Heterogeneous Responses in Financial Market: Evidence from Artificial Intelligence Learning *

Xiaoxiao Tang†, Xiwei Tang‡, and Guofu Zhou§

Abstract

We propose an AI-based framework that extends the Fama-MacBeth regression to capture individual-level heterogeneity in stock returns, building on the group-level heterogeneity approach introduced by Patton and Weller (2022). Leveraging 15 firm characteristics from Lewellen (2015), our model nearly doubles the Sharpe ratio of long-short portfolios compared to the Fama-MacBeth model. Our model is more interpretable and outperforms other machine learning models, even in high-dimensional settings with 94 characteristics. Our findings emphasize the importance of stock-level heterogeneity, particularly during recessions, challenging state-of-the-art asset pricing models that assume homogeneity.

Key Words: Artificial Intelligence Learning, Fama-MacBeth Approach, Cross-Section of Stock Returns, Out-of-Sample Prediction.

*We are grateful to Guanhao Feng, Zhe Geng, Mehdi Khorram, Haitao Mo, Junbo Wang, Yingquang Zhang, seminar participants at Boston College, Capital University of Economics and Business, Fudan University, Hunan Normal University, Hunan University, Jiangxi University of Finance and Economics, Louisiana State University, Nanjing University, Peking University, Renming University of China, Tongji University, Tsinghua University, University of Manitoba, University of Nottingham, Washington University in St. Louis, and Xi'An Jiaotong University, and conference participants at 2021 International FinTech Conference organized by Renmin University of China and 2023 Hong Kong Conference for Fintech, AI and Big Data in Business. We thank Songrun He for excellent research assistance.

†University of Texas at Dallas, Naveen Jindal School of Management. E-mail: xiaoxiao.tang@utdallas.edu.

‡University of Texas at Dallas, Department of Mathematical Sciences. E-mail: xiwei.tang@utdallas.edu.

§Washington University in St. Louis, Olin Business School. E-mail: zhou@wustl.edu.

1 Introduction

Artificial intelligence (AI) has increasing applications in all areas of science, especially in financial economics. However, the complexity and opacity of most AI frameworks, notably deep neural networks, necessitate their operation as "black boxes," which often lack interpretability and stability. Therefore, it is important to develop new-generation AI techniques with more transparency and interpretability that allow the embedding of human interventions and economic knowledge. In this paper, we propose a novel AI-empowered framework that extends the conventional Fama and MacBeth (1973) (FM) regression model, which offers rich economic interpretations. Our proposed model not only achieves superior predictive power but also reveals the pattern and potential mechanism of heterogeneous return responses to firm characteristics across individual stocks in the market.

The FM regression model is the most widely used framework for studying the risk factors that drive the cross-section of stock returns.[1] However, a significant drawback of the FM model is its assumption of a homogeneous response of stock returns to firm characteristics used as proxies for risk exposures. Market frictions can lead to heterogeneous responses among asset groups (e.g., Fama and French, 1993; Foerster and Karolyi, 1999; Griffin, 2002; Fama and French, 2008). Patton and Weller (2022) propose a formal solution to this heterogeneity problem, using a clustering approach to argue forcefully that risk premia are heterogeneous across asset groups. However, because firm characteristics are only imperfect proxies for risk exposures, the responses of stock returns to firm characteristics can further vary among individual stocks—leading to *complete* heterogeneity. Thus, the *partial* heterogeneity settings in Patton and Weller (2022) may not fully capture this pattern, as they still assume homogeneity within each group.[2] In this paper, we pro-

[1] For example, popular factor models today, such as the Fama and French (2015) five-factor model, the Hou et al. (2015) q-factor model, and the Stambaugh and Yuan (2017) mispricing-factor model, are closely tied to the FM regression framework. Additionally, numerous anomalies, such as those discussed in Hou et al. (2020) and Chen and Zimmermann (2021), are also related.

[2] Given the number of clusters is typically small, often below five, this within-group homogeneity assumption may fail to capture the diversity among individual stocks, as demonstrated in our later applications.

pose a novel extension of the FM regression, leveraging new-generation AI methodologies to allow for *complete* heterogeneity in return responses among individual stocks.

The core of our methodology lies in identifying individualized sensitivity of stock returns to firm characteristics for each firm. We demonstrate that this heterogeneity may arise from using firm characteristics as noisy proxies for risk exposures or from estimation errors in factor loadings. Our approach is built upon a cutting-edge machine learning framework: the heterogeneous learning model (Tang et al., 2021). The procedure starts with assigning each individual a set of unique coefficients in a regression manner, then imposes a fusion penalty to regularize pairwise differences between those individualized coefficients. The implementation is completely data-driven and employs the alternating direction method of multipliers algorithm (Boyd et al., 2011). In particular, the employed fusion penalization integrates information among similar individuals, constraining the number of free parameters and thus avoiding over-fitting. The flexible degrees of freedom inherent to this method offer the ability to cover any complex cross-sectional pattern, including group-wise heterogeneity, non-linearity, and interactions, among others. Nevertheless, in contrast to conventional AI techniques, under the transparent regression structure yet with AI-type high degrees of freedom, the architecture of this model is rather "flat" than "deep", which makes the proposed model more transparent and trackable.

To the best of our knowledge, our paper presents the first empirical framework that accounts for individual-level heterogeneity in the cross-section of stock returns. This significantly expands the conventional FM model and its existing extensions, incorporating the FM model as a special case. Leveraging advanced AI techniques, our approach offers a high-resolution map of firm characteristics as proxies for risk exposures across various individual stocks. Rather than adding more risk factors, our method enhances the explanatory power of cross-sectional variation in stock returns by exploiting the diversity in return responses across individual stocks. As demonstrated empirically later, our approach facilitates precise forecasting of individual stock returns. This, in turn, highlights the importance of introducing heterogeneity at the individual stock level, moving beyond cluster- or segmentation-based frameworks.

Capturing heterogeneity among individual stocks helps mitigate the errors induced by using firm characteristics as proxies for risk exposures, yielding enhanced predictive power for future stock returns and thus creating potential opportunities for profitable trading. In line with this perspective, we investigate the out-of-sample performance of our heterogeneous model by adopting the standard approach outlined in Lewellen (2015) for constructing trading strategies based on stock return predictions derived from 15 firm characteristics. The sample includes all CRSP common stocks from January 1964 to December 2022. The equal-weighted (value-weighted) long-short portfolio based on our model earns an average monthly return of 3.13% (1.57%) and has an annualized Sharpe ratio of 2.01 (0.86).[3] In contrast, predictions based on the FM model generate an equal-weighted (value-weighted) long-short portfolio with an average monthly return of 1.98% (0.69%) and an annualized Sharpe ratio of 1.18 (0.38), which are comparable with those levels in Lewellen (2015). Hence, the out-of-sample performance of our heterogeneous model almost doubles that of the standard homogeneous model. The result demonstrates that our proposed heterogeneous model yields substantial gains in economic value.

It is of interest to compare the out-of-sample performance of our method with that of the market segmentation approach, an ad hoc solution to the heterogeneous response problem. This approach groups assets based on prior knowledge (e.g., Fama and French (1993); Foerster and Karolyi (1999); Griffin (2002)) and then applies the homogeneous FM regression within each group, effectively assuming different risk premia among the market segments. Besides requiring prior knowledge, it is often challenging to group stocks by multiple characteristics. In our application, the equal-weighted (value-weighted) long-short portfolio based on the firm-size segmentation model improves upon the standard FM model with annualized Sharpe ratios from 1.18 (0.38) to 1.55 (0.57). However, this is still quite below the 2.01 (0.86) achieved by our heterogeneous approach. Segmentation models based on other criteria, such as industry, yield similar results. Therefore,

[3]We predict the stock returns at the end of each month with the available characteristics. During the following month, the long-short portfolio buys stocks in the top decile with the highest predicted returns and short-sells those in the bottom decile with the lowest predicted returns. The portfolio is rebalanced monthly. Such a trading strategy is real-time implementable and does not suffer from a look-ahead bias.

prespecified asset segmentation alone is unlikely to fully capture the heterogeneous sensitivity of stock returns to firm characteristics.

Our approach also adds to the group-heterogeneous method pioneered by Patton and Weller (2022). Their framework (the PW model) identifies clusters of assets with similar risk prices without requiring prespecified knowledge, thus allows learning from the data. However, while it captures heterogeneity across clusters, it still imposes homogeneity within each cluster. In contrast, our approach completely relaxes the homogeneity assumption, utilizing fusion regularization to allow the data to determine the optimal level of heterogeneity, effectively addressing the nuanced variations among individual stocks.

Empirically, for stock-level applications, our approach consistently outperforms the PW model, yielding significantly higher average returns and Sharpe ratios for both equal-weighted and value-weighted long-short portfolios.[4] This is because the average returns and Sharpe ratios of the long-short portfolios constructed by the PW model are comparable to those of the standard homogeneous FM model, highlighting the PW model's limitations when applied to individual stocks. These results indicate that the complete heterogeneity in our method is crucial for stock-level analysis.

On the other hand, machine learning and AI techniques have become increasingly prevalent in empirical asset pricing research and can be viewed as alternative extensions of the standard FM regression model.[5] We further compare our approach with a set of popular machine learning methods, including the tree-based methods and the neural networks, in both low-dimensional and high-dimensional settings. Consider first the 15-characteristics setting. The best machine learning model, the neural network, generates an average monthly return of 2.24% (1.00%) for an equal-weighted (value-weighted) long-short portfolio, yielding an annualized Sharpe ratio of 1.46 (0.58).

[4]We apply the PW model to estimate clusters of stocks with different slopes and use these clusters for return forecasts. Following Patton and Weller (2022), we select the number of clusters as 2, 3, and 4.

[5]Chinco et al. (2019), Freyberger et al. (2020), Gu et al. (2020), Kozak et al. (2020), Chen et al. (2023), Cong et al. (2021), Avramov et al. (2023), and Cui et al. (2024), among others, are examples of using machine learning techniques to forecast stock returns based on firm characteristics.

Though improving the standard homogeneous FM model substantially, it still underperforms our proposed heterogeneous approach, which provides an average monthly return of 3.10% (1.41%) and an annualized Sharpe ratio of 2.17 (0.81%).[6] We further test all models on a set of 94 characteristics as studied in Gu et al. (2020). The superior out-of-sample performance of our approach persists in such a high-dimensional setting, delivering the highest return and Sharpe ratio compared to all machine learning methods. These results confirm the ability of our method to deal with asset pricing models based on a large number of firm characteristics.

The superior performance of our approach sends an important message: even in the age of large-scale and complexity-favored AI, incorporating appropriate and interpretable economic structures can add substantial value to predicting stock returns beyond mere model complexity. Our empirical findings suggest that heterogeneity in return responses plays a more significant role in predicting stock returns than the non-linear and interaction effects usually captured by standard machine learning and AI techniques. Furthermore, unlike most traditional AI techniques, which often function as "black boxes," our methodology resembles an "open AI" framework, capable of visualizing the identified heterogeneous patterns of return responses, thereby shedding light on the potential cross-sectional structure of individual stocks. The results from our approach lay the groundwork for further economic analyses, such as investigating which firm characteristics contribute to the observed heterogeneity.

To determine which characteristics yield these heterogeneous responses, we conduct a test to examine the cross-sectional variation of heterogeneous responses with respect to each characteristic. Using a data-driven algorithm, we analyze the firm-wise response estimates from our heterogeneous model and test the differences between groups based on the time-series average of responses. Our results reveal that at least 11 out of the 15 characteristics show heterogeneous return responses, underscoring the critical role of incorporating heterogeneous return responses

[6]The prediction period in the comparison is shorter compared to our main analysis due to the need for training and validation samples by machine learning methods, so the statistics here are slightly different from our main analysis. See Section 4.3 for more details.

in asset-pricing models. In addition, we test the performance of models that allow heterogeneity in the response to only one firm characteristic at a time. Our findings show that models permitting heterogeneous responses for characteristics identified as more heterogeneous perform better, suggesting that accounting for heterogeneity in key characteristics enhances model performance. This improvement highlights that a better model not only offers theoretical advantages, but also translates them into tangible economic and investment value.

We further explore the time-series pattern of the cross-sectional variation of individualized firm-wise responses to each characteristic. We compute standard deviations of return responses and find that 14 out of 15 responses are less heterogeneous during expansion than during recession periods.[7] This suggests that the standard homogeneous model is less likely to hold when the market is in a bad state. Therefore, employing our heterogeneous model is particularly valuable during recessions or in a state in which the uncertainty is high.

In short, our paper contributes to a range of areas in empirical finance by extending the standard FM regression to allow complete heterogeneous responses. This extension accounts for the limitations of firm characteristics as proxies for risk exposures and the estimation errors in factor loadings. Fama and French (1993); Foerster and Karolyi (1999); Griffin (2002), among others, segment the market with prior knowledge to introduce heterogeneity, while our work extends the data-driven method of Patton and Weller (2022) to the stock level. Amidst the extensive literature on employing firm characteristics and corporate variables to predict or explain the future cross-section of expected stock returns, our method can be applied to assess the heterogeneous responses and the corresponding economic value across a wide range of applications.

Our work also adds to the empirical studies that predict the cross-section of stock returns with different firm characteristics representing risk factors. Lewellen (2015) shows the standard FM approach has exceptional out-of-sample performance relying on 15 firm characteristics. More recently, machine learning methodologies have been introduced to generate consistently superior

[7]The remaining firm characteristic exhibits similar cross-sectional variation in the stock return response during expansion compared to recession periods.

predictive abilities on stock returns (Chinco et al. (2019), Freyberger et al. (2020), Gu et al. (2020), Kozak et al. (2020), Chen et al. (2023), Cong et al. (2021), Avramov et al. (2023), and Cui et al. (2024)). Our work suggests that allowing stock-level heterogeneity also delivers significant investment gains.

The rest of this paper is organized as follows. Section 2 introduces the heterogeneous model. Section 3 tests the predictive power of the proposed model. We compare our model with other popular extensions to the FM model in Section 4. Section 5 investigates the underlying heterogeneity mechanism. Further analyses are presented in Section 6. Section 7 concludes.

2 Methodology

2.1 Heterogeneous-response model

Consider an economy with N^t stocks at each time t from $t = 1, \ldots, T$. Let r_i^t denote the return for the i-th firm at time t. Consider a J-factor model:

$$r_i^t = \alpha_i^t + \sum_{j=1}^{J} \beta_{ij}^t f_j^t + \epsilon_i^t, \tag{1}$$

where $\{f_j^t\}_{j=1}^J$ are J systematic risk factors and $\{\beta_{ij}^t\}_{j=1}^J$ are the associated factor loadings for firm i. According to the linear asset pricing rule, we have that

$$\mathbf{E}(r_i^t) = \gamma_0^t + \sum_{j=1}^{J} \gamma_j^t \cdot \beta_{ij}^t, \tag{2}$$

where γ_j^t is the risk premium for factor j at time t. Given the factor loadings $\{\beta_{ij}^t\}_{j=1}^J$, the Fama-MacBeth (FM) regression provides a simple way to estimate the risk premium from the cross-

sectional regression:

$$r_i^t = \gamma_0^t + \sum_{j=1}^{J} \gamma_j^t \beta_{ij}^t + \epsilon_i^t. \tag{3}$$

When applying the model to predict future stock returns, a conventional approach is to use lagged firm characteristics as proxies for $\{\beta_{ij}^t\}_{j=1}^J$. This leads to the standard FM regression, where asset returns exhibit a *homogeneous* and linear relationship with firm characteristics:

$$r_i^t = b_0^t + \sum_{j=1}^{J} b_j^t \cdot char_{ij}^{t-1} + \epsilon_i^t, \tag{4}$$

where $char_{ij}^{t-1}$ is the j-th characteristic for the firm i at time $t-1$, and b_j^t represents the homogeneous response of stock returns to characteristic j at time t.

Under this standard model, the sensitivity of stock returns to each characteristic, measured by b_j^t, is the same across all stocks. As a result, when two firms experience the same amount of change in their characteristics, their incremental expected returns must also be the same. Such a homogeneous response assumption is clearly just a simplification.

