Interpretable Deep Learning for Stock Returns: A Consensus-Bottleneck Asset Pricing Model *

Bong-Gyu Jang Younwoo Jeong Changeun Kim

This draft: November 14, 2025
First draft: March 5, 2025

*This paper is a revised version of Master's thesis by Changeun Kim titled "A Consensus-Bottleneck Asset Pricing Model", submitted to the Department of Industrial and Management Engineering, POSTECH, Korea. We would like to thank Hyeng Keun Koo, Kwangmin Jung, Dojoon Park (discussant), JinGi Ha, Jeonggyu Huh, and seminar participants at the 2024 Spring Joint Conference of KORMS and KIIET, 2025 APAF International Conference, 2025 KFA Fall Conference, 2025 4th Workshop on Financial Mathematics and Engineering (Pusan National University), for helpful discussions and insightful comments. This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023R1A2C2003927). Jang (E-mail: bonggyujang@postech.ac.kr) is at Department of Industrial and Management Engineering, POSTECH, Korea University Business School; Jeong (E-mail: younwoo48@postech.ac.kr) is at Graduate School of Artificial Intelligence, POSTECH; Kim (E-mail: changeun120@postech.ac.kr) is at Department of Industrial and Management Engineering, POSTECH. Correspondence concerning this article should be addressed to Changeun Kim, Department of Industrial and Management Engineering, POSTECH, Pohang 37673, Republic of Korea. E-mail: changeun120@postech.ac.kr.

Interpretable Deep Learning for Stock Returns: A Consensus-Bottleneck Asset Pricing Model

Abstract

We introduce the *Consensus-Bottleneck Asset Pricing Model* (CB-APM), a partially interpretable neural network that replicates the reasoning processes of sell-side analysts by capturing how dispersed investor beliefs are compressed into asset prices through a consensus formation process. By modeling this "bottleneck" to summarize firm- and macro-level information, CB-APM not only predicts future risk premiums of U.S. equities but also links belief aggregation to expected returns in a structurally interpretable manner. The model improves long-horizon return forecasts and outperforms standard deep learning approaches in both predictive accuracy and explanatory power. Comprehensive portfolio analyses show that CB-APM's out-of-sample predictions translate into economically meaningful payoffs, with monotonic return differentials and stable long-short performance across regularization settings. Empirically, CB-APM leverages consensus as a regularizer to amplify long-horizon predictability and yields interpretable consensus-based components that clarify how information is priced in returns. Moreover, regression and GRS-based pricing diagnostics reveal that the learned consensus representations capture priced variation only partially spanned by traditional factor models, demonstrating that CB-APM uncovers belief-driven structure in expected returns beyond the canonical factor space. Overall, CB-APM provides an interpretable and empirically grounded framework for understanding belief-driven return dynamics.

Keywords: Asset Pricing Model, Analysts' Consensus, Neural Network, Interpretable Deep Learning, Cross-Section of Stock Returns

JEL Codes: C45, C53, G0, G12, G17

1 Introduction

Empirical asset pricing has long relied on statistical modeling to explain stock returns, often within the framework of factor-based models such as those proposed by Fama and French (1993, 2015) and Carhart (1997). These models aim to enhance explanatory power by identifying systematic risk factors that drive returns. However, despite decades of research, the ability of traditional models to predict future stock returns remains constrained, particularly in out-of-sample settings (Ang and Bekaert, 2007; Campbell and Thompson, 2008; Cochrane, 2008). Moreover, it remains uncertain whether the results from the existing literature can be successfully reproduced and whether such predictors and econometric modeling methodologies can be generalized across a broader set of assets or diverse economic conditions. The proliferation of new factors—often referred to as the "factor zoo" (Cochrane, 2011)—has further complicated the landscape, raising concerns about robustness, data mining, and the true economic relevance of many proposed predictors.

To address these challenges, it is essential to explore deep inside the factor zoo to identify economically meaningful signals and evaluate their contribution to return prediction. Drawing from a number of studies on stock return predictors, [1] seminal work of Gu et al. (2020) proposes a "return prediction model" that integrates traditional asset pricing empirical frameworks and theories with the rapidly evolving field of machine learning. By utilizing a variety of machine learning algorithms including neural networks, and leveraging a high-dimensional set of predictive factors, their results significantly contribute to the literature by showing the effectiveness of nonlinear and complex modeling on empirical asset pricing. Several subsequent studies utilize the conceptual formulation of this study across diverse financial markets and assets such as bonds (Bianchi et al., 2021), cryptocurrencies (Jaquart et al., 2021; Fang et al., 2024) and foreign stock exchanges (Leippold et al., 2022). Theoretical studies have also emerged to justify the use of machine learning into empirical asset pricing. For instance, Kelly et al. (2024) illustrates how model complexity can be instrumental in achieving superior performance in cross-sectional return prediction, demonstrated through a simple example of penalized linear regression.

While return prediction models benefit from machine learning approaches due to their empirical flexibility, deep learning has also proven successful in approximating "asset pricing factor models".

[1] See Welch and Goyal (2008), Green et al. (2013), Hou et al. (2015), Harvey et al. (2016), He et al. (2017), Green et al. (2017), Gu et al. (2020), Feng et al. (2020), Freyberger et al. (2020),Bybee et al. (2023) and Jensen et al. (2023).

Expanding on the research by Kelly et al. (2019), which defines the covariance term β using the covariance of "characteristics", Feng et al. (2018) and Gu et al. (2021) employ deep neural network architecture and the resulting latent factors to model the state variables of Intertemporal Capital Asset Pricing Model (ICAPM, Merton, 1973). Chen et al. (2024) introduce a novel architecture consisting of feedforward networks and LSTMs, that are trained via minimax optimization technique similar to that of Generative Adversarial Networks (GAN, Goodfellow et al., 2020). Based on arbitrage pricing theory (APT), the proposed model successfully approximates the stochastic discount factors (SDF) and corresponding risk loadings to formulate a highly predictive asset pricing model.

Despite the strong evidence that deep learning approaches illustrate evident potential in capturing the complex topology of predictor structures, critical limitation remains: Can the results from these models be considered trustworthy? Rudin et al. (2022) highlights such critical issue with machine learning black box models,

> *Black box models often predict the right answer for the wrong reason (the "Clever Hans" phenomenon), leading to excellent performance in training but poor performance in practice.*

Recent studies in machine learning asset pricing frequently employ models that are not interpretable, which raises concerns about relying on complex machine learning algorithms in empirical asset pricing without a clear understanding of why and how these models arrive at their conclusions. Furthermore, these papers often attempt to interpret the prediction results based on the learned models and derive economic implications. However, Rudin (2019) argues that such analyses are solely based on post-hoc explanations that should be considered as fitting narratives to the outcomes. These explanations are conveniently aligned with prevailing economic theories and tend to disregard contradictory evidence, which limit the scope and applicability of the findings. For these reasons, Rudin (2019) strives to rectify researchers and practitioners to use interpretable models over black-box algorithms. Despite growing interest in interpretable machine learning and trustworthy Artificial Intelligence (AI), a notable gap persists in applying and validating these approaches within asset pricing, beyond traditional regression or decision-tree models. In particular, existing machine-learning frameworks rarely achieve both strong predictive performance and

economic interpretability. To address this gap, we propose the Consensus-Bottleneck Asset Pricing Model (CB-APM), a framework that employs a partially interpretable neural architecture to predict future stock returns while preserving clear economic structure.

Our approach builds upon two established pillars of financial economics, the rational expectations hypothesis and empirically documented relationships between analyst consensus information and asset prices. Rational expectations, proposed by Muth (1961), posit that market participants form forecasts using all available historical information. Several research find evidence of such a hypothesis from the decisions of sell-side analysts, deriving the economic implications of analysts' opinions and estimates. Subsequent empirical work demonstrates this principle in sell-side analysts' behavior. Lovell (1986) shows economic agents systematically incorporate public information into earnings forecasts, while Lim (2001) establishes predictable patterns in analysts' forecast revisions consistent with Bayesian updating. Crucially, Jegadeesh et al. (2004) identify specific style factors—including momentum, growth prospects, and trading volume—that systematically influence analysts' stock recommendations, suggesting a quantifiable link between firm characteristics and consensus formation.

These theoretical insights find strong empirical support in asset pricing studies. Research consistently demonstrates that analyst consensus information contains predictive signals for future returns. Diether et al. (2002) documents that stocks with high dispersion in analysts' earnings forecasts earn significantly lower future returns, implying consensus alignment affects risk premiums. Sorescu and Subrahmanyam (2006) further substantiates this connection, showing pronounced price reactions to revisions in analyst recommendations and earnings estimates. Such findings validate consensus information as both a measurable economic construct and a plausible mediator between firm characteristics and expected returns.

The CB-APM framework operationalizes these insights through a concept-bottleneck architecture inspired by Koh et al. (2020). Our model first predicts "analyst consensus", the aggregated estimates of sell-side analysts regarding firm fundamentals and valuations-using observable return predictors. It then employs a linear layer to map this latent consensus to expected returns, preserving interpretability in the final pricing stage. This design explicitly recognizes consensus formation as an information aggregation process while maintaining testable links to asset pricing theory. By integrating the rational expectations literature with empirical evidence on analyst behavior, CB-

APM achieves dual objectives: it delivers state-of-the-art cross-sectional return predictions while elucidating the data-generating process of risk premiums. Our approach not only validates the economic significance of consensus information but also provides a computational framework to quantify its role in price formation, a critical advancement given increasing market reliance on analyst-derived signals.

Our contributions are threefold. First, we introduce a concept-bottleneck framework that synthesizes the high-dimensional predictor set into interpretable, consensus-style expectations, providing a structured economic link between characteristics, analysts' beliefs, and expected returns. Second, we demonstrate that this architecture delivers economically large improvements in long-horizon return prediction across expanding-window evaluations. Third, we show that the learned consensus representations encode priced information that is only partially spanned by traditional factor models, offering new empirical insight into how belief heterogeneity and information aggregation shape risk premia. These contributions advance recent efforts to integrate interpretable machine learning with the core principles of empirical asset pricing.

To empirically validate the effectiveness of CB-APM, we assess its predictive performance and economic implications using a comprehensive dataset spanning from January 1994 to December 2023, consisting of 605,722 firm-month observations across 4,683 U.S. companies. The dataset integrates 114 firm-level predictors, 123 macroeconomic indicators, and 9 analysts' consensus variables including EPS forecast revisions and forecast dispersions. To account for the time dynamics of return prediction, we employ an expanding window approach, where the training dataset grows over time while keeping validation and test sets fixed. This experimental setup allows us to assess the robustness of CB-APM under evolving market conditions.

Our empirical analysis demonstrates that CB-APM delivers substantial improvements in both predictive performance and economic interpretability. First, in the cross-section of consensus and stock returns, incorporating consensus learning markedly enhances long-horizon return forecasts: CB-APM attains an out-of-sample R^2 of 10.46% for annual returns, representing a significant improvement over a standard deep learning benchmark ($R^2 = 7.63\%$), while simultaneously achieving an average R^2 of 24.21% in approximating analyst consensus variables. These gains remain robust across expanding-window evaluations, indicating stable performance across different market regimes.

Second, portfolio-level analyses establish the model's economic relevance: portfolios formed on out-of-sample CB-APM predictions exhibit strongly monotonic payoff structures, with high-minus-low spreads reaching approximately 2.3% per month for regularized specifications ($\lambda \geq 0.3$). Double-sorted portfolios based jointly on model-implied returns and analysts' earnings forecasts show that CB-APM internalizes both the informational and behavioral components of analyst expectations—replicating the return predictability. The resulting reversal structure, characterized by positive analysts' earnings forecasts spreads in low-return groups and negative spreads in high-return groups, demonstrates that the model not only refines the predictive content of analysts' beliefs but also captures state-dependent pricing adjustments.

Finally, long-short portfolios derived from the model's forecasts achieve economically significant and stable out-of-sample performance, with mean monthly log returns rising from 1.53% at $\lambda = 0$ to 2.20% at $\lambda = 0.3$ and the annualized Sharpe ratio improving from 1.10 to 1.44. These results establish a direct correspondence between predictive accuracy, cross-sectional return ordering, and risk-adjusted profitability, confirming that consensus regularization enhances not only statistical fit but also economic value.

Beyond predictive performance, we further examine whether the consensus bottleneck captures economically meaningful pricing structure. A comparative regression analysis demonstrates that the CB-APM–implied consensuses deliver substantially stronger explanatory power for annual returns than raw analyst signals: pooled OLS regressions exhibit an order-of-magnitude improvement in adjusted R^2, together with economically interpretable shifts in coefficient signs and magnitudes. These gains arise because the consensus layer synthesizes information from firm characteristics and macroeconomic conditions into belief-like representations that are simultaneously close to observable analyst forecasts and tightly aligned with priced return variation. Variables that the model reconstructs with higher fidelity display more stable and economically intuitive return sensitivities, whereas poorly reconstructed dimensions exhibit weaker economic content or sign reversals—highlighting that economic interpretability depends jointly on approximation quality and return-pricing relevance. This evidence confirms that the consensus bottleneck does not merely denoise analyst inputs but reorganizes information into latent expectations that better capture the priced component of belief dispersion.

We further evaluate the pricing relevance of these signals using Gibbons–Ross–Shanken (GRS)

tests on benchmark portfolios and portfolios formed on model-implied returns and individual consensus dimensions. Consensus-based long–short factors span meaningful components of systematic return variation but do not fully replicate the benchmark factor structure, indicating that the learned expectations are economically relevant without collapsing onto the canonical dimensions of market, size, value, momentum, profitability, or investment. Conversely, traditional factor models increasingly fail to price portfolios formed on CB-APM's predicted returns as the consensus bottleneck tightens, suggesting that the model uncovers structured forms of nonlinear or interaction-based return heterogeneity that lie outside the linear span of standard factors. Portfolios sorted on individual consensus dimensions produce modest pricing errors, consistent with the view that belief-based signals reflect compressible yet economically meaningful combinations of characteristics. Taken together, these findings show that CB-APM extracts interpretable consensus representations that contain priced information only partially captured by existing factor models, positioning the framework as a complementary approach that links analysts' heterogeneous beliefs to expected returns in a transparent and theoretically coherent manner.

Collectively, these results establish CB-APM as a novel and effective framework that integrates interpretable deep learning with foundational principles of financial economics. Unlike prior machine learning approaches that prioritize accuracy at the expense of transparency, CB-APM demonstrates that interpretable architectures can preserve theoretical grounding while achieving state-of-the-art empirical performance. By jointly modeling analysts' expectations and stock returns, our framework provides a principled means of disentangling forward-looking information embedded in firm characteristics and macroeconomic variables, yielding insights into how such information is aggregated and priced. This dual capacity—enhancing return predictability while maintaining an economically interpretable structure—constitutes the central contribution of our paper and advances the emerging literature on interpretable machine learning in finance.

The remainder of the paper is organized as follows. Section 2 reviews interpretable machine learning and situates CB-APM within this framework. Section 3 outlines the model, estimation procedure, and architecture. Section 4 describes the data and evaluation design, including the autoencoder for macroeconomic state extraction. Section 5 presents empirical results on predictive performance, macroeconomic embeddings, and portfolio-based pricing implications. Section 6 investigates the pricing content of the approximated consensuses using regression and GRS tests.

Section 7 concludes. The Internet Appendix provides additional robustness analyses and supplementary results.

2 Interpretable Artificial Intelligence

Before we further expand the discussion, it is necessary to clarify the often-confused concept of explainability and interpretability, to prevent any potential misunderstanding. Although they both focus on understanding the nature of machine learning from a human perspective, the primary difference between these two fields with a long and intense history of research lies in their focus areas.

Explainable AI, also known as XAI, focuses on the reason *why* the prediction of a model has been inferred, while interpretable AI is more interested in *how* the model is trained to find the approximate mapping from the hypothesis set. In particular, XAI does not attempt to dissect the functioning of the black box model but rather accepts the opacity of such models as it is intended to be. The fundamental assumption of XAI is that the result of such an opaque system should be strongly related to the input features, which, in the context of deep learning, is often the extent of our understanding. Therefore, researchers design an ad hoc statistical model, which is simpler compared to a model subject to explanation in most cases, to explain the relationships between the input and output of the estimated model.

Interpretable AI, in comparison, designs a model to be perceivable in human knowledge itself, without requiring further explanations. The most dominant and perhaps the most well-known example of an interpretable model is linear regression. Linear regressions are interpretable since their outputs can be directly represented as linear combinations of input features and coefficients. Conversely, Benítez et al. (1997) argue that deep neural networks are considered non-interpretable since these models typically do not provide insight into how input features are transformed through hidden layers to produce outputs.

Nevertheless the black box nature of deep neural networks, researchers may attempt to make such models understandable by introducing the concept of "disentangled representations" (See Higgins et al., 2018; Locatello et al., 2019a,b). To understand the concept of disentanglement, we can consider the simple example of a naïve feedforward neural network, which consists of multiple hid-

den layers. The hidden layers are intermediate vector representations, often interpreted as features extracted from input data. Since they are trained with a downstream task-related objective in most cases, those layers are presumed to represent the essential information within the high-dimensional and noisy input data, well enough for the successful performance on that task. However, these features exist in an "entangled" manner, meaning that each element of the feature representation is a mixture of multiple factors. This entanglement complicates the interpretability of the model, as it blurs the specific contributions of individual input features to the final output. In this context, "disentanglement" refers to separating the underlying causal factors of the data and intermediate representations into distinct and non-overlapping representations. For instance, in fixed income markets, dozens of yields across maturities can be effectively summarized by three disentangled factors, level, slope, and curvature, as illustrated in Figure 1. Each of these dimensions captures an independent source of variation in the yield curve, and together they provide an interpretable low-dimensional representation that still preserves the essential structure of the data.[2]

[Insert Figure 1 here]

Applying this concept, fully interpretable neural networks are designed so that all aspects of their structure and function are understandable to humans. This means that every layer, neuron, and connection in the network has a clear and understandable purpose related to the task at hand. These fully Interpretable models have been acknowledged for their transparency and the ease with which their decisions can be understood and trusted. However, a prevailing limitation of these models has been their generally lower predictive performance compared to their less interpretable counterparts due to the interpretability-accuracy trade-off, presented by Plate (1999). This trade-off has historically motivated researchers towards utilizing complex neural network-based models in machine learning studies, due to their superior predictive capabilities, despite their lack of interpretability.

Recent advances in interpretable AI, however, are challenging the notion that interpretability

[2]A clarification is needed to avoid conflating our framework with traditional factor models. Nelson and Siegel (1987) impose level, slope, and curvature terms to parameterize the yield curve, while Litterman (1991) show via principal components that similar dimensions emerge empirically. Both approaches reduce high-dimensional yields to a few interpretable coordinates. By contrast, in CB-APM the consensus-bottleneck is not imposed ex ante but learned endogenously, closer in spirit to Litterman–Scheinkman than to Nelson–Siegel. In this sense, whereas Nelson–Siegel estimate factors to fit the curve, our model embeds interpretability within the network architecture itself, rendering part of the neural network transparent while retaining predictive capacity.

must come at the cost of performance. Studies are now demonstrating that it is feasible to maintain the high performance of neural network structures while making them partially interpretable [3]. Partial interpretability takes an approach where a segment of the network is made interpretable, typically the layers closer to the input or output. The rest of the network may operate as a black box, allowing intervention on the learning process without significant loss in model performance.

3 Model and Methodology

3.1 Return prediction model

Similar to the Gu et al. (2020), the asset return prediction error model utilized in our work is formulated for the h-horizon forecasting problem as below,

$$R_{i,t+h} = E_t[R_{i,t+h}] + \epsilon_{i,t+h}, \tag{1}$$

where $R_{i,t+h}$ is the h-month return of asset i excess of the risk-free rate at time $t+h$, and $\epsilon_{i,t+h}$ is an error term. In this context, h is used to assign the forecasting horizon, enabling the consideration of multi-horizon predictions, allowing CB-APM to model long-term dependencies. The expected excess return in equation (1) is defined as the expectation conditional on information sets,

$$E_t[R_{i,t+h}] = E[R_{i,t+h}|I^f_{i,t}, I^m_t].$$

Here, $I^f_{i,t}$ and I^m_t are the set of firm-specific characteristics and macroeconomic predictors at time t, respectively. [4] It is important to note that consensus information is deliberately excluded from $I^f_{i,t}$.

This framework is further developed by defining the functional form of the conditional expectation as a composite function,

$$E[R_{i,t+h}|I^f_{i,t}, I^m_t] = g(f(I^f_{i,t}, I^m_t; \phi); \theta), \tag{2}$$

[3] Refer to the representative studies of Koh et al. (2020), Chen et al. (2020) for discriminative models and Chen et al. (2016), Higgins et al. (2017) for probabilistic generative models.

[4] A detailed description of the predictors comprising the information sets is provided in Section 4, and a complete list of variables is available in Internet Appendix D. The macroeconomic information set I^m_t is represented empirically by a latent vector extracted through an autoencoder trained on macroeconomic variables, as described in Section 4.2.

where the function $f(\cdot)$ and $g(\cdot)$ are smooth functions parameterized by learnable parameters θ and ϕ. The function $f(\cdot)$ is specifically designed to model the conditional expectation of analyst consensus. Then, the function $g(\cdot)$ models the expected return only using the features of approximated consensus from the previous step, creating a "concept-bottleneck" within the prediction model. This empirical design is predicated on the understanding that both researchers in empirical asset pricing and financial analysts share the objective of assessing a firm's value and discerning the factors that influence these valuations. While analysts often have access to broader datasets, including some predictive signals that may not be publicly available or included in this article, asset pricing panel data can represent a information subset in a descent quality by providing a comprehensive and quantifiable measures of firm's fundamentals and macroeconomic conditions that are crucial for the approximation of the consensus, as shown in the empirical results later on.

In mathematical form, $f(\cdot)$ approximates the analyst consensus variables, denoted as $C_{i,t}$.

$$C_{i,t} = f(I^f_{i,t}, I^m_t; \phi).$$

Let the approximated value of $C_{i,t}$ and the parameter ϕ be $\hat{C}_{i,t}$ and $\hat{\phi}$, respectively, then,

$$\hat{C}_{i,t} = f(I^f_{i,t}, I^m_t; \hat{\phi}).$$

Finally, the expected excess return is defined with function $g(\cdot)$ and the approximated $\hat{C}_{i,t}$ as below,

$$E_t[R_{i,t+h}] = g(\hat{C}_{i,t}; \theta). \tag{3}$$

As discussed in Daniel and Titman (1997), the main limitation of the return prediction error model is the absence of economic constraints. For instance, the fundamental theorem of asset pricing constrains the arbitrage opportunity, which implies that the difference between the price of an identical asset is improbable. This condition is referred to as "the law of one price" in asset pricing theory. In the cases without such condition, two different assets can have identical price despite disparate fundamental values.

However, CB-APM diverges from the approach of return prediction modeling for several reasons. Firstly, it offers greater flexibility, accommodating diverse scenarios involving analyst estimates and

future returns. Unlike factor models, which do not differentiate prices of identical risk factors, CB-APM acknowledges that similar analyst opinions across firms may yield distinct future returns. While we assume rational decision-making by analysts, as discussed in subsequent sections, it is prudent not to constrain such scenarios initially. Secondly, CB-APM facilitates a range of optimization approaches in approximating the asset pricing model. Unlike factor models, where the estimation process is mostly the extension of Fama–MacBeth regression (Fama and MacBeth, 1973) restricting the integration of the entire expected return modeling process, CB-APM allows for a more holistic training process, avoiding multiple optimization procedures. Overall, given that the consensus-bottleneck represents a novel approach in asset pricing research, we aimed to maintain the underlying framework as simple and flexible as possible.

Although it is designed as intended, given that neural networks are well-recognized as "universal approximators", the model can allow any scenarios and consequences as outcomes, that doesn't necessarily align with the economic theories. To overcome the limitation of the proposed prediction model, we apply stabilized optimization approaches proposed in the machine learning literature, such as regularization and scheduling. Such techniques are expected to function as "universal constraints", achieving both practical performances and theoretical rigor. See Internet Appendix B.3 for detailed discussions and experimental settings.

3.2 Estimation

In this section, we provide the loss function of the model that simultaneously estimates the parameters of function $f(\cdot)$ and $g(\cdot)$ from equation (2) in a single optimization step.

Given $\lambda > 0$, the model's loss function is structured as a joint optimization task, represented by a weighted sum of two distinct loss functions:

$$L = L_R + \lambda L_C, \tag{4}$$

where the "return loss" L_R is formulated as,

$$L_R(\phi, \theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (R_{i,t+h} - g(f(I_{i,t}^f, I_t^m; \phi); \theta))^2, \tag{5}$$

and the "consensus loss" L_C is formulated as,

$$L_C(\phi) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (C_{i,t} - f(I_{i,t}^f, I_t^m; \phi))^2. \tag{6}$$

L_R and L_C are cross-sectional mean squared errors (MSE) from a standard pooled OLS estimator, where λ is a hyperparameter that assigns weight to the consensus loss, providing additional flexibility into the empirical design of the model.

We also estimate a benchmark model taking $\lambda = 0$ from equation (4), which ignores learning analyst opinions by removing the consensus loss term, making the model identical to the naïve return prediction model. Although $f(\cdot)$ and $g(\cdot)$ are the model defined with a separate set of learnable parameters θ and ϕ, they can be considered as a single neural network when $\lambda = 0$ since the optimization procedures for each networks are not independent.

The strategy of jointly learning the consensus and excess return offers several advantages. Firstly, it tends to yield higher performance metrics due to the synergistic learning of interconnected variables. An alternative method might involve independent optimization, where the $f(\cdot)$ and $g(\cdot)$ are trained independently in two separate steps. However, this segmented approach often fails to capture the potential inter-dependencies between the consensus estimates and the resulting excess returns. Furthermore, since the information set $I_{i,t}^f$ and I_t^m are not included in equation (5), the training of $g(\cdot)$ entirely depend on the quality of the extracted signals in approximated consensus, which makes training with the loss function L_R extremely challenging.

