International Review of Financial Analysis 107 (2025) 104642

Contents lists available at ScienceDirect

International Review of Financial Analysis

journal homepage: www.elsevier.com/locate/irfa

Autoencoder asset pricing models and economic restrictions — international evidence☆

Lenka Nechvátalová *

Institute of Economic Studies, Faculty of Social Sciences, Charles University, Czech Republic
Institute of Information Theory and Automation, Czech Academy of Sciences, Czech Republic

ARTICLE INFO

JEL classification:
G11
G12
G15
C55

Keywords:
Machine learning
Asset pricing
Economic restrictions
Anomalies

ABSTRACT

We evaluate the performance of the Conditional Autoencoder (CAE) model by Gu et al. (2021) across U.S. and international datasets, considering economic constraints such as the exclusion of microcap and illiquid firms and the inclusion of transaction costs. The CAE model captures nonlinear relationships between returns and firm characteristics by jointly estimating latent factors and conditional betas while enforcing the no-arbitrage condition. The original study demonstrated significant reductions in out-of-sample pricing errors from both statistical and economic perspectives in the U.S. context. We validate these findings on the original U.S. dataset and show that the model generalises well to a U.S. dataset with a broader set of firm characteristics and to international markets. When economic constraints are introduced, portfolio profitability declines substantially. Profitability drops by 60%–85% when shifting from the full sample to the liquid sample before trading costs. However, after costs, only the liquid strategies remain profitable. In particular, long-only strategies on the liquid sample are the only ones to consistently outperform market benchmarks across all datasets, achieving Sharpe ratios between 0.65 and 0.78 for both equal- and value-weighted portfolios. Overall, the findings underscore both the limitations and the practical potential of the CAE model under realistic market frictions.

1. Introduction

Understanding the underlying factors that drive asset returns is a core challenge in both academic research and practical finance. While traditional factor models are well-established tools for capturing these dynamics, the Conditional Autoencoder (CAE) proposed by Gu et al. (2021) marks a notable methodological advance over traditional approaches. The CAE leverages autoencoder-based neural networks for dimension reduction, models nonlinear factor exposures from asset characteristics, and enforces the no-arbitrage condition by design. As a result, it achieves lower out-of-sample pricing errors than traditional linear models.

In this study, we evaluate how well the CAE model generalises to new data and scenarios. Our contributions are twofold: first, we assess its robustness across three datasets – the original U.S. dataset, a U.S. dataset with a broader set of firm characteristics, and an international dataset – demonstrating its applicability in diverse contexts. Second, we examine the model under economic constraints, including liquidity filters and transaction costs, to better understand its real-world implications. While portfolio profitability is generally limited under these constraints, certain portfolios remain profitable, particularly in the liquid sample. Moreover, profitability initially declines when restricting the dataset to liquid firms, but this pattern reverses after accounting for transaction costs, underscoring the importance of incorporating such considerations in empirical asset pricing.

The CAE model builds on the earlier Instrumental Principal Component Analysis (IPCA) framework of Kelly et al. (2019), which models asset returns using latent factors and linear, characteristic-based exposures. While IPCA incorporates firm information into a structured factor model, it is limited to linear relationships. The CAE extends this framework by using a neural network architecture to learn nonlinear mappings from characteristics to factor exposures, while maintaining economic structure through a no-arbitrage constraint. This added flexibility enables the CAE to better capture complex return dynamics and improve model fit. In the U.S. setting, it delivers higher total and predictive R^2 values and Sharpe ratios, outperforming traditional linear models in both statistical and economic measures.

Machine learning has been applied to asset pricing with varying emphasis on prediction, structure, and interpretability. Gu et al. (2020) evaluate a range of machine learning methods for predicting stock returns from firm characteristics. Among them, feedforward neural

☆ The financial support was provided by the Czech Science Foundation under the 24-11555S project.
* Corresponding author at: Institute of Economic Studies, Faculty of Social Sciences, Charles University, Opletalova 26, 110 00 Prague 1, Czech Republic.
E-mail address: lenka.nechvatalova@fsv.cuni.cz.

https://doi.org/10.1016/j.irfa.2025.104642
Received 12 February 2025; Received in revised form 18 July 2025; Accepted 15 September 2025
Available online 19 September 2025

networks perform best; this approach is purely predictive, without any underlying economic structure. In contrast, the CAE incorporates economic structure by modelling returns through latent factors and characteristic-based betas under a no-arbitrage condition. Its latent factors can be interpreted as time-varying linear combinations of returns, with exposures learned as nonlinear functions of firm characteristics. Extending this framework, Yang et al. (2024) introduce the Conditional Quantile Variational Autoencoder (CQVAE), which targets conditional quantiles to better capture the distributional properties of returns. Other machine learning approaches to asset pricing include Bryzgalova et al. (2020), Chen et al. (2024), and Kozak et al. (2020), among others. For a broader discussion of methodological contributions in asset pricing using machine learning and factor models, see Giglio et al. (2022).

Testing the validity of asset pricing anomalies and ensuring model consistency across datasets and market conditions are critical for robust financial research. Hou et al. (2020) analyse 447 individual anomalies and find that many fail rigorous tests, particularly after excluding microcap firms. Similarly, Avramov et al. (2023) demonstrate that excluding microcaps and distressed stocks significantly impacts the performance of machine learning-based portfolios, emphasising the role of economic constraints in practical applications. Additionally, literature on transaction costs underscores the importance of accounting for trading expenses in asset pricing models. DeMiguel et al. (2020) examine the transaction costs of multiple anomalies, while Nechvátalová (2024) and Novy-Marx and Velikov (2019) consider various cost-mitigating techniques and their performance.

While replication and economic constraints are critical for robust asset pricing research, extending analyses to international markets provides an additional layer of validation, as results from the U.S. may not always generalise to other markets. Since much of asset pricing research focuses on the U.S. Andrew Karolyi (2016), international evidence is crucial for evaluating the broader applicability of machine learning methods. Building on Gu et al. (2020) and Tobek and Hronec (2021) apply machine learning techniques to developed countries, while Hanauer and Kalsbach (2023) focus on emerging markets, demonstrating the global applicability of these approaches. Moreover, Jiang et al. (2023) show that price patterns can predict returns both in the U.S. and internationally, further validating the potential of machine learning in diverse market conditions.

In this study, we extend the CAE model's evaluation in two key ways. First, we examine its robustness across three datasets: the U.S. GKX dataset, the U.S. extended dataset with a broader set of firm characteristics, and an international dataset covering multiple developed markets. We test the CAE model's adaptability to different data environments and firm characteristics. Second, we incorporate real-world economic constraints, including liquidity filters and transaction costs, to evaluate portfolio profitability under practical trading conditions. By addressing these dimensions, we provide a comprehensive analysis of the CAE model's statistical and economic performance in diverse contexts.