To illustrate such simplification, we consider two types of characteristics that are usually used in FM regressions. First is an actual firm characteristic. Suppose that, for simplicity, size is the only risk factor. We use the lagged firm size, $size_i^{t-1}$, as the proxy for size exposure, β_i^t, for firm i to predict its return at time t. Let $\eta_i^t = \frac{\beta_i^t}{size_i^{t-1}}$ denote the proxy error, then the FM regression (3) becomes

$$r_i^t = b_0^t + (\gamma^t \eta_i^t)\, size_i^{t-1} + \epsilon_i^t, \tag{5}$$

where the corresponding slope is $\gamma^t \eta_i^t$. This slope can be heterogeneous among asset groups due to varying levels of risk premium γ^t, as acknowledged in the literature, and it can also be heterogeneous among individual stocks due to varying proxy errors η_i^t. Market frictions and segmentation may cause the price of the same factor to differ across stock groups, while the use of firm char-

acteristics as imperfect proxies for risk exposures can result in different levels of proxy error for each firm. The combination of these effects leads to completely heterogeneous responses of stock returns to firm characteristics.

This conclusion is also true for the second type of firm characteristics: estimated factor loadings, like the market beta. To see this, consider the case in which the market is the only factor in (1). To predict stock returns for firm i, the corresponding firm characteristics in the FM regression (3) is conventionally estimated from a first-stage time-series regression, $\hat{\beta}_{Mi}^{t-1}$, based on data available as of time $t-1$. Since the true market beta, β_{Mi}^{t}, is unknown, the estimated beta inherently contains an error. Let $\theta_i^t = \beta_{Mi}^t / \hat{\beta}_{Mi}^{t-1}$ denote the estimation error, which can clearly vary across assets. Then the FM regression (3) becomes

$$r_i^t = b_0^t + (\gamma^t \theta_i^t)\, \hat{\beta}_{Mi}^{t-1} + \epsilon_i^t. \tag{6}$$

Similarly, in this case, due to the existence of estimation error in factor loadings, the slope in the cross-sectional regression is also likely to be heterogeneous among stocks.

In summary, the sensitivity of stock returns to their firm characteristics generally varies across assets. The common assumption of constant slopes in (4) is made for tractability rather than based on economic rationale. Under this simplified assumption, standard regression methods can be used to estimate the slopes, which approximate the average risk premia. However, this estimated risk premia is prone to errors due to the proxy error of firm characteristics or the estimation error of factor loadings, and these errors can vary across individual stocks. Using such estimated risk premia to forecast stock returns can clearly lead to inaccurate predictions. Our proposed approach addresses this issue by allowing b_j^t to vary across assets, enhancing the precision of return predictions.

In particular, our model assumes

$$r_i^t = b_{i0}^t + \sum_{j=1}^{J} b_{ij}^t \cdot char_{ij}^{t-1} + \epsilon_i^t, \tag{7}$$

where b_{ij}^t represents the specific response of stock i's return to characteristic j. This differs from the conventional homogeneous model (4) in that it considers heterogeneous responses of firms to the risks, that is, each firm i is allowed to deliver a unique response to any firm characteristic j.

Our heterogeneous model (7) offers a flexible approach to modeling the relationship between stock returns and firm characteristics. However, this flexibility comes with a challenge: the model involves a large number of parameters. Specifically, there are $(J+1) \times N^t$ unknown parameters at each time t in (7), yet only N^t observations. How can we identify this model? The solution, as detailed below, involves introducing regularizations on the slopes to facilitate their identification and estimation.

2.2 Estimation: fusion penalty

In the conventional model (4), the homogeneous responses are estimated in a least-squares (LS) manner. However, a fully heterogeneous model as in (7) would be inestimable under a regular LS framework, as it has more parameters than data observations. To tackle this challenge, we adopt a newly developed statistical modeling method proposed by Tang et al. (2021) and Zhu et al. (2021) to estimate the heterogeneous coefficients by leveraging a fusion-type regularization. This method does not require any pre-known knowledge of the heterogeneous coefficients and directly estimates them from the data.

For each time t, let $\boldsymbol{b}_i^t = (b_{i0}^t, b_{i1}^t, b_{i2}^t, \ldots, b_{iJ}^t)^\intercal$ denote the $(J+1)$-dimensional vector of coefficients representing the individualized return responses for stocks $i = 1, \ldots, N^t$ as well as the intercepts. All unknown parameters in the heterogeneous model (7) are estimated through minimizing the following penalized loss function

$$\sum_{i=1}^{N^t} \left(r_i^t - b_{i0}^t - \sum_{j=1}^{J} b_{ij}^t \cdot char_{ij}^{t-1} \right)^2 + \lambda^t \sum_{1 \le i < i' \le N^t} f_\xi \left(\|\boldsymbol{b}_i^t - \boldsymbol{b}_{i'}^t\|_2 \right), \tag{8}$$

where λ^t is the associated penalization parameter, and $\|\cdot\|_2$ denotes the ℓ_2 norm.[8] The first term in (8) is the standard LS loss, while the second term is the adopted *fusion*-type penalty on individualized regression coefficients for regularization purposes. In particular, this penalization is imposed on pairwise differences of individualized coefficients, $\boldsymbol{b}_i^t$ and $\boldsymbol{b}_{i'}^t$, which is also known as the *total variation* regularization. Specifically, instead of using a regular ℓ_1-penalty, we adopt a minimax concave penalty as introduced by Zhang (2010),

$$f_\xi(x) = \int_0^x \left(1 - \frac{s}{\xi}\right)_+ ds, \tag{9}$$

where ξ is a tuning parameter, and $(\cdot)_+$ denotes the positive part of a real value. In principle, the imposed fusion penalty limits the freedom of the parameters involved in the regular part of the LS loss and thus enables the estimation of individualized coefficients with fewer observations.

We make a few remarks here to help understand the idea and the implementation of the proposed heterogeneous model. First, the pairwise fusion penalty imposed on differences of individualized slopes essentially encourages to combine similar individualized slopes in the LS fitting and thus reduces the degrees of freedom of model parameters. In addition, the tuning parameter λ^t controls the power to penalize the differences between individualized slopes. Intuitively, if λ^t takes a small value, the estimated individual slopes would have more degrees of freedom and thus tend to be very different from each other. If λ^t gets larger and larger, the pairwise fusion penalty eventually dominates the LS term in (8). Ultimately, the model in (8) reduces to a regular homogeneous model when λ^t is extremely large. In our study, the penalty parameter λ^t is selected in a data-driven fashion based on a modified Bayesian Information Criteria following Zhu et al. (2021).[9]

[8] The ℓ_2-norm of a vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_J)^\mathsf{T}$ is defined as $\|\boldsymbol{a}\|_2 = \left(\sum_{j=1}^J a_j^2\right)^{1/2}$.

[9] More technical details can be found in Internet Appendix A1, and the codes will be posted soon.

2.3 Return forecast

Next, we demonstrate how to apply either the standard homogeneous model or the proposed heterogeneous model to predict future stock returns in the cross-section. Specifically, based on the available information at the end of month $t^* - 1$, we aim to predict the returns for stock i in the following month t^*.

Following Lewellen (2015), among others, we estimate the expected returns in two steps. The first step is to fit models (4) or (7) with monthly stock return data and lagged firm characteristics. In particular, for each $t = 1, 2, 3, \ldots, t^* - 1$, we calculate $\hat{b}_j^t$ or $\hat{b}_{ij}^t$, and obtain time-series of coefficient estimates $\hat{\boldsymbol{b}}^t = (\hat{b}_0^t, \hat{b}_1^t, \hat{b}_2^t, \ldots, \hat{b}_J^t)^\mathsf{T}$, or $\hat{\boldsymbol{b}}_i^t = (\hat{b}_{i0}^t, \hat{b}_{i1}^t, \hat{b}_{i2}^t, \ldots, \hat{b}_{iJ}^t)^\mathsf{T}$ for $i = 1, \ldots, N^t$, corresponding to (4) or (7), respectively. In this step, the homogeneous or heterogeneous responses of stock returns to firm characteristics are estimated for each month based on the cross-sectional variation of stock returns and firm characteristics, under the corresponding model specifications.

In the second step, we aggregate return responses over certain time horizons and use the aggregated information to predict the return of stock i for the month t^*.[10] Consider the time horizon of length τ, let

$$\bar{b}_j^{t^*} = \frac{1}{\tau} \sum_{t=1}^{\tau} \hat{b}_j^{t^*-t}, \quad j = 0, \ldots, J, \tag{10}$$

for the standard homogeneous model (4), and

$$\bar{b}_{ij}^{t^*} = \frac{1}{\tau} \sum_{t=1}^{\tau} \hat{b}_{ij}^{t^*-t}, \quad i = 1, \ldots, N^{t^*},\ j = 0, 1, \ldots, J. \tag{11}$$

for the heterogeneous model (7). Clearly, $\bar{b}_j^{t^*}$ and $\bar{b}_{ij}^{t^*}$ aggregate the information of return responses

[10] The aim of this aggregation is to reduce the noise in estimation with only one-month data. However, it does yield bias due to slow time-series moving estimations, as any rolling- or expanding-window analysis does. See Section 6.2 for a further discussion of the bias-and-variance trade-off.

over horizon τ. With the aggregated information, we predict $r_i^{t^*}$ as

$$\tilde{r}_i^{t^*} = \bar{b}_0^{t^*} + \sum_{j=1}^{J} \bar{b}_j^{t^*} \cdot char_{ij}^{t^*-1}, \tag{12}$$

for the standard homogeneous model (4), and

$$\tilde{r}_i^{t^*} = \bar{b}_{i0}^{t^*} + \sum_{j=1}^{J} \bar{b}_{ij}^{t^*} \cdot char_{ij}^{t^*-1}, \tag{13}$$

for the heterogeneous model (7). The predictions are based on information available only up to month t^*-1, thus, the forecasting procedure is out-of-sample and does not suffer from a look-ahead bias.

3 Out-of-Sample Performance

In this section, we compare the out-of-sample performance of our heterogeneous model (7) with the standard homogeneous model (4). If the data indeed display heterogeneity, the predictive capacity for stock returns is expected to benefit from identifying and incorporating this variability. We start with investigating the performance of the standard long-short portfolios constructed based on the predicted returns. We then check the predictive power by regressing the true monthly returns on the predicted ones. This comparison is not trivial, as statistically, merely increasing the model complexity does not automatically lead to enhancements in out-of-sample performance. Such improvement is only achieved when the extended model successfully captures the authentic underlying information.

3.1 Data

Our sample period is from January 1964 to December 2022. We obtain monthly stock returns from the Center for Research in Security Prices (CRSP). Our sample includes all common stocks with share code 10 or 11. We rely on data from both CRSP and Compustat to calculate firm characteristics. We adopt the set of 15 representative firm characteristics from Lewellen (2015). In particular, the characteristics include CAPM beta (Betamkt), size (LogMV), value (LogBM), momentum (LagRet), long-term momentum (LagRet_L), return volatility (RV), accruals (ACC), profitability (ROA), investment (LogAG), long-term and short-term growth in shares outstanding (LogIssue and LogIssue_S, respectively), dividends (DY), turnover (ATO), debt (DP), and sales (SP). The detailed definitions of these characteristics are in the Appendix. To mimic a real-time situation, we assume that market data are public to all investors immediately, but accounting data are available four months after the end of each firm's fiscal year. For example, investors observe daily stock prices and trading volumes, but they can only acquire sales and earnings by the end of November for a firm whose fiscal year ends in July.

Table 1 provides the summary statistics of all these characteristics. The average panel size of the stocks is $4,272$. The summary statistics are comparable to those in Table 1 of Lewellen (2015). Table 2 and Figure 1 present pairwise correlations. With no surprise, we can see that a number of characteristics are highly correlated with each other in either negative or positive relations. For example, LogIssue and LogIssue_S are positively related due to the information overlap. It is important to note, however, that the presented multicollinearity usually does not appear as a concern in the out-of-sample analysis, as it does not influence the predictions (Greenberg and Parks, 1997).[11]

[11] We also consider three-factor and seven-factor models in Section 6.1, where the factors capture almost orthogonal information.

3.2 Portfolio sorting

We follow standard ways to construct investment portfolios. First, at the end of each month, we sort the stocks into deciles based on their predicted returns by (12) or (13). Then, a long-short portfolio is formed, which takes long positions in the high decile (the 10% stocks with the highest predicted returns) and short positions in the low decile (the 10% stocks with the lowest predicted returns). We rebalance the positions monthly. This long-short trading strategy only utilizes information available before the time of investing, thus it evaluates the out-of-sample performance of both models and is not subject to a look-ahead bias.

In our main analysis, we focus on the 10-year rolling estimates ($\tau = 120$ in (10) and (11)) for out-of-sample prediction analogous to Lewellen (2015). All 15 characteristics are included in both the standard homogeneous and the proposed heterogeneous model.[12] Since our data cover a time range from January 1964 to December 2022, the out-of-sample predictions of monthly returns are accordingly calculated from January 1974 to December 2022.

Table 3 summarizes the performance of both the equal- and value-weighted portfolios based on the heterogeneous model and the standard homogeneous model. The average monthly returns as well as the standard deviations over time are reported for each decile. Consider first the equal-weighted portfolios. The average realized monthly returns of the heterogeneous approach range from -0.14% to 2.99%, while those of the homogeneous model range from 0.27% to 2.26%. Obviously, the heterogeneous model yields simultaneously a much smaller return in the low decile and a much larger return in the high decile. Consequently, the spread between the top and bottom deciles (H-L) is 3.13% based on the heterogeneous model, yielding an extra return of 1.15% beyond the homogeneous model (1.98%). This suggests a much higher ability to predict stocks with extreme returns of our heterogeneous model compared to the standard homogeneous one. On the other hand, the H-L portfolio based on the heterogeneous model also exhibits lower volatility (5.41%) compared with that based on the standard homogeneous model (5.97%).

[12]The model performance with different sets of factors and alternative rolling windows are further investigated in Sections 6.1 and 6.2.

Consistent with the results in Lewellen (2015), the standard homogeneous model generates an H-L portfolio with an annualized Sharpe ratio as high as 1.18.[13] However, with a higher average return and lower volatility, the heterogeneous model yields an even higher annualized Sharpe ratio of 2.01, nearly doubling that of the standard homogeneous model. The difference is significant at the 1% level based on the heteroskedasticity and autocorrelation consistent (HAC) standard errors (Ledoit and Wolf, 2008). Given that the standard homogeneous model sets a very high standard benchmark with a high average return and Sharpe ratio, the substantial improvement suggests a significantly better predictive power of the heterogeneous model. All these results, higher average return, smaller volatility, and correspondingly higher Sharpe ratio, demonstrate the superior out-of-sample performance of the proposed model.

The results for value-weighted portfolios are consistent with those for equal-weighted portfolios. The heterogeneous model has a lower average return (0.21% v.s. 0.74%) for the low decile and a higher average return (1.77% v.s. 1.44%) for the high decile, compared with those relying on the standard homogeneous model. The volatilities of the returns are of similar levels. As a result, there is a significant increase in the annualized Sharpe ratio to 0.86, more than doubling the Sharpe ratio of the standard homogeneous model, 0.38. Notice that the Sharpe ratio of the S&P 500 index over the same sample period is 0.56, the value-weighted H-L portfolios based on the standard homogeneous FM model hardly beat the simple buy-and-hold strategy of the market. The remarkable improvement by introducing heterogeneity for value-weighted portfolios indicates the success of our model in the prediction of returns for bigger stocks.

To better understand the investment implications, we also visually summarize the results in Figure 2. The cumulative returns of the equal-weighted (value-weighted) high and low deciles based on both our heterogeneous model and the standard homogeneous model are presented in Panel A (B). We can clearly see that the cumulative returns of the high (low) decile portfolios based on our model are almost always higher (lower) than those based on the homogeneous model.

[13] We follow Lewellen (2015) to calculate the Sharpe ratio of long-short portfolios as the average realized returns divided by realized volatility. Adjusting average realized returns with average risk-free rates yields similar conclusions.

This indicates that the superior performance of our model is quite persistent and robust against the selection of sample periods. Such patterns further strengthen the average return results in Table 3, showing that dominance is exhibited almost all the time. It is also worth noting that, when forming the value-weighted portfolios, the standard homogeneous model appears to fail in identifying stocks with low future returns, while our model is quite robust. This again demonstrates our model's strong predictive power for returns of bigger stocks.

