improve over LR. Additionally, all models significantly outperform EN, except for NN. In contrast, the last column shows DM test statistics ranging from 2.07 to 13.28, showing that DMT significantly outperforms all other models. The second last column indicates that MT significantly outperforms EN and NN, but is significantly outperformed by RF and XRF, illustrating that tree-based models can perform well at forecasting probabilities.

4.3 Multi-factor Timing

Next, we evaluate the economic significance of factor timing strategies derived from each model's forecasts, using a multi-factor portfolio. Motivated by Campbell and Thompson (2008), we apply realistic constraints that restrict short-selling and leverage. For every factor, we take a long position if the model forecasts a positive excess return, otherwise we invest at the risk-free rate, yielding zero excess returns. The strategy return for factor i at time $t+1$ is expressed as

$$r_{i,t+1}^{Strategy} = I(\hat{\pi}_{i,t+1} > 0.5) r_{i,t+1}, \tag{3}$$

where $I(\cdot)$ is an indicator function that takes a value of one if the predicted probability exceeds 50% and zero otherwise. Our multi-factor strategy is then an equal-weighted portfolio of the strategy excess returns across all factors. Motivated by Moreira and Muir (2017), for each model we conduct a time series regression of the strategy on the multi-factor Buy that is an equal-weighted average of the excess returns across all factors. This spanning regression provides the strategy's alpha, beta, and R^2 relative to the multi-factor Buy.

The first row of Table 3 presents the annualized Sharpe ratio for each model, which aligns very closely with the results on the average accuracy reflected in the first row of Table 1. The linear models, LR and EN, yield Sharpe ratios of 0.84 and 0.83, respectively, underperforming the multi-factor Buy Sharpe ratio of 0.98. However, improvements over linear models are seen with the incorporation of nonlinear interactions in RF, XRF, GBT, SVM models, resulting in Sharpe ratios of 1.07, 1.01, 0.88, and 0.88 respectively. Notably, RF and XRF are the only off-the-shelf models to achieve a higher Sharpe ratio than the multi-factor Buy.

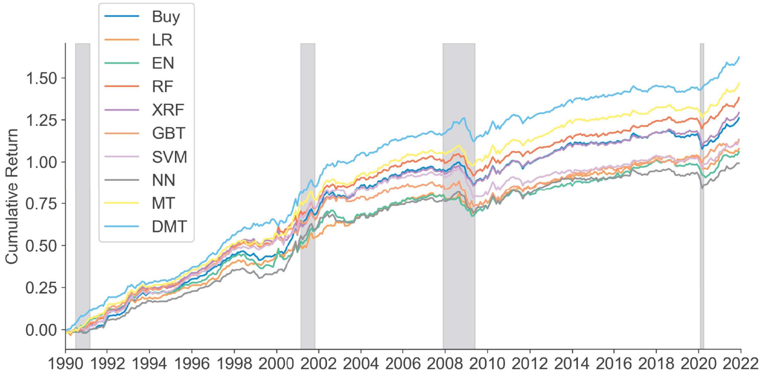
Turning to deep learning models, NN posts the lowest Sharpe ratio of 0.78, due to its

Table 3: Multi-factor Timing Strategy Performance

	Buy	LR	EN	RF	XRF	GBT	SVM	NN	MT	DMT
SR	0.98	0.84	0.83	1.07	1.01	0.88	0.88	0.78	1.14	1.26
α		0.56	0.20	0.80	0.43	0.53	0.11	-0.49	0.98	1.68
$t(\alpha)$		1.21	0.47	2.38	1.45	1.20	0.32	-1.77	3.16	4.29
β		0.60	0.72	0.86	0.89	0.66	0.79	0.86	0.81	0.75
R^2		48	62	79	84	57	74	85	80	68

This table presents the performance of the multi-factor timing strategies by model. Reported are the annualized Sharpe ratio (SR), along with the alpha (α), alpha t-statistic (t(α)), beta (β), and percentage R^2 with respect to the multi-factor buy-and-hold benchmark (Buy).