For statistical performance, we evaluate the CAE model's out-of-sample ability to capture return variation using two key metrics. The total R^2 measures how well the model explains contemporaneous return variation across all assets, reflecting their shared riskiness and covariation. The predictive R^2, in contrast, evaluates the model's ability to forecast return variation using lagged factor exposures, functioning as a robust out-of-sample metric for assessing differences in risk compensation among assets. Using a three-factor specification, we evaluate the model across three datasets, distinguishing between full and liquid subsamples. Our results demonstrate that including a broader set of firm characteristics significantly enhances the model's predictive accuracy. Notably, the U.S. extended dataset achieves total R^2 and predictive R^2 values comparable to Gu et al. (2021), underscoring the robustness of the CAE model when applied to datasets with expanded firm characteristics.

In addition to these statistical results, we analyse the model's out-of-sample economic performance by evaluating portfolio profitability under constraints. We form long-short decile portfolios based on out-of-sample stock return predictions. Consistent with Avramov et al. (2023), who document that portfolio returns and profitability decline when focusing on more liquid stocks in the U.S. market, we find the same pattern across all three datasets. However, after accounting for transaction costs, the profitability proves to be higher in the liquid sample compared to the full sample. While short-side profitability is generally concentrated in illiquid and high-transaction-cost firms, its overall impact is limited, particularly after accounting for transaction costs. Consequently, in the U.S. GKX dataset, the long-only portfolio on the liquid sample stands out as the only consistently profitable strategy after transaction costs, outperforming the S&P index with Sharpe ratios of 0.70 for the equal-weighted and 0.67 for the value-weighted portfolios.

We extend our analysis to the U.S. extended dataset, which includes a broader set of firm characteristics, and to international markets, providing diverse environments to evaluate the CAE model. The U.S. extended dataset demonstrates that the model remains effective when applied to datasets with different firm characteristics, achieving stronger portfolio performance compared to the GKX dataset. Internationally, the results mostly align with U.S. outcomes in both statistical and economic metrics. However, the long-short strategy on the liquid sample is more effective, achieving Sharpe ratios of 0.73 (equal-weighted) and 0.54 (value-weighted), both outperforming the market index. Overall, these findings demonstrate that the CAE model delivers stable predictive performance and consistent long-only profitability even after transaction costs, in both U.S. and international markets.

The rest of this paper is organised as follows: Section 2 provides methodology and dataset description. Section 3 presents empirical results for the two U.S. datasets and the international dataset, and finally, Section 4 concludes the paper.

2. Methodology and data

2.1. Conditional autoencoder model

In this section, we describe the CAE model used in our study. Introduced by Gu et al. (2021), it incorporates firm characteristics and characteristic-managed portfolios to estimate latent factors and conditional betas. The main equation, which captures the essence of the model's structure, is:

$$r_{i,t} = \beta'_{i,t-1} f_t + u_{i,t}, \tag{1}$$

where $r_{i,t}$ represents the excess return of asset i at time t, $\beta_{i,t-1}$ denotes the factor loadings, f_t represents the latent factors, and $u_{i,t}$ is the error term.

Fig. 1 shows the overall structure of the conditional autoencoder model. The left part of the network models conditional betas as a nonlinear function of firm characteristics. The right side shows how we model the latent factors as a linear function of a set of characteristic-managed portfolios x_t.

We consider two methods for calculating the portfolios x_t: the first ensures consistency with the methodology of Gu et al. (2021), while the second introduces regularisation to address collinearity and instability in high-dimensional settings. Both approaches reduce dimensionality from the number of firms to the number of characteristics and accommodate incomplete panels by relying only on non-missing observations.

1. Ordinary Least Squares (OLS): Following Gu et al. (2021), the characteristic-managed portfolios are computed as:

$$x_t = \left(Z'_{t-1} Z_{t-1}\right)^{-1} Z'_{t-1} r_t, \tag{2}$$

where Z_{t-1} is the matrix of firm characteristics at time $t-1$, and r_t is the vector of returns at time t.

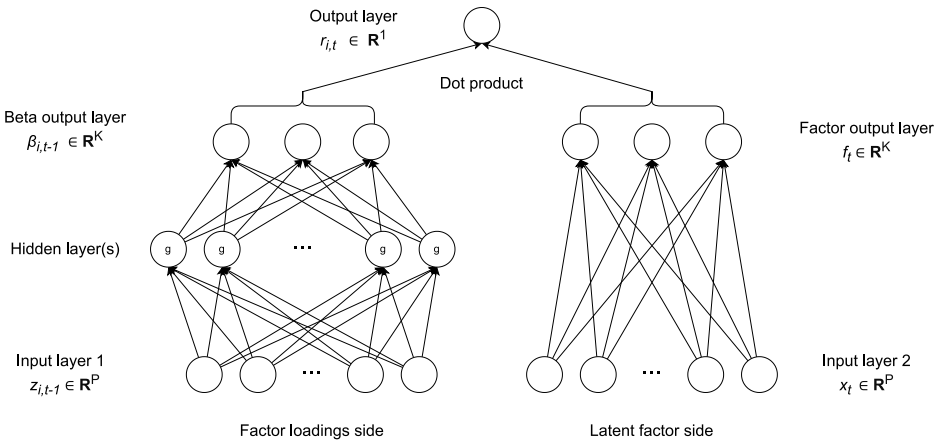

Fig. 1. Conditional autoencoder model. The left-hand side of the figure illustrates the process of obtaining factor loadings $\beta_{i,t-1}$ using firm characteristics $z_{i,t-1}$ as input. It involves one or more hidden layers with a nonlinear activation function g. The right-hand side shows how P characteristic-managed portfolios x_t are linearly transformed to derive the latent factor f_t.

2. Elastic Net: To mitigate instability caused by multicollinearity, we also estimate x_t using an elastic net regression.

$$x_t = \underset{x_t}{\operatorname{argmin}} \left(\|r_t - Z_{t-1}x_t\|^2 + \lambda_2 \|x_t\|^2 + \lambda_1 \|x_t\|_1 \right), \tag{3}$$

where λ_1 and λ_2 control the balance between sparsity and shrinkage. We set the total penalty to $\alpha = 0.0001$, with $\lambda_1 = 0.25 \cdot \alpha$ and $\lambda_2 = 0.75 \cdot \alpha$. Elastic Net is particularly beneficial when the number of characteristics is large and the cross-sectional sample is limited (e.g., after liquidity filtering). In such cases, it produces more stable estimates of x_t, which is crucial because the latent factors f_t are obtained via a linear transformation. Any instability in x_t is directly reflected in f_t, potentially distorting the factor structure and reducing both in-sample and out-of-sample performance. For this reason, we adopt the Elastic Net specification as our preferred approach.