Overall, the portfolio-sorting results indicate that the heterogeneous model significantly outperforms the standard homogeneous model in terms of stock return prediction. The results provide strong evidence regarding the extra information gained by using the heterogeneous model versus the conventional homogeneous model.

3.3 Stock return forecasts

The long-short portfolio focuses on the performance of the top and bottom deciles. To assess the general predictive accuracy, following Lewellen (2015), for each month t^*, we cross-sectionally regress the actual realized returns $(r_i^{t^*})$ on the forecasts of expected returns $(\tilde{r}_i^{t^*})$:

$$r_i^{t^*} = \kappa_0^{t^*} + \kappa_1^{t^*} \tilde{r}_i^{t^*} + \epsilon_i^{t^*}. \tag{14}$$

Compared to portfolio-sorting analysis which relies more on the ranking of the stock returns, this regression provides additional information regarding prediction accuracy at the stock level. Ideally, with a 100% precise prediction, the regression would have a slope of exactly one and an intercept of zero. While empirically, we posit that a model's accuracy in predicting future returns surpasses another one's on average when the estimated slope, $\hat{\kappa}_1^{t^*}$, is closer to one, and the estimated intercept, $\hat{\kappa}_0^{t^*}$, is closer to zero.

Table 4 reports the time-series average of the estimated slope $(\hat{\kappa}_1^{t^*})$ and intercept $(\hat{\kappa}_0^{t^*})$ in (14) for return forecasts $(\tilde{r}_i^{t^*})$ based on both the heterogeneous model (13) and the regular homogeneous regression model (12). The table also includes the Newey-West standard errors with six lags and

the resulting t-statistics.[14] The average of slope estimate $\hat{\kappa}_1^{t^*}$ is 0.852 for the heterogeneous model, compared to 0.447 for the standard homogeneous model. In addition, the heterogeneous model also yields a significantly larger t-statistic for $\hat{\kappa}_1^{t^*}$ as 20.403, compared to 8.721 for the standard homogeneous approach. Such improvement on out-of-sample prediction essentially conveys an additional message — the heterogeneous model does not only better identify stocks with extreme future returns, but it also provides more accurate predicted returns in general.

On the other hand, the intercept estimate of our heterogeneous model is not significantly different from zero, while that of the standard homogeneous model is highly significant at the 1% level. This indeed provides more evidence about the outperformance of the heterogeneous framework. All these results provide a strong indication that the new model does have extra predictive power compared to the standard model.

Overall, we show that the out-of-sample performance is improved significantly after introducing heterogeneity to the standard FM model, both economically and statistically. Next, we compare our methodology with some other more advanced models.

4 Comparison with Alternative Models

In this section, we compare our heterogeneous model with popular models that enhance the return predictive power with respect to the standard FM model. We first consider segmenting the sample based on different criteria, which is a widely used methodology to introduce heterogeneity among asset groups. We next illustrate the different levels of heterogeneity between our model and Patton and Weller (2022)'s. Finally, we compare our model to popular machine learning models which are introduced to asset pricing by Gu et al. (2020), among others.

[14] Following Stock and Watson (2011), we choose the number of lags for the Newey-West test based on the rule of thumb: $L = 0.75T^{\frac{1}{3}}$, where L is the number of lags used and T is the number of observations in the time series.

4.1 Segmented homogeneous model

Existing work in asset pricing literature that deals with heterogeneous return responses usually draws conjectured market segments based on some prior knowledge, among which size-based segmentation is the most widely investigated. The intuition is that, different market segments may lead to different levels of risk premia. To provide additional insights about the proposed heterogeneous model, we now investigate further the predictive performance of the segmented homogeneous model.

Fama and French (2008) and Lewellen (2015), among others, conduct separate homogeneous cross-sectional regressions on different subsamples split by market value at the end of the previous month. Along this direction, we check the out-of-sample predictive power of a regular homogeneous model estimated based on subsamples. Following Fama and French (2008) and Lewellen (2015), we consider two sample-splitting strategies: "tiny v.s. all-but-tiny", with market equity cut by NYSE 20th percentile, and "large v.s. all-but-large", with market equity cut by NYSE 50th percentile. Within each subsample, we run a standard cross-sectional homogeneous regression model using all 15 characteristics. The forecast of individual stock return is estimated based on the respective subsample estimates with a 10-year rolling average.

Table 5 reports the performance of decile portfolios sorted by returns predicted from different subsample specifications. For the equal-weighted portfolios, the two segmented homogeneous models achieve average monthly H-L returns of 2.19% ("large v.s. all-but-large") and 2.25% ("tiny v.s. all-but-tiny"), and annualized Sharpe ratios of 1.47 and 1.55, respectively. While for value-weighted portfolios, the average monthly returns for the two subsample-based homogeneous models are 1.18% ("large v.s. all-but-large") and 1.13% ("tiny v.s. all-but-tiny"), and the annualized Sharpe ratio values are 0.59 and 0.57, respectively. Comparing the results with those of the proposed heterogeneous model as reported in Table 3, it is clear that the latter has extra predictive power beyond the segmented homogeneous model. The results indicate that segmenting the market by firm size does not sufficiently capture the underlying heterogeneity.

On the other hand, we observe that the size-based segmented models indeed yield better performances for H-L portfolios compared to the standard homogeneous model, as reflected in higher average monthly returns and annualized Sharpe ratio values. To further explore segmentation strategies, we investigate an alternative approach based on industry sectors, another method also widely discussed in the literature. Specifically, we segment the market into eleven industry sectors according to firms' Standard Industrial Classification (SIC) codes: Agriculture, Mining, Construction, Manufacturing, Transportation, Utilities, Wholesale, Retail, Finance, Services, and Public.[15] Similarly, for each month, we run separate cross-sectional homogeneous regression models across different industry subsamples using the 15 characteristics. The forecast for individual stock returns is then estimated based on the respective industry subsample estimates, using a 10-year rolling average.

Table 6 presents the performance of decile portfolios sorted by returns predicted from the industry-based segmented model. The average monthly return and annualized Sharpe ratio are 1.58 (1.05) and 0.68 (0.43) for the equal-weighted (value-weighted) H-L portfolio, respectively. Interestingly, this industry-specific model, despite its finer market segmentation, does not outperform the size-based segmented model, nor does it surpass the standard homogeneous model.[16] These findings suggest that prespecified segmentation, even when employing widely adopted strategies or factors well-documented in the literature, does not necessarily enhance the prediction accuracy of stock returns. While some segmentation approaches may partially improve predictive power, exhaustively searching for an optimal segmentation among infinitely many options is impractical. This underscores the importance of developing a fully data-driven approach to capture the underlying cross-sectional heterogeneity more effectively.

[15] More details about this segmentation, including the corresponding SIC code ranges, are provided in the Appendix.

[16] We also test segmented models using other well-known factors, such as idiosyncratic volatility and institutional ownership, which yield similar performance. These results are provided in the Internet Appendix.

4.2 Patton and Weller (2022) model

Different from the segmented homogeneous model usually based on some prespecified knowledge, Patton and Weller (2022) propose a data-driven approach and extend the k-means clustering method to the cross-section of stocks, by detecting groups of stocks that share the same slope within a cluster but with different slopes across clusters. The Patton and Weller (PW) model effectively addresses the heterogeneity issue caused by market frictions. We next compare our framework with the PW model and illustrate the difference and additional contribution that our method brings.

Table 7 presents the performance of decile portfolios based on the PW model applied to individual stocks. In line with Patton and Weller (2022), we test its performance with different numbers of clusters (k), ranging from 2 to 4. Notably, the model with $k = 2$ delivers the most favorable out-of-sample results. The average return of the equal-weighted (value-weighted) H-L portfolio is 1% (0.48%), with a volatility of 4.15% (4.65%), yielding a Sharpe ratio of 0.84 (0.36). However, the H-L portfolios based on the PW model only marginally outperform those based on the standard homogeneous model. This highlights a limitation of the PW model when applied to individual stocks.

While both our proposed method and the PW model are data-driven, empirical evidence indicates that only our approach provides superior out-of-sample performance compared to the conventional FM model. This differential success demonstrates that, at the stock level, the heterogeneity caused by proxy and estimation errors is more significant than that resulting from market frictions. The difference is empirically reflected in different model configurations. Our model accommodates complete heterogeneity across all stocks, in contrast to the PW model, which only leverages heterogeneity at a group level and maintains homogeneity within each individual group. To provide more insights, we further illustrate this difference in Figure 3 based on the data in January 2010.[17] For each factor, we plot the distribution of individual return responses under our model, the return responses of the two clusters identified by the PW model when $k = 2$, as well as the

[17]This month is arbitrarily chosen for illustration. Results are similar for other months in our sample.

homogeneous one under the standard model.[18]

From Figure 3, it is straightforward to see that, for each characteristic, the stock responses under the PW model are restricted to two distinct values. In contrast, our heterogeneous model yields a much wider range of stock responses, allowing for substantial discrepancies among different stocks. For example, the coefficient for LagRet_L ranges from -1.14 to 1.03, while that for the two clusters are -0.18 and 0.05, respectively. As such, by pursuing individual-level heterogeneity, our model provides much higher degrees of freedom in characterizing the diversity of return responses across different stocks. This is critical for cross-sectional analysis due to the large number of individual stocks.

Furthermore, Figure 3 reveals that our model is able to capture different distribution patterns across characteristics. For example, there are unimodal distributions, such as ACC and Betamkt, bimodal distributions, such as ROA and LagRet, as well as trimodal distributions, such as LogIssue and ATO. On the other hand, the PW model enforces the same two-cluster structure on all characteristics by its mechanism. This can be insufficient to capture the characteristic-wise diversity, especially when the number of factors is large.[19] In addition, even with the same number of modes for risk-response distributions, the cluster structure may still vary across different factors. For example, the risk-response distribution for DP and RV are both bimodal. However, while the two clusters of individual return responses related to RV are of similar sizes, the clusters related to DP are significantly unbalanced. This does not fit into the PW model by its nature.

Essentially, the PW model can be seen as a partially heterogeneous model, positioned between the standard homogeneous model and our completely heterogeneous model. It primarily addresses heterogeneity arising from market frictions. Building on the group-wise heterogeneous framework, our model advances further by allowing complete heterogeneity, accommodating variations due to proxy and estimation errors. This approach provides greater empirical flexibility in capturing the

[18] We present the results based on $k = 2$ here for illustration as this case exhibits the best out-of-sample performance. Comparison with the PW model based on $k = 3$ and $k = 4$ is similar.

[19] In Internet Appendix A2, we also present results of the PW model based on different factor models and indeed find that its performance is better when the number of factors is smaller.

complex cross-sectional patterns of return responses.

4.3 Popular machine learning models

In recent years, machine learning models have attracted explosive attention in empirical asset pricing studies. Beyond the traditional regression-based approach, machine learning models usually achieve exceptional predictive power by accounting for more complex nonlinear relations. Therefore, it is intriguing and beneficial to compare our method with these prevalent models. We investigate the out-of-sample performance of a set of popular machine learning models outlined in Gu et al. (2020), including random forest (RF), gradient boosting regression trees (GBRT), and neural networks (NN) with one to five layers, in contrast to our heterogeneous approach and the benchmark standard homogeneous regression model.

We first apply all the machine learning approaches to our data. We utilize a ten-year rolling sample for training and a five-year rolling sample for validation following that in Gu et al. (2020), and both the proposed heterogeneous model and the standard homogeneous regression model are still evaluated with ten-year rolling estimates. Tables 8 and 9 report the average monthly returns, standard deviations, and annualized Sharpe ratios of decile portfolios sorted based on return forecasts of different models, equal-weighted and value-weighted, respectively. For equal-weighted (value-weighted) long-short portfolios, the machine learning models generate average monthly H-L returns ranging from 1.71% (0.65%) to 2.24% (1.00%), yielding Sharpe ratio values ranging from 1.18 (0.38) to 1.46 (0.58). Meanwhile, as a benchmark, the homogeneous FM model achieves average returns of 1.89% and 0.62%, and Sharpe ratios of 1.13 and 0.34, for equal- and value-weighted portfolios, respectively.[20] Consistent with the findings in Gu et al. (2020), the machine learning models indeed substantially enhance the predictive power of the standard FM model. In particular, among all machine learning approaches, the neural networks achieve the best results (one-layer

[20]To keep consistent with the testing sample period of machine learning models, in this comparison, the sample period for both our heterogeneous model and the standard homogeneous model covers from January 1969 to December 2022. As a result, the summary statistics of the H-L portfolios differ slightly from those reported in Section 3.2.

NN in the equal-weighted case and three-layer NN in the value-weighted case) due to its flexibility in capturing nonlinear relations and interaction effects between characteristics and return forecasts.

However, all these machine learning models still underperform our proposed heterogeneous model, which generates average monthly returns of 3.10% and 1.41%, yielding the Sharpe ratios of 2.17 and 0.81, for equal- and value-weighted portfolios, respectively. In addition to these average results, Figure 4 provides a visualization of cumulative returns of long and short portfolios for both equal-weighted (Panel A) and value-weighted (Panel B) cases. The proposed heterogeneous approach in general has the highest (lowest) cumulative return of the high (low) decile compared to all alternatives. This time-path dominance indicates that the superior performance of our method is independent of the chosen sample period.

Machine learning models are conventionally credited with superb power in capturing complex relations, which often results in great predictive performance. However, these models still adhere to a homogeneous framework, assuming uniform responses to firm characteristics across all stocks. The remarkable out-of-sample performance of our proposed model highlights the significant value of accounting for heterogeneity among individual stocks in cross-sectional analysis. Even with highly complex and "black-box" settings, their performance does not surpass that of our model, which simply incorporates proxy and estimation errors within the classic FM framework. This demonstrates that, even in an era dominated by large-scale, complexity-driven AI, integrating interpretable and economically grounded structures can substantially enhance stock return predictions beyond what is achievable through model complexity alone.

Next, we further compare the performance of our heterogeneous approach with machine learning models in a high-dimensional setting as tested in Gu et al. (2020) covering 94 characteristics. In Table 10, we list the average monthly returns for the high, low, and H-L value-weighted portfolios, as well as the annualized Sharpe ratios from Table 7 in Gu et al. (2020). To keep consistent, we also perform the same exercise with our heterogeneous model for the same 94 characteristics and sample period (from January 1987 to December 2016). The machine learning models generate average monthly H-L returns ranging from 1.17% to 1.75%, and yield annualized Sharpe ratio

values ranging from 0.81 to 1.35, still lower than those of the proposed heterogeneous model with a return of 2.21% and a Sharpe ratio of 1.91. This empirical result indicates that our proposed heterogeneous model is capable of handling a large number of characteristics, and the success of our model is persistent in such a high-dimensional scenario.

Generally, our empirical evidence suggests that economic structures can add substantial value to return forecasts, surpassing the last-generation artificial intelligence techniques. Moreover, the strength of our approach extends beyond enhanced predictive power for returns – it includes interpretability. Unlike the "black-box" nature of traditional machine learning methods, such as neural networks, our model yields comprehensive insights into the cross-sectional patterns among individual stocks. The results produced by our approach facilitate subsequent analysis to address broader economic inquiries, such as identifying potential factors influencing the observed heterogeneity pattern. We further study these issues in the next section.

5 Underlying Mechanism of Heterogeneity

With sufficient evidence of heterogeneous stock return responses among individual stocks, several important questions arise. Which characteristics exhibit heterogeneous responses? Which factors are the potential driving forces of the underlying heterogeneity? Is the heterogeneity pattern time-varying? All of these questions are interesting for understanding the underlying mechanism of the relation between risk factors and stock returns. To address these questions, we investigate the heterogeneity both in the cross-section and in the time series.

5.1 Heterogeneity test on firm-specific responses

To find out which characteristics exhibit heterogeneous responses, following Lin and Ng (2012), Bonhomme and Manresa (2015), and Patton and Weller (2022), we propose a test to detect significant heterogeneity across stocks. This helps us better understand the cross-sectional variation of

stock returns with respect to the heterogeneous model. For example, it allows us to answer whether the stocks have heterogeneous responses only to some of the factors and react homogeneously to others, or they react differently to all characteristics.

In our heterogeneous model, the cross-sectional variation of the individualized response estimates, b_{ij}^t, serves as a natural measurement for the level of heterogeneity regarding a specific risk factor j. To test if this variation is significant across stocks, we first group the stocks according to the estimated individualized responses for each characteristic. Then we utilize the analysis of variance (ANOVA) to test the potential heterogeneity of each of the firm characteristics.