Secondly, it provides deeper insights and more intuitive understanding of the underlying financial dynamics. Independently learning $f(\cdot)$ using the equation (6) is not a novel concept and aligns with the existing literature supporting the evidence of the rational expectations hypothesis. As discussed in previous sections, the set of predictor signals used in this study is regarded to contain a significant amount of information sufficient to make "rational" expectations, which simplifies the problem of approximating the opinions of individuals, compared to predicting future returns of assets. However, since analysts perform their analysis as their job, they must think and act beyond being merely "rational"; they must be "professional". Therefore, we posit that professional and successful analysts strive to make their estimates predict not only "macroeconomic" consequences but also "firm-specific" outcomes. More specifically, proficient analysts will make decisions that

better predict the future returns of a firm's stocks.

3.3 Model architecture

In this section, we provide detailed explanations the model architecture. The overall framework of CB-APM is described in Figure 2. The model consists of two main components; the consensus module and the prediction module. Each of these modules corresponds to the function $f(\cdot)$ and $g(\cdot)$ in equation (2).

[Insert Figure 2 here]

In the proposed model, the consensus module is designed as an arbitrary feedforward network, while the prediction module is restricted to a simple linear regression that receives consensus variables as inputs and yield the expected excess return. This design choice is critical for enhancing interpretability in particular. When both modules are complex feedforward networks with multiple hidden layers, the advantage of using a consensus-based approach diminishes since it creates two separate black-box models from a single black-box model.

The loss functions, as defined in equations (5) and (6), are computed using the outputs from the respective modules. Once we get the return loss from the return module, the final loss function is calculated via weighted sum of these two loss functions as described in equation (4). The backpropagation in the CB-APM is conducted in a single step, utilizing the composite loss function in equation (4), which simultaneously adjusts the weights in both the consensus and prediction modules.

For an activation function, we utilize Gaussian Error Linear Units function (GELU) as non-linearity of the neural network. The mathematical formulation of GELU is given as below.

$$GELU(x) = x \cdot P(X \leq x) = x \cdot \Phi(x),$$

where $\Phi(x)$ the cumulative distribution function for Gaussian distribution $X \sim N(0, \sigma^2)$. GELU was first introduced by Hendrycks and Gimpel (2016) as an alternative of Rectified linear units (ReLU) (Nair and Hinton, 2010). Figure 3 shows that GELU permits some small interval for negative inputs to propagate through subsequent layers. See Internet Appendix B for a review of the literature on activation functions and justification for the selection of GELU.

[Insert Figure 3 here]

The mathematical form of the model architecture is given as follows. First, let X denote the input layer, and $H^{(1)}, H^{(2)}, \ldots, H^{(n)}$ represent the hidden layers. The weight matrices connecting the layers are denoted as $W_0, W_1, \ldots, W_n^c, W_n^r$, where W_0 connects the input layer to the first hidden layer, W_1 connects the first hidden layer to the second hidden layer, and so forth, up to W_n^c connecting the n-th hidden layer to the output layer of the consensus module, and W_n^r connecting the output layer of the consensus module to the output layer of the return module. Similarly, the bias vectors are represented as $b_0, b_1, \ldots, b_n^c, b_n^r$. The computations for the hidden layers are as follows,

$$
\begin{aligned}
H^{(1)} &= GELU(W_0(I^f \oplus I^m) + b_0) \\
H^{(2)} &= GELU(W_1 H^{(1)} + b_1) \\
&\vdots \\
H^{(n)} &= GELU(W_{n-1} H^{(n-1)} + b_{n-1}).
\end{aligned}
$$

Then the output layer computation of the consensus module is given by,

$$f(I^f, I^m; \phi) = W_n^c H^{(n)} + b_n^c,$$

and the output layer computation of the return module is given by,

$$g(f(I^f, I^m; \phi); \theta) = W_n^r f(I^f, I^m; \phi) + b_n^r.$$

Note that there are no activation layers between the consensus and return modules for interpretability. Therefore, when $\lambda = 0$, the CB-APM functions as a simple feedforward network, with the number of hidden layers matching that of the consensus module. The learnable weights of CB-APM are initialized by adopting the He initialization proposed by He et al. (2015).

4 Data

4.1 Data description

In this section, we provide the brief explanations on the dataset and the sampling splitting scheme employed for empirical studies. The dataset comes from four distinct sources, which are all publicly available at the moment. Firstly, we obtain open-source asset pricing panel data from Chen and Zimmermann (2022), available to download on their website (`https://www.openassetpricing.com/`).[5] It comprises 114 firm-level predictors consisting of diverse financial metrics such as accounting figures, 13F filings, trading activities, and derivatives data.

Chen and Zimmermann (2022) also features 9 analysts' consensus variables including EPS forecast revision (*AnalystRevision*), Change in recommendation (*ChangeInRecommendation*), Change in Forecast and Accrual (*ChForecastAccrual*), Long-vs-short EPS forecasts (*EarningsForecastDisparity*), Analyst earnings per share (*FEPS*), EPS Forecast Dispersion (*ForecastDispersion*), Earnings forecast revisions (*REV6*), Analyst Value (*AnalystValue*), and Analyst Optimism (*AOP*).

Secondly, stock prices and firm sizes data are sourced from CRSP (Center for Research in Security Prices)[6], companies listed on the NYSE, Amex, and Nasdaq. This dataset is synchronized with the firm list from the panel data provided by Chen and Zimmermann (2022).

Lastly, the macroeconomic variables are obtained from FRED-MD database (McCracken and Ng, 2016) and Welch and Goyal (2008). FRED-MD consists of 115 monthly predictors that includes macroeconomic indicators reflecting the U.S. labor markets, consumption rates, monetary policies, etc. An additional set of 8 macroeconomic variables is constructed from the database maintained by Welch and Goyal (2008) on Goyal's website (`https://sites.google.com/view/agoyal145`), following Gu et al. (2020). T-bill rate is also obtained from this dataset, which is used for calculating risk premiums.

The final merged dataset consist of samples spanning from January 1994 to December 2023, with total 605,722 samples from 4,683 U.S. companies. Detailed descriptions of the dataset components and their respective sources are provided in Internet Appendix D.

[5] Data from Chen and Zimmermann (2022) undergoes several preprocessing steps including lagging, data sampling, data imputation, and rank normalization, as detailed in Internet Appendix A.

[6] Accessible via WRDS (Wharton Research Data Services).

4.2 Extracting macroeconomic state variables via autoencoder

To incorporate macroeconomic dynamics into the conditional expectation function $E_t[R_{i,t+h}]$ defined in equation (2), we encode the aggregate information set I_t^m through an autoencoder-based representation. While macroeconomic variables are often dismissed in cross-sectional asset pricing due to their perceived homogeneity across firms, we argue that macro context exerts differentiated influence through sectoral dynamics, capital structure sensitivity, and behavioral channeling. However, the sheer volume and redundancy of macroeconomic indicators, particularly those sourced from databases such as FRED-MD, pose significant challenges for model training. Including hundreds of highly correlated variables not only increases the risk of overfitting but also dilutes the learning signal by overwhelming the model with noise and irrelevant information.

Moreover, many macroeconomic series track similar phenomena at varying lags, granularities, or levels of transformation (e.g., growth rates, differences, log-levels), creating unnecessary dimensionality without proportional gains in explanatory power. This redundancy hinders both the stability and interpretability of predictive models, especially those trained on firm-level data where macro variables are shared across the entire cross-section. Reducing this high-dimensional input into a compact, informative representation is thus not only computationally efficient but also essential for isolating the latent economic regimes that meaningfully affect asset returns.

Dimensionality reduction techniques have long been used in financial modeling to address such issues. Principal Component Analysis (PCA) has served as a standard tool for extracting latent factors from large panels of macroeconomic variables (Ludvigson and Ng, 2007), while extensions such as Sparse PCA and Independent Component Analysis (ICA) have been applied to improve factor interpretability and reduce multicollinearity (Fan et al., 2016; Erichson et al., 2020). More recently, deep learning approaches—particularly autoencoders—have gained traction in the asset pricing literature for their ability to capture nonlinear interactions and extract economically meaningful latent structures from noisy, high-dimensional data (Chen et al., 2024; Gu et al., 2021). These methods have proven effective in modeling complex macro-financial dynamics that traditional linear techniques may fail to uncover.

To address these challenges, we enhance model performance by encoding the macroeconomic regime using an autoencoder, thereby providing structured, compact, and economically inter-

pretable representations that condition firm-level predictions. Instead of feeding all 115 raw macroeconomic variables from the FRED-MD database directly into the model, we train an autoencoder to learn a lower-dimensional latent representation of the macroeconomic environment at each time step. Figure 4 illustrates this process, where the encoder compresses high-dimensional macroeconomic inputs into a latent macroeconomic state vector $\mathbf{z}_t$, which is subsequently concatenated with firm-level features and passed into the CB-APM architecture. During training, the decoder reconstructs the input variables, and the network is optimized to minimize the mean squared reconstruction error. After training, only the encoder is retained to generate macroeconomic embeddings for prediction.

Formally, let the macroeconomic input at time t be $\mathbf{x}_t \in \mathbb{R}^D$, where $D = 123$. The encoder $f_\phi(\cdot)$ maps this input to a latent representation $\mathbf{z}_t \in \mathbb{R}^d$:[7]

$$\mathbf{z}_t = f_\phi(\mathbf{x}_t).$$

The decoder $g_\theta(\cdot)$ reconstructs the input:

$$\hat{\mathbf{x}}_t = g_\theta(\mathbf{z}_t),$$

and the model is trained to minimize the reconstruction loss:

$$L_{\text{AE}}(\theta, \phi) = \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{x}_t - g_\theta(f_\phi(\mathbf{x}_t))\|_2^2 .$$

After training, the encoder output $\mathbf{z}_t$ is concatenated with each firm's feature vector $\mathbf{x}_{i,t}^{\text{firm}}$ to form the model input:

$$\mathbf{x}_{i,t}^{\text{input}} = \left[\mathbf{x}_{i,t}^{\text{firm}}; \mathbf{z}_t\right].$$

Formally, the latent representation $\mathbf{z}_t$ learned through the autoencoder serves as an empirical proxy for the macroeconomic information set I_t^m introduced in equation (2). In this context, $\mathbf{z}_t$ functions as a compressed, data-driven approximation of the macroeconomic state observable to investors at time t. This design allows the CB-APM to integrate the high-dimensional macroe-

[7]Empirically, setting the latent dimension to $d = 32$ yields the best out-of-sample performance (see Internet Appendix C.4.1).

conomic information set into a tractable latent representation, ensuring that firm-level forecasts remain conditioned on a parsimonious yet informative depiction of the aggregate economic environment. By mapping I_t^m into $\mathbf{z}_t$, the model effectively operationalizes the theoretical information set within a learnable structure, thereby linking the empirical implementation of the macro encoder to the conditional expectation framework defined in equation (2).

As illustrated in Figure 4, this framework visualizes the overall data pipeline of the CB-APM, depicting how macroeconomic inputs are encoded, compressed, and subsequently integrated with firm-level characteristics for return prediction. The figure serves as a conceptual representation clarifying the interaction between the macro autoencoder and the return-prediction module. The full architecture details of the autoencoder, including hidden-layer configurations and activation functions, are provided in Internet Appendix B.3. The empirical findings underscore the importance of representing macroeconomic regimes in shaping cross-sectional return dynamics and highlight the utility of neural representation learning in extracting economically salient signals from high-dimensional macro data. At each expanding-window step, the autoencoder is trained only on macro data available up to the window end date, and the encoder is then used to compute $\mathbf{z}_t$ for that window's validation and test months, thereby preventing look-ahead bias.

[Insert Figure 4 here]

An ablation study, presented in Internet Appendix C.5.1, further confirms the contribution of this component. Removing the autoencoder from CB-APM leads to a pronounced deterioration in predictive performance—particularly under higher λ values—demonstrating that the learned macroeconomic embedding is essential to preserving both interpretability and accuracy. These results underscore that macroeconomic state conditioning is not a redundant extension but a core mechanism that stabilizes learning and improves out-of-sample generalization.

Finally, Appendix C.3 provides direct empirical evidence that the learned macroeconomic embeddings are economically revelatory. A two-dimensional projection of the 32-dimensional latent vectors[8] reveals a smooth temporal trajectory that coherently tracks major macroeconomic transitions, including distinct clusters corresponding to NBER-defined recessions such as the 2001 and 2008 downturns. Beyond these discrete regime shifts, the latent trajectory captures gradual cycli-

[8] We apply PCA to reduce the 32-dimensional latent state vectors to two dimensions only for visualization purpose.

cal and structural evolutions in the U.S. economy, reflecting shifts in growth, inflation, and monetary policy regimes. Collectively, these findings validate that the autoencoder encodes meaningful macro-financial state dynamics rather than statistical artifacts, yielding a compact and economically coherent representation that conditions firm-level return predictions within the CB-APM framework.

4.3 Expanding window approach for model evaluation

To evaluate model performance under realistic and evolving market conditions, this study employs an expanding window as a sample splitting scheme. Unlike static train-validation-test splits, the expanding window approach incrementally grows the training dataset over time while keeping the validation and testing sets fixed in size. This dynamic design mirrors the constraints of real-world applications, where future regimes are unknown and models must generalize across economic environments without the benefit of hindsight. By gradually shifting the end point of the training set forward, the expanding window simulates a time-consistent learning process that naturally adapts to structural changes in the data. As a result, this framework offers both methodological rigor and practical relevance, allowing the model to be evaluated not only on statistical metrics but also on its robustness across different economic cycles.

Figure 5 illustrates the expanding window approach for dataset partitioning, with the arrow along the bottom denoting the timeline of the window. The validation dataset spans two years, while the testing dataset spans a single year. Starting from the training set from January 1994 to December 2010, each training window ends at December of a given year, and subsequently expands by one year for the next window. This process continues sequentially, ensuring that the testing datasets do not overlap in any window. Consequently, the complete testing set spans from January, 2013 to December, 2022. [9]

[Insert Figure 5 here]

[9] The final year of the dataset (January–December 2023) is reserved solely for constructing annual stock returns, as computing these returns requires at least one full year of subsequent observations.

5 Empirical Results

5.1 Cross-section of consensuses and stock returns

This section presents empirical results on the cross-sectional prediction of stock returns and consensus variables. We evaluate predictive performance under varying forecast horizons h from equation (3) to assess the effectiveness of the consensus-bottleneck in asset pricing. Out-of-sample R^2 is used as the primary evaluation metric and is defined as:

$$R^2_{\text{return}} = 1 - \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(R_{i,t+h} - \hat{R}_{i,t+h})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T} R^2_{i,t+h}},$$

for return prediction and,

$$R^2_{\text{consensus}} = 1 - \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(C_{i,t} - \hat{C}_{i,t})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T} C^2_{i,t}},$$

for consensus prediction, where N and T denote the number of firms and time periods, respectively.

While much of the asset pricing literature emphasizes short-horizon return forecasts, sell-side analysts typically issue multi-quarter to annual forecasts. Consensus measures therefore reflect longer-term expectations about fundamentals and risk premia rather than short-term price fluctuations. Evaluating the consensus-bottleneck over horizons that align with analysts' forecast horizons is more economically relevant than using noisy short-term intervals. Accordingly, we focus on annual return prediction, consistent with prior studies on long-horizon predictability (Gu et al., 2020; Leippold et al., 2022).[10]

Table 1 reports the monthly out-of-sample R^2 values (in percentage) for both annual stock return prediction (R_{t+12}) and the approximation of analysts' consensus variables (C_t) across different values of the regularization parameter λ. The benchmark case ($\lambda = 0$), which excludes consensus learning, yields an annual return R^2 of 7.63%, serving as a baseline for evaluating the incremental benefits of integrating consensus prediction into the CB-APM framework.

[Insert Table 1 here]

Introducing consensus learning via $\lambda > 0$ leads to a pronounced improvement in return pre-

[10]The results for other forecasting horizons are provided in Internet Appendix C.1

dictability. The out-of-sample R^2 for annual returns rises steadily, peaking at 10.46% when $\lambda = 0.3$, a 37% increase relative to the benchmark. While larger λ values beyond 0.3 result in a gradual decline in R^2, it is noteworthy that even at $\lambda = 1.0$, the return forecasting accuracy remains above the benchmark case (9.37% versus 7.63%), demonstrating that the integration of consensus information provides robust predictive gains across all tested settings.

The consensus approximation results provide further insight into this regularization effect. Among the nine consensus variables, *Analyst Earnings per Share* dominates, achieving an R^2 of 71.43% at $\lambda = 1.0$, followed by strong performance in *EPS Forecast Dispersion* and *Analyst Optimism*. These results corroborate empirical findings that earnings estimates and their associated dispersion contain salient information about future returns (Diether et al., 2002; Jegadeesh et al., 2004). By contrast, *Change in Recommendation* exhibits persistently negative R^2, consistent with prior evidence of limited incremental predictive content in recommendation changes once earnings revisions are accounted for.

The consensus average R^2 increases monotonically from 7.33% at $\lambda = 0.1$ to 24.21% at $\lambda = 1.0$, indicating that the model becomes progressively better at reconstructing analyst consensus as λ grows. However, the modest decline in return R^2 beyond $\lambda = 0.3$ reflects the trade-off inherent in joint optimization; while higher λ emphasizes consensus approximation, return forecasting benefits most when consensus serves as an auxiliary concept rather than the dominant objective.

Figure 6 complements Table 1 by visualizing these trends. The left panel shows how return predictability improves sharply with the introduction of consensus learning, peaks around $\lambda = 0.3$-0.4, and then tapers slightly while remaining above the benchmark even at $\lambda = 1.0$. The right panel demonstrates the monotonic improvement in consensus approximation with increasing λ, eventually plateauing near 24%. Together, these panels illustrate the trade-off, where moderate λ balances return prediction and consensus learning most effectively, while larger λ values shift focus toward consensus reconstruction.

[Insert Figure 6 here]

To further examine robustness, Figure 7 presents results from the expanding window evaluation, comparing a naïve neural network ($\lambda = 0$) to the best-performing CB-APM model ($\lambda = 0.42$). The naïve network exhibits volatile performance, including negative R^2 in early periods (e.g.,

2013) and isolated spikes (e.g., 2020). By contrast, CB-APM delivers consistently positive return R^2 across nearly all periods. Notably, improvements are most pronounced in periods where the naïve model already performed well (e.g., 2020–2021), indicating that CB-APM amplifies return-predictive signals in favorable regimes while mitigating underperformance in weaker ones (e.g., 2013 and 2017).

[Insert Figure 7 here]

The consensus approximation results (red bars) in the right panel further underscore CB-APM's stability. Consensus R^2 remains high across all testing periods, demonstrating that the concept-bottleneck captures analysts' aggregated expectations in a time-consistent manner. This stability contributes to return forecasting accuracy without overfitting to specific regimes, yielding gains across both tranquil periods (e.g., 2014–2018) and turbulent intervals such as the 2020 pandemic shock.

Collectively, these results validate the core design of CB-APM that by incorporating consensus learning as a concept-bottleneck enhances return prediction while retaining interpretability. The model's ability to achieve robust gains across different market environments underscores both its practical relevance under realistic, expanding-window evaluation and its theoretical grounding in analyst-driven information aggregation.

While the out-of-sample R^2 metrics directly capture forecasting accuracy, they do not reveal how the joint loss function in equation (4) balances return prediction and consensus approximation during training. To address this, Internet Appendix C.2 provides additional evidence on the optimization dynamics of CB-APM by reporting the in-sample MSE, which demonstrates that, at short horizons, increasing λ introduces the expected trade-off between predictive accuracy and consensus reconstruction, whereas at longer horizons the two objectives reinforce each other, yielding what we term an interpretability-accuracy amplification effect.

5.2 Portfolio-based pricing validation

We conduct further empirical analysis of the CB-APM by examining its economic implications through portfolio-level tests. While the preceding sections evaluated the model's predictive and explanatory power using out-of-sample R^2 metrics, these statistical measures alone do not reveal

whether the predicted returns merely reflect transitory noise. Portfolio-based analyses provide a more direct and economically interpretable assessment of model performance by linking cross-sectional predictions to realized investment payoffs. In particular, if the CB-APM successfully extracts a priced component of expected returns from the consensus structure, portfolios formed on its predictions should yield monotonic and persistent return differentials across quantiles.

Our portfolio analysis proceeds in three steps. First, we perform single-sort tests that rank stocks by CB-APM-predicted annual returns to evaluate the model's raw cross-sectional discriminating power. These tests quantify whether higher model-implied expected returns translate into higher realized payoffs and whether the strength of this relationship varies with the degree of consensus regularization. Second, we conduct double-sort analyses that jointly sort stocks by both predicted returns and consensus variables to examine how the model's inferred expectations interact with, and potentially refine, traditional analyst forecasts. Finally, we form long-short portfolios based on out-of-sample CB-APM predictions to evaluate their risk-adjusted performance relative to benchmark strategies and to assess the model's practical value from an asset-management perspective.

These portfolio-level analyses allow us to connect the statistical accuracy of the CB-APM to its economic relevance. By translating predictive signals into realized return differentials, we can determine whether the consensus-bottleneck representation captures genuinely priced information—consistent with rational risk compensation—or reflects transitory deviations unrelated to systematic risk premia. The following subsections detail the construction of these portfolio tests and discuss their empirical results.

5.2.1 Portfolio sorts on approximated consensuses and expected returns

For each month in the out-of-sample evaluation period, the CB-APM produces annual return forecasts for all stocks. Based on these out-of-sample predictions, stocks are ranked by their expected returns and assigned to ten value-weighted decile portfolios, ranging from the lowest (decile 1) to the highest (decile 10) predicted-return group. Portfolio constituents and weights are updated monthly as new forecasts become available, ensuring that portfolio formation relies exclusively on information observable at the prediction date. The realized monthly returns of each decile are then computed over the subsequent month, thereby evaluating the model's ex-ante forecasts in

a strictly out-of-sample setting.

[Insert Table 2 here]

The single-sort portfolio results in Table 2 reinforce the predictive validity of the CB-APM framework in the cross-section of returns. Average realized returns increase monotonically from the lowest to the highest predicted-return decile, with the bottom portfolios consistently yielding negative returns and the top portfolios earning approximately 1.3% per month. The resulting high-minus-low (H–L) spreads range from 1.64% for the naïve neural network ($\lambda = 0$) to around 2.3% for regularized CB-APM specifications ($\lambda \geq 0.3$). This progressive widening of the return differential highlights the model's ability to produce more economically meaningful and stable return rankings as the degree of consensus regularization increases. Beyond the level effects, the distribution of decile returns also becomes smoother and more monotonic as λ rises, suggesting that the bottleneck constraint mitigates noise in the model-implied expected returns.

The patterns in portfolio payoffs align closely with the out-of-sample performance metrics reported in Table 1. While the predictive R^2 for stock returns peaks around 10% and remains relatively stable across higher λ values, the R^2 for consensus variable approximation improves dramatically—from roughly 7% at $\lambda = 0.1$ to over 24% at $\lambda = 1.0$. This joint evidence implies that better recovery of analysts' consensus structure translates into more reliable expected-return forecasts. In other words, the improvement in cross-sectional pricing performance—as captured by the H–L spread—parallels the enhanced interpretability and generalization observed in the consensus approximation task. Together, the results indicate that the consensus-bottleneck regularization enables the model to balance flexibility and economic discipline, yielding forecasts that are both interpretable and empirically potent in explaining the cross-section of returns.

To further examine the pricing content of the CB-APM forecasts, we conduct double-sorting analyses jointly based on the model-implied expected returns and the analyst-based variable, *Analyst earnings per share* (*FEPS*). Specifically, at each month in the out-of-sample evaluation period, all stocks are first grouped into quintiles according to their CB-APM-approximated *FEPS* levels, and within each *FEPS* group, stocks are independently sorted into quintiles based on their CB-APM-predicted annual returns. This procedure produces 5×5 portfolios that are rebalanced monthly, ensuring that both the sorting and performance evaluation rely exclusively on information

available at the prediction date. The resulting portfolio returns thus reflect a purely out-of-sample assessment of the model's joint ordering power. The bottom and rightmost rows of each panel report the high-minus-low (H–L) spreads along the predicted-return and consensus dimensions, respectively, quantifying the incremental explanatory power of the CB-APM forecasts conditional on *FEPS*.

[Insert Table 3 here]

The results in Table 3 reveal that the CB-APM produces economically meaningful return spreads across both sorting dimensions while uncovering nontrivial interactions between model-implied expected returns and analysts' earnings expectations. The *Analyst earnings per share* (*FEPS*) measure corresponds to the most recent consensus forecast of next-fiscal-year earnings per share reported in I/B/E/S, serving as a standardized proxy for analysts' expectations about firms' future profitability. Empirical evidence from Cen and Zhang demonstrates that FEPS carries substantial predictive power for future stock returns: firms with higher forecasted earnings consistently earn higher subsequent returns even after controlling for market, size, value, and momentum factors. This return premium is concentrated among smaller and neglected firms and persists over long horizons without subsequent reversal, suggesting that it cannot be attributed to conventional risk exposures. Rather, Cen and Zhang interpret the FEPS effect as arising from systematic errors in expectations—investors tend to overvalue firms when analysts' earnings forecasts are pessimistic vice versa. Accordingly, while FEPS reflects valuable information about expected profitability, it also embeds behavioral biases and mispricing components. The double-sort design therefore provides a natural setting to assess whether the CB-APM's consensus-bottleneck isolates the priced component of these expectations, refining the raw analyst forecasts into a signal that aligns more closely with risk-adjusted return differentials.

Across all λ settings, mean realized returns increase monotonically from the lower-left corner (low *FEPS*, low predicted return) to the upper-right corner (high *FEPS*, high predicted return), indicating strong joint ordering power. Within each *FEPS* quintile, portfolios sorted on CB-APM predictions exhibit clear monotonicity, with H–L spreads ranging from roughly 0.9% to 2.5% per month. These spreads peak around intermediate regularization strengths ($\lambda = 0.3$-0.6), consistent with the notion that moderate consensus constraints yield an optimal trade-off between model

flexibility and economic discipline. For small λ, excessive flexibility introduces predictive noise, while for large λ (> 0.8), over-regularization attenuates cross-sectional differentiation.

More importantly, the cross-sectional pattern along the *FEPS* dimension exhibits a distinct reversal structure that sheds light on how the CB-APM internalizes analysts' earnings expectations. The H–L spreads of *FEPS* are positive among stocks with low CB-APM-predicted returns but become negative among those with high predicted returns. This inversion implies that, conditional on the model's expected-return signal, firms with high analyst-forecasted earnings (high *FEPS*) outperform in low-return segments but underperform in high-return segments. In parallel, the magnitude of the expected-return H–L spread declines systematically as one moves from low to high *FEPS* quintiles, indicating that CB-APM's return differentiation is strongest where analyst optimism is weakest.