The conditional betas $\beta_{i,t-1}$ (left side of Fig. 1) are derived through a series of transformations of the firm characteristics $z_{i,t-1}$:

$$z_{i,t-1}^{(0)} = z_{i,t-1}, \tag{4}$$

$$z_{i,t-1}^{(l)} = g\left(b^{(l-1)} + W^{(l-1)} z_{i,t-1}^{(l-1)}\right), \quad l = 1, \ldots, L_\beta, \tag{5}$$

$$\beta_{i,t-1} = b^{(L_\beta)} + W^{(L_\beta)} z_{i,t-1}^{(L_\beta)}. \tag{6}$$

Here, $z_{i,t-1}^{(l)}$ represents the firm characteristics transformed through lth layer of the neural network with the nonlinear activation function g. In the empirical part, we use the Rectified Linear Unit (ReLU) as the nonlinear activation function g and one to three hidden layers.

The latent factor f_t (right side of Fig. 1) is obtained from the characteristic-managed portfolios x_t:

$$x_t^{(0)} = x_t, \tag{7}$$

$$f_t = \widetilde{b}^{(0)} + \widetilde{W}^{(0)} x_t^{(0)}. \tag{8}$$

Here, $x_t^{(0)}$ is the initial input, which is the characteristic-managed portfolios, and Eq. (8) shows the linear transformation used to transform x_t into the latent factor f_t. The final output from the CAE model is obtained by taking the dot product of the conditional betas $\beta_{i,t-1}$ and the latent factors f_t.

2.2. Data

To evaluate the robustness and generalisability of our approach, we employ three datasets: the original U.S. equity dataset introduced by Gu et al. (2020, 2021) (U.S. GKX), an extended U.S. dataset, and an international dataset covering 23 developed markets. The datasets span up to 2018, with U.S. data beginning in 1957 and non-U.S. data starting in 1980 for the international dataset. We use the three-month U.S. Treasury bill rate as the risk-free rate across all datasets to maintain consistency, as all returns are denominated in U.S. dollars. Compared to the original study by Gu et al. (2021), our analysis includes two additional years of data. These datasets include a broad range of firm characteristics, enabling us to test the performance of the model across diverse settings and on unseen data.

U.S. GKX equity dataset

For the U.S. GKX dataset construction, we use individual stock returns from the CRSP/Compustat Merged Database from the Center for Research in Security Prices, with a dataset spanning from 1957 to 2018.

We incorporate 94 firm characteristics as used and provided by Gu et al. (2020, 2021), comprising 61 annual, 13 quarterly, and 20 monthly updated characteristics. To avoid forward-looking bias, the characteristics are appropriately lagged. For a comprehensive list of characteristics and additional details, refer to Gu et al. (2020).

Extended U.S. and international datasets

The extended U.S. and international datasets cover the period from 1957 to 2018 for the U.S. and 1980 to 2018 for ex-U.S. countries. The extended U.S. dataset, like the GKX dataset, covers the U.S. market but is independently constructed with a broader selection of firm characteristics. The international dataset covers equity data from the U.S. and 22 other developed countries.[1] The countries are grouped into four regions: U.S., Europe, Japan, and Asia Pacific. U.S. data is sourced from the CRSP/Compustat Merged Database, while equity data for the remaining countries is obtained from Refinitiv Datastream. The preprocessing methodology, as outlined in Nechvátalová (2024), addresses the challenges of working with international data from Datastream.

Both datasets share the same set of 153 firm characteristics, expanding the analysis beyond the original U.S. GKX dataset and enabling us to test whether the CAE model generalises to new firm characteristics and across international markets. Incorporating some overlapping firm characteristics while also introducing additional variables, these datasets include 93 fundamental, 43 market friction, and 11 I/B/E/S

[1] Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom, Japan, Australia, New Zealand, Hong Kong, and Singapore.

characteristics. To ensure consistency between the U.S. and international datasets, quarterly fundamental characteristics are excluded due to their limited availability internationally, with fundamental data relying solely on annual observations. For comprehensive implementation details, refer to Tobek and Hronec (2021).

Liquidity filters

For the U.S. GKX dataset in the full sample, no additional filtering is applied to ensure consistency with the dataset used by Gu et al. (2021). In contrast, for the extended U.S. and international datasets, the full sample excludes observations with a price below 1(0.10 for the Asia Pacific) at the end of the previous month. In addition, for CRSP data in these datasets, we retain only common stocks, identified by share codes 10 and 11, while the GKX dataset includes all share codes.

To obtain the liquid subsample for all three datasets, we apply a liquidity filter that excludes microcap stocks, removing small, illiquid firms that are costly or impractical to trade. We sort firms by market capitalisation each month and exclude the smallest firms until the excluded firms' combined market capitalisation equals 5% of the region's total. Similarly, firms with the lowest trading volumes are excluded until their combined volume equals 5% of the region's total. Firms lacking trading volume data are excluded if they rank in the lowest 10% by market capitalisation. Non-U.S. stocks must have a market capitalisation exceeding the lowest decile of NYSE firms for the given month to ensure comparability with U.S. stocks. For CRSP data, we retain only common shares. Firms in the liquid subsample must meet a price threshold of $1 at the end of the previous month, or $0.10 for Asia Pacific stocks.

For all three datasets, both the full and liquid samples undergo the same preprocessing steps to ensure consistency and mitigate the impact of outliers. More specifically, missing firm characteristics are replaced by the cross-sectional median of that characteristic within the region. To stabilise learning and reduce the influence of outliers, we cross-sectionally rank-normalise all characteristics to lie within the range (−1, 1) for each region. Additionally, excess returns are clipped at −3 and 3 to further limit the impact of severe outliers.

Table 1 provides descriptive statistics for the three datasets, presenting both the full and liquid samples. The full sample of the U.S. GKX dataset includes an average of over 6900 firms per month, while the liquid subsample averages around 1200 firms per month. The extended U.S. full sample includes around 5300 firms per month, due to additional filters on price and share type not used in the GKX dataset. The liquid subsample for the extended U.S. dataset contains the same firm-month observations as the GKX dataset because the same filtering criteria were applied. Internationally, the full dataset averages nearly 17,000 firms per month, which decreases to about 3000 firms in the liquid subsample.

2.3. Model estimation

To estimate our model, we divide our dataset into training, validation and testing datasets while maintaining the temporal ordering of the data. We train the model with various hyperparameter configurations on the training set and evaluate them on the validation set. The best set of hyperparameters is based on the mean square error over the validation sample. The testing sample is used to obtain out-of-sample predictions on previously unseen data.