In particular, for each stock i, based on our heterogeneous response model, we have available a $T \times J$ matrix of individualized response estimates: $\boldsymbol{B}_i = (b_{ij}^t)_{1\leq t\leq T, 1\leq j\leq J}$. We perform the k-means algorithm to classify the stocks into two groups, $\mathcal{I}_1$ and $\mathcal{I}_2$, according to $\boldsymbol{B}_i$. At each time t for characteristic j, we then aggregate the return responses $\boldsymbol{b}_i$ within each group, obtaining two time series of group-average responses:

$$\boldsymbol{b}_{j,\mathcal{I}_k} = \frac{1}{N_k} \sum_{i\in\mathcal{I}_k} \boldsymbol{b}_{ij},$$

for $k = 1$ and 2, where N_k is the number of stocks in group k. The heterogeneity on responses to characteristic j is then tested by:

$$H_0 : \boldsymbol{b}_{j,\mathcal{I}_1} = \boldsymbol{b}_{j,\mathcal{I}_2} \quad v.s. \quad H_a : \boldsymbol{b}_{j,\mathcal{I}_1} \neq \boldsymbol{b}_{j,\mathcal{I}_2}, \tag{15}$$

which leads to an ANOVA-type test on the monthly estimates of group-average return responses.

Notice that this test is not equivalent to the exact test of the existence of heterogeneity. In fact, if there are at least two individual firms that exhibit different responses to factor j, then we can claim that firms react differently to this factor. Therefore, rejecting the test, i.e., concluding that average return responses are different between two groups of stocks, is stronger than the existence of heterogeneity. That is to say, when we are able to reject the above test for factor j, we detect

heterogeneity. However, if we fail to reject it, then it is still possible that firms react differently to this factor.

Next, we apply the test to each of the 15 firm characteristics. Table 11 reports the time-series average of the group-average return responses for both groups, their differences, as well as all respective t-statistics. The sample size (number of firms) for the two groups are $3,463$ and 526, respectively, leading to unbalanced groups.[21] Test results show that return responses exhibit heterogeneity to at least 11 out of 15 firm characteristics. Among all of the 11 characteristics, the most significant heterogeneity is associated with LogMV, LagRet_L, and RV, with corresponding t-statistic values of 12.593, 6.002, and -9.204, respectively. In addition, LagRet, DY, ATO, ACC, ROA, LogAG, LogIssue, and DP also have significant heterogeneity, with t-statistic values typically larger than 2.58, leading the rejection of the test at the 1% level.[22] However, the test is unable to detect heterogeneity for the rest four factors. As discussed earlier, failing to reject does not necessarily mean that there is no heterogeneity.

Comparing return responses between the two groups, we see two interesting and distinct patterns. First, there are six factors that have average return responses in the same direction (while significantly different), including LogAG, LogIssue, LogIssue_S, ATO, LogBM, and Betamkt. For example, the average return responses for ATO are -0.022 (t-stat$=-3.040$) and -0.157 (t-stat$=-5.963$) in the two groups, respectively, with a difference of 0.136 (t-stat$=4.019$). The uniform direction of the slopes suggests that the related characteristic shares a consistent relation trend with stock returns for both groups. Yet, the main distinction is in the magnitude.

On the other hand, some other characteristics have different patterns in heterogeneous return responses with opposite signs. For example, the average return response estimates for LogMV are 0.002 (t-stat$=6.884$) and -0.021 (t-stat$=-16.833$) for the two groups, respectively, with a significant difference of 0.022 (t-stat$=12.595$). Similar patterns are also observed for ACC, DY,

[21] We examine the properties of each group in detail in Section 5.3.

[22] Among these 11 characteristics, only the heterogeneity on DP is marginally significant with a t-statistic of 1.767. All others are significant at the 1% level.

LagRet_L, and RV. Note that for all these factors, the average return responses are still significant within each group, yet they have different trends in relation to stock returns. Such patterns are potentially more interesting as they could stem from intricate and profound economic nuances. The introduction of our model becomes crucial in this context. The homogeneous model essentially evaluates an average effect of the relation between characteristics and stock returns. In scenarios where both groups display the same sign, the homogeneous model might somewhat capture this association. However, in the second and more interesting case, a homogeneous model might fail completely, as the negative and positive heterogeneous responses could be washed out substantially by taking an average.

5.2 Importance of heterogeneity in each characteristic

To further examine the importance of heterogeneity in stock return responses to different characteristics, we introduce an alternative framework that allows only one characteristic to exhibit heterogeneity at a time. In this model, firm-specific responses are heterogeneous for a single characteristic, while the other 14 characteristics remain homogeneous. This approach enables us to isolate the impact of heterogeneity for each factor and assess the relative importance of heterogeneity across different characteristics

Specifically, for each factor j^*, we estimate responses of stock returns to firm characteristics using a modified objective function

$$\sum_{i=1}^{N^t}\left(r_i^t - \sum_{0\le j\neq j^*\le J} b_j^t\cdot char_{ij}^{t-1} - b_{ij^*}^t\cdot char_{ij^*}^{t-1}\right)^2 + \lambda^t \sum_{1\le i<i'\le N^t} f_\xi\left(\mid b_{ij^*}^t - b_{i'j^*}^t \mid\right),$$

where the firm-specific slopes and the corresponding fusion penalty are imposed only on factor j^*. The responses to all other factors $j \neq j^*$, are assumed to be homogeneous (b_j^t) and are estimated in a standard least-squares fashion. We then apply the method described in Section 2.3 to predict stock returns next month, and subsequently sort and construct portfolios based on these predictions.

The performance of H-L portfolios based on predictions from the 15 single-factor heterogeneous models is presented in Table 12. The average monthly returns and Sharpe ratios of the equal-weighted (value-weighted) portfolios for the 15 models range from 2.46% to 2.97% (0.74% to 1.31%) and 1.42 to 1.80 (0.45 to 0.76), respectively. Note that the average returns and Sharpe ratios for H-L based on the standard homogeneous-response model, which serves as a benchmark here, are 1.98% and 1.18 (0.69% and 0.38) for the equal-weighted (value-weighted) H-L portfolios (see Table 4), respectively. These results clearly indicate that introducing heterogeneity, even for just one factor, enhances the prediction accuracy of cross-sectional stock returns.

The H-L portfolio results summarized in Table 12 also provides an overview about the relative importance of heterogeneity across different factors. Upon comparing the performances of these 15 H-L portfolios, we observe the most improvement gained from imposing heterogeneity on factors LogMV, LagRet_L, RV, and ROA, with outstanding Sharpe ratios of 1.71 (0.76), 1.61 (0.50), 1.80 (0.71), and 1.64 (0.69), for equal-weighted (value-weighted) portfolios, respectively. Notably, the relative importance revealed by the H-L portfolio performance largely agrees with the test results in Table 11. For instance, the characteristics LogMV, LagRet_L, and RV, which exhibit the highest t-statistics, also have the best H-L portfolio performances with Sharpe ratios exceeding 1.6 (equal-weighted). Conversely, factors with insignificant t-statistics correspond to lower Sharpe ratios. This consistency confirms our model's ability to identify important factors that exhibit greater heterogeneity. The results of this analysis also demonstrate the model's flexibility in accommodating different modelings of heterogeneity.

5.3 Exploring determinants in heterogeneity

In addition to the detection of the heterogeneous pattern, another important question is what mechanism drives the heterogeneity. In this subsection, we explore along this direction and aim to identify pivotal factors that may provide more insights into the hidden mechanism underlying the heterogeneity pattern.

Specifically, we examine whether certain firm characteristics exhibit systematic differences across groups. Indeed, the characteristics with distinct distributions between groups are likely to be potential factors that drive the underlying mechanism of heterogeneity. Table 13 reports the time-series average of mean values for all characteristics in both groups as well as their respective standard deviations over time. We test the difference in these means and also report the t-statistics in the table.

The two groups are significantly different in 12 out of 15 characteristics. This shows that the underlying mechanism that drives heterogeneity is rather complicated and might be determined by a high dimension of information variables. The result suggests that the conventional ad hoc remedies in the literature, which usually rely on sorting by one single firm characteristic to segment the stocks, are unlikely to model well the heterogeneity.

Interestingly, the characteristics that we fail to find differences in the mean are not necessarily those that we fail to detect heterogeneous responses in Section 5.1. For example, while the two groups have significantly different CAPM beta (t-stat$= -9.065$), there seems to be no significant difference in the average return responses to it.[23] Although stocks in different groups react significantly differently to LagRet, they are not significantly different in momentum itself (t-stat$= 0.439$). The inconsistency in these two tests demonstrates exactly the complexity of the heterogeneity mechanism – the characteristic itself may not be the reason that drives its own response heterogeneity.

5.4 Time-series pattern in heterogeneity

To further shed light on the underlying mechanism, we also check the time-series pattern of the cross-sectional heterogeneity level. To proxy the degree of heterogeneity on return responses, for each month t, we calculate the cross-sectional standard deviation of the estimated individualized

[23] Note that, as we discuss in Section 5.1, no difference in the average return responses between groups does not lead to no heterogeneous responses at all.

responses associated with factor j,

$$\sigma(\hat{b}_j^t) = \frac{1}{N^t - 1} \sum_{i=1}^{N^t} (\hat{b}_{ij}^t - \hat{b}_{\cdot j}^t)^2, \quad j = 1, \ldots, J, \tag{16}$$

where $\hat{b}_{\cdot j}^t = \frac{1}{N^t} \sum_{i=1}^{N^t} \hat{b}_{ij}^t$, is the cross-sectional average of the individualized return responses at time t. This cross-sectional standard deviation serves as a natural proxy to measure the variation level of heterogeneous responses with respect to each specific characteristic. We then explore the time-series patterns of the heterogeneity levels for all 15 firm characteristics by dividing our prediction sample period from January 1974 to December 2022 into two subperiods: recession and expansion, according to the NBER recession indicator. The time-series average of $\sigma(\hat{b}_j^t)$ for each subperiod is reported in Table 14. We also conducted a test to compare the difference in $\sigma(\hat{b}_j^t)$ between recession and expansion, and include the corresponding t-statistics in the table.

The results of Table 14 show that 14 out of the 15 factors experience significantly higher levels of heterogeneity during recession than expansion. DY is the only firm characteristic that has higher cross-sectional variation in return response during non-recession periods, but the difference is not significant (t-stat$= -0.59$). This indicates that the standard homogeneous model is less likely to hold when the overall market is in a bad state. However, during market downtimes, accurate stock return prediction is much more desirable to investors. Therefore, the introduction of our heterogeneous model is extremely valuable during the market recession.

6 Further Analyses

In this section, we further test the performance of the proposed model with different sets of factors and alternative rolling windows.

6.1 Alternative factor models

In this subsection, we explore the performance of the heterogeneous model with different sets of predictors, in terms of both the portfolio-sorting and out-of-sample regressions.

Specifically, in addition to the main model with all 15 predictors, we follow Lewellen (2015) to consider two more settings, one with three predictors including logMV, LogBM, and LagRet, and another one with seven predictors further adding ROA, LogAG, ACC, and LogIssue. Reducing the number of factors in the model essentially tests the predictive power of our heterogeneous model when the available information is limited. Also note that the three-factor, seven-factor, and 15-factor models are in a nested manner, which allows us to investigate the partial contribution when adding characteristics sequentially.

We first check the decile portfolios constructed based on the forecasts from the three-factor and seven-factor models, both using ten-year-rolling-average estimates ($\tau = 120$ in (10) and (11)) for regression parameters analogous to the 15-predictor model as reported in Table 3. Following the same layout as before, Table 15 reports the average monthly returns as well as the standard deviations over time for each decile, both equal- and value-weighted.

Consider first the three-factor model, average realized returns of the equal-weighted portfolios based on the heterogeneous approach range from 0.17% to 2.72% monthly, yielding a spread of 2.55% for the H-L portfolio, while those of the homogeneous model ranges from 0.66% to 2.32%, yielding a spread of 1.66% for its H-L. Compared to the homogeneous model, our heterogeneous model yields smaller returns in lower deciles and larger returns in higher deciles as well as lower volatility for the H-L portfolio. Consequently, the heterogeneous model generates an annualized Sharpe ratio of 1.80 for the H-L portfolio, an increase of 0.64 compared to the 1.16 generated by the standard homogeneous model. The difference is again significant at the 1% level with respect to the HAC standard errors. The value-weighted portfolios exhibit the same pattern with a Sharpe ratio of 0.63 by the heterogeneous model and a Sharpe ratio of 0.37 by the standard homogeneous model.

Analogously, for the equal-weighted portfolios of the seven-factor model, the H-L portfolio based on our heterogeneous model again generates a higher average return (2.91% v.s. 2.18%) and a significantly higher Sharpe ratio (2.07 v.s. 1.32). The value-weighted portfolios exhibit similar patterns. Overall, all three factor models convey the same information: our heterogeneous model delivers stronger investment implications than the standard homogeneous model.

Another interesting observation is made by comparing the heterogeneous model with three factors and the standard homogeneous models with all 15 factors. Surprisingly, the heterogeneous approach based on three factors still has a better predictive power with a higher portfolio return of 2.55% and a higher Sharpe ratio of 1.80 than those of the homogeneous approach based on 15 factors (1.98% and 1.18, respectively), despite the extra twelve predictors added to the regular model. Strikingly, the performance of the three-factor heterogeneous model is even better than all ML models based on 15 factors reported in Tables 8 and 9. This essentially suggests that the potential critical heterogeneity information cannot be accessed by simply adding extra factors.

It is worth of noting that both heterogeneous and homogeneous effects refer to the factors or characteristics that contribute to explain the cross-sectional variation of stock returns. The superior performance of the three-factor heterogeneous model, compared to the 15-factor homogeneous and machine learning models, indicates that accounting for heterogeneity provides a more substantial improvement in the predictive power of cross-sectional stock returns than merely adding more factors to a homogeneous model. Additionally, when comparing the 15-factor heterogeneous model to the three-factor heterogeneous model, we also observe a significant improvement. This finding underscores that incorporating additional "informative" factors with heterogeneous effects yields better predictive performance than models that eliminate these factors in a homogeneous manner.

Next, We investigate the prediction accuracy regression (14) with respect to different factor models. The results are presented in Table 16. Similar to the case with the 15-factor model, the heterogeneous model is always associated with an average slope estimate $(\hat{\kappa}_1^{t^*})$ closer to one, and an average intercept estimate $(\hat{\kappa}_0^{t^*})$ closer to zero. In particular, the average slope estimate is 0.716 (heterogeneous) v.s. 0.489 (homogeneous) for the three-factor model, and the average slope

estimate is 0.787 (heterogeneous) v.s. 0.576 (homogeneous) for the seven-factor model. Moreover, for both factor models, the intercept with respect to the heterogeneous approach is not significantly different from zero. Both the slope and intercept results show that our heterogeneous model yields more precise predictions regardless of the factor model considered.

6.2 Alternative rolling windows

As discussed in Section 2.3, when we predict next-month returns, we average the return response estimates over a certain time window to reduce noise. However, if the adopted window is too long, then the prediction may incorporate too much historical information that may not be relevant under the current market condition, leading to a bias in the prediction. In this subsection, we evaluate this bias-and-variance trade-off and check the effects of different rolling window lengths (τ in (10) and (11)) on the model forecast ability.

Table 17 presents the H-L portfolio results based on forecasts using (10) or (11) with 1-, 3-, 5-year rolling averages (corresponding to $\tau = 12$, $\tau = 36$, and $\tau = 60$, respectively) as well as the cumulative averages aggregating all available past return responses. We present the average monthly returns, volatilities, and annualized Sharpe ratios for both equal- and value-weighted portfolios within the heterogeneous as well as the homogeneous framework. Three panels in the table summarize the results for the three-factor, seven-factor, and 15-factor models, respectively.