This reversal pattern complements and extends the empirical regularities documented by Cen and Zhang, who show that firms with higher forecasted earnings per share earn persistently higher future returns, while market reactions to low-*FEPS* forecasts exhibit larger expectation errors and stronger optimism bias. Consistent with this evidence, the CB-APM identifies similar expectation-driven asymmetries but goes further by internalizing them within its consensus-bottleneck. Specifically, the model learns to separate the information about future profitability embedded in earnings forecasts from the heterogeneous market responses they trigger. As a result, the cross-sectional reversals observed in the double-sorted portfolios reflect not a contradiction but an adjustment: the CB-APM retains the predictive content of *FEPS* while capturing how market optimism and pessimism jointly shape return outcomes.

Together, these findings demonstrate that the CB-APM forecasts complement rather than replicate the informational content of *FEPS*. The model's consensus-bottleneck captures the latent component of analysts' expectations that aligns with priced risk, while filtering out the noise associated with systematic optimism. The resulting reversal and attenuation patterns in Table 3 thus provide direct evidence that the CB-APM transforms forecasted earnings information into a state-dependent pricing signal that refines, rather than contradicts, the analysts' consensus view.

5.2.2 Long-short portfolio performance

We construct the long-short portfolio as follows. The first step involves generating monthly predicted annual returns for each stock within the universe from CB-APM. These predicted returns are then ranked from highest to lowest and sorted into deciles based on their values. Subsequently, a long portfolio is formed by purchasing the top 10% of stocks with the highest predicted returns, while concurrently establishing a short portfolio by selling the bottom 10% of stocks with the lowest predicted returns. Weighting of the stocks within each portfolio is executed based on the size of the firm, ensuring that larger firms are assigned higher weights. Then the long-short portfolio is rebalanced every month to uphold the desired exposure and maintain alignment with the initial strategy.

The long-short construction directly operationalizes the cross-sectional ordering evidence reported in Tables 2. The monotonic increase in realized returns across predicted-return deciles translates naturally into economically significant long-short spreads.

To evaluate the risk-adjusted performance of the CB-APM portfolio, we compute seven portfolio metrics: monthly mean log return, standard deviation, cumulative log return, annualized Sharpe ratio, maximum one-month loss, maximum drawdown, and turnover rate. Monthly mean and cumulative returns quantify the overall profitability of the model, while the Sharpe ratio measures risk-adjusted performance by relating expected excess returns to return volatility. Maximum one-month loss and maximum drawdown capture downside risk by quantifying the worst historical losses, both in single periods and cumulatively. Finally, portfolio turnover measures the degree of portfolio rebalancing activity, which is directly linked to transaction costs and practical implementability.

Maximum drawdown (Max DD) is defined as the largest cumulative loss from a historical peak in portfolio wealth:

$$\text{Max DD} = \max_{t \in T} \left(1 - \frac{W_t}{\max_{\tau \leq t} W_\tau} \right), \qquad W_t = \prod_{\tau=1}^{t} (1 + R_\tau),$$

where W_t denotes cumulative portfolio wealth at time t. This measure captures the worst peak-to-trough decline experienced over the sample period.

Portfolio turnover is calculated as,

$$\text{Turnover } = \frac{1}{T_r} \sum_{t=1}^{T_r} \left(\sum_{i=1}^{N} \left| w_{i,t+1} - \frac{w_{i,t}\,(1 + R_{i,t+1})}{1 + \sum_{j=1}^{N} w_{j,t} R_{j,t+1}} \right| \right),$$

where $w_{i,t}$ denotes the portfolio weight of asset i at time t, $R_{i,t}$ is its arithmetic monthly return, and $T_r \subset T$ denotes the set of rebalancing dates.

Portfolio positions are formed using CB-APM's out-of-sample return forecasts, allowing the portfolio tests to evaluate genuine real-time predictability over a long-horizon target.

[Insert Table 4 here]

The portfolio performance results in Table 4 mirror the statistical improvements in predictive and explanatory performance documented in Table 1. As the hyperparameter λ increases to moderate values around 0.3–0.4, both out-of-sample return R^2 and consensus-approximation accuracy rise sharply, and this improvement translates directly into superior realized portfolio returns. Mean monthly log returns climb from 1.53% at $\lambda = 0$ to 2.20% at $\lambda = 0.3$, while the annualized Sharpe ratio concurrently increases from 1.10 to 1.44. This near one-to-one correspondence between predictive power and portfolio profitability substantiates the economic value of the consensus-bottleneck: the same mechanism that refines predictive signal extraction in-sample also enhances risk-adjusted returns out-of-sample.

Beyond moderate λ values, both predictive and portfolio metrics exhibit mild flattening, as excessive weighting on consensus reconstruction ($\lambda > 0.4$) marginally reduces return R^2 and diminishes economic gains. This pattern implies a practical upper bound to interpretability regularization, beyond which the model overemphasizes consensus consistency at the expense of direct return optimization. Nonetheless, even at high λ values, performance remains consistently above the benchmark, confirming that consensus learning contributes persistently to economically meaningful predictability rather than statistical overfitting.

Risk profiles exhibit a moderate but economically intuitive trade-off between profitability and downside exposure. As λ increases to 0.3–0.4, maximum one-month losses rise slightly relative to the naïve network ($\lambda = 0$), while remaining of similar magnitude at $\lambda = 0.3$, which yields the highest Sharpe ratio. Maximum drawdowns, by contrast, are consistently lower than those

of the S&P 500 benchmark—staying below 21% versus the market's 25%—indicating that CB-APM's consensus-regularized predictions generate smoother long-term wealth trajectories. The modest increase in short-horizon losses is more than compensated by the substantial improvement in mean return and Sharpe ratio, implying enhanced efficiency on a risk-adjusted basis. Overall, the co-movement of predictive R^2, Sharpe ratios, and drawdown behavior captures an economically meaningful balance between return amplification and risk containment, reflecting the emergence of stable, consensus-aligned risk premia rather than transient noise-fitting effects.

Portfolio turnover remains high—approximately 60% per month—which is consistent with the characteristics of complex nonlinear architectures. This observation aligns with the findings of Gu et al. (2020), suggesting that neural-network-based return predictors typically produce higher turnover than linear or tree-based models due to their greater sensitivity to small shifts in cross-sectional signals. While Kelly et al. (2024) argue that out-of-sample predictive R^2 and Sharpe ratios of characteristics-sorted portfolios may not always constitute decisive evidence of pricing relevance, the convergence of both statistical and economic measures in CB-APM suggests that its latent consensus components capture systematically priced information that conventional deep learning frameworks fail to isolate. Together, these results affirm that CB-APM's consensus-bottleneck not only improves explanatory power but also yields tangible, risk-adjusted portfolio benefits, linking interpretability and profitability within a unified empirical asset-pricing framework.

[Insert Figure 9 here]

Figure 9 visualizes the cumulative out-of-sample performance of CB-APM long-short portfolios across different regularization strengths λ. All neural-network portfolios substantially outperform the S&P 500 buy-and-hold benchmark (black dashed line), demonstrating that the model's predictive signals translate into economically meaningful excess returns. The naïve network ($\lambda = 0$, purple line) already yields notable outperformance relative to the market, yet introducing the consensus-bottleneck regularization ($\lambda > 0$) substantially elevates cumulative returns. Portfolio performance improves sharply up to $\lambda \approx 0.3$, after which cumulative returns remain at a comparably high level with minor oscillations across subsequent λ values. The best-performing specification at $\lambda = 1.0$ represents a continuation of this high-return plateau rather than a strict monotonic gain, highlighting the robustness of CB-APM's economic performance across a wide range of regularization

intensities. This stability suggests that consensus regularization consistently enhances the model's predictive and economic relevance without overfitting to a narrow hyperparameter regime.

The figure further highlights the temporal robustness of CB-APM's performance. Even during adverse market conditions—notably the 2020 downturn—consensus-regularized portfolios experience smaller and more rapidly recovered drawdowns relative to both the market and the unregularized model, reflecting smoother wealth accumulation and improved resilience to macro shocks. The consistent separation between the consensus-based portfolios and the S&P 500 benchmark indicates that the learned consensus representations capture priced information that is both persistent and broadly exploitable.

6 Dissecting Approximated Consensuses

The CB-APM framework is designed not only to forecast risk premia but also to provide a transparent interpretation of how firm- and macro-level information maps into priced return variation. Unlike most machine-learning predictors—which typically compress characteristics into opaque nonlinear transformations—the CB-APM architecture explicitly separates two economic mechanisms: (i) a nonlinear mapping that synthesizes the high-dimensional information set into consensus-like latent expectations, and (ii) a final linear stage that maps these expectations into forecasts of future returns. This structural decomposition allows the consensus layer to be interpreted as a set of economically meaningful conditional expectations, while the final linear layer mirrors the role of factor loadings in a traditional cross-sectional model.

[Insert Figure 8 here]

Figure 8 visualizes the estimated prediction-layer coefficients at ($\lambda = 1$), computed using expanding training windows.[11] Each coefficient reflects the model's inferred sensitivity of expected returns to a given consensus element, while the color shading indicates the corresponding out-of-sample R^2 for consensus approximation. Because the prediction module is linear, these coefficients admit a familiar interpretation: they reveal the direction and magnitude with which each consensus

[11] We focus on $\lambda = 1$ because it delivers the highest out-of-sample R^2 for consensus approximation. Analyzing the most accurate consensus-reconstruction specification provides the clearest window into how CB-APM translates analyst information into interpretable pricing components.

dimension influences expected returns, analogous to factor loadings in conventional asset-pricing regressions.

Several patterns emerge. First, sentiment-oriented variables such as *Analyst Optimism* load positively and persistently, indicating that firms with stronger analyst sentiment are assigned higher expected-return forecasts. In contrast, variables reflecting recommendation changes or forecast revisions often load negatively, suggesting that optimistic updates embed short-lived overreaction that subsequently reverses. Second, consensus dimensions that the model reconstructs more accurately—particularly dispersion- and accrual-related variables—tend to receive larger-magnitude coefficients. This alignment between approximation quality and economic relevance implies that the CB-APM's interpretive layer concentrates information in the dimensions where analyst signals are both reliably reconstructable and strongly predictive of return heterogeneity.

These observations underscore an important conceptual feature: the interpretable consensus layer can be evaluated independently of the model's nonlinear feature-extraction stage. Once the consensus mapping is estimated, the subsequent linear relation

$$\hat{R}_{i,t+h} = a + \boldsymbol{b}^\top \hat{C}_{i,t}$$

can be analyzed using the same tools employed to study traditional factor models. This allows us to examine, in a transparent and economically interpretable manner, whether the learned consensus dimensions behave like priced sources of return variation or simply capture information-based heterogeneity unrelated to systematic risk.

To formalize this connection, we align our empirical strategy with standard asset-pricing methodology and implement two complementary tests. First, we estimate pooled panel OLS regressions of annual stock returns on either raw analyst consensus variables or their CB-APM–inferred counterparts, thereby quantifying the incremental explanatory content gained through the consensus-bottleneck transformation. Second, we examine whether factor-mimicking portfolios—constructed from the consensus dimensions via decile sorts—span the stochastic discount factor by applying the Gibbons–Ross–Shanken (GRS) test for mean–variance efficiency (Gibbons et al., 1989). These analyses serve a dual purpose: they link the interpretability of CB-APM's consensus layer to established empirical asset-pricing tools, and they enable a direct assessment of whether the machine-inferred

beliefs embody priced economic content beyond what is observable from raw analyst forecasts.

6.1 Comparative regression analysis

Having established that CB-APM's interpretable layer produces economically meaningful consensus coefficients, we examine whether variations in the CB-APM–implied consensus translate into priced differences in expected annual returns, following the empirical design of standard asset-pricing regressions. These regressions are not intended as structural pricing tests; rather, they serve as diagnostic tools that evaluate whether the model's consensus representations capture priced variation more effectively than the raw analyst signals.

Table 5 reports pooled OLS regressions that relate future annual stock returns to either raw analyst consensus variables or CB-APM-implied consensus at ($\lambda = 1$). We estimate the following pooled panel regression:

$$R_{i,t+h} = a^{(C)} + \boldsymbol{b}^{(C)\top} C_{i,t} + \varepsilon^{(C)}_{i,t+h} \quad \text{or} \quad R_{i,t+h} = a^{(\hat{C})} + \boldsymbol{b}^{(\hat{C})\top} \hat{C}_{i,t} + \varepsilon^{(\hat{C})}_{i,t+h}. \tag{7}$$

For the analysis, we set $h = 12$ to focus on annual return predictability, thereby aligning the return horizon with prior empirical studies discussed in the preceding sections. The model is estimated on the stacked cross-section of firm-month (i, t) observations to obtain a time-invariant coefficient vector $\widehat{\boldsymbol{b}}$. Inference is based on heteroskedasticity-robust covariance estimation tailored to the dependent variable's structure. To handle an overlapping long-horizon return, we compute Driscoll–Kraay (kernel HAC) standard errors with a Bartlett kernel and an eleven-month bandwidth, which are robust to heteroskedasticity, cross-sectional dependence, and and the serial correlation induced by overlapping observations (Driscoll and Kraay, 1998; Newey and West, 1986; Hodrick, 1992).

Panel A reports coefficient estimates, t-statistics, and a variable-level fit measure; Panel B summarizes the intercept and overall adjusted R^2. The comparison isolates the incremental explanatory content obtained when analyst information is first synthesized by CB-APM's consensus-bottleneck and then mapped linearly into expected returns.

[Insert Table 5 here]

The pooled OLS regression using raw analyst consensus variables yields limited explanatory power, with an adjusted R^2 of just 0.40%. Most predictors exhibit weak statistical significance;

for example, *EPS forecast revision* and *Earnings forecast revisions* produce t-statistics of -1.31 and -0.62, respectively, with neither variable exhibiting meaningful predictive content. Although *Change in recommendation* ($t = 11.54$) and *Change in Forecast and Accrual* ($t = 8.57$) are statistically significant at the 1% level, their estimated effects are modest in magnitude, and the overall model fit remains poor.

By contrast, the regression using CB-APM-inferred consensus achieves a substantially higher adjusted R^2 of 8.35%, representing more than a twentyfold improvement in explanatory power. A few key coefficients also reverse in sign relative to their raw counterparts. For instance, the coefficient on *Change in recommendation* shifts from $+0.0307$ ($t = 11.54$) to -3.9080 ($t = -6.67$), while *EPS Forecast Dispersion* turns from $+0.0076$ ($t = 0.59$) to -0.6263 ($t = -4.87$). In addition, *Analyst earnings per share* becomes strongly positive and significant ($t = 4.46$), whereas *Analyst Value* becomes significantly negative ($t = -2.75$). These shifts suggest that CB-APM extracts transformed representations that encode economically distinct pricing content beyond the raw analyst signals.

While the CB-APM-inferred consensus variables deliver substantial improvements in explanatory power, caution is warranted in interpreting their individual coefficients. As shown in Table 5, consensus variables with low predictor-level approximation R^2 values often exhibit coefficient patterns that diverge from those estimated using raw consensus inputs. For example, *Change in recommendation*, which has one of the lowest approximation R^2 values, exhibits a pronounced sign reversal, while *Change in Forecast and Accrual* weakens substantially in magnitude. This pattern highlights that the CB-APM approximations are not one-to-one reconstructions of analyst beliefs but rather encode transformed features with distinct pricing implications.

Consequently, interpretability must be grounded in a dual-lens framework. The consensus-level approximation R^2, previously reported in Table 1, reflects the degree to which a model-inferred variable aligns with its human-interpretable counterpart, whereas the coefficient estimate from the return regression captures the economic relevance of that signal. Coefficients associated with well-approximated variables are more directly interpretable as refinements of analyst expectations, whereas those tied to poorly reconstructed signals likely reflect alternative representations or reweightings learned by the model. Thus, proper interpretation requires joint consideration of both approximation fidelity and return sensitivity rather than treating coefficients in isolation.

Importantly, this improvement follows directly from the design of the framework, which trains the approximated consensus layer under a joint objective that simultaneously targets return prediction and consensus reconstruction. By doing so, the model synthesizes information from a wide set of firm-level characteristics and macroeconomic variables into consensus features that retain risk-relevant content while reducing noise. Although the reported t-statistics primarily capture in-sample explanatory strength and are not intended for direct investment use due to inherent look-ahead bias, they underscore that the approximated consensus simultaneously explains both realized analyst consensus and future returns. This dual property makes the learned consensus features a rich source of information that merits closer examination beyond forecasting alone.

Taken together, the results indicate that CB-APM's consensus module extracts signals that are both more informative and more economically meaningful than raw analyst inputs. Although the reported regressions are estimated in-sample and do not directly measure out-of-sample predictive accuracy, they nonetheless support the model's central objective: to learn interpretable latent representations that jointly capture analyst expectations and priced return variation. Proper interpretation of these results requires concurrent evaluation of (i) approximation R^2, which gauges the alignment between model-inferred signals and observable analyst variables, and (ii) coefficient sign and magnitude, which reflect the economic relevance of each signal for cross-sectional return prediction.

6.2 Pricing error of test assets

The preceding analysis establishes that the CB-APM's interpretable consensus layer captures return-relevant structure in the cross-section of individual stocks. We now examine whether these signals possess asset-pricing content when evaluated through the lens of linear factor models. Specifically, we assess (i) whether the latent representations learned by CB-APM[12] can serve as risk factors capable of pricing standard benchmark portfolios, and (ii) whether traditional factor models can price the return patterns implied by the CB-APM's predictions and consensus-based characteris-

[12]It is important to point out that the consensus representations learned by the CB-APM are not designed to approximate the span of the stochastic discount factor (SDF). Rather, the architecture learns a set of conditional expectation operators that map firm characteristics and macroeconomic conditions into consensus-like forecasts of future fundamentals. These latent expectations summarize belief-based or information-based heterogeneity, not compensated sources of systematic risk. Consequently, the CB-APM should not be expected to replicate the factor structure implicit in linear SDF models; instead, its consensus layer provides an economically interpretable decomposition of expected returns that is complementary to—rather than a substitute for—the traditional factor space.

tics. To do so, we employ the multivariate Gibbons–Ross–Shanken (GRS) test (Gibbons et al., 1989), which jointly evaluates whether the intercepts ($\boldsymbol{\alpha}$) in time-series regressions are statistically different from zero.

We consider three sets of standard test portfolios widely used in empirical asset pricing: the Fama–French 25 portfolios sorted on size and book-to-market ratio (5×5), the 25 portfolios sorted on size and momentum, and the 30 value-weighted industry portfolios. These portfolios span well-known sources of cross-sectional variation linked to value, momentum, and industry structure, and thus provide a benchmark for evaluating alternative factor models. As reference models, we estimate the CAPM, the Fama–French three-factor model (FF3), the Carhart four-factor model, the Fama–French five-factor model (FF5), and the Fama–French six-factor model (FF6). All models are estimated using monthly excess returns, and all results are reported in-sample to maintain comparability with the standard evaluation framework in the factor-pricing literature.

A central element of the empirical design is the construction of tradable portfolios that proxy for the consensus signals extracted by the CB-APM. Because the model produces firm-level consensus measures rather than aggregate time-series factors, we translate each consensus dimension into a value-weighted long–short portfolio by sorting firms into deciles and taking the return spread between the highest and lowest deciles. This approach parallels the construction of empirically traded factors such as HML or UMD and yields a set of zero-investment portfolios whose returns reflect cross-sectional variation in the corresponding consensus dimension. These portfolios provide a tractable representation of the CB-APM signals within a traditional factor-pricing framework and permit a direct comparison with benchmark linear factor models using GRS tests.

Importantly, the CB-APM portfolios used in these tests are constructed from the same model configuration employed in the empirical return-forecasting exercise. That is, we apply the trained CB-APM—optimized to forecast annual excess returns—to generate firm-level predicted returns and consensus representations, which are then used to form sorted portfolios and factor-mimicking returns. This design ensures coherence across empirical sections: the factor-pricing analysis evaluates the economic content of the very signals that the CB-APM learns to use for long-horizon prediction.

We conduct three complementary GRS exercises. First, we evaluate whether the CB-APM factor-mimicking portfolios can jointly price the benchmark 25– and 30–portfolio test assets. Suc-

cessful pricing performance would indicate that the consensus-based signals span systematic risks similar to those captured by traditional factors. Second, we form decile portfolios based on CB-APM predicted returns and test whether standard factor models can explain their realized returns. This analysis assesses whether the return patterns generated by the model are incremental to the span of existing factors. Third, we construct decile portfolios sorted on each individual consensus dimension and examine whether traditional models can price these portfolios. This final exercise isolates which consensus channels are most and least aligned with traditional factor structures.

Each specification is evaluated using the GRS F-statistic, its associated p-value, and mean absolute and root-mean-squared pricing errors. All results are computed in-sample, consistent with empirical asset-pricing conventions in which factor-pricing tests focus on explaining cross-sectional return patterns rather than forecasting performance. Together, these exercises provide a comprehensive assessment of whether the consensus representations learned by the CB-APM contain distinct factor-pricing information or whether their explanatory power is largely captured by established benchmark models.

Tables 6–8 present a comprehensive set of in-sample Gibbons–Ross–Shanken (GRS) tests evaluating the pricing performance of the CB-APM relative to conventional factor models. Across all tests, the GRS F-statistic assesses the joint null hypothesis that all pricing errors ($\boldsymbol{\alpha}$) are zero, such that lower F-statistics and higher p-values indicate superior mean–variance efficiency. The accompanying mean absolute and root-mean-squared (RMS) alphas summarize the magnitude of mispricing across the corresponding test assets.

[Insert Table 6 here]

Panels A–C of Table 6 evaluate whether the CB-APM's consensus-based factor-mimicking portfolios can price the returns of the Fama–French 25 size–book-to-market portfolios, the 25 size–momentum portfolios, and the 30 industry portfolios. Across these benchmarks, the CB-APM factors deliver GRS statistics that are broadly comparable to those of standard models, but they remain somewhat higher than the FF5 and FF6 specifications. Mean and RMS pricing errors are likewise modest yet consistently larger than those generated by traditional factor structures. These results indicate that the consensus-based factors span meaningful components of systematic return variation, but not the full set captured by benchmark style factors. This is consistent with evidence

that only a limited number of characteristic-based factors are strongly priced in the cross-section, while many signals are redundant or weakly informative (Kozak et al., 2020). Within this environment, the CB-APM factors behave as an additional block of characteristic-sorted portfolios that contributes incremental explanatory variation without supplanting the canonical factor structure.

[Insert Table 7 here]

Table 7 examines whether traditional factor models can jointly price decile portfolios formed on the CB-APM's predicted return scores. When the consensus bottleneck is weak (small λ), conventional factor models achieve moderate GRS statistics and economically small pricing errors, suggesting that a substantial portion of the model's predictive content overlaps with standard style factors. As λ increases, however, the GRS statistics rise sharply and the joint null of zero pricing errors is rejected uniformly. This monotonic deterioration indicates that stronger reliance on the consensus bottleneck induces expected-return patterns that increasingly depart from the linear span of market, size, value, momentum, and profitability/investment factors. Conceptually, this aligns with evidence that machine-learning models often extract nonlinear or interaction-based transformations of firm characteristics that extend beyond linear factor structures (Freyberger et al., 2020; Gu et al., 2020). In particular, the CB-APM with a tight consensus constraint appears to generate forecasts that incorporate structured forms of return heterogeneity that are difficult to reconcile with the standard factor space.

[Insert Table 8 here]

Table 8 evaluates portfolios formed on the individual consensus signals at $\lambda = 1.0$. Several dimensions—most prominently *Analyst Value*, *Analyst Optimism*, and *Analyst Earnings per Share*—produce relatively low GRS statistics and economically small pricing errors, suggesting strong alignment between these inferred consensus measures and established factor structures. Forecast-based and dispersion-based dimensions (such as *EPS forecast dispersion* and related revisions) exhibit somewhat larger pricing errors, but even here the magnitudes remain concentrated in the range of a few basis points per month. These patterns reinforce the idea that much of the predictive information contained in analyst-derived consensus measures can be represented through low-dimensional combinations of characteristics, often with sparse or localized influence (Chinco

et al., 2019), while still accommodating nonlinear interactions and heterogeneous partitions (Bryzgalova et al., 2025). The CB-APM's consensus variables therefore fit naturally within the broader empirical finding that return-relevant structure can be extracted by compressing high-dimensional characteristics into well-organized representations.

The three sets of GRS tests reveal how the CB-APM relates to the traditional factor space. First, consensus-based factor-mimicking portfolios do not fully price the classic benchmark portfolios, which indicates that the latent consensus dimensions do not function as close substitutes for the core priced factors. Second, the ability of traditional factor models to price CB-APM–generated portfolios deteriorates as the consensus bottleneck becomes more stringent, implying that the model's predictive signals progressively move outside the span of standard linear characteristics. This behavior is consistent with the broader view that, while the priced dimension of the stochastic discount factor is relatively low, flexible methods can uncover structured forms of return heterogeneity that improve portfolio efficiency (Cong et al., 2025) without reproducing the canonical factors directly. Third, portfolios sorted on individual consensus dimensions exhibit moderate but non-negligible pricing errors, suggesting that the learned signals contain meaningful information about expected returns but do not themselves constitute a new stand-alone factor system.

Overall, the evidence positions the CB-APM as a complementary asset-pricing framework: it enhances cross-sectional return prediction by compressing analysts' heterogeneous beliefs into interpretable consensus signals that partially overlap with—but do not collapse onto—the priced dimensions emphasized in modern work on characteristic-based factor representations (e.g., Cochrane, 2011). At the same time, the results indicate that the CB-APM does not merely denoise or reweight analyst inputs. Instead, it isolates structured and economically relevant components of analyst-derived information that are priced in the cross-section. The model therefore reveals that analyst-based information contains priced elements that conventional factor models only partially span, and that the consensus bottleneck organizes these elements into an interpretable and economically coherent representation. This places the CB-APM within a growing line of research showing that belief-based or characteristic-based signals can be synthesized into low-dimensional, economically meaningful components without reproducing the canonical factor structure directly.