To avoid overfitting, we follow Gu et al. (2021) in using l_1 LASSO penalisation and taking an ensemble of the ten best models based on the validation sample. The objective function to optimise the parameters of the model is:

$$\mathcal{L}(\theta;\cdot) = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\|r_{i,t} - \beta'_{i,t-1}f_t\right\|^2 + \lambda\sum_j|\theta_j| \tag{9}$$

where the second part of the objective function is the regularisation term, applying the LASSO penalty to the network's weights. The regularisation strength parameter λ is determined using the validation sample during hyperparameter tuning. We implement early stopping to terminate training if the validation loss increases, preventing overfitting. The Adam algorithm (Kingma & Ba, 2014), an extension of stochastic gradient descent with adaptive learning rates, is used for optimisation. Additionally, batch normalisation (Ioffe & Szegedy, 2015) is applied to each hidden layer to stabilise and accelerate training.

In our analysis, we implement an expanding window approach, sequentially training and evaluating our models over various time periods. This method allows us to obtain more precise estimates by incorporating more recent data. For the training of our first model, we divide the dataset into a 20-year training period (1957–1976), a 10-year validation period (1977–1986), and a 31-year out-of-sample testing period (1987–2018). To reduce computational costs, we refit the model every five years, extending the training sample by five years each time while keeping the validation sample size fixed.

In our analysis, we keep the number of latent factors $K = 3$. We perform a hyperparameter search to select a suitable model. The number of hidden layers on the factor loadings side is either one, two, or three, with 32 neurons in the first hidden layer, 16 in the second, and 8 in the third. The learning rates checked were 0.01, 0.001, 0.0001, and 0.00001. The strength of the LASSO regularisation λ can be 0, 0.0001, 0.00001 or 0.000001. The batch size is either 64 or 128.

3. Empirical results

3.1. Statistical performance

To evaluate the statistical performance of the model on the out-of-sample testing data, we use the total and predictive R^2 metrics proposed by Kelly et al. (2019).

The total R^2 measures how well the model explains contemporaneous factor realisations, reflecting its ability to capture individual stock riskiness. It is defined as:

$$R^2_{\text{total}} = 1 - \frac{\sum_{(i,t)\in\text{OOS}}(r_{i,t} - \hat{\beta}'_{i,t-1}\hat{f}_t)^2}{\sum_{(i,t)\in\text{OOS}} r^2_{i,t}}, \tag{10}$$

where OOS denotes the out-of-sample testing subsample for a given model, comprising data not used for training or hyperparameter selection.

In contrast, the predictive R^2 evaluates the model's forecasting ability, quantifying its accuracy in predicting future individual excess stock returns and capturing variations in risk compensation across the panel. This metric is defined as:

$$R^2_{\text{pred}} = 1 - \frac{\sum_{(i,t)\in\text{OOS}}(r_{i,t} - \hat{\beta}'_{i,t-1}\hat{\lambda}_{t-1})^2}{\sum_{(i,t)\in\text{OOS}} r^2_{i,t}}, \tag{11}$$

where $\hat{\lambda}_{t-1}$ is the mean of the estimated latent factor $\hat{f}$ up to month $t-1$.

Importantly, while R^2_{total} focuses on the contemporaneous explanatory power of the model, R^2_{pred} explicitly relies on factor exposures estimated with data available only up to $t-1$. This ensures that predictions are based solely on prior information, making R^2_{pred} a robust measure of true predictive performance.

Table 2 presents the total and predictive R^2 values for the U.S. GKX, U.S. extended, and international datasets using the elastic net approach to obtain characteristic-managed portfolios x_t and three latent factors K. These values are calculated on the out-of-sample testing datasets, comprising the years 1987–2018 for the two U.S. datasets and 1995–2018 for the international sample.

Using the same dataset and methodological framework as Gu et al. (2021), we obtain a total R^2 of 13.27% and a predictive R^2 of 0.21% for the full-sample U.S. GKX dataset under our default five-year retraining schedule. While these results reflect our best effort to replicate their approach, the R^2 values differ from those reported by Gu et al. (2021), who find a slightly lower total R^2 (12.5%) but more than double the

Table 1
Descriptive statistics.

Statistic	Full sample				Liquid sample			
	Ex. Return	MC	# Firms	TC	Ex. Return	MC	# Firms	TC
Panel A: U.S. dataset (GKX)								
Mean	0.71	2002.10	6925	2.75	0.69	9044.87	1181	0.66
Std	18.43	11 710.91	1057	4.44	12.12	26 949.41	228	0.90
25%	−6.39	36.12	5801	0.28	−5.03	1026.31	977	0.06
50%	−0.09	149.23	6786	1.31	0.66	2491.60	1138	0.26
75%	6.24	685.98	7574	3.94	6.37	6406.05	1275	0.99
Panel B: U.S. dataset (extended)								
Mean	0.77	2207.98	5248	3.05	0.69	9044.87	1181	0.66
Std	19.66	13 307.68	1206	4.64	12.12	26 949.41	228	0.90
25%	−7.21	31.54	3998	0.34	−5.03	1026.31	977	0.06
50%	−0.29	136.14	5375	1.61	0.66	2491.60	1138	0.26
75%	6.97	707.38	6206	4.57	6.37	6406.05	1275	0.99
Panel C: International dataset								
Mean	0.83	1615.32	16 912	3.81	0.45	7704.26	2991	0.76
Std	20.67	9919.25	1214	6.45	11.80	21 668.82	436	1.09
25%	−6.65	25.07	16 418	0.71	−5.47	937.42	2645	0.15
50%	−0.30	99.77	17 093	2.09	0.29	2230.28	2997	0.39
75%	6.12	478.40	17 832	5.00	6.07	5855.19	3207	0.86

Panel A reports statistics for the U.S. GKX dataset, Panel B for the extended U.S. dataset, and Panel C for the international dataset, which includes the U.S. and 22 developed countries. The table reports the monthly mean, standard deviation, and the 25th, 50th, and 75th percentiles for excess returns (in %), market capitalisation (in millions of dollars), number of firms per month, and estimated transaction costs (in %). Transaction costs are estimated via closing quoted spread (Chung & Zhang, 2014), with missing values filled using the volatility-over-volume method (Fong et al., 2018) and a 5% rate applied to any residual missing values. Statistics are reported separately for the full sample (left) and the liquid subsample (right).

predictive R^2 (0.52%). By comparison, Yang et al. (2024), who also adopt a five-year retraining frequency and directly replicate the CAE model as a benchmark for their own work, report a total R^2 of 11.49% and a predictive R^2 of 0.28%, with the latter more closely matching our results.

To account for the remaining differences in predictive R^2 relative to Gu et al. (2021), we examine several aspects of the implementation: retraining frequency, sample period length, and hyperparameter selection. To isolate the impact of retraining frequency, we re-estimate the model using annual retraining, as in Gu et al. (2021). This adjustment raises the predictive R^2 for the U.S. GKX full sample (elastic net) to 0.31%, narrowing the gap to their reported figure, while the total R^2 remains largely unchanged (see Table A.7). We also evaluate portfolio performance under annual retraining and find it closely mirrors the five-year case; hence, we report portfolio results only for the five-year specification in the sections that follow. Furthermore, these differences cannot be explained by our longer out-of-sample evaluation period: results remain nearly identical when restricting the evaluation to end in 2016 (see Tables A.6 and A.7). A likely remaining source of difference is that Gu et al. (2021) do not report their hyperparameter settings, leaving room for implementation variation. In contrast, Yang et al. (2024) report theirs in full, allowing us to match their setup more closely, and ultimately obtain results that are much closer to theirs.