Overall, with different factor models and different rolling windows, the heterogeneous model almost always yields H-L portfolios with higher returns, smaller standard deviations, and higher Sharpe ratios than the standard homogeneous model does, for both equal- and value-weighted portfolios. Within each specification, average monthly returns in general decrease with respect to the length of rolling windows. For example, for the heterogeneous seven-factor model, the average monthly returns are 4.43%, 3.26%, 3.07%, and 2.55% for the 1-, 3-, 5-year, and cumulative rolling windows, respectively. The prediction accuracy decreases as we include more historical information, which introduces more bias. On the other hand, we note that the volatility of portfolio

returns decreases as the rolling window expands. This is also expected, as the variance of the averaged slope estimates usually reduces with a longer rolling window. As a result, the Sharpe ratio exhibits no particular pattern as the rolling window changes. This is in line with the well-known bias-variance-trade-off for a general predictive performance in machine learning literature.

Comparing the results to Table 3, the forecasts with 10-year rolling estimates tend to give the best prediction performance in general. However, Table 17 indicates that both the heterogeneous model and the standard homogeneous model are quite robust to the rolling window selections in terms of Sharpe ratios as a balance of bias and variances.[24]

7 Conclusion

In this paper, we propose an AI-based extension of the widely used Fama-MacBeth regression by allowing stock returns to respond differently to firm characteristics, generalizing group-level heterogeneity originated by Patton and Weller (2022) to individual-level heterogeneity. Our new model accounts for heterogeneity arising from proxy errors or estimation errors when firm characteristics are used as imperfect proxies for risk exposures. In comparison with the growing use of machine learning methods, our AI-based extension allows for more flexibility beyond the nonlinearity and interaction effects that machine learning models usually capture.[25]

Empirically, applying our approach to a set of 15 firm characteristics as used by Lewellen (2015), we find that the equal-weighted (value-weighted) long-short portfolio has an annualized Sharpe ratio of 2.01 (0.86), doubling that of the standard homogeneous model. In addition, we find that our model outperforms not only the ad hoc segmentation methods but also the prominent machine learning models. The latter improves the traditional Fama-MacBeth regression by incor-

[24]We also repeat the accuracy regression test (14) for all different specifications, and find that the results are both quantitatively and qualitatively similar, which are thus omitted for brevity.

[25]Recent applications of machine learning in finance primarily concentrate on elucidating the cross-section of expected stock returns. Notable studies in this domain include Chinco et al. (2019); Rapach et al. (2019); Freyberger et al. (2020); Gu et al. (2020); Kozak et al. (2020); Bryzgalova et al. (2020); Chen et al. (2023); Cong et al. (2021); Han et al. (2023); Liu et al. (2021); Dong et al. (2022); Avramov et al. (2023); Cui et al. (2024).

porating various interaction and nonlinear effects from a data modeling perspective. In contrast, we rely essentially on the Fama-MacBeth regression while capturing the heterogeneous effects. Interestingly, the better performance of our approach indicates that machine learning models can be beaten once one imposes suitable economic structures on the problem, suggesting that economic insights are still important in the age of machine learning and artificial intelligence. We also propose a test to identify which firm characteristics exhibit heterogeneous responses and find that heterogeneity is significant for at least 11 out of 15 characteristics. Furthermore, we find that heterogeneous responses are more pronounced during recession periods.

Since the Fama-MacBeth regression is the fundamental methodology in finance, for both empirical asset pricing and empirical corporate, we expect wide applications of our heterogeneous-response model. For instance, our method can be applied to examine currencies, mutual funds, and hedge funds. Additionally, it can be employed to study investment strategies and the impact of various factors.

Appendix

A1. The 15 firm characteristics in Lewellen (2015)

- CAPM beta (Betamkt) is market beta estimated from weekly returns from month $t-36$ to month $t-1$.
- Size (LogMV) is the log market value of equity at the end of month $t-1$;
- Value (LogBM) is the log book value of equity minus the log market value of equity at the end of the month $t-1$;
- Momentum (LagRet) is the stock return between month $t-12$ to month $t-2$;
- Long-term momentum (LagRet_L) is the log stock return from month $t-36$ to month $t-13$;
- Return volatility (RV) is the standard deviation of daily returns from month $t-12$ to month $t-1$;
- Accruals (ACC) is the change in non-cash net working capital minus depreciation in the prior fiscal year;
- Profitability (ROA) is income before extraordinary items divided by average total assets in the prior fiscal year;
- Investment (LogAG) is the log growth in total assets in the prior fiscal year;
- Growth in shares outstanding (LogIssue) is the log growth in split-adjusted shares outstanding from month $t-36$ to month $t-1$;
- Short-term growth in shares outstanding (LogIssue_S) is the log growth in split-adjusted shares outstanding from month $t-12$ to month $t-1$;

- Dividends (DY) is the dividends per share over the prior 12 months divided by price at the end of month $t-1$;
- Turnover (ATO) is the average monthly turnover from month $t-12$ to month $t-1$;
- Debt (DP) is short-term plus long-term debt divided by market value at the end of month $t-1$;
- Sales (SP) is the sales in the prior fiscal year divided by market value at the end of month $t-1$.

A2. Industry segmentation with standard industrial classification (SIC) codes

The 11 specific industries and corresponding SIC code ranges are provided as follows.

- Agriculture: 1 - 999
- Mining: 1000 - 1499
- Construction: 1500 - 1799
- Manufacturing: 2000 - 3999
- Transportation: 4000 - 4899
- Utilities: 4900 - 4999
- Wholesale: 5000 - 5199
- Retail: 5200 - 5999
- Finance: 6000 - 6799
- Services: 7000 - 8999
- Public: 9000 - 9999

References

Avramov, D., Cheng, S., and Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, 69(5):2587–2619.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

Bryzgalova, S., Pelger, M., and Zhu, J. (2020). Forest through the trees: Building cross-sections of stock returns. Working paper, available at SSRN 3493458.

Chen, A. Y. and Zimmermann, T. (2021). Open source cross-sectional asset pricing. *Critical Finance Review, Forthcoming*.

Chen, L., Pelger, M., and Zhu, J. (2023). Deep learning in asset pricing. *Management Science, Forthcoming*.

Chinco, A., Clark-Joseph, A. D., and Ye, M. (2019). Sparse signals in the cross-section of returns. *The Journal of Finance*, 74(1):449–492.

Cong, L. W., Tang, K., Wang, J., and Zhang, Y. (2021). Alphaportfolio: Direct construction through deep reinforcement learning and interpretable AI. Working paper, available at SSRN 3554486.

Cui, L., Feng, G., and Hong, Y. (2024). Regularized gmm for time-varying models with applications to asset pricing. *International Economic Review*, 65(2):851–883.

Dong, X., Li, Y., Rapach, D. E., and Zhou, G. (2022). Anomalies and the expected market return. *The Journal of Finance*, 77(1):639–681.

Fama, E. and French, K. (1993). Common risk factors in the return on stocks and bonds. *Journal of Financial Economics*, 33(1):3—56.

Fama, E. F. and French, K. R. (2008). Dissecting anomalies. *The Journal of Finance*, 63(4):1653–1678.

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

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, 81(3):607–636.

Foerster, S. R. and Karolyi, G. A. (1999). The effects of market segmentation and investor recognition on asset prices: Evidence from foreign stocks listing in the united states. *The Journal of Finance*, 54(3):981–1013.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33:2326–2377.

Greenberg, E. and Parks, R. P. (1997). A predictive approach to model selection and multicollinearity. *Journal of Applied Econometrics*, 12(1):67–75.

Griffin, J. M. (2002). Are the fama and french factors global or country specific? *The Review of Financial Studies*, 15(3):783–803.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Han, Y., He, A., Rapach, D. E., and Zhou, G. (2023). Cross-sectional expected returns: New fama-macbeth regressions in the era of machine learning. Working paper, available at SSRN 3185335.

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

Hou, K., Xue, C., and Zhang, L. (2020). Replicating anomalies. *The Review of Financial Studies*, 33(5):2019–2133.

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

Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance*, 15(5):850–859.

Lewellen, J. (2015). The cross section of expected stock returns. *Critical Finance Review*, 4:1–44.

Lin, C.-C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1(1):42–55.

Liu, Y., Zhou, G., and Zhu, Y. (2021). Maximizing the Sharpe ratio: A genetic programming approach. Working paper, available at SSRN 3726609.

Patton, A. J. and Weller, B. (2022). Risk price variation: The missing half of empirical asset pricing. *The Review of Financial Studies*, 35(11):5127–5184.

Rapach, D. E., Strauss, J. K., Tu, J., and Zhou, G. (2019). Industry return predictability: A machine learning approach. *Journal of Financial Data Science*, 1(3):9–28.

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

Stock, J. H. and Watson, M. W. (2011). Introduction to econometrics, 3rd international edition.

Tang, X., Xue, F., and Qu, A. (2021). Individualized multidirectional variable selection. *Journal of the American Statistical Association*, pages 1–17.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhu, X., Tang, X., and Qu, A. (2021). Longitudinal clustering for heterogeneous binary data. *Statistica Sinica*, 31(2):603–624.

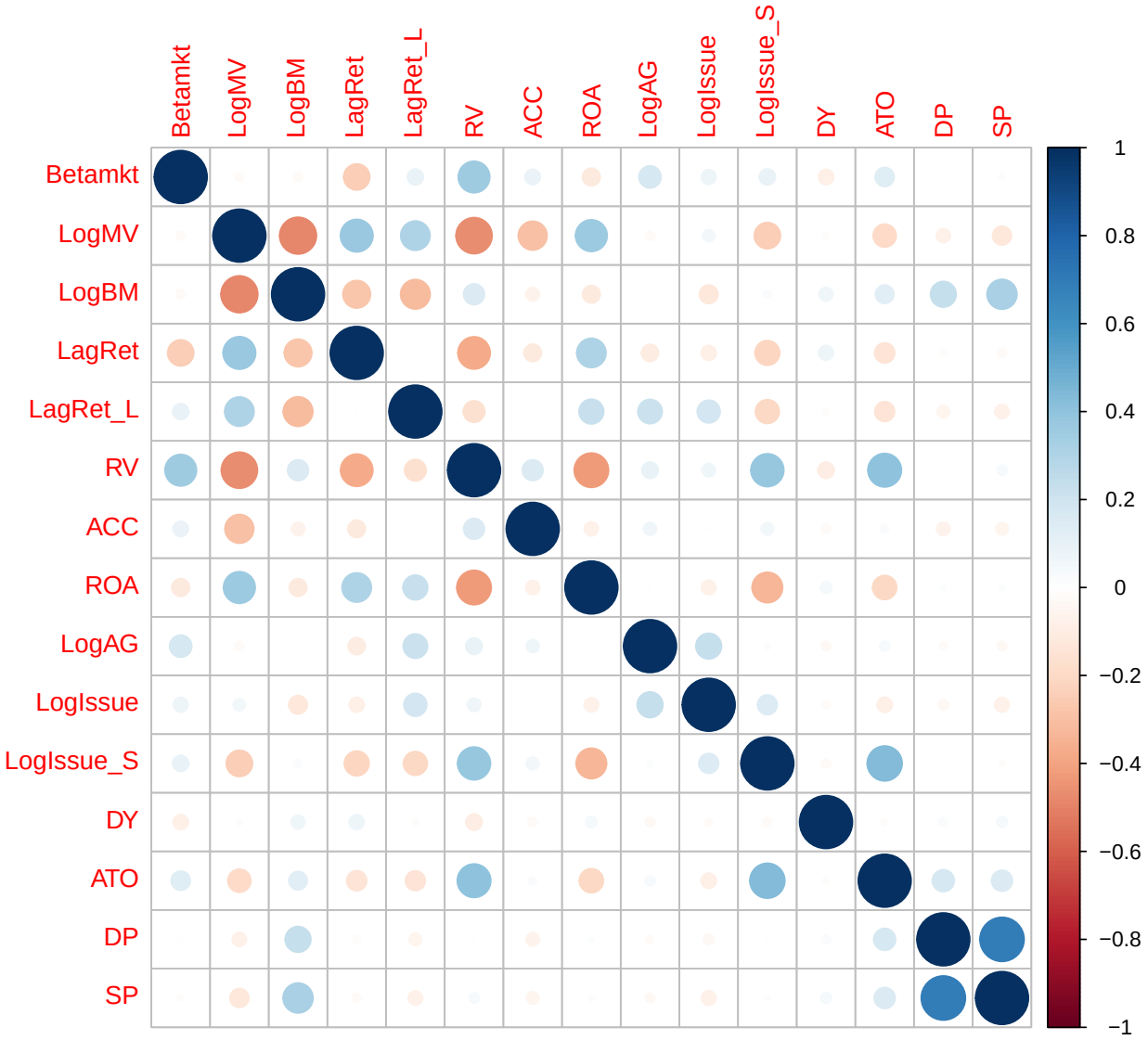


Figure 1: Correlation Heat Map of Firm Characteristics

This figure presents correlation heat map between 15 firm characteristics. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. Detailed definitions of firm characteristics are listed in the Appendix.

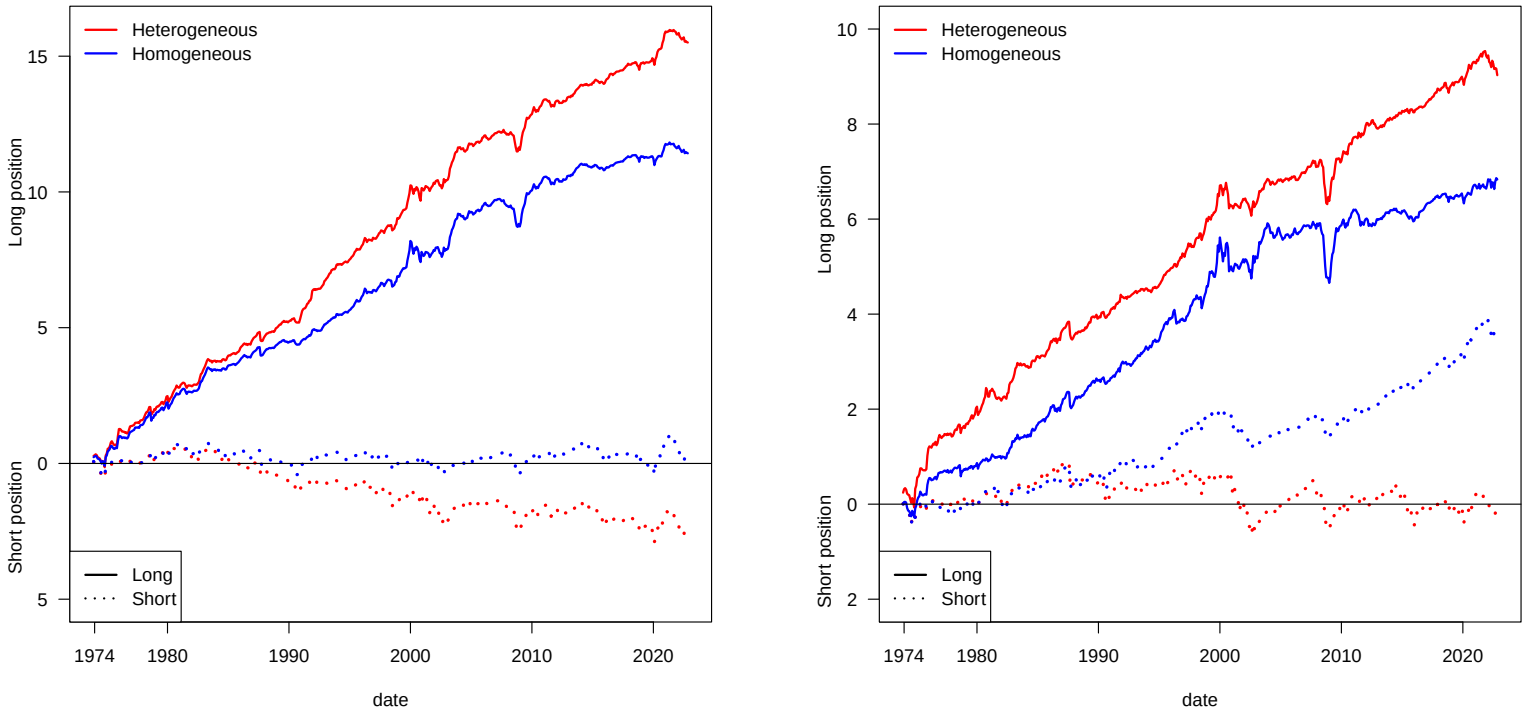


Figure 2: Cumulative Log Portfolio Returns for Heterogeneous Model and the Standard Homogeneous Model

This figure plots the cumulative log returns of the long (solid curves) and short (dashed curves) portfolios constructed based on predicted expected returns according to (12) for the standard homogeneous model and (13) for the heterogeneous model. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. The first ten years are left out for estimating rolling-window average of the prediction parameters. We use 15 firm characteristics in Lewellen (2015) to predict stock returns. Detailed definitions of firm characteristics used in the prediction are listed in the Appendix. We predict returns at the end of each month and sort stocks into deciles for trading during the next month. The decile portfolios are combined by equal weights (Panel A) or weights proportional to stock sizes at the end of the previous month (Panel B) and rebalanced monthly.