7 Conclusion

This study introduces the Consensus-Bottleneck Asset Pricing Model (CB-APM), a novel framework that integrates interpretable AI with empirical asset pricing. By embedding a concept-bottleneck architecture into a neural network, CB-APM not only achieves state-of-the-art predictive accuracy in cross-sectional stock return forecasts but also provides transparent insights into the role of analysts' consensus in shaping risk premiums. Our empirical results demonstrate that interpretability and performance are not inherently conflicting. CB-APM outperforms conventional deep learning benchmarks in long-horizon forecasts while preserving a clear, economically grounded structure. By linking machine learning's predictive capabilities with the theoretical underpinnings of financial economics, and by demonstrating that interpretable AI can yield both statistical and economic validity, this work offers a blueprint for building models that are both high-performing and aligned with established asset pricing principles.

The success of CB-APM highlights three key implications for empirical finance. First, interpretable neural architectures can reconcile the flexibility of machine learning with economic reasoning, enabling researchers to assess whether models capture meaningful risk factors rather than spurious correlations. Second, embedding interpretability directly within model design fosters transparency and trust, addressing the skepticism that often surrounds "black-box" methods in high-stakes financial applications. Third, by explicitly modeling analysts' consensus as a latent mediator between firm characteristics and returns, CB-APM sheds new light on how information aggregation mechanisms influence asset prices, aligning closely with rational expectations theory and empirical evidence on analyst behavior.

Future research can extend this framework in several promising directions. Incorporating additional economically meaningful bottlenecks—such as macroeconomic regimes or investor sentiment—could further disentangle the sources of risk premiums and strengthen the theoretical interpretability of model outputs. Addressing practical constraints, such as data latency in analyst consensus measures or improving computational efficiency for large-scale implementation, would enhance CB-APM's applicability in real-world investment contexts. More broadly, as the "factor zoo" continues to grow, interpretable frameworks like CB-APM will be instrumental in bridging data-driven discovery with economic theory, offering a structured approach to understanding how high-dimensional

predictors translate into priced information. By demonstrating that interpretable AI can achieve both predictive accuracy and theoretical coherence, this study lays the groundwork for a new generation of financially grounded machine learning models, advancing the study of asset pricing in both academic research and practical decision-making.

References

Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993.

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

Mark M Carhart. On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82, 1997.

Andrew Ang and Geert Bekaert. Stock return predictability: Is it there? *The Review of Financial Studies*, 20(3):651–707, 2007.

John Y Campbell and Samuel B Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531, 2008.

John H Cochrane. The dog that did not bark: A defense of return predictability. *The Review of Financial Studies*, 21(4):1533–1575, 2008.

John H Cochrane. Presidential address: Discount rates. *The Journal of finance*, 66(4):1047–1108, 2011.

Ivo Welch and Amit Goyal. A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508, 2008.

Jeremiah Green, John RM Hand, and X Frank Zhang. The supraview of return predictive signals. *Review of Accounting Studies*, 18:692–730, 2013.

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

Campbell R Harvey, Yan Liu, and Heqing Zhu. ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68, 2016.

Zhiguo He, Bryan Kelly, and Asaf Manela. Intermediary asset pricing: New evidence from many asset classes. *Journal of Financial Economics*, 126(1):1–35, 2017.

Jeremiah Green, John RM Hand, and X Frank Zhang. The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies*, 30(12): 4389–4436, 2017.

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

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370, 2020.

Joachim Freyberger, Andreas Neuhierl, and Michael Weber. Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377, 2020.

Leland Bybee, Bryan Kelly, and Yinan Su. Narrative asset pricing: Interpretable systematic risk factors from news text. *The Review of Financial Studies*, 36(12):4759–4787, 2023.

Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen. Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518, 2023.

Daniele Bianchi, Matthias Büchner, and Andrea Tamoni. Bond risk premiums with machine learning. *The Review of Financial Studies*, 34(2):1046–1089, 2021.

Patrick Jaquart, David Dann, and Christof Weinhardt. Short-term bitcoin market prediction via machine learning. *The journal of finance and data science*, 7:45–66, 2021.

Fan Fang, Waichung Chung, Carmine Ventre, Michail Basios, Leslie Kanthan, Lingbo Li, and Fan Wu. Ascertaining price formation in cryptocurrency markets with machine learning. *The European Journal of Finance*, 30(1):78–100, 2024.

Markus Leippold, Qian Wang, and Wenyu Zhou. Machine learning in the chinese stock market. *Journal of Financial Economics*, 145(2):64–82, 2022.

Bryan Kelly, Semyon Malamud, and Kangying Zhou. The virtue of complexity in return prediction. *The Journal of Finance*, 79(1):459–503, 2024.

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

Guanhao Feng, Jingyu He, Nicholas G Polson, and Jianeng Xu. Deep learning in characteristics-sorted factor models. *Journal of Financial and Quantitative Analysis*, pages 1–36, 2018.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450, 2021.

Robert C Merton. An intertemporal capital asset pricing model. *Econometrica: Journal of the Econometric Society*, pages 867–887, 1973.

Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Management Science*, 70(2):714–750, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

John F Muth. Rational expectations and the theory of price movements. *Econometrica: journal of the Econometric Society*, pages 315–335, 1961.

Michael C Lovell. Tests of the rational expectations hypothesis. *The American Economic Review*, 76(1):110–124, 1986.

Terence Lim. Rationality and analysts' forecast bias. *The journal of Finance*, 56(1):369–385, 2001.

Narasimhan Jegadeesh, Joonghyuk Kim, Susan D Krische, and Charles MC Lee. Analyzing the analysts: When do recommendations add value? *The journal of finance*, 59(3):1083–1124, 2004.

Karl B Diether, Christopher J Malloy, and Anna Scherbina. Differences of opinion and the cross section of stock returns. *The journal of finance*, 57(5):2113–2141, 2002.

Sorin Sorescu and Avanidhar Subrahmanyam. The cross section of analyst recommendations. *Journal of Financial and Quantitative Analysis*, 41(1):139–168, 2006.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

José Manuel Benítez, Juan Luis Castro, and Ignacio Requena. Are artificial neural networks black boxes? *IEEE Transactions on neural networks*, 8(5):1156–1164, 1997.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019a.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. *Advances in neural information processing systems*, 32, 2019b.

Charles R Nelson and Andrew F Siegel. Parsimonious modeling of yield curves. *Journal of business*, pages 473–489, 1987.

Robert Litterman. Common factors affecting bond returns. *Journal of fixed income*, pages 54–61, 1991.

Tony A Plate. Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using gaussian process models. *Behaviormetrika*, 26:29–50, 1999.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

Kent Daniel and Sheridan Titman. Evidence on the characteristics of cross sectional variation in stock returns. *the Journal of Finance*, 52(1):1–33, 1997.

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

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Andrew Y. Chen and Tom Zimmermann. Open source cross-sectional asset pricing. *Critical Finance Review*, 27(2):207–264, 2022.

Michael W McCracken and Serena Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

Sydney C Ludvigson and Serena Ng. The empirical risk–return relation: A factor analysis approach. *Journal of financial economics*, 83(1):171–222, 2007.

Jianqing Fan, Yuan Liao, and Weichen Wang. Projected principal component analysis in factor models. *Annals of statistics*, 44(1):219, 2016.

N Benjamin Erichson, Peng Zheng, Krithika Manohar, Steven L Brunton, J Nathan Kutz, and Aleksandr Y Aravkin. Sparse principal component analysis via variable projection. *SIAM Journal on Applied Mathematics*, 80(2):977–1002, 2020.

Ling Cen and Jie Zhang. Forecasted earnings per share and the cross section of expected stock returns. Technical report.

Michael R Gibbons, Stephen A Ross, and Jay Shanken. A test of the efficiency of a given portfolio. *Econometrica: Journal of the Econometric Society*, pages 1121–1152, 1989.

John C Driscoll and Aart C Kraay. Consistent covariance matrix estimation with spatially dependent panel data. *Review of economics and statistics*, 80(4):549–560, 1998.

Whitney K Newey and Kenneth D West. A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix. 1986.

Robert J Hodrick. Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *The Review of Financial Studies*, 5(3):357–386, 1992.

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

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

Svetlana Bryzgalova, Markus Pelger, and Jason Zhu. Forest through the trees: Building cross-sections of stock returns. *The Journal of Finance*, 80(5):2447–2506, 2025.

Lin William Cong, Guanhao Feng, Jingyu He, and Xin He. Growing the efficient frontier on panel trees. *Journal of Financial Economics*, 167:104024, 2025.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus), 2016.

Jonathan T. Barron. Continuously differentiable exponential linear units, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012a.

PyTorch: Reducelronplateau - pytorch 2.3.0 documentation. `https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html`.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr, 2013.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, 2012b.

Matthias Feurer and Frank Hutter. Hyperparameter optimization. *Automated machine learning: Methods, systems, challenges*, pages 3–33, 2019.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Table 1: Out-of-sample R^2 for stock return and consensus approximations.
This table reports monthly R^2 (%) of annual stock return estimation and analysts' consensus variable approximation for different λ settings. The out-of-sample R^2 is computed by concatenating realized values across all testing periods and comparing them with model predictions, thereby capturing performance over the full evaluation horizon. The Overall Results panel reports (i) the average R^2 across all consensus variables and (ii) the R^2 for annual stock return predictions based on the specified λ regularization parameter.

	Consensus Variables									Overall Results	
λ	EPS Forecast Revision	Change in Rec-ommend-ation	Change in Forecast & Accrual	Long vs short EPS Forecasts	Analyst Earnings per Share	EPS Forecast Dispersion	Earnings Forecast Revisions	Analyst Value	Analyst Optimisim	Consensus Average	Stock Returns
0	-	-	-	-	-	-	-	-	-	-	7.63
0.1	1.40	−0.24	1.52	3.04	22.60	11.70	5.14	9.72	11.11	7.33	9.49
0.2	2.45	−0.24	2.34	4.71	40.13	21.19	7.65	17.06	19.63	12.77	10.11
0.3	3.21	−0.25	2.92	5.91	50.96	27.25	9.81	22.22	24.97	16.33	10.46
0.4	3.67	−0.24	3.35	6.72	56.73	30.52	11.45	25.42	28.21	18.43	10.44
0.5	4.01	−0.23	3.67	7.36	61.11	33.03	12.69	28.16	30.73	20.06	10.23
0.6	4.3	−0.19	3.9	7.92	64.31	34.97	13.71	30.27	32.7	21.32	9.9
0.7	4.53	−0.18	4.14	8.33	66.82	36.4	14.48	31.99	34.13	22.29	9.77
0.8	4.71	−0.16	4.31	8.66	68.55	37.4	15.08	33.34	35.32	23.02	9.6
0.9	4.85	−0.17	4.5	8.89	70.39	38.44	15.73	34.65	36.53	23.76	9.51
1.0	4.97	−0.16	4.62	9.12	71.43	39.06	16.18	35.45	37.24	24.21	9.37

Note: Results are reported for a sampled subset of λ settings due to redundancy.

Table 2: Realized monthly returns of out-of-sample single-sorted portfolios across λ.
Each panel reports mean monthly realized returns (in percentage points) for monthly rebalanced decile portfolios, formed by sorting stocks on CB-APM-predicted annual returns. The bottom row (H–L) represents the spread between the highest- and lowest-decile portfolios.

λ	0.0	0.1	0.2	0.3	0.4	0.5
Low	−0.36	−0.70	−0.78	−0.96	−0.88	−0.92
2	−0.27	−0.24	−0.26	−0.24	−0.25	−0.23
3	0.06	−0.03	−0.11	−0.03	−0.16	−0.16
4	0.14	0.13	0.22	0.20	0.30	0.31
5	0.20	0.39	0.41	0.33	0.35	0.36
6	0.30	0.36	0.51	0.52	0.52	0.36
7	0.47	0.52	0.48	0.55	0.41	0.49
8	0.64	0.72	0.61	0.62	0.79	0.79
9	0.77	0.78	0.88	0.91	0.85	0.93
High	1.28	1.31	1.27	1.34	1.32	1.30
H–L	1.64	2.00	2.06	2.30	2.20	2.21

λ	0.6	0.7	0.8	0.9	1.0
Low	−0.91	−0.93	−0.92	−0.96	−0.94
2	−0.29	−0.29	−0.33	−0.27	−0.31
3	−0.03	0.05	0.03	−0.06	−0.07
4	0.29	0.22	0.30	0.25	0.27
5	0.41	0.41	0.35	0.39	0.40
6	0.32	0.27	0.29	0.37	0.34
7	0.50	0.49	0.44	0.44	0.51
8	0.72	0.85	0.85	0.84	0.79
9	0.93	0.79	0.87	0.89	0.85
High	1.29	1.38	1.35	1.35	1.40
H–L	2.20	2.31	2.27	2.31	2.34

Table 3: Realized monthly returns of out-of-sample double-sorted portfolios across λ.
Each panel reports mean monthly realized returns (in percentage points) for monthly rebalanced 5×5 portfolios sorted by the approximated *Analyst earning per share* ($E[FEPS]$, rows) and predicted annual returns ($E[R]$, columns), independently. H–L denotes the high-minus-low spread across the corresponding dimension.

Panel: $\lambda = 0.1$

	$E_t[R_{i,t+h}]$					
$E_t[FEPS_{i,t}]$	Low	2	3	4	High	H–L
Low	−0.65	−0.43	0.04	0.68	1.47	2.12
2	−0.31	0.03	0.38	0.30	0.63	0.94
3	−0.27	0.39	0.35	0.40	0.69	0.96
4	−0.11	0.17	0.45	0.33	0.86	0.98
High	0.28	0.45	0.46	0.70	0.81	0.53
H–L	0.93	0.88	0.41	0.02	−0.66	−1.59

Panel: $\lambda = 0.2$

	$E_t[R_{i,t+h}]$					
$E_t[FEPS_{i,t}]$	Low	2	3	4	High	H–L
Low	−0.73	−0.41	−0.04	0.67	1.54	2.27
2	−0.34	0.01	0.38	0.54	0.62	0.95
3	−0.27	0.27	0.47	0.45	0.72	0.99
4	−0.14	0.27	0.24	0.45	0.85	0.99
High	0.19	0.44	0.56	0.62	0.73	0.54
H–L	0.92	0.85	0.60	−0.05	−0.81	−1.73

Panel: $\lambda = 0.3$

	$E_t[R_{i,t+h}]$					
$E_t[FEPS_{i,t}]$	Low	2	3	4	High	H–L
Low	−0.90	−0.24	−0.04	0.61	1.66	2.55
2	−0.41	0.05	0.30	0.64	0.62	1.03
3	−0.40	0.25	0.35	0.50	0.78	1.18
4	−0.19	0.18	0.32	0.68	0.83	1.03
High	0.11	0.39	0.60	0.65	0.75	0.64
H–L	1.01	0.63	0.64	0.04	−0.90	−1.91

Panel: $\lambda = 0.4$

	$E_t[R_{i,t+h}]$					
$E_t[FEPS_{i,t}]$	Low	2	3	4	High	H–L
Low	−0.83	−0.40	−0.05	0.67	1.61	2.44
2	−0.40	0.03	0.30	0.57	0.71	1.11
3	−0.31	0.30	0.35	0.44	0.81	1.11
4	−0.19	0.18	0.38	0.57	0.89	1.08
High	0.14	0.36	0.51	0.72	0.73	0.59
H–L	0.97	0.76	0.57	0.05	−0.88	−1.86

Panel: $\lambda = 0.5$

	$E_t[R_{i,t+h}]$					
$E_t[FEPS_{i,t}]$	Low	2	3	4	High	H–L
Low	−0.77	−0.29	−0.01	0.34	1.68	2.45
2	−0.59	0.03	0.37	0.58	0.80	1.39
3	−0.26	0.24	0.42	0.52	0.86	1.12
4	−0.17	0.17	0.40	0.62	0.83	1.00

(cont'd on next page)

Table 3: Realized monthly returns of out-of-sample double-sorted portfolios across λ (cont'd).

High	0.02	0.34	0.49	0.68	0.77	0.75
H–L	0.79	0.63	0.49	0.34	−0.91	−1.70

Panel: $\boldsymbol{\lambda = 0.6}$

	$E_t[R_{i,t+h}]$					
$E_t[FEPS_{i,t}]$	Low	2	3	4	High	H–L
Low	−0.75	−0.31	−0.03	0.54	1.56	2.31
2	−0.55	0.02	0.38	0.55	0.89	1.44
3	−0.29	0.36	0.30	0.51	0.90	1.19
4	−0.25	0.20	0.39	0.60	0.81	1.06
High	0.04	0.42	0.48	0.62	0.73	0.69
H–L	0.79	0.73	0.51	0.08	−0.83	−1.62

Panel: $\boldsymbol{\lambda = 0.7}$

	$E_t[R_{i,t+h}]$					
$E_t[FEPS_{i,t}]$	Low	2	3	4	High	H–L
Low	−0.88	−0.27	0.12	0.58	1.51	2.39
2	−0.56	0.12	0.31	0.65	0.79	1.35
3	−0.23	0.26	0.27	0.50	0.94	1.17
4	−0.27	0.27	0.40	0.53	0.80	1.07
High	−0.04	0.43	0.46	0.70	0.72	0.75
H–L	0.84	0.69	0.33	0.11	−0.79	−1.63

Panel: $\boldsymbol{\lambda = 0.8}$

	$E_t[R_{i,t+h}]$					
$E_t[FEPS_{i,t}]$	Low	2	3	4	High	H–L
Low	−0.81	−0.27	0.09	0.56	1.57	2.38
2	−0.53	0.06	0.29	0.69	0.85	1.38
3	−0.23	0.20	0.33	0.51	0.92	1.15
4	−0.26	0.22	0.42	0.50	0.80	1.06
High	−0.05	0.45	0.46	0.61	0.75	0.80
H–L	0.76	0.72	0.37	0.04	−0.81	−1.58

Panel: $\boldsymbol{\lambda = 0.9}$

	$E_t[R_{i,t+h}]$					
$E_t[FEPS_{i,t}]$	Low	2	3	4	High	H–L
Low	−0.93	−0.08	−0.10	0.64	1.62	2.55
2	−0.62	0.16	0.32	0.55	0.92	1.54
3	−0.26	0.27	0.32	0.54	0.93	1.19
4	−0.30	0.16	0.49	0.57	0.76	1.05
High	−0.10	0.39	0.51	0.60	0.73	0.83
H–L	0.82	0.47	0.61	−0.05	−0.90	−1.72

Panel: $\boldsymbol{\lambda = 1.0}$

	$E_t[R_{i,t+h}]$					
$E_t[FEPS_{i,t}]$	Low	2	3	4	High	H–L
Low	−0.94	−0.16	−0.01	0.66	1.63	2.57
2	−0.64	0.23	0.31	0.57	0.90	1.53
3	−0.21	0.24	0.40	0.53	0.91	1.11
4	−0.42	0.21	0.41	0.67	0.71	1.13
High	−0.06	0.41	0.47	0.57	0.73	0.79
H–L	0.89	0.57	0.47	−0.09	−0.90	−1.79

Table 4: Out-of-sample portfolio performance of CB-APM long-short portfolios.
This table reports performance metrics for value-weighted CB-APM long-short portfolios under different hyperparameter λ. Mean ($\bar{r}$) and standard deviation ($\sigma(r)$) are computed from monthly log returns, and cumulative log return ($\sum_t r_t$) is aggregated over the full sample period. The Sharpe ratio ($\bar{R}/\sigma(R)$) is annualized using the standard $\sqrt{12}$ scaling, assuming a zero risk-free rate. Maximum one-month loss ($-\min(R)$) and maximum drawdown (Max DD) are expressed in percentage terms, while Turnover denotes the average monthly portfolio turnover. The S&P 500 index serves as a benchmark.

λ	$\bar{r}$	$\sigma(r)$	$\sum_t r_t$	$\bar{R}/\sigma(R)$	$-\min(R)$	Max DD	Turnover
0	0.0153	0.0528	1.8318	1.0997	9.8654	12.7337	58.2867
0.1	0.0187	0.0573	2.2488	1.2697	11.8965	12.8505	58.1035
0.2	0.0194	0.0600	2.3292	1.2630	14.9458	14.9458	58.8336
0.3	0.0220	0.0605	2.6347	1.4375	12.7858	13.4161	60.9016
0.4	0.0209	0.0632	2.5125	1.3051	18.6285	19.2519	61.0962
0.5	0.0211	0.0632	2.5325	1.3169	18.2476	20.1880	60.6515
0.6	0.0210	0.0636	2.5156	1.2992	18.6622	20.1824	60.9962
0.7	0.0219	0.0643	2.6332	1.3535	19.8222	21.2800	60.2769
0.8	0.0215	0.0631	2.5858	1.3496	18.3169	19.0805	60.6148
0.9	0.0220	0.0636	2.6423	1.3727	18.8771	19.7140	61.2379
1.0	0.0223	0.0642	2.6709	1.3766	18.9305	19.1723	60.7656
S&P 500	0.0083	0.0428	0.9903	0.7028	12.5119	24.7695	–

Note: r_t and R_t denote log and arithmetic returns, respectively, where $r_t = \ln(1 + R_t)$ and $R_t = e^{r_t} - 1$. Metrics based on r (e.g., $\bar{r}$, $\sigma(r)$, $\sum_t r_t$) are computed in log-return space for time additivity, whereas those based on R (e.g., $\bar{R}/\sigma(R)$, $-\min(R)$, and Max DD) use arithmetic returns to ensure interpretability in percentage terms. Turnover is defined as the average absolute change in portfolio weights between rebalancing dates. All portfolios are value-weighted to reflect firm-size heterogeneity.

Table 5: OLS regressions with raw versus model-inferred consensus.
This table reports pooled OLS regressions in which the dependent variable is the *annual* stock return $R_{i,t+12}$. We compare specifications that use raw analyst consensus variables to those that use CB-APM-inferred consensus estimates at ($\lambda = 1$), evaluated on the longest training set from the expanding-window procedure. The CB-APM consensus corresponds to the averaged output of an ensemble of models. Panel A reports coefficient estimates, t-statistics, and predictor-level R^2 for each variable, while Panel B summarizes the intercept, its standard error, and the overall in-sample adjusted R^2.

Panel A. Coefficients and t-statistics

	Raw Consensus		Approximated Consensus		
Variable	Coef.	t-stat.	R^2 (%)	Coef.	t-stat.
EPS forecast revision	−0.0049	−1.31	4.97	0.1164	0.38
Change in recommendation	0.0307	11.54 ***	−0.16	−3.9080	−6.67 ***
Change in Forecast and Accrual	0.0398	8.57 ***	4.62	0.3781	1.53
Long-vs-short EPS forecasts	0.0002	0.04	9.12	−0.0940	−0.87
Analyst earnings per share	−0.0240	−1.63	71.43	0.3174	4.46 ***
EPS Forecast Dispersion	0.0076	0.59	39.06	−0.6263	−4.87 ***
Earnings forecast revisions	−0.0068	−0.62	16.18	−0.3328	−1.82 *
Analyst Value	0.0146	0.88	35.45	−0.1364	−2.75 ***
Analyst Optimism	0.0216	2.37 **	37.24	0.1250	1.55

Panel B. Model summary statistics

	Raw Consensus	Approximated Consensus
Intercept	−0.0064	0.0005
SE of Intercept	0.0321	0.0159
In-Sample *adj*-R^2 (%)	0.40	8.35

Note: *** significance at the 1% level; ** significance at the 5% level; * significance at the 10% level. Standard errors are computed using the Driscoll–Kraay (kernel HAC) estimator with a Bartlett kernel and an eleven-month bandwidth, robust to heteroskedasticity, cross-sectional dependence, and the serial correlation induced by overlapping returns.

Table 6: GRS tests for CB-APM factor-mimicking portfolios versus standard factor models. This table reports in-sample Gibbons–Ross–Shanken (GRS) tests of mean–variance efficiency for competing factor models.The test assets include (i) the 25 size–book-to-market portfolios, (ii) the 25 size–momentum portfolios, and (iii) the 30 value-weighted industry portfolios. CB-APM factors are constructed by sorting stocks into deciles on each consensus dimension and taking the value-weighted return spread between the top and bottom deciles, yielding tradable factor-mimicking portfolios.Different values of λ correspond to CB-APM models trained under varying strengths of the consensus bottleneck.Each panel reports the GRS F-statistic, p-value, mean absolute and RMS pricing errors (monthly and annualized), and the number of factors (K). All statistics use monthly excess returns.

Panel A. 25 portfolios formed on size and book-to-market ratio

Factor Model	GRS F	p-value	Mean$\|\alpha\|$ (M)	Mean$\|\alpha\|$ (A)	RMSα (M)	RMSα (A)	K
CB-APM (λ=0.1)	4.091	0.00	0.0075	0.0946	0.0077	0.0969	9
CB-APM (λ=0.5)	4.069	0.00	0.0074	0.0925	0.0076	0.0947	9
CB-APM (λ=1.0)	4.042	0.00	0.0102	0.1292	0.0103	0.1307	9
CAPM	4.196	0.00	0.0016	0.0196	0.0021	0.0252	1
FF3	4.392	0.00	0.0014	0.0163	0.0018	0.0216	3
Carhart4	4.000	0.00	0.0013	0.0152	0.0016	0.0195	4
FF5	3.620	0.00	0.0013	0.0153	0.0016	0.0197	5
FF6	3.410	0.00	0.0012	0.0143	0.0015	0.0180	6

Panel B. 25 portfolios formed on size and momentum

Factor Model	GRS F	p-value	Mean$\|\alpha\|$ (M)	Mean$\|\alpha\|$ (A)	RMSα (M)	RMSα (A)	K
CB-APM (λ=0.1)	2.389	0.00	0.0075	0.0942	0.0078	0.0987	9
CB-APM (λ=0.5)	2.676	0.00	0.0073	0.0922	0.0078	0.0988	9
CB-APM (λ=1.0)	3.115	0.00	0.0104	0.1327	0.0106	0.1356	9
CAPM	2.270	0.00	0.0028	0.0329	0.0035	0.0411	1
FF3	2.329	0.00	0.0028	0.0331	0.0036	0.0424	3
Carhart4	2.135	0.00	0.0016	0.0193	0.0019	0.0228	4
FF5	2.070	0.00	0.0022	0.0257	0.0028	0.0332	5
FF6	1.970	0.00	0.0010	0.0127	0.0015	0.0179	6

Panel C. 30 industry portfolios

Factor Model	GRS F	p-value	Mean$\|\alpha\|$ (M)	Mean$\|\alpha\|$ (A)	RMSα (M)	RMSα (A)	K
CB-APM (λ=0.1)	1.517	0.05	0.0068	0.0849	0.0072	0.0905	9
CB-APM (λ=0.5)	1.461	0.06	0.0066	0.0827	0.0070	0.0883	9
CB-APM (λ=1.0)	1.644	0.02	0.0092	0.1166	0.0095	0.1201	9
CAPM	1.233	0.19	0.0023	0.0278	0.0031	0.0368	1
FF3	1.572	0.03	0.0026	0.0305	0.0032	0.0379	3
Carhart4	1.583	0.03	0.0024	0.0283	0.0029	0.0347	4
FF5	1.774	0.01	0.0030	0.0350	0.0039	0.0452	5
FF6	1.666	0.02	0.0026	0.0311	0.0035	0.0405	6

Note: The GRS F-statistic tests the null hypothesis that all pricing errors ($\boldsymbol{\alpha}$) are jointly zero. Mean$|\alpha|$ and RMSα denote mean absolute and root-mean-squared pricing errors, reported in monthly (M) and annualized (A) terms. p-values are rounded to two decimal places; values below 0.005 appear as 0.00.