In the U.S. GKX dataset, the predictive R^2 is slightly higher for the liquid subsample (0.27%) compared to the full sample (0.21%). In contrast, the U.S. extended dataset delivers higher predictive R^2 values, recording 0.54% for the full sample and 0.38% for the liquid subsample. The full sample value aligns closely with that reported in Gu et al. (2021). These results show that the extended dataset's broader set of characteristics enhances predictability across both liquid and full samples. The international dataset shows predictive R^2 values of 0.64% for the full sample and 0.25% for the liquid subsample. These results indicate the model's ability to capture systematic patterns in returns across diverse markets, though the differences in R^2 can reflect variations in market structure, stock characteristics, and dataset composition rather than intrinsic differences in model performance.

The total R^2 values for the liquid subsamples are roughly double those of the full samples across all three datasets, indicating stronger systematic factor relationships and reduced noise in more liquid stocks. However, when looking at predictive R^2, the pattern differs: in the U.S. extended and international datasets, predictive R^2 is lower for the liquid subsample compared to the full sample. The U.S. GKX dataset stands out, with predictive R^2 slightly higher in the liquid subsample, potentially due to its broader share coverage and smaller set of firm characteristics compared to the other two datasets. This pattern in the U.S. extended and international datasets suggests that liquid stocks, being more homogenous and less volatile in terms of characteristics, offer fewer cross-sectional signals for monthly return prediction.

While predictive R^2 values in our setting may appear modest, this is typical in the literature on the cross-section of expected returns, where predictive R^2 values tend to be very small due to the high level of noise in monthly stock returns. Moreover, predictive R^2 captures average squared error across all assets but does not directly reflect the model's ability to rank expected returns, a capability that is more relevant for portfolio construction. As such, even when point predictions are noisy, the model may still identify profitable opportunities if it gets the relative ordering of returns right. We observe that even with low predictive R^2 values, the model produces economically relevant portfolio outcomes, as shown in the next section. Supporting this perspective, Kelly et al. (2024) show in a simpler single-asset market timing context that predictive R^2 can be a misleading proxy for economic value: models may yield substantial timing benefits even when R^2 is close to zero or negative, due to the distorting effect of forecast variance on the statistic.

In Table A.6, we compare results from the OLS and Elastic Net models, as well as from an evaluation period ending in 2016. OLS performs well only in the U.S. GKX full sample – the original setting of Gu et al. (2021) – but breaks down elsewhere. In the GKX liquid subsample, instability arises from collinearity driven by the smaller cross-section and overlap in the rank-normalised values of beta and beta squared. In the extended and international datasets, a similar issue emerges due to the larger characteristic set (153 variables), which increases the likelihood of multicollinearity. In all cases, this leads to extreme values of x_t, which propagate through the model and distort the latent factor estimates, resulting in poor overall fit under OLS. Elastic Net mitigates these issues and consistently improves performance.

We also assess the sensitivity of model performance to hyperparameter settings and ensemble size. Results in Tables A.8 and A.9, based

Table 2
Out-of-Sample R^2 results.

Dataset	Sample	R^2_{pred}	R^2_{total}
U.S. (GKX)	Full	0.21	13.27
	Liquid	0.27	26.27
U.S. (extended)	Full	0.54	13.37
	Liquid	0.38	24.95
International	Full	0.64	8.88
	Liquid	0.25	19.68

This table reports R^2_{total} and R^2_{pred} values obtained from the elastic net model for individual stocks. The results are based on the out-of-sample period spanning 1987 to 2018 for both the U.S. (GKX) and U.S. (extended) datasets and 1995 to 2018 for International data. Results are presented separately for the full and liquid subsamples. All values are expressed in percentages.

on the U.S. GKX liquid dataset (first model from the expanding-window procedure), confirm that predictive and total R^2 remain stable across a range of well-performing configurations. Ensemble averaging further improves and stabilises predictive accuracy, with gains levelling off beyond 10 models. Comparable results were obtained across subsequent time splits and datasets, supporting the generalisability and reliability of our findings.

3.2. Economic performance

To evaluate the economic performance of our models, we construct long-short decile portfolios based on out-of-sample predicted returns obtained from the elastic net model with three latent factors.

In addition to long-short portfolios, we separately analyse the performance of the long-only and short-only sides of the portfolio. This provides detailed insights into the performance of each leg of the portfolio and is relevant for scenarios with shorting constraints. We report results for both equal-weighted and value-weighted portfolios.

To estimate transaction costs, we utilise the closing quoted spread as described by Chung and Zhang (2014). Missing observations are filled using the volatility-over-volume method proposed by Fong et al. (2018), with any remaining gaps assigned a transaction cost rate of 5%. Portfolio returns with transaction costs are calculated iteratively to account for these estimated firm-month-specific transaction costs (Nechvátalová, 2024). Portfolio turnover is defined as the sum of absolute trade sizes divided by the gross exposure. For instance, a turnover rate of 200% in a given month indicates that all positions were closed and replaced with new ones.

The performance evaluation focuses on monthly mean returns, annualised Sharpe ratios, and turnover rates. Results are reported for both the full and liquid subsamples, with and without transaction costs, across three datasets: U.S. GKX, U.S. extended, and international.

3.2.1. Evidence from the U.S. GKX dataset

The U.S. GKX out-of-sample evaluation period spans from 1987 to 2018 and includes U.S. stocks with 94 firm characteristics. Table 3 presents the performance metrics for the U.S. GKX dataset. Without transaction costs, long-short portfolios on the full sample achieve monthly mean returns of 3.49% and 2.19%, with Sharpe ratios of 2.22 and 1.16 for the equal-weighted and value-weighted portfolios, respectively. However, if we account for transaction costs, these portfolios are not profitable. As we restrict our portfolio to the liquid subsample, this reduces returns to 1.28% (equal-weighted) and 1.15% (value-weighted). After including transaction costs, the portfolios on the liquid subsample remain marginally profitable, with Sharpe ratios of 0.25 (equal-weighted) and 0.33 (value-weighted).

The limited profitability of the long-short portfolio is primarily driven by the underperformance of the short side, particularly after transaction costs. As a result, only long-only portfolios on the liquid subsample remain profitable. After transaction costs, the equal-weighted portfolio achieves a mean monthly return of 1% and a Sharpe ratio of 0.7, while the value-weighted portfolio delivers a mean monthly return of 0.89% and a Sharpe ratio of 0.67. For comparison, a buy-and-hold S&P strategy yields a mean monthly return of 0.64% and a Sharpe ratio of 0.53.