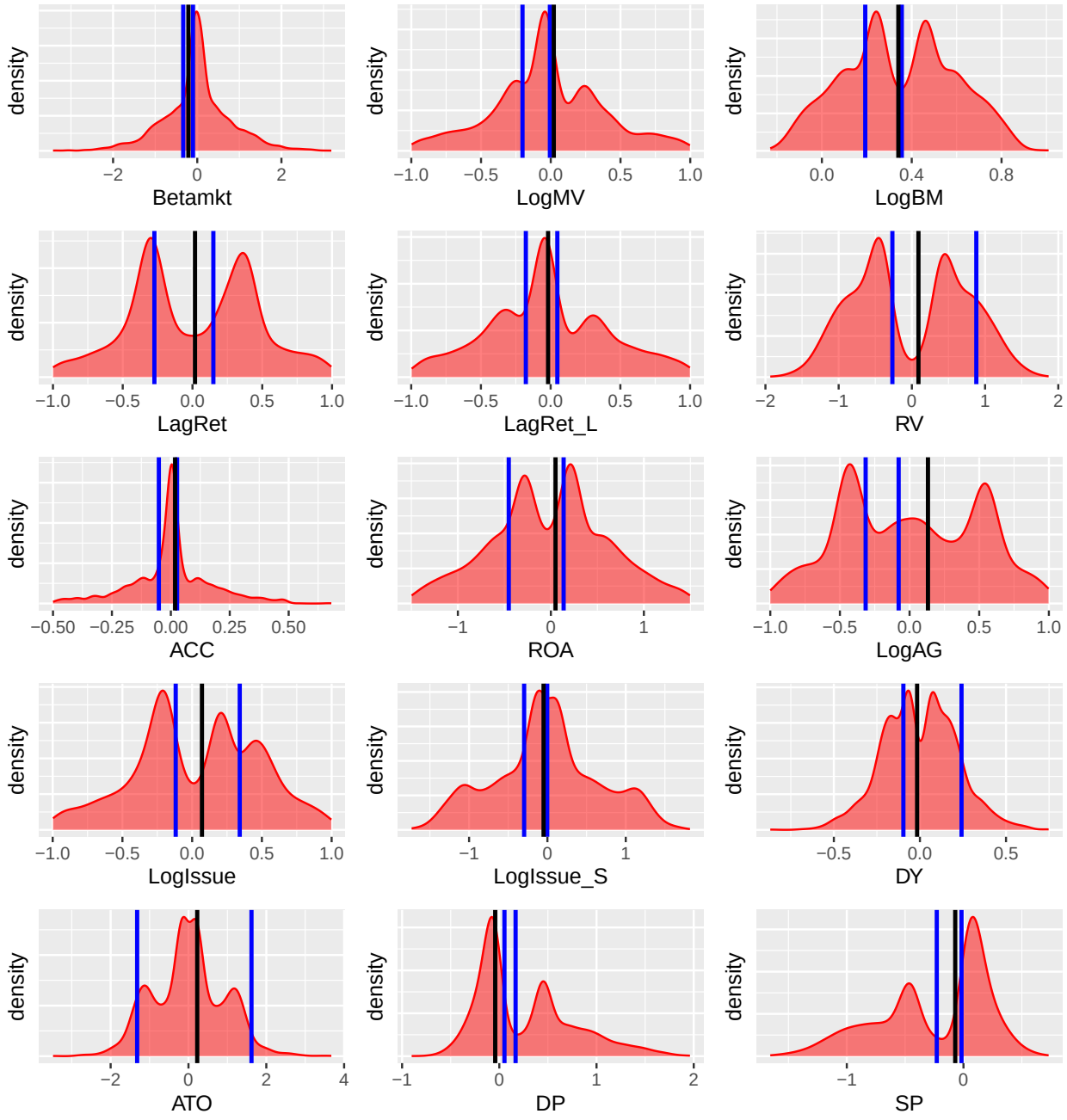
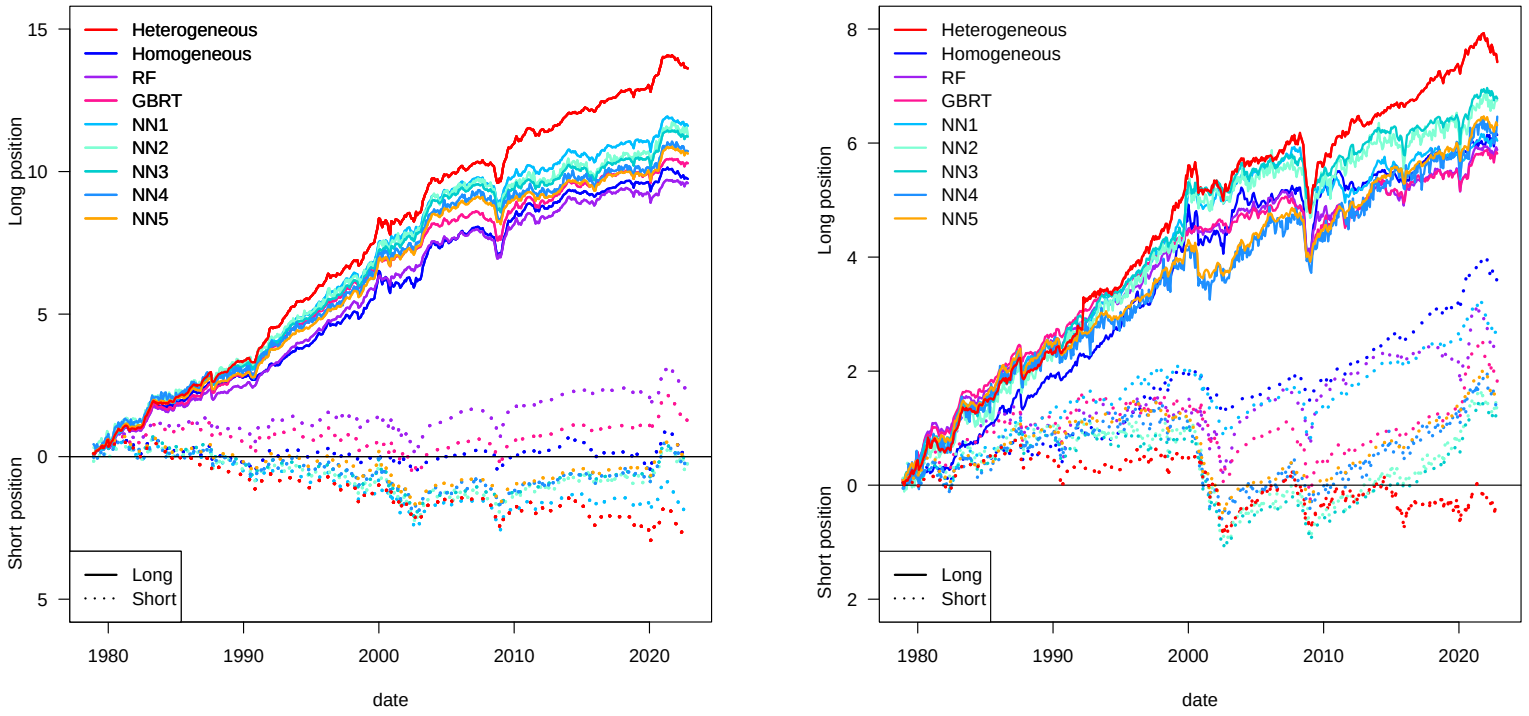


Figure 3: Distribution of return responses

This figure plots the distribution of individual return responses under the heterogeneous model, the return responses of the two clusters identified by the Patton and Weller (2022) approach when $k = 2$ (blue lines), and the homogeneous one (black line) under the standard model. The sample includes all CRSP common stocks (share code = 10 or 11) in January 2010. We use 15 firm characteristics in Lewellen (2015) to predict stock returns. Detailed definitions of firm characteristics used in the prediction are listed in the Appendix.



Panel A: Equal-Weighted

Panel B: Value-Weighted

Figure 4: Cumulative Log Portfolio Returns for Heterogeneous and Machine Learning Models

This figure plots the cumulative log returns of the long (solid curves) and short (dashed curves) portfolios constructed based on predicted expected returns by the heterogeneous, homogeneous, and machine learning models. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. We use 15 firm characteristics in Lewellen (2015) to predict stock returns. Detailed definitions of firm characteristics used in the prediction are listed in the Appendix. We predict returns at the end of each month and sort stocks into deciles for trading during the next month. The decile portfolios are combined by equal weights (Panel A) or weights proportional to stock sizes at the end of the previous month (Panel B) and rebalanced monthly. The popular machine learning models are the random forest (RF), the gradient boosting regression tree (GBRT), 1-layer neural network (NN-1), 2-layer neural network (NN-2), 3-layer neural network (NN-3), 4-layer neural network (NN-4), and 5-layer neural network (NN-5).

Table 1: Summary Statistics of Firm Characteristics.

This table reports time-series averages of monthly cross-sectional summary statistics of 15 firm characteristics. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. Detailed definitions of firm characteristics are listed in the Appendix.

Char	Mean	Std	5%	25%	50%	75%	95%	# Obs
Return(%)	1.251	14.766	-17.740	-5.720	0.251	6.637	22.579	4,195
Betamkt	0.941	0.568	0.116	0.547	0.894	1.284	1.944	4,243
LogMV	4.809	2.011	1.687	3.348	4.713	6.185	8.243	4,223
LogBM	-0.374	1.088	-2.033	-0.927	-0.351	0.141	1.226	4,090
LagRet	0.140	0.571	-0.484	-0.154	0.064	0.312	0.976	4,185
LagRet_L	0.073	0.618	-1.007	-0.233	0.127	0.430	0.974	3,775
RV	0.032	0.021	0.012	0.019	0.027	0.040	0.068	4,243
ACC	-0.011	0.059	-0.047	-0.004	0.000	0.000	0.002	3,484
ROA	0.016	0.746	-0.312	-0.005	0.032	0.068	0.144	4,238
LogAG	0.097	0.289	-0.249	-0.014	0.070	0.175	0.542	4,205
LogIssue	0.128	0.363	-0.128	-0.005	0.027	0.167	0.704	3,901
LogIssue_S	0.042	0.190	-0.049	-0.001	0.004	0.028	0.270	4,242
DY	0.022	0.066	0.000	0.001	0.010	0.032	0.073	4,179
ATO	0.097	0.315	0.009	0.030	0.062	0.112	0.271	3,946
DP	1.159	3.730	0.000	0.069	0.318	0.930	4.163	4,208
SP	2.883	6.129	0.126	0.585	1.278	2.742	10.067	4,228

Table 2: Correlations between Firm Characteristics

This table reports pairwise correlations between 15 firm characteristics. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. Detailed definitions of firm characteristics are listed in the Appendix.

	Betamkt	LogMV	BM	LagRet	LogRet	SD	ACC	ROA	LogAG	LogIssue	LogIssue_S	DY	ATO	DP	SP
Betamkt	1.00														
LogMV	-0.03	1.00													
LogBM	-0.03	-0.49	1.00												
LagRet	-0.24	0.38	-0.27	1.00											
LagRet_L	0.09	0.31	-0.31	-0.01	1.00										
SD	0.36	-0.47	0.16	-0.37	-0.17	1.00									
ACC	0.09	-0.30	-0.07	-0.11	0.00	0.15	1.00								
ROA	-0.11	0.36	-0.11	0.31	0.22	-0.42	-0.07	1.00							
LogAG	0.18	-0.03	0.00	-0.11	0.21	0.10	0.06	0.01	1.00						
LogIssue	0.08	0.05	-0.12	-0.08	0.19	0.07	0.01	-0.08	0.24	1.00					
LogIssue_S	0.09	-0.25	0.02	-0.22	-0.20	0.39	0.06	-0.33	0.02	0.14	1.00				
DY	-0.08	-0.02	0.07	0.08	-0.02	-0.10	-0.03	0.05	-0.03	-0.02	-0.03	1.00			
ATO	0.13	-0.19	0.12	-0.15	-0.14	0.40	0.02	-0.21	0.04	-0.08	0.44	-0.02	1.00		
DP	-0.01	-0.07	0.24	-0.02	-0.06	0.00	-0.07	0.01	-0.02	-0.04	0.00	0.03	0.17	1.00	
SP	-0.01	-0.13	0.32	-0.03	-0.07	0.04	-0.06	0.01	-0.03	-0.08	-0.01	0.04	0.16	0.70	1.00

Table 3: Portfolios Sorted on Out-of-Sample Return Forecasts

This table reports monthly average returns (Avg), volatilities (Vol), and annualized Sharpe ratios (SR) for deciles of stocks sorted by predicted expected returns according to (12) for the standard homogeneous model and (13) for the heterogeneous model. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. The first ten years are left out for estimating the rolling-window average of the prediction parameters. We use 15 firm characteristics in Lewellen (2015) to predict stock returns. Detailed definitions of firm characteristics used in the prediction are listed in the Appendix. We predict returns at the end of each month and sort stocks into deciles for trading during the next month. The decile portfolios are combined by equal weights (Equal-Weighted) or weights proportional to stock sizes at the end of the previous month (Value-Weighted) and rebalanced monthly. We also report monthly average return, volatility, and annualized Sharpe ratio for a trading strategy (H-L) that longs the high (H) decile and shorts the low (L) decile.

	Heterogeneous Model						Homogeneous Model					
	Equal-Weighted			Value-Weighted			Equal-Weighted			Value-Weighted		
	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR
Low(L)	-0.14	8.08	-0.06	0.21	6.58	0.11	0.27	7.49	0.13	0.74	5.54	0.47
1	0.64	6.73	0.33	0.52	5.45	0.33	0.92	6.12	0.52	0.96	4.82	0.69
2	0.98	6.07	0.56	0.76	4.91	0.54	1.13	5.71	0.69	1.11	4.81	0.80
3	1.14	5.62	0.71	1.05	4.75	0.76	1.19	5.61	0.73	1.14	4.93	0.80
4	1.30	5.41	0.83	0.96	4.49	0.74	1.28	5.59	0.79	1.15	5.08	0.79
5	1.39	5.21	0.92	1.00	4.46	0.78	1.37	5.67	0.83	1.19	5.22	0.79
6	1.48	5.31	0.97	1.09	4.64	0.81	1.46	5.88	0.86	1.25	5.61	0.77
7	1.67	5.47	1.06	1.14	4.95	0.80	1.64	6.12	0.93	1.25	5.76	0.75
8	1.93	6.21	1.08	1.34	5.30	0.88	1.89	6.68	0.98	1.34	6.52	0.71
High(H)	2.99	8.34	1.24	1.77	6.80	0.90	2.26	7.85	1.00	1.44	7.36	0.68
H-L	3.13	5.41	2.01	1.57	6.33	0.86	1.98	5.81	1.18	0.69	6.31	0.38

Table 4: Return Forecast Accuracy

This table reports the time-series average (Avg) of estimates $\hat{\kappa}_1^{t^*}$ and $\hat{\kappa}_0^{t^*}$ in (14), where $\tilde{r}_i^{t^*}$ is estimated by (12) for the standard homogeneous model and (13) for the heterogeneous model. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. The first ten years are left out for estimating the rolling-window average of the prediction parameters. We use 15 firm characteristics in Lewellen (2015) to predict stock returns. Detailed definitions of firm characteristics used in the prediction are listed in the Appendix. The table also reports Newey-West standard errors (SE) with six lags and the corresponding t-test statistics (t-stat).