Table 7: GRS tests for traditional factor models on CB-APM decile portfolios.
This table reports in-sample Gibbons–Ross–Shanken (GRS) tests applied to decile portfolios constructed from CB-APM predicted return scores. Each λ corresponds to a distinct CB-APM specification that generates the test assets. For each λ, we report the GRS F-statistic, its p-value, and mean absolute and root-mean-squared pricing errors (monthly and annualized). All models are estimated using monthly excess returns.

Factor Model	λ	GRS F	p-value	Mean$\lvert\alpha\rvert$ (M)	RMSα (M)	RMSα (A)
CAPM	0.0	1.9886	0.0466	0.0052	0.0056	0.0701
FF3	0.0	1.8914	0.0603	0.0054	0.0058	0.0722
Carhart4	0.0	1.9192	0.0565	0.0061	0.0065	0.0815
FF5	0.0	1.8657	0.0649	0.0052	0.0056	0.0694
FF6	0.0	1.8376	0.0700	0.0058	0.0061	0.0765
CAPM	0.2	4.2663	0.0001	0.0059	0.0066	0.0812
FF3	0.2	4.1061	0.0002	0.0062	0.0068	0.0840
Carhart4	0.2	3.9454	0.0003	0.0069	0.0074	0.0923
FF5	0.2	4.1928	0.0001	0.0060	0.0066	0.0815
FF6	0.2	3.9469	0.0003	0.0065	0.0071	0.0878
CAPM	0.4	4.6838	0.0000	0.0063	0.0071	0.0873
FF3	0.4	4.4962	0.0001	0.0065	0.0073	0.0904
Carhart4	0.4	4.3423	0.0001	0.0072	0.0079	0.0986
FF5	0.4	4.5035	0.0001	0.0064	0.0071	0.0880
FF6	0.4	4.2818	0.0001	0.0069	0.0076	0.0942
CAPM	0.6	3.7890	0.0004	0.0065	0.0070	0.0866
FF3	0.6	3.6982	0.0005	0.0067	0.0073	0.0898
Carhart4	0.6	3.5643	0.0007	0.0073	0.0079	0.0983
FF5	0.6	3.9175	0.0003	0.0066	0.0071	0.0873
FF6	0.6	3.7039	0.0005	0.0070	0.0076	0.0939
CAPM	0.8	4.0003	0.0002	0.0066	0.0072	0.0882
FF3	0.8	4.1132	0.0002	0.0068	0.0075	0.0920
Carhart4	0.8	4.1205	0.0002	0.0074	0.0081	0.1006
FF5	0.8	4.3215	0.0001	0.0067	0.0073	0.0896
FF6	0.8	4.2176	0.0001	0.0071	0.0078	0.0963
CAPM	1.0	3.9004	0.0003	0.0067	0.0073	0.0902
FF3	1.0	3.6067	0.0006	0.0069	0.0076	0.0936
Carhart4	1.0	3.7242	0.0005	0.0075	0.0082	0.1022
FF5	1.0	3.8619	0.0003	0.0067	0.0074	0.0913
FF6	1.0	3.8339	0.0004	0.0072	0.0079	0.0980

Note: Each λ denotes a distinct CB-APM configuration used to generate decile-sorted test portfolios. The GRS F-statistic tests the joint null hypothesis that all pricing errors ($\boldsymbol{\alpha}$) are zero. Mean$\lvert\alpha\rvert$ and RMSα are reported in monthly (M) and annualized (A) terms. p-values are rounded to two decimal places; values below 0.005 appear as 0.00.

Table 8: GRS tests for traditional factor models applied to portfolios sorted on CB-APM approximated consensus signals.
This table reports Gibbons–Ross–Shanken (GRS) test statistics for value-weighted decile portfolios formed on each dimension of the CB-APM's approximated consensus at $\lambda = 1.0$. For each consensus dimension, portfolio returns are regressed on the CAPM, Fama–French three-factor model, Carhart four-factor model, Fama–French five-factor model, and Fama–French six-factor model. Reported are the GRS F-statistic, corresponding p-value, and mean absolute and root-mean-squared pricing errors, shown in both monthly and annualized units.

Model	GRS F	p-value	Mean$\|\alpha\|$ (M)	Mean$\|\alpha\|$ (A)	RMSα (M)	RMSα (A)
EPS forecast revision						
CAPM	1.3805	0.21	0.0046	0.0564	0.0047	0.0584
FF3	1.3699	0.21	0.0050	0.0617	0.0051	0.0636
Carhart4	1.4237	0.19	0.0058	0.0714	0.0060	0.0743
FF5	1.3368	0.23	0.0048	0.0597	0.0050	0.0615
FF6	1.3854	0.21	0.0055	0.0682	0.0057	0.0705
Change in recommendation						
CAPM	2.4101	0.02	0.0058	0.0717	0.0062	0.0761
FF3	2.2623	0.02	0.0061	0.0752	0.0065	0.0798
Carhart4	2.4376	0.01	0.0069	0.0851	0.0072	0.0894
FF5	2.1439	0.03	0.0059	0.0728	0.0063	0.0778
FF6	2.2909	0.02	0.0066	0.0811	0.0069	0.0857
Change in Forecast and Accrual						
CAPM	1.5123	0.15	0.0046	0.0563	0.0048	0.0588
FF3	1.5732	0.13	0.0049	0.0607	0.0051	0.0633
Carhart4	1.5667	0.13	0.0056	0.0700	0.0059	0.0737
FF5	1.5269	0.15	0.0048	0.0590	0.0050	0.0614
FF6	1.5191	0.15	0.0054	0.0673	0.0057	0.0703
Long-vs-short EPS forecasts						
CAPM	2.5764	0.01	0.0042	0.0522	0.0048	0.0588
FF3	2.4627	0.01	0.0045	0.0561	0.0050	0.0621
Carhart4	2.8118	0.01	0.0054	0.0673	0.0059	0.0731
FF5	2.3544	0.02	0.0044	0.0541	0.0048	0.0597
FF6	2.7049	0.01	0.0050	0.0624	0.0055	0.0688
Analyst earnings per share						
CAPM	0.9325	0.51	0.0043	0.0527	0.0047	0.0587
FF3	1.1561	0.33	0.0046	0.0566	0.0050	0.0625
Carhart4	1.2215	0.29	0.0054	0.0673	0.0059	0.0729

(cont'd on next page)

Table 8: GRS tests for traditional factor models on CB-APM consensus-sorted portfolios. (cont'd)

Model	GRS F	p-value	Mean$\|\alpha\|$ (M)	Mean$\|\alpha\|$ (A)	RMSα (M)	RMSα (A)
FF5	1.1588	0.33	0.0044	0.0541	0.0048	0.0597
FF6	1.1697	0.33	0.0050	0.0624	0.0055	0.0681
EPS Forecast Dispersion						
CAPM	1.4355	0.18	0.0050	0.0618	0.0053	0.0662
FF3	1.4461	0.18	0.0051	0.0635	0.0056	0.0693
Carhart4	1.4664	0.17	0.0057	0.0706	0.0063	0.0791
FF5	1.3880	0.20	0.0050	0.0617	0.0054	0.0670
FF6	1.3785	0.21	0.0054	0.0675	0.0060	0.0749
Earnings forecast revisions						
CAPM	2.3250	0.02	0.0047	0.0585	0.0050	0.0616
FF3	2.5170	0.01	0.0051	0.0631	0.0053	0.0659
Carhart4	2.4525	0.01	0.0058	0.0724	0.0061	0.0761
FF5	2.3764	0.02	0.0050	0.0614	0.0052	0.0638
FF6	2.3054	0.02	0.0056	0.0695	0.0058	0.0724
Analyst Value						
CAPM	1.0717	0.39	0.0042	0.0520	0.0044	0.0547
FF3	1.0577	0.41	0.0045	0.0555	0.0047	0.0580
Carhart4	1.0581	0.41	0.0054	0.0671	0.0056	0.0689
FF5	1.0187	0.44	0.0043	0.0526	0.0045	0.0551
FF6	1.0238	0.43	0.0050	0.0621	0.0052	0.0639
Analyst Optimism						
CAPM	1.1976	0.31	0.0042	0.0521	0.0046	0.0573
FF3	1.1288	0.35	0.0045	0.0556	0.0048	0.0600
Carhart4	1.0821	0.39	0.0054	0.0672	0.0057	0.0706
FF5	1.0605	0.40	0.0043	0.0527	0.0046	0.0573
FF6	1.0211	0.44	0.0050	0.0623	0.0053	0.0658

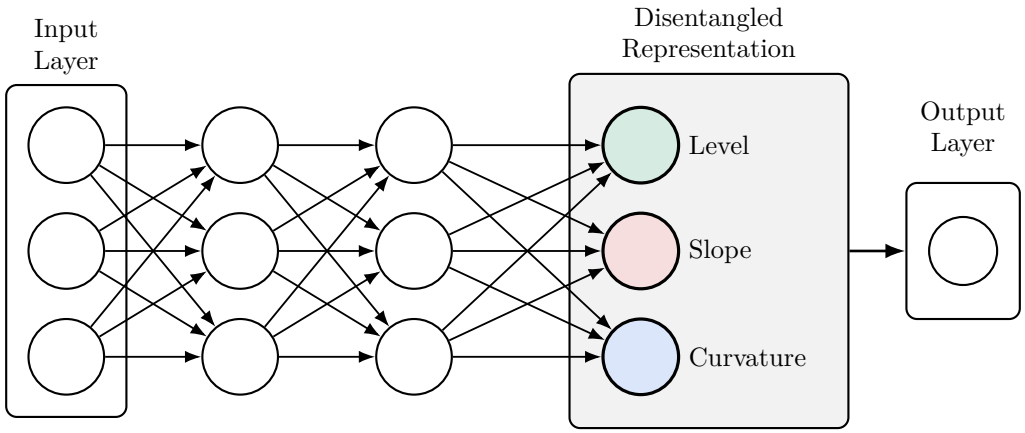


Figure 1: Disentangled representations of neural network.
This schematic illustrates how a high-dimensional input is compressed into a small set of latent representations that correspond to interpretable concepts. The example shown mirrors the yield curve decomposition in fixed income, where dozens of yields can be summarized by three factors: level, slope, and curvature (Nelson and Siegel, 1987). The disentangled representation isolates these dimensions, which are then mapped by the output layer into the final prediction, here defined as the yield curve.

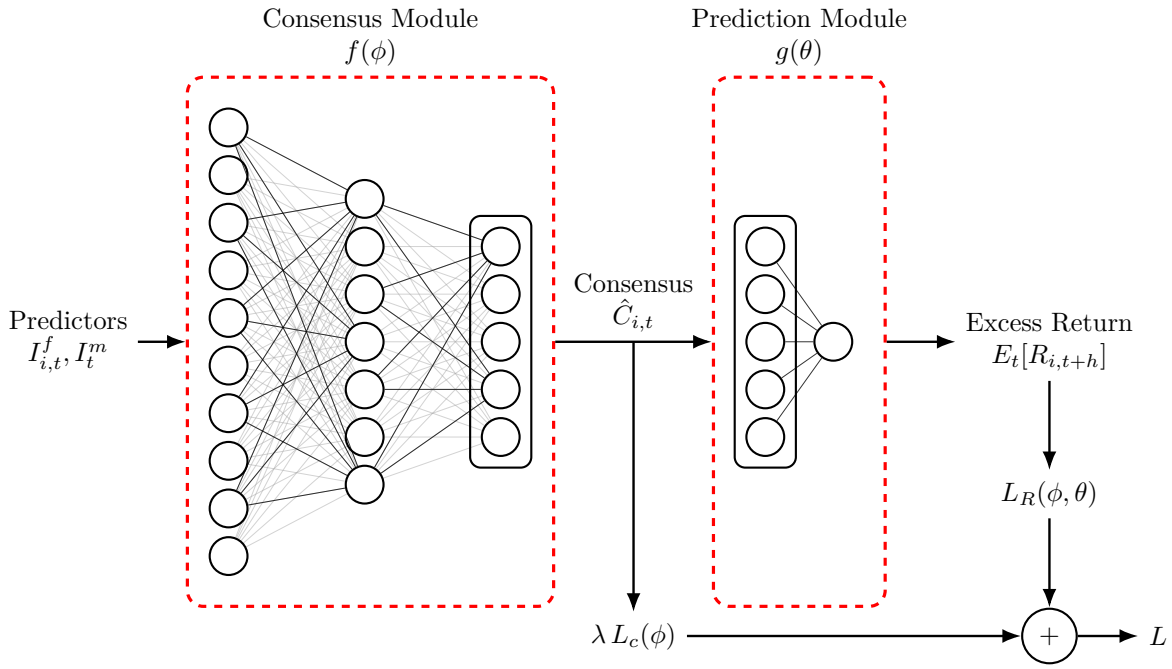


Figure 2: Architecture of the Consensus-Bottleneck Asset Pricing Model (CB-APM).
The model is composed of two modules, the consensus module $f(\phi)$ (left) and the prediction module $g(\theta)$ (right). The consensus module compresses firm-specific predictors $I^f_{i,t}$ and macroeconomic variables I^m_t into a lower-dimensional consensus vector $\hat{C}_{i,t}$ through a feedforward neural network. This bottleneck enforces interpretability by design, as each coordinate of $\hat{C}_{i,t}$ is treated as a consensus concept. The prediction module then maps these consensus variables into expected excess returns $E_t[R_{i,t+h}]$ using a linear layer. The return loss $L_R(\phi, \theta)$ and the consensus loss $L_c(\phi)$ are optimized jointly using the weighted sum $L = \lambda L_c + L_R$, ensuring that the consensus layer is both predictive of returns and interpretable.

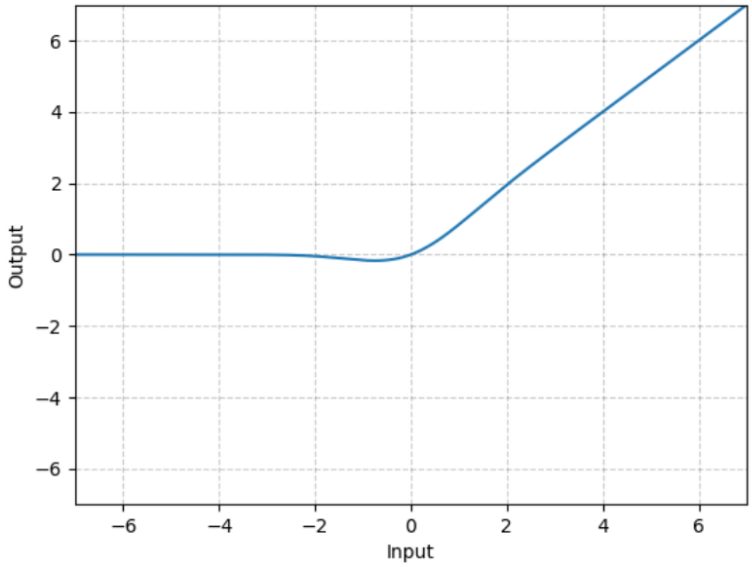


Figure 3: Gaussian Error Linear Unit (GELU) activation function.
GELU is a smooth nonlinear activation that combines properties of the ReLU and sigmoid functions.

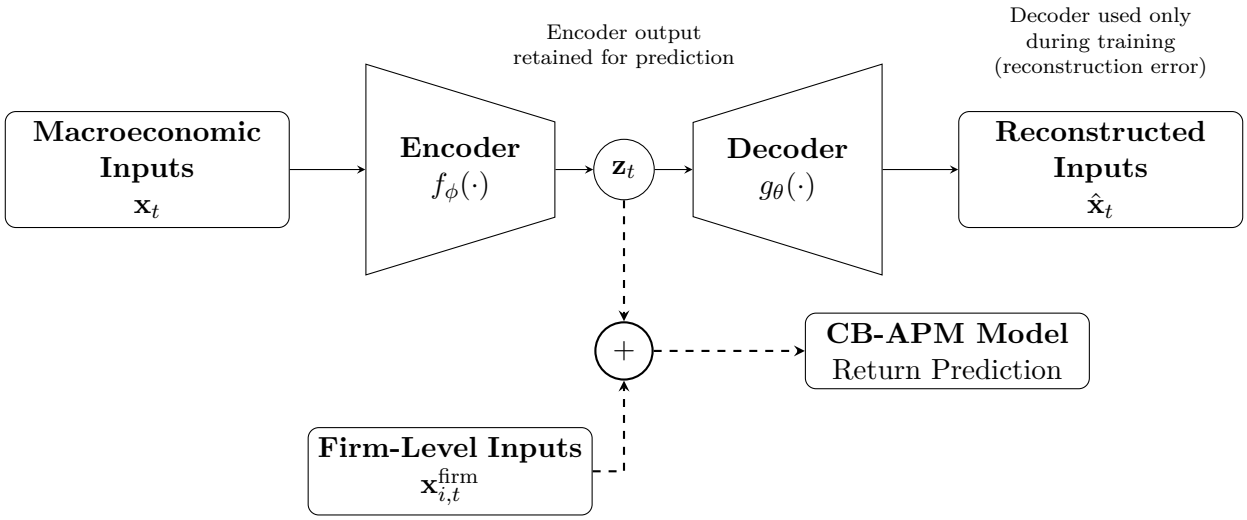


Figure 4: Autoencoder-based macroeconomic embedding.
The encoder narrows horizontally to compress high-dimensional macroeconomic inputs into a latent state $\mathbf{z}_t$, concatenated with firm-level features for return prediction. The decoder is used only during training for reconstruction loss.

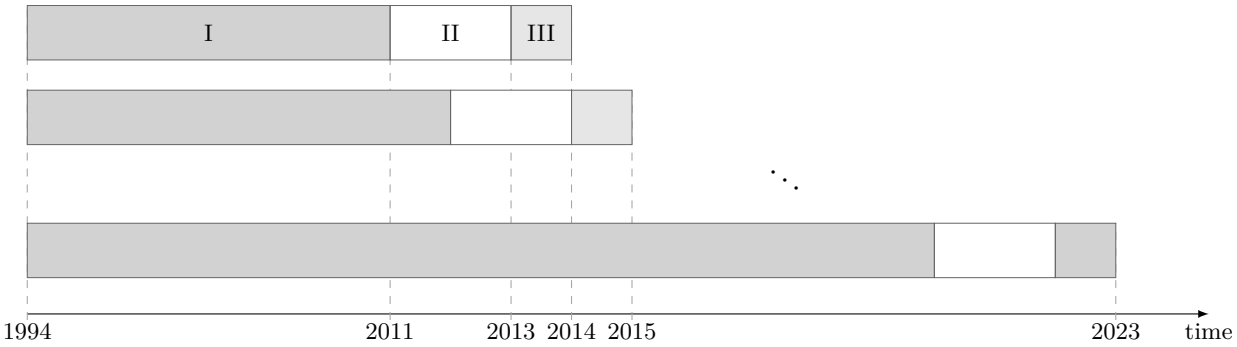


Figure 5: Expanding window evaluation.
This figure illustrates the expanding-window procedure used for model evaluation. At each iteration, the available data are divided into three subsets: I (training set), II (validation set), and III (test set). The training set expands over time, while the validation and test sets are fixed in length at two years and one year, respectively.

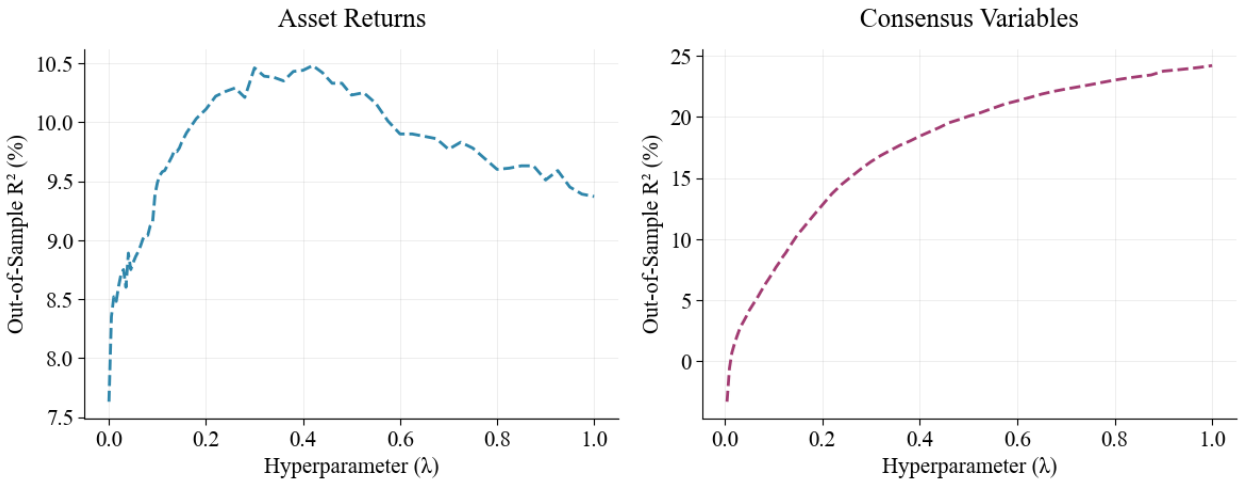


Figure 6: Out-of-sample R^2 of return predictions and consensus approximations.
This figure presents monthly R^2 of annual stock return estimation (left) and average R^2 of analysts' consensus variable approximation (right) across the entire evaluation sets for different λ settings. Return predictability improves sharply when consensus learning is introduced, peaking around $\lambda = 0.3$-0.4, and remains above the benchmark even at $\lambda = 1.0$. Consensus approximation accuracy increases monotonically with λ.

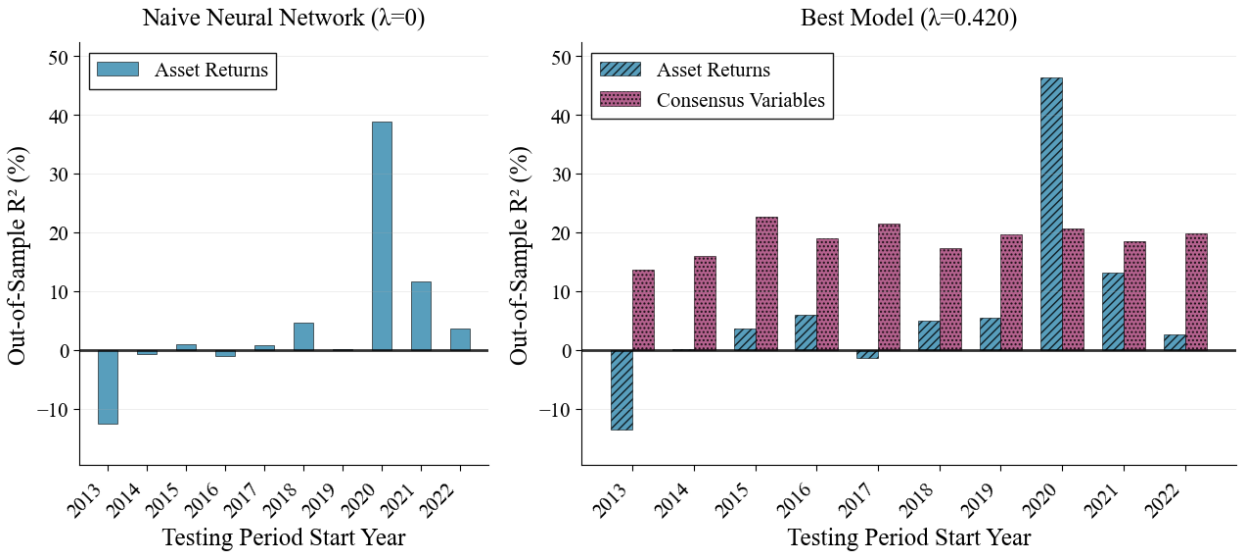


Figure 7: Out-of-sample R^2 by testing period under expanding window evaluation.
This figure reports monthly R^2 of annual stock return and consensus prediction by testing period, based on an expanding-window evaluation. The left panel presents results for a naïve neural network without consensus learning ($\lambda = 0$), while the right panel shows results for the best-performing model ($\lambda = 0.42$).

