarterly	Kama	2009	JBFA
Sales growth	sgr	Compustat	Annual	Lakonishok, Shleifer & Vishny	1994	JF
Sales-to-cash	salecash	Compustat	Annual	Ou & Penman	1989	JAE
Sales-to-inventory	saleinv	Compustat	Annual	Ou & Penman	1989	JAE
Sales-to-price	sp	Compustat	Annual	Barbee, Mukherji, & Raines	1996	FAJ
Sales-to-receivables	salerec	Compustat	Annual	Ou & Penman	1989	JAE
Secured debt indicator	securedind	Compustat	Annual	Valta	2016	JFQA
Share turnover	turn	CRSP	Monthly	Datar, Naik & Radcliffe	1998	JFM
Sin stocks	sin	Compustat	Annual	Hong & Kacperczyk	2009	JFE
Tax income-to-book income	tb	Compustat	Annual	Lev & Nissim	2004	TAR
Volatility of liquidity (dollar trading vol.)	std_dolvol	CRSP	Monthly	Chordia, Subrahmanyam & Anshuman	2001	JFE
Volatility of liquidity (share turnover)	std_turn	CRSP	Monthly	Chordia, Subrahmanyam, & Anshuman	2001	JFE
Working capital accruals	acc	Compustat	Annual	Sloan	1996	TAR

Table 9 continued . . .

Table 9 continued . . .

. . .	**Code name**	**Source**	**Freq.**	**Author(s)**	**Year**	**Jnl.**
Zero trading days	zerotrade	CRSP	Monthly	Liu	2006	JFE

Panel B2: Macro-level variables	**Code name**	**Source**	**Freq.**	**Author(s)**	**Year**	**Jnl.**
Book-to-market ratio	b/m	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Default yield spread	dfy	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Dividend-price ratio	dp	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Earnings-price ratio	eq	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Net equity expansion	ntis	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Stock variance	svar	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Term spread	tms	Amit Goyal	Monthly	Welch & Goyal	2008	RFS
Treasury bill rate	tbl	Amit Goyal	Monthly	Welch & Goyal	2008	RFS

Table 10: Performance comparison, monthly forecast frequency: long training, rank transformation. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. For Panel A, the forecast horizon is one month and for Panel B, it is one year. In both panels, forecasts are issued at the end of each month. The out-of-sample testing period starts in January 1996 and ends in November 2018. The features are rank-scaled as described in Appendix A.5. The machine learning results are obtained using the long training scheme depicted in Figure 2.

Panel A: one-month horizon

		R^2_{oos}×100	Std Dev	p-val.	SR
Theory-Based	MW	0.2	3.2	0.154	0.30
	KT	−1.8	6.9	0.704	0.30
Machine Learning	ENet	0.5	3.5	0.073	0.65
	ANN	0.4	3.4	0.053	0.34
	GBRT	−0.8	4.3	0.300	0.37
	RF	−0.8	4.8	0.294	0.17

Panel B: one-year horizon

		R^2_{oos}×100	Std Dev	p-val.	SR
Theory-Based	MW	8.8	16.3	0.051	0.37
	KT	3.1	47.6	0.694	0.37
Machine Learning	ENet	6.9	22.5	0.174	0.49
	ANN	8.1	22.1	0.097	0.63
	GBRT	9.7	23.1	0.086	0.49
	RF	9.6	43.3	0.361	0.67

Table 11: Performance comparison, one-month horizon, monthly forecast frequency: theory-based vs. machine learning approaches vs. hybrid approach, rank transformation. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches, the four machine learning models, and a hybrid approach in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*). The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. The one-month horizon forecasts are issued at the end of each month. The out-of-sample testing period starts in January 1998 and ends in November 2018. The features are rank-scaled as described in Appendix A.5. The machine learning results are obtained using the short training scheme depicted in Figure 3.

		R^2_{oos}×100	Std Dev	p-val.	SR
Theory-Based	MW	0.1	3.4	0.206	0.32
	KT	−2.0	7.2	0.739	0.32
Machine Learning	ENet	−0.1	2.8	0.277	0.26
	ANN	−0.1	2.9	0.163	0.04
	GBRT	−2.5	5.3	0.914	0.17
	RF	−4.7	8.3	0.898	−0.06
ML with theory features	ENet	−0.1	2.8	0.277	0.26
	ANN	−0.2	3.0	0.214	0.15
	GBRT	−8.5	15.9	0.926	0.19
	RF	−5.7	9.8	0.943	−0.11

Table 12: Performance comparison, one-year horizon, monthly forecast frequency: theory-based vs. machine learning approaches vs. hybrid approaches, rank transformation. The table reports predictive R^2, their standard deviation and statistical significance, and the annualized Sharpe ratios (SR) implied by Martin and Wagner's (2019) and Kadan and Tang's (2020) theory-based approaches and the four machine learning models. Results of two hybrid approaches, one in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*), and another in which the machine learning models are trained to account for the approximation residuals of MW (*Theory assisted by ML*), are also reported. The standard deviation of the $R^2_{oos,s} \times 100$ (Std Dev) is calculated based on the annual test samples. The SR refer to a zero-investment strategy long in the portfolio of stocks with the highest excess return prediction and short in the portfolio of stocks with the lowest excess return prediction. The p-values are associated with a test of the null hypothesis that the respective forecast has no explanatory power over the zero forecast, $\mathbb{E}(R^2_{oos,s}) \leq 0$. All results refer to a one-year forecast horizon and use the out-of-sample testing period January 1998 to December 2017. All forecasts are issued monthly (end-of-month). The features are rank-scaled as described in Appendix A.5. The machine learning results are obtained using the short training scheme depicted in Figure 3.

		R^2_{oos}×100	Std Dev	p-val.	SR
Theory-Based	MW	9.1	17.1	0.072	0.37
	KT	3.1	49.9	0.706	0.37
Machine Learning	ENet	4.3	25.3	0.388	0.49
	ANN	11.5	22.2	0.048	0.50
	GBRT	6.5	30.9	0.521	0.39
	RF	15.0	35.4	0.186	0.59
ML with theory features	ENet	4.3	25.3	0.385	0.49
	ANN	11.1	23.5	0.096	0.45
	GBRT	6.1	32.8	0.596	0.42
	RF	14.0	35.7	0.236	0.57
Theory assisted by ML	ENet	8.6	31.4	0.331	0.47
	ANN	11.2	27.7	0.183	0.45
	GBRT	6.2	38.7	0.548	0.40
	RF	13.0	42.4	0.320	0.58