	Heterogeneous Model		Homogeneous Model	
	$\hat{\kappa}_1^{t^*}$	$\hat{\kappa}_0^{t^*}$	$\hat{\kappa}_1^{t^*}$	$\hat{\kappa}_0^{t^*}$
Avg	0.852	0.004	0.447	0.006
SE	1.013	0.065	1.384	0.058
t-stat	20.403	1.491	8.721	3.076

Table 5: Portfolio Sorting with Size-Segmented Homogeneous Model

This table reports monthly average returns (Avg), volatilities (Vol), and annualized Sharpe ratios (SR) for deciles of stocks sorted by predicted expected returns according to (12) with market segmentation based on the firm market values. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. The first ten years are left out for estimating the rolling-window average of the prediction parameters. For each month, we split the stocks into two subsamples: "tiny v.s. all-but-tiny", with NYSE 20th percentile as the market equity cut-off, and "large v.s. all-but-large", with NYSE 50th percentile as the market equity cutoff. Then we predict the next-month return within each subsample by (12). We use 15 firm characteristics in Lewellen (2015) to predict stock returns. Detailed definitions of firm characteristics used in the prediction are listed in the Appendix. We predict returns at the end of each month and sort stocks into deciles for trading during the next month. The decile portfolios are combined by equal weights (Equal-Weighted) or weights proportional to stock sizes at the end of the previous month (Value-Weighted) and rebalanced monthly. We also report monthly average return, volatility, and annualized Sharpe ratio for a trading strategy (H-L) that longs the high (H) decile and shorts the low (L) decile.

	Large v.s. All-but-large						Tiny v.s. All-but-tiny					
	Equal-Weighted			Value-Weighted			Equal-Weighted			Value-Weighted		
	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR
Low (L)	0.22	7.02	0.11	0.46	6.33	0.25	0.24	6.99	0.12	0.32	6.63	0.17
2	0.89	5.88	0.52	0.71	5.07	0.49	0.81	6.07	0.46	0.61	5.31	0.39
3	1.01	5.45	0.64	0.91	4.68	0.67	1.05	5.72	0.64	0.85	5.06	0.58
4	1.23	5.27	0.81	0.97	4.46	0.75	1.17	5.41	0.75	0.86	4.67	0.64
5	1.26	5.29	0.83	0.99	4.48	0.77	1.25	5.21	0.83	1.07	4.52	0.82
6	1.39	5.41	0.89	1.09	4.61	0.82	1.42	5.19	0.95	1.16	4.62	0.87
7	1.48	5.67	0.90	1.23	5.11	0.83	1.47	5.35	0.95	1.22	4.71	0.90
8	1.63	6.01	0.94	1.21	5.28	0.79	1.56	5.81	0.93	1.29	5.12	0.87
9	1.93	6.63	1.01	1.41	6.39	0.76	1.91	6.53	1.01	1.31	6.03	0.75
High (H)	2.41	7.87	1.06	1.64	7.82	0.73	2.49	8.06	1.07	1.45	7.81	0.64
H-L	2.19	5.17	1.47	1.18	6.98	0.59	2.25	5.03	1.55	1.13	6.92	0.57

Table 6: Portfolio Sorting with Industry-Segmented Homogeneous Model

This table reports monthly average returns (Avg), volatilities (Vol), and annualized Sharpe ratios (SR) for deciles of stocks sorted by predicted expected returns according to (12) with market segmentation based on different industries. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. The market is segmented to eleven industries based on the firms' standard industrial classification (SIC) codes. The first ten years are left out for estimating the rolling-window average of the prediction parameters. For each month, we split the stocks into eleven different industries. Then we predict the next-month return within each industry by (12) using the 15 firm characteristics in Lewellen (2015) to predict stock returns. Detailed definitions of firm characteristics used in the prediction and the industry segmentation based on the SIC codes are listed in the Appendix. We predict returns at the end of each month and sort stocks into deciles for trading during the next month. The decile portfolios are combined by equal weights (Equal-Weighted) or weights proportional to stock sizes at the end of the previous month (Value-Weighted) and rebalanced monthly. We also report monthly average return, volatility, and annualized Sharpe ratio for a trading strategy (H-L) that longs the high (H) decile and shorts the low (L) decile.

	Equal-weighted			Value-weighted		
	Avg	Vol	SR	Avg	Vol	SR
L	0.55	7.47	0.25	0.62	5.87	0.37
1	0.87	5.91	0.51	0.88	4.88	0.63
2	1.12	5.71	0.68	1.07	4.85	0.77
3	1.21	5.66	0.74	1.07	5.01	0.74
4	1.33	5.62	0.82	1.13	4.98	0.78
5	1.40	5.66	0.86	1.08	5.20	0.72
6	1.47	5.84	0.87	1.16	5.29	0.76
7	1.60	6.14	0.90	1.16	5.896	0.68
8	1.85	6.70	0.95	1.14	6.31	0.62
H	2.13	7.75	0.95	1.30	6.61	0.68
H-L	1.58	5.21	1.05	0.68	5.45	0.43

Table 7: Portfolio Sorting with Patton and Weller (2022)

This table reports monthly average returns (Avg), volatilities (Vol), and annualized Sharpe ratios (SR) for deciles of stocks sorted by predicted expected returns based on the model proposed in Patton and Weller (2022) using 15 factors suggested by Lewellen (2015), with the number of clusters k varying from 2 to 4. Detailed definitions of firm characteristics used in the prediction are listed in the Appendix. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. The first ten years are left out for estimating the rolling-window average of the prediction parameters. We predict returns at the end of each month and sort stocks into deciles for trading during the next month. The decile portfolios are combined by equal weights (Equal-Weighted) or weights proportional to stock sizes at the end of the previous month (Value-Weighted) and rebalanced monthly. We report monthly average return, volatility, and annualized Sharpe ratio for a trading strategy (H-L) that longs the high (H) decile and shorts the low (L) decile.

	$k=2$						$k=3$						$k=4$					
	Equal-Weighted			Value-Weighted			Equal-Weighted			Value-Weighted			Equal-Weighted			Value-Weighted		
	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR
Low(L)	0.95	6.00	0.55	0.79	5.12	0.54	1.25	6.02	0.72	1.03	4.87	0.74	1.19	6.06	0.68	1.01	4.83	0.73
1	0.82	4.82	0.59	0.79	4.49	0.61	1.06	5.33	0.69	0.88	4.72	0.64	1.02	5.39	0.66	0.81	4.79	0.59
2	1.02	4.74	0.74	0.99	4.17	0.82	1.13	4.96	0.79	0.92	4.34	0.74	1.05	4.93	0.74	0.91	4.30	0.73
3	1.14	5.00	0.79	0.94	4.35	0.75	1.17	4.96	0.82	0.95	4.33	0.76	1.16	4.97	0.81	0.91	4.34	0.73
4	1.22	5.22	0.81	0.89	4.39	0.70	1.27	5.23	0.84	0.96	4.36	0.76	1.28	5.11	0.87	1.01	4.39	0.80
5	1.39	5.66	0.85	0.95	4.81	0.69	1.38	5.53	0.86	0.88	4.72	0.65	1.38	5.58	0.86	0.85	4.78	0.62
6	1.48	6.24	0.82	1.05	5.26	0.69	1.55	6.20	0.87	1.14	5.14	0.77	1.57	6.16	0.89	1.10	5.16	0.74
7	1.64	7.06	0.80	1.13	6.18	0.64	1.54	6.71	0.79	1.21	5.91	0.71	1.63	6.85	0.83	1.26	6.00	0.73
8	1.82	8.37	0.75	1.16	7.29	0.55	1.59	7.55	0.73	1.04	6.81	0.53	1.62	7.45	0.76	1.06	6.73	0.55
High(H)	1.95	8.79	0.77	1.27	7.44	0.59	1.69	7.42	0.79	1.27	5.80	0.76	1.72	7.43	0.80	1.29	5.80	0.77
H-L	1.00	4.15	0.84	0.48	4.65	0.36	0.43	2.86	0.53	0.23	2.93	0.28	0.53	2.83	0.65	0.28	2.94	0.33

Table 8: Equal-Weighted Portfolio Sorting with Machine Learning Models

This table reports monthly average returns (Avg), volatilities (Vol), and annualized Sharpe ratios (SR) for deciles of stocks sorted by predicted expected returns based on the heterogeneous, homogeneous, and machine learning models. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. For machine learning models, the first ten years are left out as the training sample and the next five years are used as the validation sample. The sample period for heterogeneous and homogeneous models is from January 1969 to December 2022. The first ten years are left out for estimating the rolling-window average of the prediction parameters. We use 15 firm characteristics in Lewellen (2015) to predict stock returns. Detailed definitions of firm characteristics used in the prediction are listed in the Appendix. We predict returns at the end of each month and sort stocks into deciles for trading during the next month. The decile portfolios are equal-weighted and rebalanced monthly. The machine learning models are the random forest (RF), the gradient boosting regression tree (GBRT), 1-layer neural network (NN-1), 2-layer neural network (NN-2), 3-layer neural network (NN-3), 4-layer neural network (NN-4), and 5-layer neural network (NN-5).

	Heterogeneous			Homogeneous			RF		
	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR
Low(L)	-0.20	8.21	-0.08	0.26	7.57	0.12	0.34	8.07	0.33
2	0.60	6.74	0.31	0.89	6.10	0.51	1.04	6.90	0.52
3	0.93	6.05	0.53	1.09	5.63	0.67	1.12	5.84	0.67
4	1.09	5.52	0.69	1.13	5.45	0.72	1.26	5.43	0.80
5	1.24	5.28	0.81	1.21	5.39	0.78	1.25	5.45	0.80
6	1.33	5.02	0.92	1.31	5.45	0.83	1.18	5.46	0.75
7	1.41	5.02	0.98	1.39	5.64	0.85	1.25	5.32	0.82
8	1.60	5.14	1.08	1.57	5.85	0.93	1.30	5.46	0.82
9	1.86	5.77	1.12	1.78	6.38	0.97	1.15	5.45	0.73
High(H)	2.90	7.82	1.29	2.14	7.61	0.97	2.05	6.55	1.08
H-L	3.10	4.95	2.17	1.89	5.78	1.13	1.71	5.07	1.18

	GBRT			NN1			NN2		
	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR
Low(L)	0.39	7.91	0.24	-0.06	8.18	-0.03	0.02	8.26	-0.04
2	0.97	6.74	0.50	0.79	6.52	0.42	0.80	6.54	0.42
3	1.15	6.12	0.65	1.04	5.89	0.61	1.00	5.86	0.59
4	1.27	5.54	0.80	1.16	5.30	0.76	1.24	5.19	0.83
5	1.26	5.54	0.79	1.28	5.12	0.87	1.30	5.18	0.87
6	1.23	5.50	0.77	1.33	5.08	0.91	1.29	5.07	0.88
7	1.28	5.32	0.83	1.36	5.15	0.92	1.39	5.15	0.93
8	1.25	5.30	0.81	1.45	5.14	0.98	1.49	5.18	0.99
9	1.25	5.35	0.81	1.54	5.44	0.98	1.56	5.49	0.99
High(H)	2.17	6.46	1.17	2.27	7.22	1.09	2.26	7.25	1.17
H-L	1.78	5.12	1.20	2.34	5.25	1.54	2.24	5.31	1.46

	NN3			NN4			NN5		
	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR
Low(L)	0.33	8.53	0.13	0.26	8.24	0.14	0.29	8.18	0.12
2	0.83	6.60	0.44	0.87	6.58	0.46	0.83	6.59	0.43
3	1.03	5.89	0.60	1.05	5.80	0.63	1.05	5.78	0.63
4	1.13	5.43	0.72	1.02	5.49	0.65	1.12	5.49	0.71
5	1.20	5.16	0.80	1.19	5.10	0.81	1.16	5.17	0.78
6	1.24	5.09	0.84	1.23	5.06	0.84	1.23	5.08	0.84
7	1.31	5.02	0.91	1.33	5.09	0.90	1.34	5.11	0.91
8	1.34	5.03	0.93	1.38	5.12	0.94	1.46	5.12	0.99
9	1.57	5.43	1.00	1.61	5.58	1.00	1.63	5.52	1.02
High(H)	2.39	6.94	1.19	2.29	6.96	1.14	2.27	6.92	1.14
H-L	2.06	5.04	1.42	2.02	5.08	1.38	1.98	5.12	1.34

Table 9: Value-Weighted Portfolio Sorting with Machine Learning Models

This table reports monthly average returns (Avg), volatilities (Vol), and annualized Sharpe ratios (SR) for deciles of stocks sorted by predicted expected returns based on the heterogeneous, homogeneous, and machine learning models. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. For machine learning models, the first ten years are left out as the training sample and the next five years are used as the validation sample. The sample period for heterogeneous and homogeneous models is from January 1969 to December 2022. The first ten years are left out for estimating the rolling-window average of the prediction parameters. We use 15 firm characteristics in Lewellen (2015) to predict stock returns. Detailed definitions of firm characteristics used in the prediction are listed in the Appendix. We predict returns at the end of each month and sort stocks into deciles for trading during the next month. The decile portfolios are combined by weights proportional to stock sizes at the end of the previous month and rebalanced monthly. The machine learning models are the random forest (RF), the gradient boosting regression tree (GBRT), 1-layer neural network (NN-1), 2-layer neural network (NN-2), 3-layer neural network (NN-3), 4-layer neural network (NN-4), and 5-layer neural network (NN-5).

	Heterogeneous			Homogeneous			RF		
	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR
Low(L)	0.20	6.69	0.10	0.83	5.51	0.52	0.60	7.61	0.27
2	0.54	5.47	0.34	1.00	4.79	0.72	1.06	6.59	0.56
3	0.77	4.87	0.55	1.12	4.75	0.81	0.96	5.29	0.63
4	1.06	4.72	0.78	1.13	4.87	0.80	1.11	5.11	0.75
5	0.96	4.39	0.75	1.16	5.00	0.81	1.03	5.02	0.71
6	0.98	4.33	0.78	1.18	5.11	0.80	0.95	5.05	0.65
7	1.04	4.46	0.81	1.23	5.61	0.76	1.07	4.78	0.78
8	1.11	4.79	0.80	1.26	5.77	0.76	1.05	5.23	0.69
9	1.33	5.08	0.91	1.34	6.56	0.71	1.09	5.32	0.71
High(H)	1.62	6.31	0.89	1.45	7.43	0.67	1.29	5.78	0.77
H-L	1.41	6.08	0.81	0.62	6.41	0.34	0.69	5.72	0.42

	GBRT			NN1			NN2		
	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR
Low(L)	0.62	7.47	0.29	0.69	6.18	0.39	0.64	6.16	0.36
2	1.01	6.22	0.56	0.92	5.19	0.61	0.92	5.14	0.62
3	0.85	5.20	0.57	1.08	4.79	0.78	1.16	4.81	0.84
4	1.09	4.94	0.77	1.09	4.53	0.84	1.03	4.58	0.78
5	1.03	5.05	0.71	1.23	4.57	0.94	1.23	4.48	0.95
6	1.11	5.08	0.76	1.10	4.90	0.78	1.10	5.01	0.76
7	1.14	5.04	0.78	1.14	5.12	0.77	1.17	5.07	0.80
8	1.09	4.93	0.77	1.15	5.16	0.77	1.14	5.16	0.77
9	1.03	5.31	0.67	1.26	5.39	0.81	1.36	5.34	0.88
High(H)	1.27	5.69	0.77	1.36	6.43	0.73	1.35	6.41	0.73
H-L	0.65	5.89	0.38	0.67	5.60	0.42	0.71	5.77	0.43

	NN3			NN4			NN5		
	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR
Low(L)	0.47	6.75	0.22	0.50	6.62	0.31	0.49	6.58	0.26
2	0.76	5.51	0.48	0.90	5.57	0.56	0.96	5.56	0.60
3	0.83	5.08	0.57	0.99	5.01	0.68	0.91	4.96	0.63
4	1.02	4.75	0.75	1.14	4.70	0.84	1.04	4.71	0.77
5	0.98	4.75	0.72	1.01	4.91	0.71	1.01	4.95	0.70
6	1.10	4.80	0.79	0.94	4.89	0.67	1.06	4.92	0.75
7	1.12	4.80	0.81	1.25	4.98	0.87	1.22	5.06	0.84
8	1.22	5.16	0.82	1.20	5.01	0.83	1.23	5.06	0.84
9	1.45	5.58	0.90	1.50	5.38	0.97	1.46	5.41	0.94
High(H)	1.47	6.87	0.77	1.41	6.10	0.80	1.40	6.11	0.79
H-L	1.00	5.97	0.58	0.91	5.84	0.56	0.90	5.71	0.57

Table 10: Portfolio Sorting with Machine Learning Models using 94 Characteristics

This table lists the average monthly returns of the high decile (High), low decile (Low), and long-short portfolios (H-L), as well as annualized Sharpe ratios (SR) for H-L from Table 7 in Gu et al. (2020). Portfolios are constructed based on return predictions by machine learning models. All portfolios are value-weighted. This table also reports the same statistics for corresponding portfolios constructed based on the heterogeneous model. The sample period is from January 1987 to December 2016. The machine learning models are the random forest (RF), the gradient boosting regression tree (GBRT), 1-layer neural network (NN-1), 2-layer neural network (NN-2), 3-layer neural network (NN-3), 4-layer neural network (NN-4), and 5-layer neural network (NN-5).