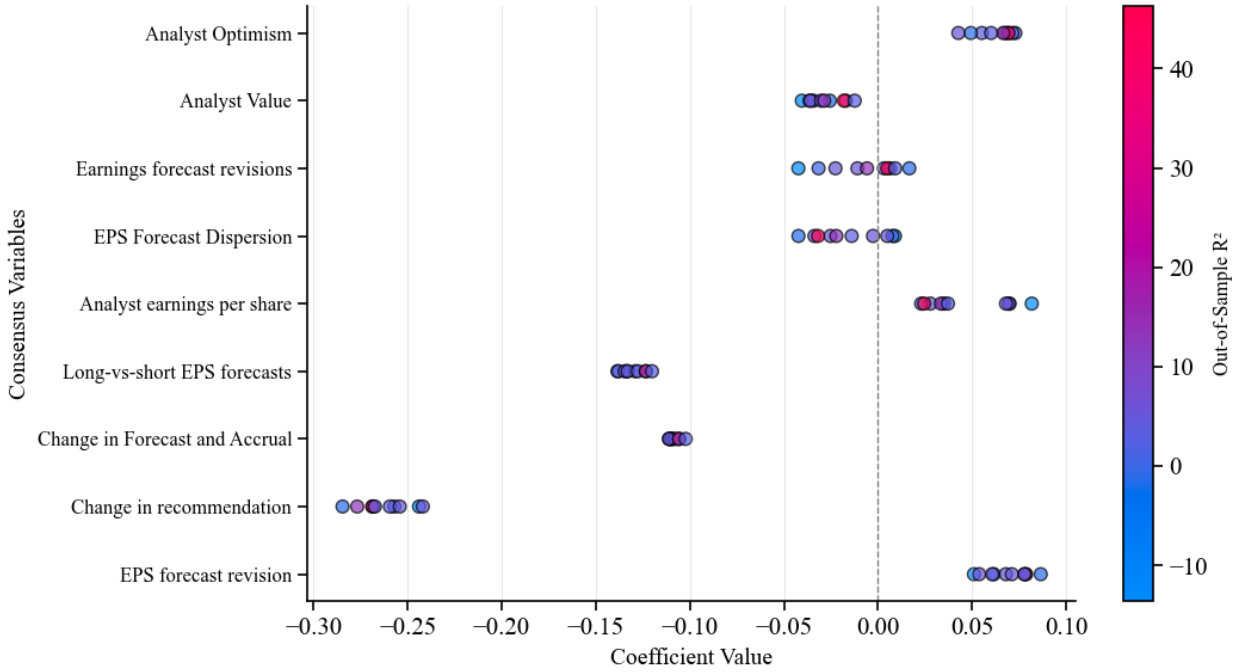


Figure 8: Estimated coefficients for consensus variables.
Prediction module coefficient estimates at ($\lambda = 1$), plotted across expanding training windows. Each point denotes a coefficient for one consensus variable in a given split, colored by its out-of-sample R^2. *Note:* The y-axis displays model-derived consensus variables, not the raw consensus values.

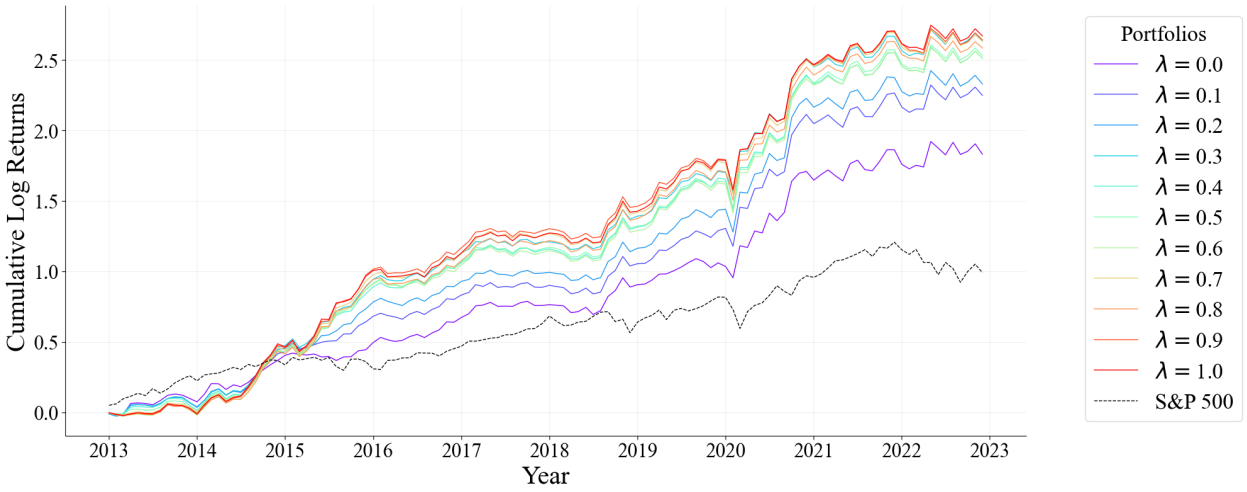


Figure 9: Out-of-sample cumulative returns of long-short decile portfolios.
The figure plots cumulative log returns of value-weighted long-short decile portfolios formed from annual return forecasts, rebalanced monthly using out-of-sample predictions. Each line corresponds to a different hyperparameter λ, with the S&P 500 index buy-and-hold strategy (dashed) as a benchmark. The naïve neural network ($\lambda = 0$) outperforms the S&P 500 benchmark, while CB-APM models with $\lambda > 0$ deliver substantially higher performance than the naïve specification, underscoring the added value of consensus learning.

Internet Appendices to "Interpretable Deep Learning for Stock Returns: A Consensus-Bottleneck Asset Pricing Model" *

Bong-Gyu Jang Younwoo Jeong Changeun Kim

*This paper is a revised version of Master's thesis by Changeun Kim titled "A Consensus-Bottleneck Asset Pricing Model", submitted to the Department of Industrial and Management Engineering, POSTECH, Korea. We would like to thank Hyeng Keun Koo, Kwangmin Jung, Dojoon Park (discussant), JinGi Ha, Jeonggyu Huh, and seminar participants at the 2024 Spring Joint Conference of KORMS and KIIET, 2025 APAF International Conference, 2025 KFA Fall Conference, 2025 4th Workshop on Financial Mathematics and Engineering (Pusan National University), for helpful discussions and insightful comments. This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023R1A2C2003927). Jang (E-mail: bonggyujang@postech.ac.kr) is at Department of Industrial and Management Engineering, POSTECH, Korea University Business School; Jeong (E-mail: younwoo48@postech.ac.kr) is at Graduate School of Artificial Intelligence, POSTECH; Kim (E-mail: changeun120@postech.ac.kr) is at Department of Industrial and Management Engineering, POSTECH. Correspondence concerning this article should be addressed to Changeun Kim, Department of Industrial and Management Engineering, POSTECH, Pohang 37673, Republic of Korea. E-mail: changeun120@postech.ac.kr.

The Internet Appendix is organized as follows. Section A discusses data preprocessing procedures applied prior to estimation and evaluation. Section B provides details on the implementation and architectural choices of the neural network used throughout this study. Section C reports additional empirical results for robustness and supplementary insights. Section D presents the list of variables comprising the dataset used in this study.

A Data Preprocessing

The quality and temporal consistency of input data are fundamental to the empirical validity of the CB-APM framework. Because our model relies on a rich set of firm-level and macroeconomic variables to approximate analysts' consensus and forecast future returns, ensuring that the data accurately reflect the information available to investors at each point in time is essential. Accordingly, this section outlines the complete preprocessing pipeline applied before model training. The procedures include (i) lagging variables to eliminate look-ahead bias, (ii) sampling and filtering firms and predictors to balance coverage and data quality, (iii) imputing missing observations through economically informed methods, and (iv) normalizing the data to harmonize variable scales. Together, these steps construct a temporally aligned, cross-sectionally coherent, and numerically stable dataset that serves as the foundation for the empirical analysis.

A.1 Data delay lagging

Publicly available monthly asset-pricing data are typically released with reporting delays, which can introduce look-ahead bias. This issue arises because the recorded date of the data often reflects when the metric was calculated rather than when it became available. Such discrepancies can mislead researchers about the actual availability of the data at a given time. Several papers point out this problem including Chen and Zimmermann (2022), and we follow their recommended practices for handling such discrepancies from data delays.

Initially, we check the data frequency of firm-level characteristics as detailed in Table D.1. Recognizing that consensus variables are provided irregularly, we conservatively assume an annual frequency. Then, quarterly data are lagged by three months and annual data by six months, respectively. All firm-level predictors are lagged prior to constructing the learning dataset to ensure

temporal alignment between available information and subsequent returns. This approach ensures that the data utilized for predicting future returns are sufficiently historical, thereby minimizing the risk of inadvertently using information that would not have been available at the forecast horizon.

A.2 Data Sampling

This appendix details the construction of the learning dataset employed in the CB-APM estimation. The preprocessing follows a systematic six-step procedure implemented in the `get_data` function, which integrates firm-level predictors, macroeconomic variables, and risk-free rates into a unified panel suitable for model training. Each step ensures data quality, temporal consistency, and the preservation of meaningful cross-sectional information, as summarized below.

A.2.1 Firm screening

Because I/B/E/S analyst consensus variables are relatively sparse compared to other firm-level characteristics, filling in missing observations without preliminary filtering would yield an artificial dataset dominated by interpolated or substituted values. To avoid such distortion, firms without any valid analyst consensus data are excluded from the investable universe at the outset. Among the 17,743 stocks in the full Chen and Zimmermann (2022) dataset, we retain only 4,683 companies for which the complete set of analyst-related variables is available for a sufficiently long history to ensure stable estimation. After applying this screening, the resulting sample contains a total of 605,722 firm-month observations.

This exclusion criterion deliberately sacrifices sample size to ensure that the model learns directly from genuine analyst opinions rather than imputed proxies. Although previous studies in machine-learning-based asset pricing (e.g., Gu et al., 2020; Chen et al., 2024) generally tolerate higher sparsity levels, the stricter sampling rule adopted here is essential for faithfully training the consensus module.

A.2.2 Variable selection

The original dataset from Chen and Zimmermann (2022) includes 161 firm-level predictors with significant long-short portfolio t-statistics exceeding 4 in absolute value. While these variables have been validated for statistical predictability, many suffer from low firm coverage and short sample

histories. Such sparsity weakens the ability of the consensus module to capture variation across firms, since consensus approximation relies on observing cross-firm differences over comparable horizons.

Accordingly, we retain 114 predictors after applying the following criteria:

1. variables with missing-value rates exceeding 20% across the firm panel are removed;
2. variables with insufficient historical coverage (sample starting year after January 1994) are excluded.

The resulting set of firm-level characteristics provides a balanced trade-off between data completeness and information diversity, ensuring that each firm contributes a meaningful set of observations to both the consensus and return-prediction modules.

After all preprocessing steps, the final dataset comprises:

1. 4,683 firms with nonmissing analyst consensus data,
2. 114 firm-level predictors and 123 macroeconomic indicators (including 115 from FRED-MD and 8 from Welch and Goyal, 2008),
3. a total of 605,722 firm-month observations spanning January 1994 to December 2023.

This refined panel forms the empirical foundation for all model estimation and evaluation procedures described in Section 3.

A.3 Data imputation

Although the majority of studies neglect the importance of data imputation methods and simply handle missing values by substituting a cross-sectional mean or median (Green et al., 2017; Gu et al., 2020), we adopt a distinct approach to ensure the integrity of the concept information set, which is crucial in bottleneck modeling.

For variables representing firm characteristics, we primarily employ the Last Observation Carried Forward (LOCF) method. This technique assumes that the most recent observation remains valid until updated information becomes available, thereby maintaining temporal continuity over

short gaps. This approach also mirrors real-world information flow, as the last observation reflects the data actually observed by investors until the next public update.

However, for variables that represent growth rates or changes in firm characteristics, [13] we apply time-series mean imputation. This method captures the inherent continuity and trend in firm-specific dynamics, producing more realistic estimates than cross-sectional averaging. Relying repeatedly on the last observed value for growth-related factors could falsely imply persistence or monotonic trends, misrepresenting the inherently dynamic nature of such variables.

When firms lack historical observations entirely—as in the case of newly listed firms or those undergoing restructuring—we revert to imputing missing values using the cross-sectional mean computed within the same month. Although less ideal, this fallback preserves dataset integrity without introducing excessive bias from outdated or anomalous firm histories.

For analyst consensus data, which include earnings forecasts and investment recommendations, we adopt a two-pronged strategy based on data availability. Because analyst estimates evolve gradually in response to changing fundamentals rather than shifting abruptly, we apply linear interpolation between adjacent data points to capture smooth temporal adjustments. When neither past nor future observations are available (e.g., for firms with sparse coverage), missing entries are filled using the cross-sectional mean of all firms in the same month.

This multi-stage imputation strategy preserves both temporal coherence and cross-sectional comparability, ensuring that the constructed concept information set remains economically interpretable and suitable for the CB-APM framework.

A.4 Data normalization

In asset pricing research, just as in other fields that utilize numerical data, the presence of outliers can significantly distort model outputs, necessitating the standardization of data prior to model integration. Specifically, in cross-sectional asset pricing, the relative position of a metric within the spectrum of similar data points across firms is often more informative than the metric's absolute value. Consequently, aligning with methodologies employed in recent studies (Kelly et al., 2019; Gu et al., 2020; Freyberger et al., 2020; Gu et al., 2021), we compute the rank percentage of firm-level data cross-sectionally, subsequently scaling these ranks to the interval of $[-1, 1]$. This

[13] Variables marked with asterisks in Table D.1 of Internet Appendix D

transformation not only mitigates the influence of outliers but also facilitates more meaningful comparisons across firms. For macroeconomic data, min-max normalization is applied to ensure compatibility with the firm-level data scale. This method adjusts macroeconomic indicators to the same $[-1, 1]$ interval. To promote numerical stability during joint optimization and to align the scale of all model components, we also apply the same rank transformation to consensus variables, even though they primarily serve as target outputs.

B Implementing Neural Network for Asset Pricing

B.1 Activation functions

Rectified Linear Unit (ReLU) is frequently chosen in various machine learning applications due to its computational simplicity and efficiency, as also evidenced by its usage in empirical asset pricing research employing deep learning architectures such as Gu et al. (2020) and Chen et al. (2024). In contrast to activations such as SoftMax or Sigmoid functions, ReLU has demonstrated comparable performance while offering faster computational speed. Moreover, ReLU is particularly valued for its ability to address the gradient vanishing problem by consistently producing non-negative gradients for positive inputs. The Figure B.1 illustrates a graphical form of ReLU function.

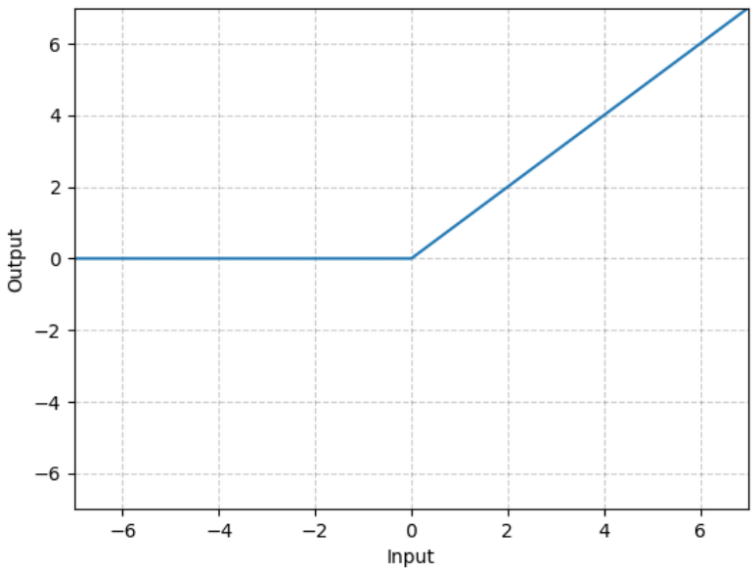


Figure B.1: Rectified Linear Unit (ReLU) activation function.

As shown in above, ReLU deactivates all negative inputs by setting the value to zero. This

characteristic of ReLU occasionally leads to the "dying ReLU" problem, where majority of layers become deactivated during training. This phenomenon occurs when the gradient's absolute value is high or when the bias is significantly negative, causing the activation to become zero and remain stuck in that state as their values are not updated for the rest of the training time. While this issue rarely arises in other applications and may even be considered a strength of ReLU due to its support for sparse learning, dying ReLU can pose a serious challenge in empirical asset pricing since higher learning rates and batch sizes are often employed to facilitate convergence to the global optimum. The problem can be mitigated by introducing a small amount of gradients on the negative side to prevent neurons from becoming completely inactive, achieved through the use of functions such as LeakyReLU, Exponential Linear Units (ELU, Clevert et al., 2016) or Continuously Differentiable Exponential Linear Units (CELU, Barron, 2017).

GELU serves as a notable example of such activation functions, while it is also the most widely adopted activation function in recent deep learning architectures due to its additional benefits. Firstly, GELU has a zero mean and unit variance for inputs drawn from a Gaussian distribution, ensuring the stability of activations and gradients throughout the network. Secondly, the GELU function is smooth and non-monotonic, which contributes to its stability during training while enabling it to capture more complex patterns in the data.

B.2 Stabilized learning

Stabilizing neural network training is essential for robust estimation, particularly when deploying deep learning models in high-stakes decision-making contexts. The following sections detail the regularization methods and machine learning techniques we employ to enhance the stability of the learning process.[14] Despite that this topic is related to experimental factors rather than based on theoretical backgrounds, it is still worthy to discuss since that not only providing these information is crucial for reproducing the results, but also ensures a clear understanding of deep learning techniques. This is particularly important in fields like financial economics, where such concepts may not yet be widely understood or adopted. As such, this discussion is not only relevant but vital for integrating advanced computational methods into financial economic research effectively.

Early stopping is a regularization technique commonly used in deep learning algorithms to pre-

[14] Detailed parameter settings are provided in Table B.1.

vent overfitting and improve generalization performance. During the training process, the model's performance on a validation set is monitored after each epoch. If the validation performance starts to decline or no longer improves, training is halted early, preventing the model from further fitting to noise in the training data. By stopping training before the model becomes overfitted, early stopping helps to achieve better generalization performance on unseen data. Early stopping is a simple yet effective method for improving the robustness and generalization ability of deep learning models, particularly in empirical asset pricing where the amount of training data is limited.

Adaptive Moment Estimation (ADAM) is an optimization algorithm proposed by Kingma and Ba (2017), which is commonly used in practice of training deep learning models. It combines ideas from both momentum optimization and RMSprop (Hinton et al., 2012a), making it well-suited for optimizing non-convex objective functions commonly encountered in neural network training. ADAM maintain the first moment and the second moment of the gradients as two separate moving averages. These moving averages are used to adaptively update the parameters of the model during training. ADAM automatically adjusts the learning rate for each parameter based on the magnitude of the gradients and the accumulated past gradients, allowing it to converge quickly and efficiently in practice.

Learning rate scheduling is a technique used to dynamically adjust the learning rate during training to improve optimization performance. Instead of using a fixed learning rate throughout the training process, learning rate scheduling gradually decrease the learning rate over time, allowing the model to fine-tune its parameters more effectively as training progresses. By annealing the learning rate, learning rate scheduling helps to prevent the optimization process from getting stuck in local minima, which happens surprisingly often as discussed in the next section. Although ADAM inherently adjusts the learning rate, combining it with learning rate scheduling further optimizes performance by refining the initial learning rate. For training CB-APM, we use the ReduceLROnPlateau scheduler provided in PyTorch (See the official document provided by PyTorch for more details Red), which monitors validation performance and reduces the learning rate by a specified factor.

Gradient clipping is a technique introduced by Pascanu et al. (2013) to prevent the exploding gradient problem during the training of neural networks. This problem arises when the gradients become too large, causing numerical instability and hindering the convergence of the optimization

algorithm. Gradient clipping limits the magnitude of the gradients to a predefined threshold. By capping the gradient values, gradient clipping helps stabilize the training process and improves the convergence of the model, allowing for more stable and efficient training of neural networks.

Ensemble learning combines the predictions of multiple individual models to improve overall performance. Such learning scheme is inspired by tree-based models such as Random forests, where outputs of multiple tree estimators are aggregated to generate final prediction results. The most common approach is to compute an average of model outputs, which is also adopted for this work. Specifically, in the case of CB-APM, ensemble learning is applied to both consensuses and individual stock returns. This entails training the entire model multiple times, with the final approximations of analysts' consensus and future returns derived as the average of each model's output.

Layer normalization is a technique designed by Ba et al. (2016) that is used in deep learning to normalize the activations of neurons within each layer of a neural network. Unlike batch normalization (Ioffe and Szegedy, 2015), which normalizes across the entire batch of data, layer normalization computes the mean and standard deviation of the inputs along the hidden layer for each individual training example. This normalization process ensures that the activations of neurons have a mean of zero and a standard deviation of one, which helps stabilize the training process and accelerates convergence.

Dropout is a stochastic regularization technique introduced by Hinton et al. (2012b) that helps prevent overfitting in neural networks by randomly setting a proportion of neurons to zero during each training iteration. This dropout process effectively removes certain connections between neurons, forcing the network to learn more robust and generalizable features. During training, dropout is applied to the input and hidden layers of a neural network with a specified dropout probability, and each neuron in the selected layers is randomly dropped out with the specified probability. Dropout is applied independently to each training example, ensuring that different subsets of neurons are dropped out during each iteration. By randomly dropping out neurons, dropout prevents the network from relying too heavily on any individual neuron or feature, forcing it to learn more redundant representations. During inference, dropout is turned off, and the full network is used to make predictions. However, the weights of the network are usually scaled down by the dropout probability at inference time to account for the increased number of active neurons during training.

B.3 Model hyperparameter

A proper hyperparameter setting is well known to be a key for getting successful performance results in various machine learning applications (Feurer and Hutter, 2019). The most common approach is a process called "hyperparameter optimization", where optimal hyperparameters are chosen from a candidate set automatically by solving an optimization problem of either minimizing or maximizing validation metric. However, this kind of optimization approach can cause specific problems in cross-sectional asset pricing.

In numerous applications involving regression problems, the mean squared error (MSE) is commonly selected as the objective function for hyperparameter tuning because it directly measures the model's predictive accuracy. Yet, in empirical asset pricing, achieving precise predictions of future returns is universally recognized as a pipe dream. Consequently, researchers in this field often use alternative metrics to evaluate the performance of asset pricing models. For instance, Kelly et al. (2024) demonstrate that predictive models can exhibit negative R^2 values yet still deliver positive Sharpe ratios in long-short portfolios. This finding encourages the prioritization of portfolio performance metrics over traditional regression metrics.

Because that reachable level of positive out-of-sample R^2 is nearly 0%, which is significantly low compared to other prediction tasks, models can fall into the trap of converging towards the historical mean, a well-documented local optimum. Welch and Goyal (2008) empirically show that simply taking an average of excess returns can beat regression models with predictive factors. Although the promising developments in employing various factors and modeling approaches over the years of research, such alternative solution can still take over the predictive accuracy of complex models, depending on the model hyperparameter settings and the chosen validation data window. Moreover, even in scenarios where using the historical mean as an expected return might appear statistically optimal, such results hold limited practical economic value, since we cannot apply the results to the investment strategies directly.

Given these considerations, designing an appropriate hyperparameter optimization problem for asset pricing is important, which we leave it as an attractive and also challenging future research topic. In this paper, we rather summarize the hyperparameter "choice" that produces reasonable results as practical guidelines for finding a rational setting. The list of all the hyperparameters

under considerations are provided in Table B.1.

Table B.1: Hyperparameter settings for CB-APM and Autoencoder.

Hyperparameter	Description	Setting
Panel A: CB-APM		
Model		
# hidden layers	Number of hidden layers in consensus module	2
# nodes	Nodes per hidden layer in consensus module	64, 32
Ensemble size	Number of models used for ensembling	10
Learning		
Batch size	Mini-batch size for stochastic optimization	5,000
Learning rate	Initial step size (Adam optimizer)	0.001
Weight decay	ℓ_2 penalty (Adam)	0.005
Scheduling		
Scheduler patience	Epochs without val. improvement before LR decay	2
Scheduler factor	Multiplicative LR decay factor	0.2
Regularization		
Early stopping patience	Epochs without val. improvement before stopping	5
Gradient clip value	Max absolute gradient (global clipping)	1.0
Dropout probability	Dropout probability per linear layer	0.5
Panel B: Autoencoder		
Model		
# hidden layers	Hidden layers in encoder and decoder	2
# nodes	Nodes per hidden layer in encoder and decoder	128, 64
Latent dimension	Dimension of macro latent state $\mathbf{z}_t$	32
Learning		
Batch size	Mini-batch size	1
Learning rate	Initial step size (Adam optimizer)	0.00005
Regularization		

(cont'd on next page)

Table B.1: Hyperparameter settings for CB-APM and Autoencoder (cont'd).

Hyperparameter	Description	Setting
Early stopping patience	Epochs without val. improvement before stopping	2500
Dropout probability	Dropout probability per linear layer	0.2

Although various hyperparameters contribute to predictive performance, the CB-APM framework generally exhibits robustness to modest changes in most settings. By contrast, the choice of batch size is markedly more influential for model performance in our application. This distinction arises from two structural differences in the data used by each component. First, the CB-APM consensus module is trained on high-dimensional panel data with a large cross-sectional dimension (over 600,000 firm-monthly observations), making a relatively large batch size (5,000) computationally efficient while ensuring stable gradient estimates. In contrast, the autoencoder is trained on macroeconomic variables with a very limited number of monthly observations (fewer than 1,000 in total), which more closely resemble low-frequency time-series data. Regarding that the autoencoder does not explicitly exploit temporal dependence, the scarcity of observations motivates a batch size of 1, effectively adopting a stochastic gradient regime that maximizes the diversity of parameter updates. These settings reflect the interaction between data structure (panel versus macroeconomic series) and data availability, and they are critical for achieving stable training dynamics and avoiding overfitting in each model.

C Additional Results

C.1 Forecast horizon analysis

To complement the primary analysis of annual return prediction, we extend our empirical evaluation of CB-APM to alternative horizons, including monthly, quarterly, and semiannual forecasts. This exercise serves two purposes, first, to investigate the model's robustness across varying temporal horizons, and second, to examine whether the observed interpretability-accuracy amplification at long horizons persists in shorter-term predictions.

Tables C.1–C.3 and Figures C.1–C.4 present the out-of-sample R^2 results and their period-

Table C.1: Out-of-Sample R^2 for Stock Return and Consensus Approximations.
This table reports monthly R^2(%) of monthly stock return estimation and analysts' consensus variable approximation over the entire evaluation sets for different λ settings.