The turnover of the value-weighted long-short strategy trained on the full sample is 150% when tested on the full sample and 145% on the liquid sample, consistent with Avramov et al. (2023). In contrast, the model trained and tested on the liquid subsample has a turnover of 87%, significantly lower than the full-sample-trained model. This difference suggests that the full-sample-trained model capitalises on short-term opportunities in smaller, less liquid firms, while the liquid-sample-trained model emphasises more stable, persistent return patterns, reducing the need for frequent rebalancing.

The results evaluated on the full sample align with those of Gu et al. (2021) for both equal-weighted and value-weighted portfolios, demonstrating consistency in model performance. In contrast, Avramov et al. (2023) report significantly worse full-sample metrics for value-weighted portfolios of CAE with two hidden layers and five latent factors. Their reported Sharpe ratio of 0.78 is considerably lower than the 1.45 reported by Gu et al. (2021) for a comparable model, and their mean returns are one percentage point lower than those achieved by our model with only three latent factors. Despite these differences, Avramov et al. (2023) also observe the common pattern of decreasing profitability in more liquid samples without accounting for transaction costs.

3.2.2. Evidence from the U.S. extended dataset

Table 4 presents the portfolio performance metrics for the U.S. extended dataset. The out-of-sample evaluation period is the same as for the GKX dataset (1987–2018) but includes a broader set of 153 firm characteristics, offering an alternative dataset to evaluate model performance within the U.S. market. Returns and Sharpe ratios are higher compared to the GKX dataset for both equal-weighted and value-weighted portfolios, as well as for long-short portfolios and their individual long and short legs. Using the extended dataset, the portfolios outperform the S&P buy-and-hold benchmark, achieving a Sharpe ratio of 0.60 for long-short and 0.77 for long-only value-weighted portfolios after accounting for transaction costs. The short side of the portfolio performs slightly better compared to the GKX dataset, resulting in a smaller gap between the long-short and long-only portfolios.

Overall, the results show a similar pattern to the U.S. GKX dataset but with stronger performance across most portfolios. This improved performance could be attributed to the larger and more diverse set of firm characteristics used to train the model, which provides richer information and reduces noise.

3.2.3. Evidence from the international dataset

Table 5 presents portfolio performance metrics for the international dataset, which includes the U.S. and 22 other developed countries with 153 firm characteristics. The out-of-sample evaluation period spans from 1995 to 2018. Overall, performance aligns closely with the two U.S. datasets. Although the Sharpe ratios of portfolios are higher before transaction costs, the greater impact of transaction costs results in after-cost performance similar to that of the U.S. data. The short sides of the portfolios perform better than in the U.S., enabling long-short portfolios to achieve Sharpe ratios of 0.73 (equal-weighted) and 0.54 (value-weighted). Unlike the U.S. cases, where long-short strategies often underperformed benchmarks, these portfolios outperform the buy-and-hold strategy of the MSCI World Index, which delivers a monthly mean return of 0.45% and a Sharpe ratio of 0.37 over the same period. Long-only portfolios show comparable performance, with Sharpe ratios of 0.67 for both equal-weighted and value-weighted portfolios, narrowing the gap between the long-short and long-only strategies.

In the case of the international dataset, training the model on the liquid subsample is better than on the full sample, particularly after accounting for transaction costs. This is likely due to the larger number

Table 3
Performance metrics for Equal and Value weight portfolios in the U.S. dataset (GKX).

Training	Testing	Without transaction costs			With transaction costs			Turnover
		Mean	Std	Sharpe	Mean	Std	Sharpe	
Panel A: Long-Short (Equal-weight)								
Full	Full	3.49	5.46	2.22	−2.12	5.47	−1.34	134.08
Full	Liquid	1.28	6.81	0.65	0.48	6.75	0.25	131.39
Liquid	Liquid	0.82	8.27	0.34	0.49	8.25	0.21	81.71
Panel B: Long side (Equal-weight)								
Full	Full	2.79	9.02	1.07	−0.38	8.61	−0.15	138.34
Full	Liquid	1.41	5.70	0.86	0.96	5.66	0.59	143.22
Liquid	Liquid	1.16	4.88	0.83	0.99	4.87	0.70	88.50
Panel C: Short side (Equal-weight)								
Full	Full	0.70	8.36	0.29	−1.74	8.27	−0.73	129.83
Full	Liquid	−0.13	8.96	−0.05	−0.48	8.92	−0.18	119.56
Liquid	Liquid	−0.34	9.52	−0.12	−0.50	9.50	−0.18	74.91
Panel D: Long-Short (Value-weight)								
Full	Full	2.19	6.58	1.16	−1.06	6.49	−0.57	149.80
Full	Liquid	1.15	6.16	0.65	0.57	6.13	0.33	145.23
Liquid	Liquid	0.64	7.63	0.29	0.44	7.62	0.20	86.65
Panel E: Long side (Value-weight)								
Full	Full	1.92	7.31	0.91	0.21	7.15	0.10	165.00
Full	Liquid	1.35	5.40	0.87	1.01	5.38	0.65	157.21
Liquid	Liquid	1.01	4.62	0.76	0.89	4.61	0.67	94.48
Panel F: Short side (Value-weight)								
Full	Full	0.27	7.75	0.12	−1.28	7.67	−0.58	134.60
Full	Liquid	−0.20	7.82	−0.09	−0.43	7.80	−0.19	133.24
Liquid	Liquid	−0.37	8.79	−0.14	−0.46	8.77	−0.18	78.82

This table presents monthly mean returns, standard deviations, annualised Sharpe ratios, and turnover rates for the U.S. GKX dataset from 1987 to 2018. Results are shown for equal-weighted (Panels A, B, C) and value-weighted portfolios (Panels D, E, F), including long-short, long-only, and short-only decile portfolios, both with and without transaction costs. The Training and Testing columns specify whether models are trained and evaluated on the full or liquid subsamples. All values are expressed as percentages, except for the Sharpe ratio.

Table 4
Performance metrics for Equal and Value weight portfolios in the extended U.S. dataset.