Method	High	Low	H-L	SR
FM-H	2.21	-0.39	2.60	1.91
RF	1.53	-0.09	1.62	0.98
GBRT	1.17	0.18	0.99	0.81
NN1	1.52	-0.29	1.81	1.17
NN2	1.38	-0.54	1.92	1.16
NN3	1.69	-0.43	2.12	1.20
NN4	1.75	-0.52	2.26	1.35
NN5	1.46	-0.51	1.97	1.15

Table 11: Group-Average Individualized Return Responses

This table reports the time-series average of the group-average individualized return responses (Avg) and the t-test statistics (t-stat) with Newey-West adjusted standard errors. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. Detailed definitions of firm characteristics are listed in the Appendix. We divide stocks into two groups by their estimated individualized return responses in (7). We also report the average and the associated t-test statistics for the difference between groups.

	Group 1 (G1)		Group 2 (G2)		Diff (G1-G2)	
Char	Avg	t-stat	Avg	t-stat	Avg	t-stat
	N_1=3,463		N_2=526			
Betamkt	0.002	2.163	0.004	1.787	-0.002	-0.683
LogMV	0.002	6.884	-0.021	-16.833	0.022	12.595
LogBM	0.003	8.696	0.005	3.109	-0.002	-0.995
LagRet	0.012	11.207	0.000	0.026	0.012	3.997
LagRet_L	0.003	6.333	-0.012	-5.893	0.015	6.002
RV	-0.891	-25.082	0.566	5.452	-1.457	-9.204
ACC	0.018	1.796	-0.387	-2.593	0.404	2.463
ROA	0.026	9.961	-0.010	-0.894	0.036	2.877
LogAG	-0.003	-3.840	-0.020	-4.756	0.017	3.641
LogIssue	-0.003	-3.592	-0.019	-3.874	0.016	2.881
LogIssue_S	0.002	1.126	0.011	1.056	-0.009	-0.891
DY	0.056	4.984	-0.166	-3.490	0.222	4.355
ATO	-0.022	-3.040	-0.157	-5.693	0.136	4.019
DP	0.000	1.685	-0.001	-1.564	0.001	1.767
SP	0.000	-0.050	0.000	1.088	0.000	-0.938

Table 12: Portfolio Sorting with Single-Factor Heterogeneous Models

This table presents the average monthly returns of the high decile (High), low decile (Low), and long-short portfolios (H-L), along with the volatilities (Vol) and annualized Sharpe ratios (SR) of the H-L portfolios. Portfolios are constructed based on return predictions by 15 single-factor heterogeneous models, respectively. In each model, one specific factor is assumed to have heterogeneous responses and all other factors have homogeneous responses. The decile portfolios are combined by equal weights (Equal-Weighted) or weights proportional to stock sizes at the end of the previous month (Value-Weighted) and rebalanced monthly. The sample period is from January 1964 to December 2022.

	Equal-Weighted					Value-Weighted				
Heterogeneous Factor	Low	High	H-L	Vol	SR	Low	High	H-L	Vol	SR
Betamkt	0.06	2.53	2.47	5.95	1.44	0.58	1.77	1.19	6.77	0.61
LogMV	-0.02	2.69	2.72	5.51	1.71	0.33	1.48	1.15	5.25	0.76
LogBM	0.03	2.63	2.60	5.76	1.56	0.34	1.54	1.20	5.89	0.70
LagRet	-0.04	2.50	2.54	5.91	1.49	0.40	1.65	1.25	5.98	0.72
LagRet_L	0.08	2.60	2.52	5.44	1.61	0.55	1.43	0.88	6.07	0.50
RV	-0.00	2.97	2.97	5.71	1.80	0.55	1.84	1.29	6.26	0.71
ACC	0.01	2.49	2.48	5.97	1.44	0.62	1.73	1.11	5.49	0.70
ROA	-0.02	2.64	2.66	5.61	1.64	0.48	1.60	1.12	5.59	0.69
LogAG	0.01	2.57	2.57	5.76	1.54	0.52	1.26	0.74	5.71	0.45
LogIssue	0.01	2.49	2.48	5.93	1.45	0.59	1.84	1.25	6.60	0.65
LogIssue_S	0.01	2.46	2.45	5.98	1.42	0.62	1.81	1.18	6.65	0.62
DY	-0.00	2.49	2.49	6.01	1.43	0.59	1.86	1.27	6.81	0.65
ATO	0.00	2.48	2.48	5.76	1.49	0.49	1.59	1.10	6.81	0.56
DP	0.01	2.54	2.53	5.93	1.48	0.57	1.88	1.31	6.54	0.70
SP	0.02	2.50	2.48	5.89	1.46	0.58	1.84	1.26	6.58	0.66

Table 13: Group-Average Firm Characteristics

This table reports the averages (Avg) and standard deviations (SD) of firm characteristics for different stock groups. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. Detailed definitions of firm characteristics are listed in the Appendix. We divide stocks into two groups by their estimated individualized return responses in (7). We also report the average (Avg) and the associated t-test statistics (t-stat) for the differences in the means of characteristics between groups.

	Group 1 (G1)		Group 2 (G2)		Diff (G1-G2)	
Char	Avg	SD	Avg	SD	Avg	t-stat
	N_1=3,463		N_2=526			
Betamkt	0.986	0.163	1.073	0.202	-0.087	-9.065
LogMV	5.158	0.969	4.083	0.894	1.075	19.851
LogBM	-0.397	0.287	-0.393	0.372	-0.003	-0.174
LagRet	0.144	0.224	0.137	0.338	0.007	0.439
LagRet_L	0.100	0.278	-0.072	0.398	0.171	7.764
RV	0.030	0.008	0.043	0.014	-0.013	-13.433
ACC	-0.014	0.066	-0.004	0.022	-0.009	-5.500
ROA	0.026	0.567	-0.057	0.276	0.082	8.575
LogAG	0.095	0.246	0.079	0.312	0.035	4.868
LogIssue	0.109	0.047	0.181	0.115	-0.073	-6.626
LogIssue_S	0.032	0.016	0.057	0.040	-0.025	-6.698
DY	0.022	0.010	0.014	0.010	0.008	15.199
ATO	0.097	0.135	0.121	0.171	-0.024	-4.602
DP	2.499	4.648	0.697	2.055	1.801	1.794
SP	3.703	5.803	0.372	3.665	3.330	3.026

Table 14: Heterogeneity Level During Recessions and Expansions

This table reports the time-series average of cross-sectional standard deviations of estimated individualized return responses for 15 characteristics during recession and expansion. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. Detailed definitions of firm characteristics are listed in the Appendix. We identify recession and expansion periods according to the NBER recession indicator. The table also reports the t-test statistics (t-stat) of the difference between standard deviations of each return response during different subperiods.

Char	Recession	Expansion	t-stat
Betamkt	1.60	1.03	3.23
LogMV	0.35	0.28	1.94
LogBM	0.78	0.50	3.61
LagRet	1.97	1.11	4.70
LagRet_L	1.07	0.85	2.05
RV	6.29	5.12	1.79
ACC	1.43	0.73	2.84
ROA	6.27	3.40	3.48
LogAG	1.89	1.53	2.09
LogIssue	1.62	1.26	1.91
LogIssue_S	3.42	2.65	1.69
DY	9.27	9.83	-0.59
ATO	5.93	3.85	3.35
DP	0.16	0.07	3.52
SP	0.03	0.02	2.30

Table 15: Portfolios Sorting with Three- and Seven-Factor Models

This table reports monthly average returns (Avg), volatilities (Vol), and annualized Sharpe ratios (SR) for deciles of stocks sorted by predicted expected returns according to (12) for the standard homogeneous model and (13) for the heterogeneous model. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. The first ten years are left out for estimating the rolling-window average of the prediction parameters. We consider a seven-factor model (Panel A) and a three-factor model (Panel B) as described in Section 6.1. Detailed definitions of firm characteristics used in the prediction are listed in the Appendix. We predict returns at the end of each month and sort stocks into deciles for trading during the next month. The decile portfolios are combined by equal weights (Equal-Weighted) or weights proportional to stock sizes at the end of the previous month (Value-Weighted) and rebalanced monthly. We also report monthly average return, volatility, and annualized Sharpe ratio for a trading strategy (H-L) that longs the high (H) decile and shorts the low (L) decile.

	Heterogeneous Model						Homogeneous Model					
	Equal-Weighted			Value-Weighted			Equal-Weighted			Value-Weighted		
	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR
Panel A: Three-Factor Model												
Low(L)	0.17	7.14	0.08	0.38	6.94	0.20	0.66	6.11	0.37	0.95	5.13	0.64
1	0.75	5.93	0.44	0.71	5.68	0.43	0.89	5.87	0.52	1.06	5.07	0.73
2	1.00	5.47	0.63	0.88	5.08	0.60	1.04	5.69	0.63	1.09	4.97	0.76
3	1.10	5.34	0.71	0.98	4.95	0.68	1.06	5.74	0.64	1.17	5.11	0.79
4	1.21	5.43	0.77	0.97	4.91	0.69	1.20	5.81	0.72	1.21	5.25	0.80
5	1.37	5.52	0.86	1.01	4.75	0.74	1.22	5.79	0.73	1.26	5.51	0.79
6	1.45	5.59	0.90	0.95	4.77	0.69	1.33	5.82	0.79	1.24	5.65	0.76
7	1.56	5.75	0.94	1.00	4.65	0.75	1.57	5.83	0.94	1.50	5.87	0.88
8	1.86	6.32	1.02	1.03	4.92	0.73	1.89	6.16	1.06	1.53	6.57	0.81
High(H)	2.72	7.45	1.26	1.34	5.56	0.82	2.32	7.17	1.12	1.61	7.82	0.73
H-L	2.55	4.90	1.80	0.96	5.08	0.63	1.66	4.97	1.16	0.66	6.12	0.37
Panel B: Seven-Factor Model												
Low(L)	-0.11	7.69	-0.05	0.30	6.13	0.17	0.19	6.84	0.10	0.82	5.05	0.56
1	0.75	6.41	0.40	0.74	5.68	0.45	0.94	6.03	0.54	1.06	4.66	0.78
2	1.09	5.73	0.66	0.87	5.10	0.59	1.14	5.74	0.69	1.23	4.63	0.92
3	1.19	5.52	0.75	0.99	4.89	0.70	1.23	5.70	0.75	1.16	4.79	0.84
4	1.33	5.49	0.84	1.02	4.67	0.76	1.29	5.79	0.77	1.25	4.97	0.87
5	1.44	5.51	0.91	0.99	4.72	0.73	1.36	5.82	0.81	1.24	5.25	0.82
6	1.52	5.65	0.93	1.05	4.58	0.80	1.50	5.97	0.87	1.35	5.82	0.80
7	1.70	5.95	0.99	1.03	4.47	0.80	1.63	6.15	0.92	1.25	6.07	0.71
8	1.96	6.37	1.07	1.24	5.09	0.84	2.01	6.56	1.06	1.40	6.48	0.76
High(H)	2.80	8.03	1.21	1.37	5.83	0.81	2.38	7.70	1.07	1.37	7.16	0.66
H-L	2.91	4.87	2.07	1.07	5.00	0.75	2.18	5.44	1.32	0.56	5.55	0.35

Table 16: Return Forecast Accuracy with Three- and Seven-Factor Models

This table reports the time-series average (Avg) of estimates $\hat{\kappa}_1^{t^*}$ and $\hat{\kappa}_0^{t^*}$ in (14), where $\tilde{r}_i^{t^*}$ is estimated by (12) for the standard homogeneous model and (13) for the heterogeneous model. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. The first ten years are left out for estimating the rolling-window average of the prediction parameters. We consider a three-factor model (Panel A) and a seven-factor model (Panel B) as described in Section 6.1. Detailed definitions of firm characteristics used in the prediction are listed in the Appendix. The table also reports Newey-West standard errors (SE) with six lags, the corresponding t-test statistics (t-stat), and the time-series average of the regression R^2.

	Heterogeneous Model		Homogeneous Model	
	$\hat{\kappa}_1^{t^*}$	$\hat{\kappa}_0^{t^*}$	$\hat{\kappa}_1^{t^*}$	$\hat{\kappa}_0^{t^*}$
Panel A: Three-Factor Model				
Avg	0.716	0.003	0.489	0.006
SE	0.084	0.002	0.136	0.014
t-stat	10.20	1.64	5.46	2.51
Panel B: Seven-Factor Model				
Avg	0.787	0.004	0.576	0.005
SE	0.048	0.002	0.071	0.002
t-stat	13.52	1.36	8.72	1.89

Table 17: Portfolio Sorting with Different Aggregation Horizons

This table reports monthly average returns (Avg), volatilities (Vol), and annualized Sharpe ratios (SR) for long-short portfolios constructed based on predicted expected returns according to (12) for the standard homogeneous model and (13) for the heterogeneous model. The sample includes all CRSP common stocks (share code = 10 or 11) from January 1964 to December 2022. The first ten years are left out for estimating the rolling-window average of the prediction parameters. We consider a three-factor model (Panel A), a seven-factor model (Panel B), and a 15-factor model (Panel C) as described in Sections 3.1 and 6.1. Detailed definitions of firm characteristics are listed in the Appendix. We also consider rolling windows of 1-year, 3-year, 5-year, and cumulative in (10) and (11) to aggregate return responses. We predict returns at the end of each month and sort stocks into deciles for trading during the next month. The decile portfolios are combined by equal weights (Equal-Weighted) or weights proportional to stock sizes at the end of the previous month (Value-Weighted) and rebalanced monthly. The long-short portfolio is constructed by longing the high decile and shorting the low decile.

	Heterogeneous Model						Homogeneous Model					
Rolling	Equal-Weighted			Value_Weighted			Equal-Weighted			Value_Weighted		
Window	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR	Avg	Vol	SR
Panel A: Three-Factor Model												
1-Year	3.07	6.33	1.71	1.45	5.91	0.84	1.72	6.80	0.88	1.13	7.41	0.53
3-Year	2.83	5.54	1.77	1.31	5.16	0.88	1.41	5.70	0.86	0.71	6.21	0.39
5-Year	2.61	5.32	1.69	1.07	5.09	0.73	1.43	5.40	0.92	0.72	6.36	0.39
Cumulative	2.27	4.68	1.68	1.12	5.35	0.73	1.58	5.21	0.94	0.80	6.32	0.40
Panel B: Seven-Factor Model												
1-Year	4.43	6.45	2.08	1.44	5.82	0.86	2.37	7.00	1.17	1.12	6.67	0.58
3-Year	3.26	5.38	2.10	1.23	5.11	0.84	1.88	5.30	1.22	0.72	5.62	0.44
5-Year	3.07	5.15	2.07	1.12	5.40	0.72	1.95	5.13	1.32	0.71	5.62	0.43
Cumulative	2.55	4.85	1.82	1.25	5.15	0.84	2.02	5.29	1.33	0.62	5.29	0.40
Panel C: 15-Factor Model												
1-Year	4.21	7.23	2.02	3.00	6.91	0.92	2.14	8.35	0.89	1.07	8.23	0.45
3-Year	3.61	5.92	2.11	1.69	6.12	0.96	1.56	6.35	0.85	0.79	6.92	0.39
5-Year	3.39	5.55	2.10	1.66	6.03	0.95	1.85	6.03	1.06	0.82	6.24	0.46
Cumulative	3.04	4.74	2.06	1.37	5.70	0.84	1.91	5.84	1.12	0.72	6.09	0.41