Figure 4: Cumulative Returns of Multi-factor Timing



This figure plots the cumulative log returns of multi-factor timing strategies by model over the out-of-sample period from January 1990 to December 2021. Each model's returns are scaled to have the same volatility as the multi-factor buy-and-hold benchmark (Buy). NBER recessions are shown in grey.

lack of economic structure causing the model to overfit to factor-specific noise. However, by imposing economic structure, MT achieves a Sharpe ratio of 1.14, which is higher than all off-the-shelf models and the multi-factor Buy. Additionally, DMT, which further incorporates time series dynamics, achieves the highest Sharpe ratio of 1.26, surpassing the Sharpe ratios of the multi-factor Buy by 29% and leading off-the-shelf model (RF) by 18%. This is a larger improvement than the six factor volatility-managed portfolio in Moreira and Muir (2017), which records a 6% Sharpe ratio improvement over the multi-factor Buy. Our Sharpe ratio of 1.26 also greatly exceeds the Sharpe ratios of 0.71 and 0.73 from the PCA models of Haddad et al. (2020) and Kagkadis et al. (2023), respectively.

The second row of Table 3 displays the annualized alpha earned by each model. Only RF, MT, and DMT achieve alphas that are statistically significant at the 5% level. RF earns an alpha of 0.8% (t-stat of 2.38), which is the best risk-adjusted performance by an off-the-shelf model. NN posts a negative alpha of -0.49% (t-stat of -1.77) due to its propensity to overfit. However, MT earns an alpha of 0.98% (t-stat of 3.16) from imposing economic structure, and DMT boosts the alpha to 1.68% (t-stat of 4.29) from further incorporating time series dynamics. DMT also records a low beta of 0.75 and an R^2 of only 68%, suggesting that the model is capable of avoiding months with factor losses. These results underscore DMT's efficacy in generating superior risk-adjusted returns, underscoring the importance of incorporating economic structure and time series dynamics into deep learning models.

Figure 4 illustrates the cumulative returns of each model and the multi-factor Buy, where each model's returns are scaled to have the same volatility as the multi-factor Buy. RF and XRF earn the highest cumulative returns among the off-the-shelf models, surpassing the multi-factor Buy, but noticeably underperforming DMT and MT. LR, EN, GBT, NN and SVM share similar cumulative returns and underperform the multi-factor Buy.

Examining deep learning models, DMT achieves the highest cumulative return throughout the entire sample, with a minor drawdown during the GFC. Interestingly, DMT performs well in the post-GFC sample and avoids the large drawdown experienced by the multi-factor

Buy and other models during the Covid crisis. MT exhibits the second highest cumulative return, and also performs well in the post-GFC sample. In contrast, NN exhibits the lowest cumulative return due to its propensity to overfit to the noise of individual factors. Taken together, these results demonstrate that incorporating economic structure and time series dynamics is necessary for factor timing with deep learning.

4.4 Single-factor Timing

Next, we break down the multi-factor timing performance of each model by analyzing the performance of strategy (3) for each individual factor. Table 4 reports each model's annualized Sharpe ratio, along with the alpha, alpha t-statistic, beta, and R^2 from a time series regression of the strategy on the respective single-factor Buy. The first panel shows that none of the off-the-shelf models earn a higher Sharpe ratio than MKT's Sharpe ratio of 0.6. However, DMT stands out by significantly outperforming the market with a Sharpe ratio and alpha of 0.89 and 4.78% (t-stat of 4.17), respectively. This demonstrates that incorporating economic structure and time series dynamics is important for market timing. MT earns the same Sharpe ratio as MKT, with a positive but insignificant alpha of 0.11% (t-stat of 0.25). XRF is the best-performing off-the-shelf model, but its strategy is identical to the market as indicated by its beta of 1. All other off-the-shelf models exhibit lower Sharpe ratios than MKT and negative alphas.