Training	Testing	Without transaction costs			With transaction costs			Turnover
		Mean	Std	Sharpe	Mean	Std	Sharpe	
Panel A: Long-Short (Equal-weight)								
Full	Full	4.37	6.15	2.46	−0.74	5.85	−0.44	137.41
Full	Liquid	1.50	5.23	0.99	0.71	5.17	0.47	134.84
Liquid	Liquid	1.29	6.29	0.71	0.70	6.24	0.39	106.22
Panel B: Long side (Equal-weight)								
Full	Full	3.26	9.88	1.14	0.32	9.44	0.12	137.18
Full	Liquid	1.56	6.44	0.84	1.15	6.40	0.62	141.25
Liquid	Liquid	1.56	5.39	1.00	1.21	5.36	0.78	114.19
Panel C: Short side (Equal-weight)								
Full	Full	1.11	7.89	0.49	−1.05	7.78	−0.47	137.64
Full	Liquid	−0.06	8.07	−0.03	−0.44	8.03	−0.19	128.43
Liquid	Liquid	−0.27	8.43	−0.11	−0.52	8.40	−0.21	98.25
Panel D: Long-Short (Value-weight)								
Full	Full	2.43	6.68	1.26	0.50	6.42	0.27	154.17
Full	Liquid	1.44	4.96	1.01	0.85	4.89	0.60	144.52
Liquid	Liquid	1.09	5.77	0.65	0.67	5.70	0.41	110.84
Panel E: Long side (Value-weight)								
Full	Full	1.99	8.51	0.81	0.87	8.35	0.36	161.20
Full	Liquid	1.44	5.92	0.84	1.12	5.90	0.66	150.10
Liquid	Liquid	1.41	5.15	0.95	1.14	5.13	0.77	117.84
Panel F: Short side (Value-weight)								
Full	Full	0.44	7.57	0.20	−0.37	7.48	−0.17	147.15
Full	Liquid	0.00	6.87	0.00	−0.27	6.83	−0.14	138.95
Liquid	Liquid	−0.32	7.19	−0.15	−0.47	7.17	−0.23	103.84

This table presents monthly mean returns, standard deviations, annualised Sharpe ratios, and turnover rates for the U.S. extended dataset from 1987 to 2018. Results are shown for equal-weighted (Panels A, B, C) and value-weighted portfolios (Panels D, E, F), including long-short, long-only, and short-only decile portfolios, both with and without transaction costs. The Training and Testing columns specify whether models are trained and evaluated on the full or liquid subsamples. All values are expressed as percentages, except for the Sharpe ratio.

Table 5
Performance metrics for Equal and Value weight portfolios in the International dataset.

Training	Testing	Without transaction costs			With transaction costs			Turnover
		Mean	Std	Sharpe	Mean	Std	Sharpe	
Panel A: Long-Short (Equal-weight)								
Full	Full	4.83	3.94	4.26	−1.90	3.73	−1.76	131.19
Full	Liquid	1.64	4.31	1.32	0.74	4.30	0.59	129.59
Liquid	Liquid	1.57	4.27	1.27	0.89	4.23	0.73	113.45
Panel B: Long side (Equal-weight)								
Full	Full	3.70	7.19	1.78	−0.27	6.83	−0.14	131.35
Full	Liquid	1.34	6.27	0.74	0.92	6.25	0.51	134.96
Liquid	Liquid	1.41	5.47	0.89	1.06	5.46	0.67	119.53
Panel C: Short side (Equal-weight)								
Full	Full	1.13	5.90	0.67	−1.63	5.65	−1.00	131.03
Full	Liquid	0.31	6.28	0.17	−0.18	6.23	−0.10	124.22
Liquid	Liquid	0.15	6.49	0.08	−0.17	6.46	−0.09	107.38
Panel D: Long-Short (Value-weight)								
Full	Full	2.07	4.66	1.54	0.32	4.66	0.24	149.82
Full	Liquid	1.33	4.54	1.01	0.58	4.59	0.44	139.09
Liquid	Liquid	1.16	4.39	0.92	0.68	4.36	0.54	116.46
Panel E: Long side (Value-weight)								
Full	Full	1.58	6.94	0.79	0.75	6.91	0.38	161.87
Full	Liquid	1.22	6.05	0.70	0.90	6.04	0.51	144.87
Liquid	Liquid	1.27	5.19	0.85	1.00	5.18	0.67	119.84
Panel F: Short side (Value-weight)								
Full	Full	0.49	6.12	0.28	−0.43	6.04	−0.25	137.77
Full	Liquid	0.11	5.95	0.06	−0.31	5.92	−0.18	133.31
Liquid	Liquid	−0.11	5.94	−0.06	−0.32	5.92	−0.19	113.09

This table presents monthly mean returns, standard deviations, annualised Sharpe ratios, and turnover rates for the international dataset from 1995 to 2018. Results are shown for equal-weighted (Panels A, B, C) and value-weighted portfolios (Panels D, E, F), including long-short, long-only, and short-only decile portfolios, both with and without transaction costs. The Training and Testing columns specify whether models are trained and evaluated on the full or liquid subsamples. All values are expressed as percentages, except for the Sharpe ratio.

of observations for the liquid subsample, which provides more robust opportunities for extracting persistent return patterns.

While the CAE model achieves strong predictive R^2 values, its economic performance is less favourable compared to alternative approaches. Gu et al. (2020) employ a feedforward neural network to directly predict returns from firm characteristics, prioritising predictive accuracy. In contrast, the CAE model introduces latent factors and imposes a no-arbitrage condition, aligning with economic theory but potentially limiting its economic performance due to its design. As reported by Tobek and Hronec (2021), who apply the same model to international datasets, the long-short portfolios trained on the liquid subsample achieve Sharpe ratios of 0.74 (equal-weighted) and 0.94 (value-weighted) in the U.S. market after transaction costs. Internationally, the model performs even better, with Sharpe ratios of 0.95 for both equal-weighted and value-weighted portfolios, significantly exceeding the CAE model's performance in similar configurations. These results highlight the trade-offs between adhering to economic theory and prioritising predictive accuracy, with practical implications for asset pricing model selection.

4. Conclusion

This study evaluates the CAE model across three distinct datasets: the original U.S. sample, a U.S. dataset with a broader set of firm characteristics, and an international sample of developed markets. We assess the model's robustness and its practical relevance under realistic trading constraints.

The extended U.S. dataset, with its broader set of firm characteristics, yields improved statistical and economic performance compared to the original GKX sample. These results are comparable to those reported by Gu et al. (2021). Internationally, the model performs similarly well, with more balanced long-short performance driven by stronger returns on the short side. Although the exclusion of microcap and illiquid firms lowers raw portfolio returns, these strategies yield better net performance after accounting for transaction costs, underscoring the value of focusing on investments that are more cost-efficient to execute.

Overall, these results demonstrate that the model's statistical and economic performance is stable across varying data environments. At the same time, the economic performance of CAE-based portfolios is strongly affected by transaction costs and liquidity constraints. Accordingly, careful consideration of trading constraints is essential when applying machine learning models in asset pricing.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used ChatGPT in order to improve the readability and language of the manuscript. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Appendix. Supplementary tables

See Tables A.6–A.9.

Data availability

The authors do not have permission to share data.

Table A.6
Extended Out-of-Sample R^2 results.