	Consensus Variables									Overall Results	
λ	EPS Forecast Revision	Change in Rec-ommend-ation	Change in Forecast & Accrual	Long vs short EPS Forecasts	Analyst Earnings per Share	EPS Forecast Dispersion	Earnings Forecast Revisions	Analyst Value	Analyst Optimisim	Consensus Average	Stock Returns
0	-	-	-	-	-	-	-	-	-	-	0.72
0.1	4.17	−0.01	3.86	7.78	59.12	30.98	10.62	26.77	28.65	19.10	0.32
0.2	4.89	0.05	4.65	9.09	68.94	37.31	13.96	33.20	35.40	23.05	0.40
0.3	5.19	0.08	5.00	9.41	73.01	39.99	15.52	36.27	38.20	24.74	0.41
0.4	5.34	0.09	5.17	9.66	75.35	41.42	16.50	38.15	40.05	25.75	0.40
0.5	5.39	0.10	5.21	9.81	76.81	42.62	17.26	39.64	41.26	26.46	0.37
0.6	5.42	0.10	5.26	9.92	77.82	43.47	17.82	40.68	42.19	26.96	0.24
0.7	5.42	0.09	5.25	9.89	78.50	44.09	18.31	41.52	42.84	27.32	0.16
0.8	5.42	0.09	5.26	9.95	79.06	44.62	18.66	42.16	43.34	27.62	0.06
0.9	5.47	0.07	5.27	10.04	79.50	44.98	18.97	42.65	43.83	27.86	0.02
1.0	5.46	0.07	5.27	10.07	79.81	45.34	19.27	43.06	44.20	28.06	−0.03

Note: Consensus average is calculated as the mean of consensus variable R^2.

Table C.2: Out-of-Sample R^2 for Stock Return and Consensus Approximations.
This table reports monthly R^2(%) of quarterly stock return estimation and analysts' consensus variable approximation over the entire evaluation sets for different λ settings.

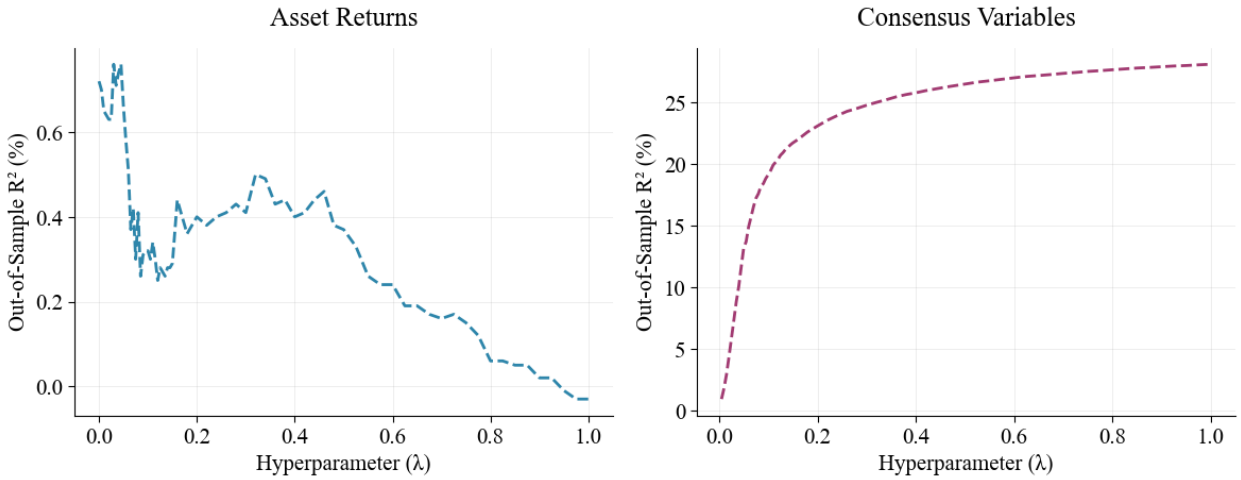
	Consensus Variables									Overall Results	
λ	EPS Forecast Revision	Change in Rec-ommend-ation	Change in Forecast & Accrual	Long vs short EPS Forecasts	Analyst Earnings per Share	EPS Forecast Dispersion	Earnings Forecast Revisions	Analyst Value	Analyst Optimisim	Consensus Average	Stock Returns
0	-	-	-	-	-	-	-	-	-	-	3.37
0.1	2.73	−0.05	2.51	5.22	43.70	22.34	7.44	18.61	20.37	13.65	3.68
0.2	4.12	−0.06	3.78	8.03	62.85	33.52	12.02	28.67	30.80	20.41	3.49
0.3	4.60	−0.02	4.36	8.90	68.32	36.79	14.21	32.33	34.58	22.67	3.35
0.4	4.83	−0.01	4.64	9.33	71.16	38.54	15.44	34.53	36.82	23.92	3.39
0.5	5.03	0.01	4.90	9.65	73.37	39.88	16.29	36.36	38.43	24.88	3.66
0.6	5.20	0.02	5.08	9.88	74.87	40.95	16.91	37.73	39.61	25.58	3.86
0.7	5.30	0.05	5.21	9.99	75.87	41.70	17.43	38.71	40.53	26.09	4.01
0.8	5.36	0.07	5.28	10.05	76.64	42.26	17.85	39.52	41.21	26.47	3.97
0.9	5.44	0.06	5.34	10.12	77.37	42.80	18.20	40.36	41.99	26.85	3.84
1.0	5.48	0.07	5.39	10.17	77.94	43.35	18.52	41.05	42.51	27.16	3.95

Note: Results are reported for a sampled subset of λ settings due to redundancy.

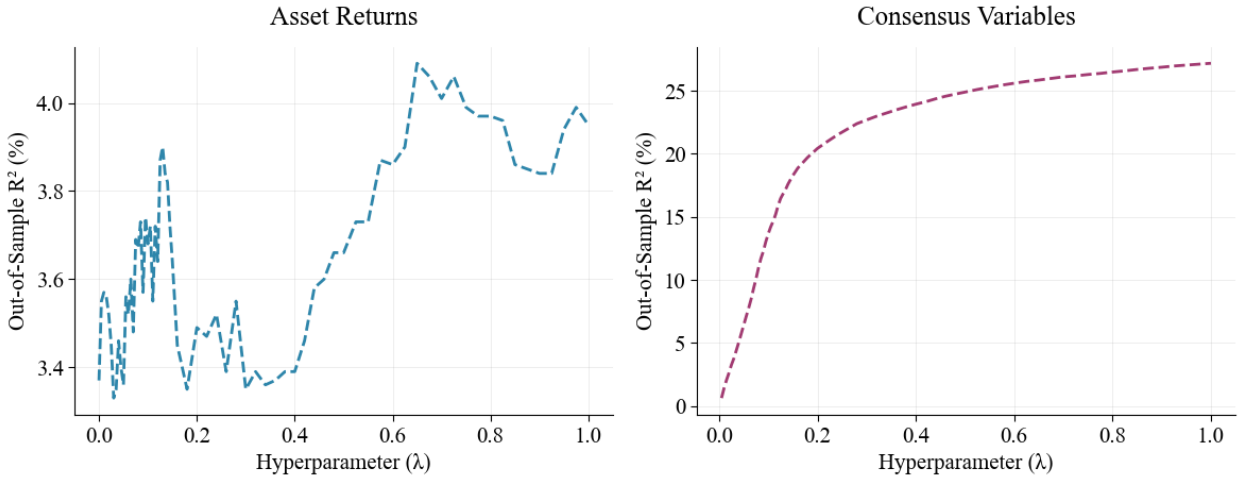
Table C.3: Out-of-Sample R^2 for Stock Return and Consensus Approximations.
This table reports monthly $R^2(\%)$ of semiannual stock return estimation and analysts' consensus variable approximation over the entire evaluation sets for different λ settings.

	Consensus Variables									Overall Results	
λ	EPS Forecast Revision	Change in Rec-ommend-ation	Change in Forecast & Accrual	Long vs short EPS Forecasts	Analyst Earnings per Share	EPS Forecast Dispersion	Earnings Forecast Revisions	Analyst Value	Analyst Optimisim	Consensus Average	Stock Returns
0	-	-	-	-	-	-	-	-	-	-	4.34
0.1	1.77	−0.12	1.95	4.02	29.53	15.52	6.17	12.53	13.59	9.44	6.87
0.2	3.12	−0.10	3.09	6.33	50.97	27.36	9.42	22.19	24.40	16.31	6.91
0.3	3.89	−0.08	3.67	7.64	60.95	32.78	11.98	27.60	29.94	19.82	6.56
0.4	4.34	−0.08	4.09	8.51	66.32	35.91	13.73	30.92	33.20	21.88	6.31
0.5	4.60	−0.05	4.38	8.94	69.33	37.58	14.82	33.12	35.20	23.10	6.19
0.6	4.79	−0.07	4.62	9.17	71.53	38.80	15.69	34.89	36.77	24.02	6.26
0.7	4.95	−0.05	4.83	9.34	73.20	39.84	16.34	36.21	38.05	24.75	6.18
0.8	5.11	−0.05	4.99	9.53	74.42	40.56	16.90	37.27	39.08	25.31	5.99
0.9	5.21	−0.04	5.10	9.66	75.42	41.23	17.31	38.20	39.92	25.78	5.93
1.0	5.32	−0.02	5.18	9.76	76.15	41.67	17.75	38.89	40.56	26.14	5.95

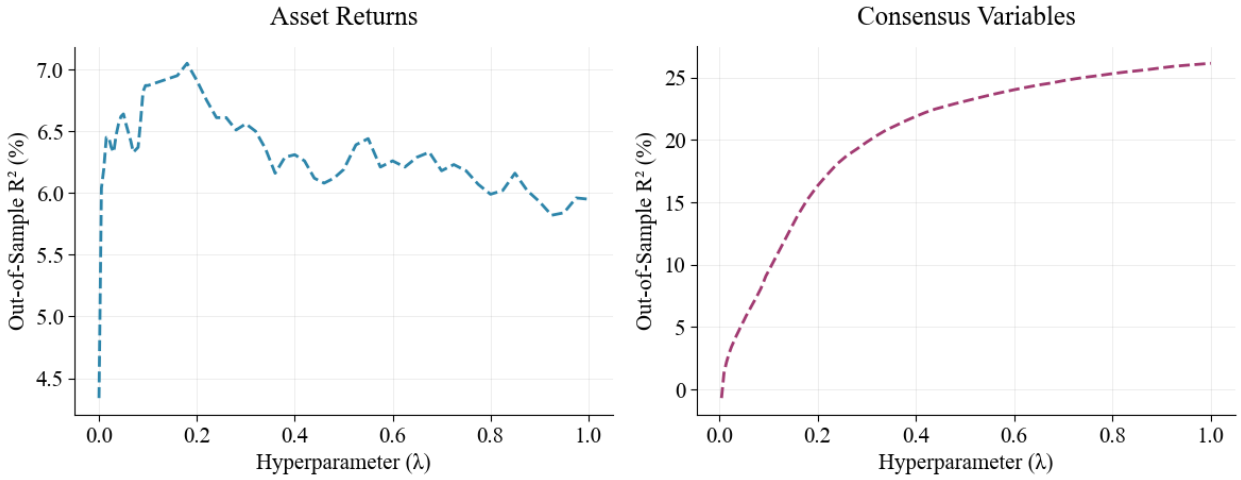
Note: Results are reported for a sampled subset of λ settings due to redundancy.



(a) Monthly horizon



(b) Quarterly horizon



(c) Semiannual horizon

Figure C.1: Out-of-sample R^2 of return predictions and consensus approximations. Panel (a), (b), and (c) presents results for the monthly, quarterly, and semiannual returns respectively.

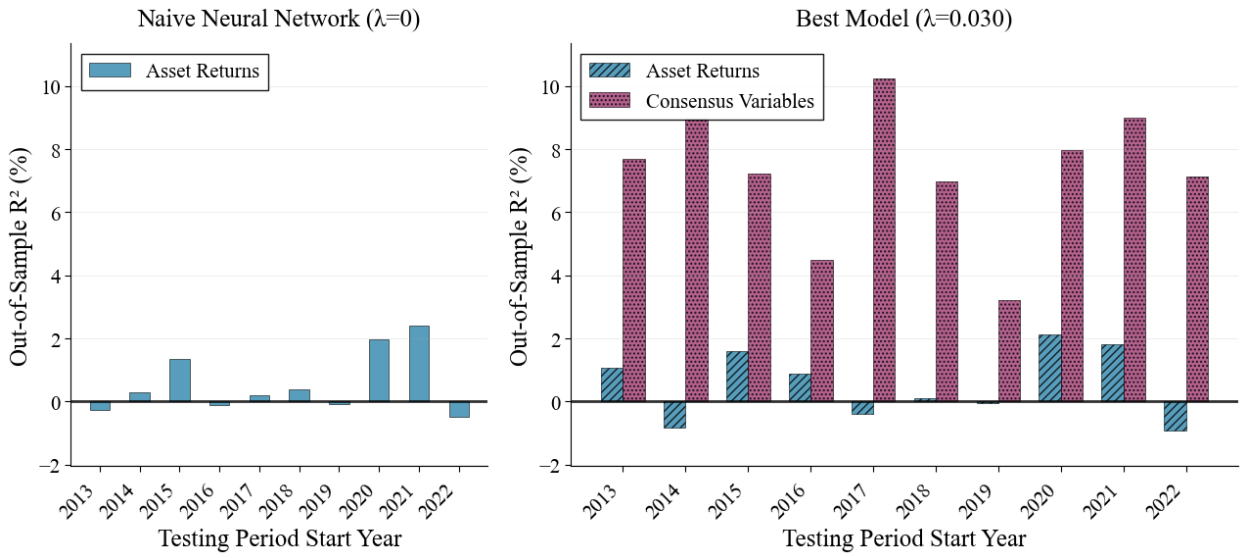


Figure C.2: Out-of-sample R^2 by testing period under expanding window evaluation.
This figure reports monthly R^2 of monthly stock return and consensus prediction by period for naïve neural network ($\lambda = 0$, left) and the best-performing model ($\lambda = 0.03$, right.

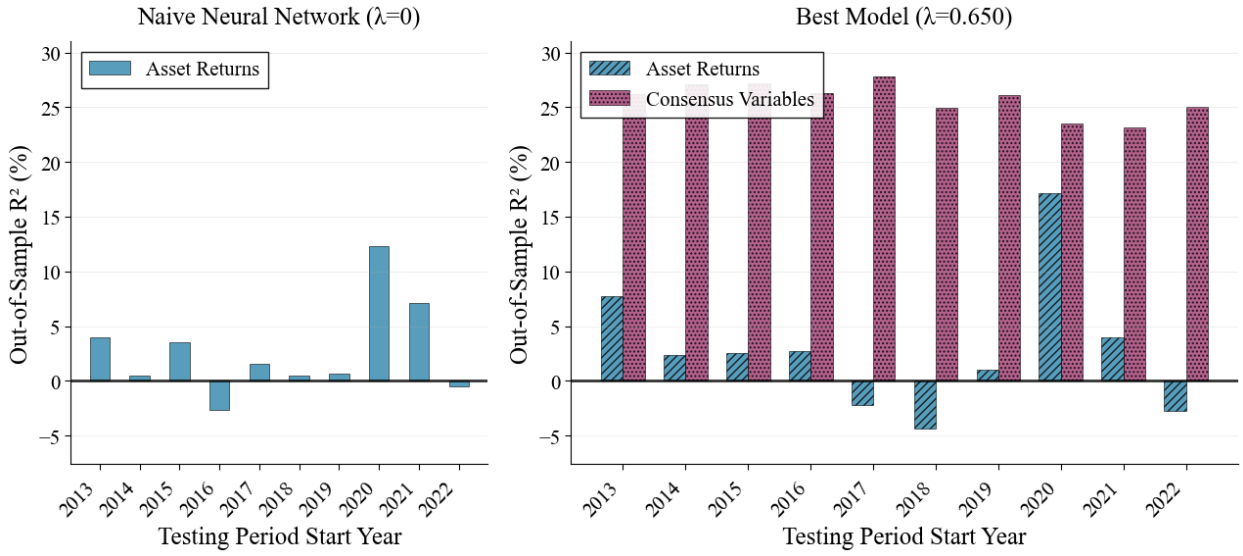


Figure C.3: Out-of-sample R^2 by testing period under expanding window evaluation.
This figure reports monthly R^2 of quarterly stock return and consensus prediction by period for naïve neural network ($\lambda = 0$, left) and the best-performing model ($\lambda = 0.65$, right.

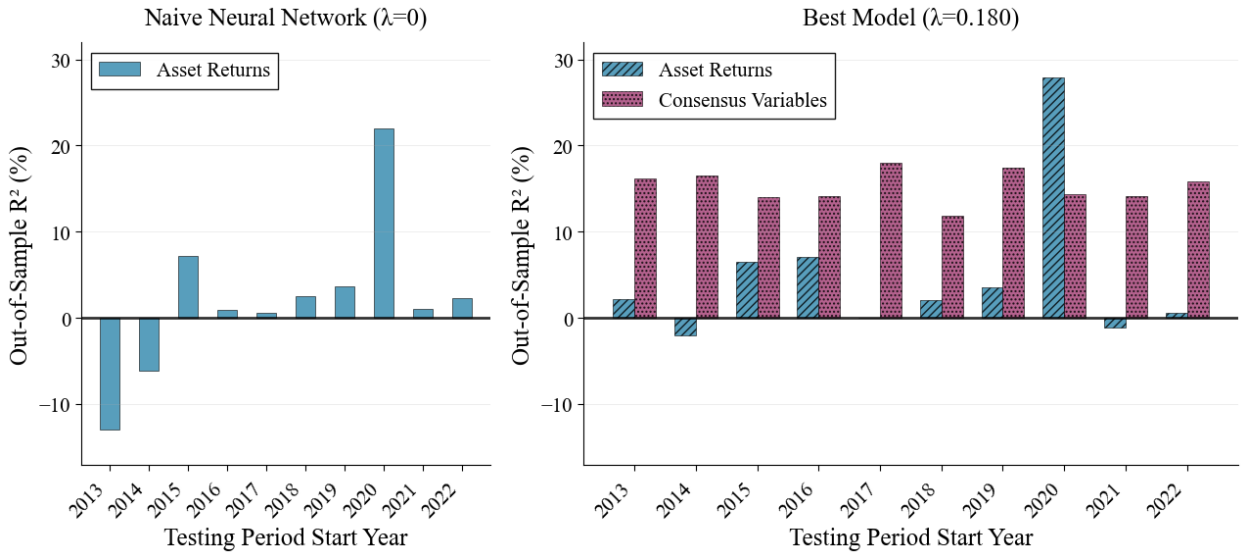


Figure C.4: Out-of-sample R^2 by testing period under expanding window evaluation.
This figure reports monthly R^2 of semiannual stock return and consensus prediction by period for naïve neural network ($\lambda = 0$, left) and the best-performing model ($\lambda = 0.18$, right.

by-period decomposition for these horizons. For shorter horizons such as one month, return predictability remains marginal, with R^2 values close to zero and even slightly negative at higher values of λ, consistent with the well-documented difficulty of forecasting near-term returns. In this regime, increasing λ intensifies the interpretability-accuracy trade-off: while consensus approximation improves monotonically, return R^2 declines, suggesting that allocating more weight to consensus modeling diverts representational capacity away from short-horizon return-specific signals.

In contrast, quarterly and semiannual horizons exhibit intermediate behavior between the monthly and annual cases. For these horizons, the inclusion of consensus learning yields positive predictive gains without incurring the sharp performance penalty seen in monthly forecasting. Notably, the semiannual horizon begins to display a pattern closer to that of the annual horizon, with joint optimization reinforcing both consensus approximation and return predictability.

These results collectively underscore the horizon-dependent effectiveness of CB-APM. At longer horizons (semiannual and annual), the integration of consensus learning acts as an economically grounded regularizer, anchoring predictions to persistent, macro-fundamental drivers that dominate long-term returns. By contrast, for near-term horizons dominated by transitory noise and market microstructure effects, interpretability constraints impose structural rigidity that impairs predictive

accuracy. This divergence aligns with the theoretical intuition that analysts' consensus reflects slow-moving fundamentals, making it more complementary to long-horizon forecasting than to short-term return prediction.

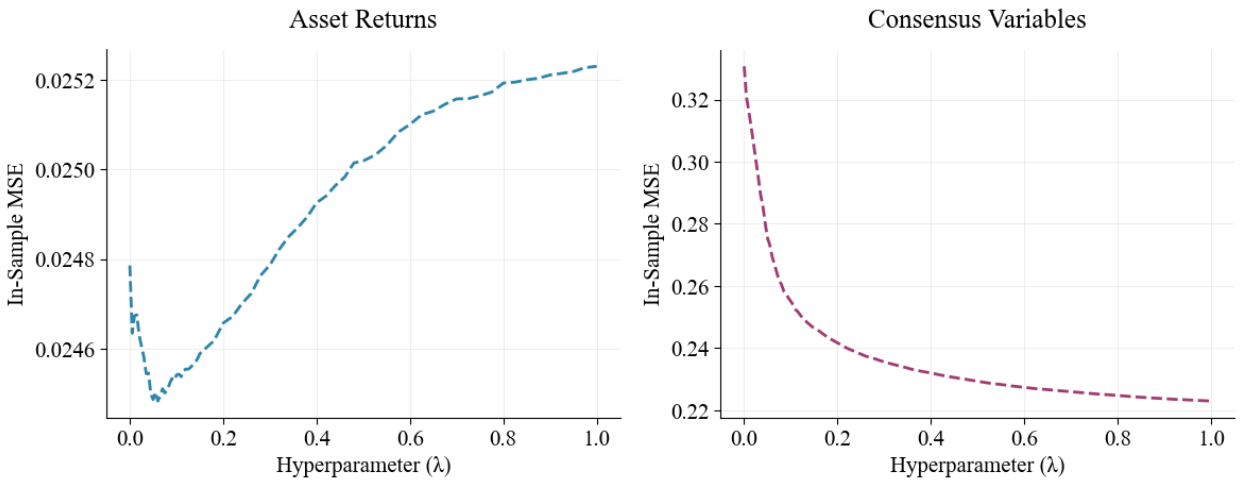
From a practical standpoint, this evidence suggests that CB-APM is particularly well-suited for medium- to long-term investment horizons, where its interpretable architecture not only improves accuracy but also aligns predictions with economically meaningful signals. Conversely, for short-term horizons, where price dynamics are less tied to fundamentals, purely data-driven models may retain an edge in capturing relatively high-frequency fluctuations.

C.2 Properties of the joint optimization

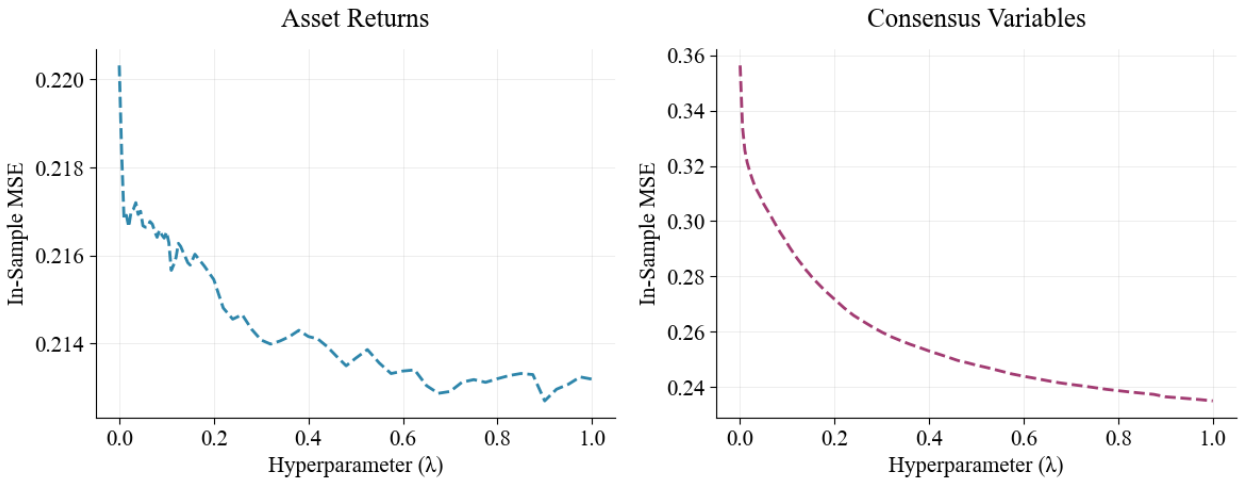
Given that CB-APM is trained using a joint loss function as defined in equation (4), a weighted sum of return prediction loss and consensus approximation loss, it is essential to verify that the model is learning in line with its design. While out-of-sample R^2 is the primary metric for evaluating forecasting performance, it does not reveal how the model balances its dual objectives during training or whether the intended interaction between predictive accuracy and interpretability materializes. In particular, because CB-APM explicitly incorporates a hyperparameter λ to control the trade-off between these two objectives, examining the in-sample MSE dynamics is crucial for understanding how different λ settings shape the model's optimization behavior. This analysis is especially important in our context, as results for the annual forecasting horizon (Section 5.1) suggest that CB-APM may improve both interpretability and performance simultaneously, deviating from the classical interpretability-accuracy trade-off often documented in machine learning applications, as described in Koh et al. (2020).

Figure C.5 illustrates the in-sample MSE dynamics of the CB-APM under varying values of the hyperparameter λ, separately for monthly and annual forecasting horizons. Each panel presents two curves: the left axis depicts the in-sample MSE of stock return predictions, while the right axis reports the average in-sample MSE for consensus variable approximation. The first panel corresponds to the monthly horizon (Figure C.5a), and the second panel presents the results for the annual horizon (Figure C.5b).

The observed patterns reveal a striking divergence between the two horizons. For monthly returns, the stock return MSE exhibits a U-shaped trajectory: it initially decreases slightly for



(a) Monthly horizon



(b) Annual horizon

Figure C.5: In-sample MSE of return and consensus approximations.
This figure plots the in-sample mean squared error (MSE) of stock return (left) and consensus approximation (right) for different λ settings. Panel (a) reports results for the monthly return, and Panel (b) for the annual return. These plots illustrate how forecasting horizons and λ values govern the trade-off between predictive accuracy and consensus reconstruction in the joint loss function.

small values of λ but increases steadily thereafter, suggesting a trade-off between return prediction accuracy and consensus approximation performance. This pattern aligns with the theoretical role of λ in the loss function in equation (4), which explicitly prioritizes consensus approximation as its value increases. Placing greater weight on L_C directs more representational capacity of the network toward modeling analyst consensus at the expense of direct return prediction. This

is consistent with the interpretability-accuracy trade-off widely documented in the interpretable machine learning literature (e.g., Rudin, 2019), where models constrained to capture auxiliary structure or explanatory variables tend to sacrifice marginal predictive performance in favor of enhanced interpretability or alignment with economic reasoning.