The second panel shows that all models, except NN, beat SMB's lackluster Sharpe ratio of 0.16. RF and LR are the best performers with Sharpe ratios of 0.38 and 0.36, respectively, with statistically significant alphas of 2.19% (t-stat of 2.55) and 1.88% (t-stat of 2.03). Additionally, MT and DMT earn Sharpe ratios of 0.35 and 0.22, respectively, suggesting that the benefits of MTL are somewhat limited for the size factor.

The third panel shows that all models, except XRF, SVM, and NN, outperform HML's low Sharpe ratio of 0.11. This suggests that the flexibility of these nonlinear models leads to overfitting on the value factor. Interestingly, LR is the best model with a Sharpe ratio of even higher-dimensional spaces driving factor risk premia. More importantly, we illustrate the benefits of DMT in capturing these sophisticated nonlinear interactions, which result in superior out-of-sample performance. This reinforces the effectiveness of DMT in leveraging these multidimensional relationships to optimize predictions, providing a holistic understanding of the intricate web of variables that influence factor risk premia.

5 Conclusion

Understanding factor risk premia is a central topic in financial economics and the burgeoning factor investing industry. An extensive literature studies factor predictability by employing static and linear single-factor models with a limited set of predictors. These approaches have yielded inconsistent outcomes, leaving the feasibility of factor timing as a topic of considerable debate. Furthermore, these traditional models are inadequate in handling a zoo of predictors that are influential for factor timing, leading to inconclusive conclusions about the key drivers of factor predictability.

In this paper, we introduce deep neural networks that incorporate economically motivated restrictions, tailored to address the main challenges of factor timing. We develop a dynamic multi-task deep learning model to forecast six well-known factors, using 123 macroeconomic and 149 financial predictors, in the 57 year period from January 1965 to December 2021. Empirically, we demonstrate several results that enhance our knowledge of factor timing. First, we show that incorporating economic structure and time series dynamics significantly improves the predictive accuracy of factor timing models. Second, our results reveal that deep learning models with economic structure produce significant economic gains in a multi-factor portfolio, which is further enhanced by incorporating time series dynamics. Third, we find that integrating multi-task learning with time series dynamics can yield consistent economic gains across all factors relative to the buy-and-hold benchmark. Fourth, we document that the most important variables for factor timing, include tail risk and variables

belonging to the price trends, leverage, and profitability categories. Fifth, we demonstrate that nonlinear interactions among these influential variables are important for effective factor timing. Overall, the improved factor timing yielded from our dynamic multi-factor deep learning approach paves the way for a more reliable investigation of the economic mechanisms driving factor risk premia, and underscores the value of incorporating economically motivated restrictions into deep learning models for factor investing.

References

Antoniou, C., Doukas, J. A., and Subrahmanyam, A. (2013). Cognitive dissonance, sentiment, and momentum. *Journal of Financial and Quantitative Analysis*, 48(1):245–275.

Arnott, R., Harvey, C. R., Kalesnik, V., and Linnainmaa, J. (2019). Alice's adventures in factorland: Three blunders that plague factor investing. *Journal of Portfolio Management*, 45(4):18–36.

Asness, C., Chandra, S., Ilmanen, A., and Israel, R. (2017). Contrarian factor timing is deceptively difficult. *The Journal of Portfolio Management*, 43(5):72–87.

Asness, C. S. (2016). Invited editorial comment: The siren song of factor timing aka "smart beta timing" aka "style timing". *The Journal of Portfolio Management*, 42(5):1–6.

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

Baba Yara, F., Boons, M., and Tamoni, A. (2021). Value return predictability across asset classes and commonalities in risk premia. *Review of Finance*, 25(2):449–484.

Bender, J., Sun, X., Thomas, R., and Zdorovtsov, V. (2018). The promises and pitfalls of factor timing. *The Journal of Portfolio Management*, 44(4):79–92.