Dataset	Sample	Full period		Up to 2016	
		R^2_{pred}	R^2_{total}	R^2_{pred}	R^2_{total}
Panel A: OLS model					
U.S. (GKX)	Full	0.20	12.73	0.24	12.96
	Liquid	−2.13e+23	−2.51e+27	−1.96e+23	−2.57e+27
U.S. (extended)	Full	0.56	−4.13	0.58	13.52
	Liquid	0.40	−37.13	0.39	24.81
International	Full	0.64	−31.11	0.63	9.09
	Liquid	0.28	−270.72	0.28	19.71
Panel B: Elastic Net model					
U.S. (GKX)	Full	0.21	13.27	0.24	13.52
	Liquid	0.27	26.27	0.26	26.49
U.S. (extended)	Full	0.54	13.37	0.56	13.57
	Liquid	0.38	24.95	0.37	25.16
International	Full	0.64	8.88	0.63	9.13
	Liquid	0.25	19.68	0.25	19.93

The reported R^2_{total} and R^2_{pred} values for individual stocks are based on out-of-sample periods spanning 1987 to 2018 for both the U.S. GKX and U.S. extended datasets, and 1995 to 2018 for International data. Results are presented separately for the full and liquid subsamples. Values for the full period are shown in the left columns, while the right columns display results for an additional evaluation period ending in 2016, chosen for comparability with Gu et al. (2021). All values are expressed in percentages.

Table A.7
Out-of-Sample R^2 results with increased retraining frequency (Annual).

Dataset	Full period		Up to 2016	
	R^2_{pred}	R^2_{total}	R^2_{pred}	R^2_{total}
U.S. (GKX), full sample, OLS	0.19	12.72	0.22	12.96
U.S. (GKX), full sample, elastic net	0.31	13.29	0.34	13.53
U.S. (GKX), liquid sample, elastic net	0.29	26.46	0.29	26.69
U.S. (extended), liquid sample, elastic net	0.43	25.29	0.42	25.51

Reported R^2_{total} and R^2_{pred} values for individual stocks are based on out-of-sample periods from 1987 to 2018. The left columns show results for the full period, while the right columns cover an evaluation period ending in 2016. All results shown here use annual model retraining, enabling more frequent updates than the five-year baseline. Values are expressed in percentages.

Table A.8
Sensitivity of CAE model performance to hyperparameter settings.

lr	λ	Layers	Batch size	Validation loss	Test r^2_{pred}	Test r^2_{total}
0.001	1e−06	32.16.8	64	0.007703	0.172	24.005
0.01	1e−06	32.16	128	0.007711	0.085	24.535
0.001	1e−06	32	64	0.007711	0.213	24.166
0.001	1e−06	32	128	0.007715	0.221	23.702
0.001	1e−06	32.16.8	128	0.007716	0.213	23.824
0.01	1e−06	32	64	0.007719	0.004	24.346
0.001	1e−05	32	128	0.007720	0.147	23.757
0.001	1e−05	32	64	0.007722	0.035	23.611
0.001	1e−05	32.16	64	0.007722	0.197	23.436
0.01	1e−06	32.16.8	64	0.007724	0.211	24.450
...	...	...	...	...	...	...
1e−05	0.001	32.16.8	128	0.010817	0.124	0.124
1e−05	0.001	32.16	128	0.010817	0.124	0.124
0.001	0.001	32	128	0.010820	0.080	0.080
0.001	0.001	32	64	0.010820	0.070	0.070
0.01	0.001	32	64	0.010823	0.062	0.062

This table reports results for the CAE model trained on the U.S. GKX liquid dataset using the first data split: validation spans ten years ending in 1986, and testing covers 1987–2018. Characteristic-managed portfolios are estimated using the elastic net approach. Each row corresponds to a specific configuration of learning rate (lr), LASSO penalty (λ), network architecture, and batch size. Validation loss and out-of-sample R^2 values (in percentages) are reported. Only the best and worst configurations are shown.

Table A.9
Performance of ensemble models with varying ensemble sizes.

Ensemble size	Validation loss	Test r^2_{pred}	Test r^2_{total}
1	0.007800	0.172	24.005
2	0.007785	0.176	24.608
3	0.007735	0.253	24.979
4	0.007731	0.269	24.852
5	0.007727	0.271	24.807
6	0.007717	0.282	24.982
7	0.007716	0.293	24.939
8	0.007718	0.291	24.878
9	0.007724	0.299	24.797
10	0.007727	0.300	24.817
11	0.007725	0.299	24.910
12	0.007729	0.300	24.952
13	0.007729	0.298	24.960
14	0.007731	0.296	24.957
15	0.007729	0.300	24.938
...	...	...	...
25	0.007734	0.283	24.815

This table reports validation loss and test-set R^2 values for ensemble models trained on the U.S. GKX liquid dataset using the first data split (validation spans ten years ending in 1986, testing spans 1987–2018). Each ensemble consists of the top-performing models (based on validation loss), with the ensemble size indicating the number of models combined.

References

Andrew Karolyi, G. (2016). Home bias, an academic puzzle. *Review of Finance*, *20*(6), 2049–2078.

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

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

Chen, L., Pelger, M., & Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, *70*(2), 714–750.

Chung, K. H., & Zhang, H. (2014). A simple approximation of intraday spreads using daily data. *Journal of Financial Markets*, *17*, 94–120.

DeMiguel, V., Martin-Utrera, A., Nogales, F. J., & Uppal, R. (2020). A transaction-cost perspective on the multitude of firm characteristics. *The Review of Financial Studies*, *33*(5), 2180–2222.

Fong, K. Y., Holden, C. W., & Tobek, O. (2018). Are volatility over volume liquidity proxies useful for global or US research? *Kelley School of Business Research Paper*, (17–49).

Giglio, S., Kelly, B., & Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, *14*(1), 337–368.

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

Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, *222*(1), 429–450.

Hanauer, M. X., & Kalsbach, T. (2023). Machine learning and the cross-section of emerging market stock returns. *Emerging Markets Review*, *55*, Article 101022.

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

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Jiang, J., Kelly, B., & Xiu, D. (2023). (Re-) imag (in) ing price trends. *The Journal of Finance*, *78*(6), 3193–3249.

Kelly, B., Malamud, S., & Zhou, K. (2024). The virtue of complexity in return prediction. *The Journal of Finance*, *79*(1), 459–503.

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

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

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

Nechvátalová, L. (2024). Multi-horizon equity returns predictability via machine learning. *Czech Journal of Economics and Finance (Finance a Úvěr)*, *74*(2), 142–190.

Novy-Marx, R., & Velikov, M. (2019). Comparing cost-mitigation techniques. *Financial Analysts Journal*, *75*(1), 85–102.

Tobek, O., & Hronec, M. (2021). Does it pay to follow anomalies research? machine learning approach with international evidence. *Journal of Financial Markets*, *56*, Article 100588.

Yang, X., Zhu, Z., Li, D., & Zhu, K. (2024). Asset pricing via the conditional quantile variational autoencoder. *Journal of Business & Economic Statistics*, *42*(2), 681–694.