In sharp contrast, the annual horizon exhibits what we term an interpretability-accuracy amplification effect. Here, increasing λ monotonically reduces the in-sample return MSE, even as the consensus approximation error steadily improves. Rather than trading off predictive accuracy for interpretability, joint learning of consensus variables appears to reinforce the return prediction objective at longer horizons. This result is particularly noteworthy in the context of financial forecasting, where long-horizon returns are notoriously noisy and difficult to predict using traditional methods. The amplification effect implies that, for CB-APM, jointly learning analyst expectations—serving as a structured, economically meaningful regularizer—can improve the model's capacity to extract signal for long-horizon returns.

This divergence between short- and long-horizon dynamics underscores an important methodological implication of interpretable neural networks in finance. For short-horizon return prediction, forcing the model to align with analyst consensus imposes additional structure that constrains flexibility, thereby introducing a predictable accuracy penalty. However, at longer horizons, the alignment between professional analyst forecasts and fundamental asset value drivers becomes more pronounced, such that the inclusion of consensus loss improves return prediction by anchoring the learning process on more persistent, macroeconomically relevant signals. This finding suggests that interpretable architectures such as CB-APM may be particularly well-suited for applications where the economic rationale underlying predictions is inherently long-term, a domain where conventional "black-box" approaches often fail to yield stable or economically meaningful forecasts.

From a broader perspective, these results provide empirical evidence that interpretability and predictive accuracy in financial neural networks need not be inherently conflicting objectives. Instead, their relationship depends critically on the forecast horizon and the economic structure embedded in the auxiliary interpretable signals. By demonstrating that interpretability constraints can, under appropriate conditions, enhance rather than undermine predictive performance, this study introduces a novel perspective on the role of interpretable modeling in financial machine learning. In particular, the amplification effect observed in annual return forecasts represents, to

our knowledge, the first documented case in which interpretability constraints directly contribute to superior performance in a realistic asset pricing task. This insight opens new avenues for research on designing financially grounded, interpretable deep learning models that exploit economically motivated auxiliary tasks to improve both transparency and forecasting efficacy.

C.3 Structure of the learned macroeconomic representations

While CB-APM achieves interpretability primarily through its consensus-bottleneck, it also relies on macroeconomic embeddings learned by an autoencoder as part of its input structure. Because these embeddings are learned in an unsupervised manner and directly influence return prediction, it is critical to empirically verify that they capture meaningful economic structure rather than spurious patterns. Thus, this section focuses on analyzing the autoencoder's latent representation through visualization and dimensionality reduction techniques. This analysis does not aim to provide instance-level explanations of model predictions but instead validates that the latent macroeconomic state aligns with established business cycle dynamics. In doing so, we complement CB-APM's built-in interpretability with evidence that its macroeconomic component operates transparently and in an economically coherent manner.

Figure C.6 illustrates the two-dimensional principal component projection of the 32-dimensional latent state vectors produced by the macroeconomic autoencoder, color-coded by month and annotated with January observations for selected years. This visualization highlights how the autoencoder successfully encodes macroeconomic conditions into a smooth, low-dimensional manifold that evolves coherently over time. The trajectory of the latent vectors follows a clear temporal progression, demonstrating that the learned embedding captures the gradual transitions and structural shifts in the U.S. macroeconomic environment across the sample period.

A notable feature of this representation is its ability to distinguish major economic regimes. The grey dashed ovals in Figure C.6 correspond to periods classified as recessions by the National Bureau of Economic Research (NBER); the early 2000s recession (2001Q1–Q4) and the Global Financial Crisis (2007Q4–2009Q2). During these intervals, the latent vectors exhibit marked departures from their preceding trajectories, forming clusters that are distinct from surrounding expansionary phases. This pattern indicates that the autoencoder embedding effectively internalizes macroeconomic shocks and regime shifts, producing representations that align with well-established business

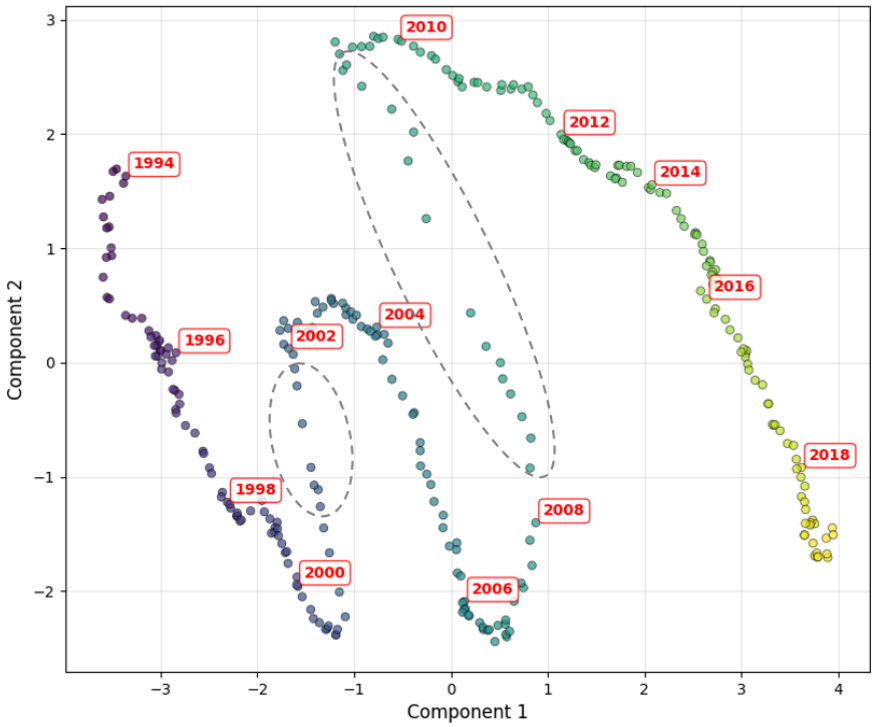


Figure C.6: PCA projection of autoencoder latent state variables.
PCA projection of in-sample 32-dimensional autoencoder latent state vectors into two dimensions, colored by month and annotated with red labels for January of select years. The grey dashed ovals mark NBER recession periods (2001Q1–Q4 and 2007Q4–2009Q2).

cycle chronologies without direct supervision from recession labels.

Beyond capturing these discrete regime shifts, the latent trajectory also reflects continuous macroeconomic evolution during non-recessionary periods. The progression from the early 1990s through the late 2010s shows a gradual unfolding in the latent space, with local curvature corresponding to cyclical fluctuations and persistent structural changes, such as those associated with the post-2008 recovery and subsequent expansion. This smooth temporal ordering suggests that the latent factors not only encode discrete downturns but also represent broader secular dynamics in economic conditions, including shifts in growth, inflation, and monetary policy regimes.

Recent deep factor models make it clear that neural networks can extract a parsimonious set of latent factors from high-dimensional financial and macroeconomic data; these latent variables then drive improved asset-pricing and predictive performance (see, for example, Feng et al., 2018; Gu et al., 2021; Chen et al., 2024). Our macroeconomic autoencoder performs a closely related function for aggregate time-series data: it distills hundreds of macro indicators into a smooth latent

trajectory that aligns with well-known business-cycle chronologies and regime shifts. While many prior studies emphasize quantitative performance metrics and offer only limited visual exploration of their latent factors, the clear temporal patterns in Figure C.6 demonstrate that CB-APM's autoencoder uncovers economically meaningful state dynamics from complex data.

These findings validate the autoencoder's role in distilling high-dimensional macroeconomic data into an economically meaningful latent state. By learning unsupervised representations that exhibit both temporal coherence and sensitivity to regime changes, the model effectively embeds the prevailing macroeconomic environment into a compact form that can be integrated into return prediction. This latent structure provides a powerful mechanism for conditioning asset pricing on macroeconomic context: it transmits shared, time-varying information to the cross-section of firms while mitigating redundancy and noise inherent in raw macroeconomic predictors. Importantly, this approach aligns with our broader CB-APM framework by ensuring that firm-level predictions are informed by a parsimonious yet rich representation of the macro-financial backdrop.

C.4 Further robustness checks

C.4.1 Sensitivity of autoencoder performance to latent dimensionality

An additional robustness check examines the sensitivity of CB-APM's performance to the choice of latent dimension in the autoencoder used for macroeconomic feature compression. While the main body of the paper reports results based on a 32-dimensional latent representation, we also experimented with smaller latent spaces of 8 and 16 dimensions, and larger latent spaces of 64 dimensions. The motivation for this analysis is straightforward. Too small latent space may discard valuable information embedded in macroeconomic predictors, while too large a latent space risks retaining noise and reducing the regularization benefits of dimensionality reduction, as discussed in Hinton and Salakhutdinov (2006).

The comparative results across latent dimensionalities highlight a clear information–bottleneck trade-off in the macroeconomic autoencoder. Increasing the latent dimension generally improves the model's ability to reconstruct analysts' consensus variables: approximation R^2 values rise monotonically for most consensus categories as D increases from 8 to 64, reflecting the greater capacity of higher-dimensional embeddings to capture the underlying macroeconomic structure. However,

Table C.4: Out-of-Sample R^2 for Stock Return and Consensus Approximations for Different Autoencoder Dimensions.
This table reports monthly $R^2(\%)$ of annual stock return estimation and analysts' consensus variable approximation over the entire evaluation sets for different λ settings.

		Consensus Variables									Overall Results	
Auto-encoder dim	λ	EPS Forecast Revision	Change in Rec-ommend-ation	Change in Forecast & Accrual	Long vs short EPS Forecasts	Analyst Earnings per Share	EPS Forecast Dispersion	Earnings Forecast Revisions	Analyst Value	Analyst Optimisim	Consensus Average	Stock Returns
$D = 8$	0	-	-	-	-	-	-	-	-	-	-	−5.82
	0.3	−1.36	−2.55	0.58	2.91	43.51	27.05	10.05	14.94	19.66	12.75	−7.86
	0.6	1.3	−1.84	1.84	3.66	59.24	34.15	13.91	24.17	30.66	18.57	−9.71
	0.9	2.25	−1.5	2.67	5.00	66.51	37.09	15.71	29.26	35.59	21.40	−9.70
$D = 16$	0	-	-	-	-	-	-	-	-	-	-	0.29
	0.3	3.05	−0.3	2.81	5.5	51.36	28.07	9.85	21.95	23.72	16.22	3.70
	0.6	4.24	−0.24	3.98	7.54	64.22	35.10	13.73	30.07	31.91	21.17	3.40
	0.9	4.79	−0.17	4.52	8.59	69.76	38.23	15.67	34.10	36.10	23.51	2.73
$D = 32$	0	-	-	-	-	-	-	-	-	-	-	4.34
	0.3	3.21	−0.25	2.92	5.91	50.96	27.25	9.81	22.22	24.97	16.33	10.46
	0.6	4.3	−0.19	3.9	7.92	64.31	34.97	13.71	30.27	32.70	21.32	9.9
	0.9	4.85	−0.17	4.5	8.89	70.39	38.44	15.73	34.65	36.53	23.76	9.51
$D = 64$	0	-	-	-	-	-	-	-	-	-	-	1.29
	0.3	3.81	−0.22	3.54	7.17	58.10	32.13	12.85	27.17	27.71	19.14	4.44
	0.6	5.02	−0.15	4.55	9.09	69.20	38.66	16.33	34.46	35.16	23.59	3.98
	0.9	5.49	−0.11	5.03	9.73	73.97	41.05	17.92	37.69	39.02	25.53	3.29

Note: Results are reported for a sampled subset of λ settings due to redundancy.

these gains come with diminishing marginal benefits and introduce the risk of over-parameterization. Very small latent spaces (e.g., $D = 8$) underfit the macroeconomic state, leading to weaker consensus approximation and substantially lower return R^2. Conversely, very large embeddings (e.g., $D = 64$) improve consensus reconstruction but begin to attenuate the regularization benefits of compression, slightly weakening return predictability in line with the classical bias–variance trade-off in autoencoder architectures (Hinton and Salakhutdinov, 2006). The 32-dimensional specification achieves a favorable balance where it captures most of the consensus-relevant macroeconomic variation while maintaining sufficient regularization for stable long-horizon return forecasting. For this reason, the main empirical analysis adopts $D = 32$ as the benchmark latent dimensionality.

C.4.2 Comparison of state variables from principal components and autoencoder (TBD)

C.5 Ablation studies

To better understand the mechanisms that drive the performance of CB-APM, we conduct a series of ablation studies. An ablation study refers to systematically removing or modifying key model components to evaluate their incremental contribution to predictive accuracy and interpretability. This approach is widely adopted in machine learning research to clarify the role of specific architectural choices,[15] and has recently been extended to interpretable models such as concept-bottleneck architectures (Koh et al., 2020). In empirical asset pricing, where models often involve high-dimensional predictors and complex nonlinear interactions, ablation studies provide a transparent way to disentangle whether observed performance gains stem from meaningful economic mechanisms or from generic model flexibility.

In the context of CB-APM, ablation studies allow us to assess the value of two key design features. First, we evaluate whether dimensionality reduction of macroeconomic predictors via an autoencoder provides genuine improvements in signal extraction compared to using the raw, redundant set of macroeconomic variables. Second, we examine the role of joint optimization of consensus approximation and return prediction. [16] In particular, we consider both extreme cases: when the model ignores consensus learning altogether ($\lambda = 0$), and when it focuses exclusively on

[15] See Gao et al. (2019) and Devlin et al. (2019) for representative examples in the deep learning literature.

[16] *To be done and reported in the next version of the paper.*

consensus approximation without return prediction ($\lambda \to \infty$). These tests enable us to evaluate whether the consensus-bottleneck provides unique value beyond replicating analysts' forecasts, and whether simultaneous optimization is critical for linking consensus formation to expected returns.

C.5.1 Effect of macroeconomic feature compression

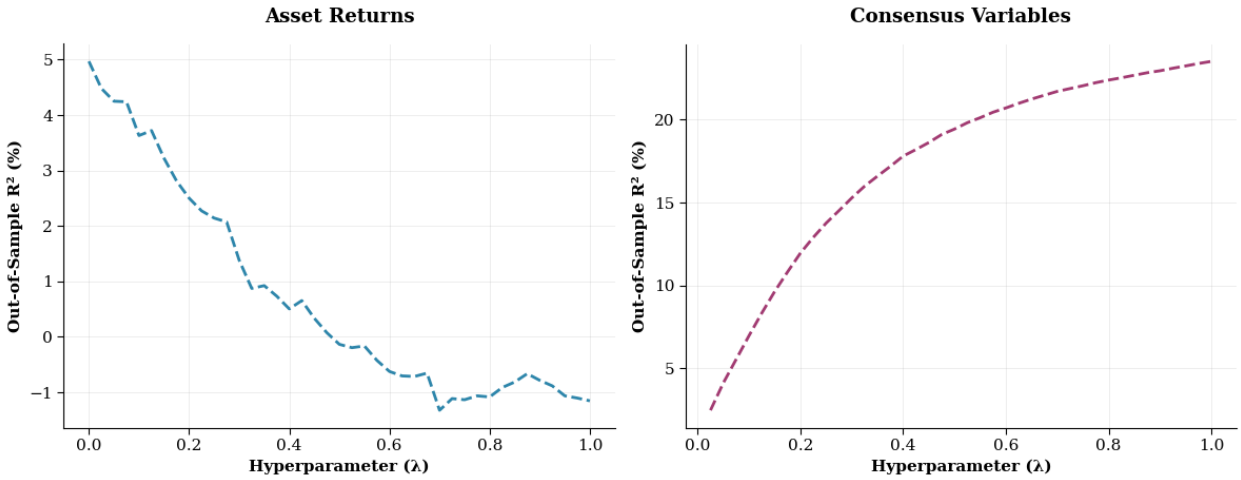


Figure C.7: Out-of-Sample R^2 without Macroeconomic State Embeddings.
This figure reports monthly R^2 of annual stock return estimation (left) and average R^2 of analysts' consensus variable approximation (right) when macroeconomic state variables are not embedded via the autoencoder.

To evaluate the contribution of macroeconomic state embeddings to CB-APM's performance, we conduct an ablation study by re-estimating the model without the autoencoder component. Figure C.7 reports the out-of-sample R^2 for annual stock return prediction (left) and consensus variable approximation (right) across varying values of the hyperparameter λ.

The results show that excluding the autoencoder leads to a sharp deterioration in return predictability. Without macroeconomic embeddings, the out-of-sample R^2 for annual returns declines steadily with increasing λ, ultimately falling below zero for moderate-to-high values of the regularization parameter. This pattern contrasts starkly with the baseline CB-APM, where joint learning with macroeconomic state variables amplifies long-horizon predictive performance. These findings highlight the critical role of macroeconomic context in anchoring the consensus-bottleneck and enhancing its informativeness for return forecasting.

Importantly, consensus approximation remains largely unaffected in this ablated model, as

shown in the right panel of Figure C.7. While the model continues to reconstruct analysts' consensus variables with reasonable accuracy, the absence of macroeconomic embeddings severs an important informational channel linking consensus to return-relevant fundamentals. This divergence underscores the complementary function of the autoencoder that by distilling high-dimensional macroeconomic signals into latent state variables, it enriches the consensus layer with persistent economy-wide information, thereby mitigating noise in firm-level predictors and improving the model's capacity to extract long-horizon risk premiums.

C.5.2 Role of joint optimization in consensus learning (TBD)

A further component analysis evaluates the role of joint optimization in CB-APM, where the model simultaneously learns to approximate analyst consensus and to predict future returns. While the main body of the paper focused on the case $\lambda = 0$, where the model collapses to a pure return prediction architecture, it is equally informative to consider the opposite extreme. When $\lambda \to \infty$, the model is trained solely to replicate contemporary consensus variables without any direct return forecasting objective. This setting allows us to assess whether the architectural design is effective in extracting meaningful consensus representations from firm and macro characteristics, even in the absence of return information. By contrasting these extreme cases with the baseline joint optimization, we can isolate the incremental contribution of consensus learning to return prediction, and provide stronger evidence that the bottleneck mechanism does more than merely fit analysts' forecasts ex post.

D Detailed Data Description

D.1 Firm-level predictors

Table D.1: Descriptions of firm-level predictors from Chen and Zimmermann (2022).

No.	Acronym	Firm-level Predictor	Authors	Year	Journal	Frequency
1	AbnormalAccruals	Abnormal Accruals	Xie	2001	AR	Annual
2	Accruals	Accruals	Sloan	1996	AR	Annual
3	AM	Total assets to market	Fama and French	1992	JF	Monthly
4	AnnouncementReturn	Earnings announcement return	Chan, Jegadeesh and Lakonishok	1996	JF	Quarterly
5	AssetGrowth*	Asset growth	Cooper, Gulen and Schill	2008	JF	Annual
6	BetaLiquidityPS	Pastor-Stambaugh liquidity beta	Pastor and Stambaugh	2003	JPE	Monthly
7	betaVIX	Systematic volatility	Ang et al.	2006	JF	Monthly
8	BM	Book to market	Stattman	1980	Other	Annual
9	BMdec	Book to market using December ME	Fama and French	1992	JPM	Half
10	BookLeverage	Book leverage (annual)	Fama and French	1992	JF	Annual
11	BPEBM	Leverage component of BM	Penman, Richardson and Tuna	2007	JAR	Monthly
12	Cash	Cash to assets	Palazzo	2012	JFE	Quarterly
13	CashProd	Cash Productivity	Chandrashekar and Rao	2009	WP	Monthly
14	CBOperProf	Cash-based operating profitability	Ball et al.	2016	JFE	Annual
15	CF	Cash flow to market	Lakonishok, Shleifer, Vishny	1994	JF	Monthly
16	cfp	Operating Cash flows to price	Desai, Rajgopal, Venkatachalam	2004	AR	Monthly
17	ChEQ*	Growth in book equity	Lockwood and Prombutr	2010	JFR	Annual
18	ChInv*	Inventory Growth	Thomas and Zhang	2002	RAS	Annual

Table D.1: Descriptions of firm-level predictors from Chen and Zimmermann (2022) (cont'd).

No.	Acronym	Firm-level Predictor	Authors	Year	Journal	Frequency
19	ChInvIA	Change in capital inv (ind adj)	Abarbanell and Bushee	1998	AR	Monthly
20	ChNNCOA	Change in Net Noncurrent Op Assets	Soliman	2008	AR	Annual
21	ChNWC	Change in Net Working Capital	Soliman	2008	AR	Annual
22	ChTax	Change in Taxes	Thomas and Zhang	2011	JAR	Quarterly
23	ConvDebt	Convertible debt indicator	Valta	2016	JFQA	Annual
24	CoskewACX	Coskewness using daily returns	Ang, Chen and Xing	2006	RFS	Monthly
25	DelBreadth	Breadth of ownership	Chen, Hong and Stein	2002	JFE	Quarterly
26	DelCOA	Change in current operating assets	Richardson et al.	2005	JAE	Annual
27	DelCOL	Change in current operating liabilities	Richardson et al.	2005	JAE	Annual
28	DelEqu	Change in equity to assets	Richardson et al.	2005	JAE	Annual
29	DelFINL	Change in financial liabilities	Richardson et al.	2005	JAE	Annual
30	DelLTI	Change in long-term investment	Richardson et al.	2005	JAE	Annual
31	DelNetFin	Change in net financial assets	Richardson et al.	2005	JAE	Annual
32	DivInit	Dividend Initiation	Michaely, Thaler and Womack	1995	JF	Annual
33	DivOmit	Dividend Omission	Michaely, Thaler and Womack	1995	JF	Annual
34	dNoa	change in net operating assets	Hirshleifer, Hou, Teoh, Zhang	2004	JAE	Annual
35	DolVol	Past trading volume	Brennan, Chordia, Subra	1998	JFE	Monthly
36	EarningsConsistency	Earnings consistency	Alwathainani	2009	BAR	Annual
37	EarningsStreak	Earnings surprise streak	Loh and Warachka	2012	MS	Quarterly
38	EarningsSurprise	Earnings Surprise	Foster, Olsen and Shevlin	1984	AR	Quarterly
39	EBM	Enterprise component of BM	Penman, Richardson and Tuna	2007	JAR	Monthly

Table D.1: Descriptions of firm-level predictors from Chen and Zimmermann (2022) (cont'd).

No.	Acronym	Firm-level Predictor	Authors	Year	Journal	Frequency
40	EntMult	Enterprise Multiple	Loughran and Wellman	2011	JFQA	Monthly
41	EP	Earnings-to-Price Ratio	Basu	1977	JF	Monthly
42	EquityDuration	Equity Duration	Dechow, Sloan and Soliman	2004	RAS	Annual
43	ExchSwitch	Exchange Switch	Dharan and Ikenberry	1995	JF	Monthly
44	grcapx	Change in capex (two years)	Anderson and Garcia-Feijoo	2006	JF	Annual
45	grcapx3y	Change in capex (three years)	Anderson and Garcia-Feijoo	2006	JF	Annual
46	Herf	Industry concentration (sales)	Hou and Robinson	2006	JF	Monthly
47	HerfBE	Industry concentration (equity)	Hou and Robinson	2006	JF	Monthly
48	hire*	Employment growth	Bazdresch, Belo and Lin	2014	JPE	Annual
49	IdioVol3F	Idiosyncratic risk (3 factor)	Ang et al.	2006	JF	Monthly
50	IdioVolAHT	Idiosyncratic risk (AHT)	Ali, Hwang, and Trombley	2003	JFE	Monthly
51	Illiquidity	Amihud's illiquidity	Amihud	2002	JFM	Monthly
52	IndIPO	Initial Public Offerings	Ritter	1991	JF	Monthly
53	IndMom	Industry Momentum	Grinblatt and Moskowitz	1999	JF	Monthly
54	IntMom	Intermediate Momentum	Novy-Marx	2012	JFE	Monthly
55	Investment	Investment to revenue	Titman, Wei and Xie	2004	JFQA	Monthly
56	InvestPPEInv	change in ppe and inv/assets	Lyandres, Sun and Zhang	2008	RFS	Annual
57	iomom_cust	Customers momentum	Menzly and Ozbas	2010	JF	Monthly
58	iomom_supp	Suppliers momentum	Menzly and Ozbas	2010	JF	Monthly
59	Leverage	Market leverage	Bhandari	1988	JF	Monthly
60	LRreversal	Long-run reversal	De Bondt and Thaler	1985	JF	Monthly

Table D.1: Descriptions of firm-level predictors from Chen and Zimmermann (2022) (cont'd).

No.	Acronym	Firm-level Predictor	Authors	Year	Journal	Frequency
61	MaxRet	Maximum return over month	Bali, Cakici, and Whitelaw	2011	JFE	Monthly
62	Mom12m	Momentum (12 month)	Jegadeesh and Titman	1993	JF	Monthly
63	Mom12mOffSeason	Momentum without the seasonal part	Heston and Sadka	2008	JFE	Monthly
64	Mom6m	Momentum (6 month)	Jegadeesh and Titman	1993	JF	Monthly
65	Mom6mJunk	Junk Stock Momentum	Avramov et al	2007	JF	Monthly
66	MomOffSeason	Off season long-term reversal	Heston and Sadka	2008	JFE	Monthly
67	MomSeason	Return seasonality years 2 to 5	Heston and Sadka	2008	JFE	Monthly
68	MomSeasonShort	Return seasonality last year	Heston and Sadka	2008	JFE	Monthly
69	NetDebtFinance	Net debt financing	Bradshaw, Richardson, Sloan	2006	JAE	Annual
70	NetEquityFinance	Net equity financing	Bradshaw, Richardson, Sloan	2006	JAE	Annual
71	NOA	Net Operating Assets	Hirshleifer et al.	2004	JAE	Annual
72	OPLeverage	Operating leverage	Novy-Marx	2011	ROF	Annual
73	Price	Price	Blume and Husic	1973	JF	Monthly
74	PriceDelayRsq	Price delay r square	Hou and Moskowitz	2005	RFS	Annual
75	RDIPO	IPO and no RD spending	Gou, Lev and Shi	2006	JBFA	Annual
76	RDS	Real dirty surplus	Landsman et al.	2011	AR	Annual
77	RealizedVol	Realized (Total) Volatility	Ang et al.	2006	JF	Monthly
78	ResidualMomentum	Momentum based on FF3 residuals	Blitz, Huij and Martens	2011	JEmpFin	Monthly
79	ReturnSkew	Return skewness	Bali, Engle and Murray	2015	Book	Monthly
80	ReturnSkew3F	Idiosyncratic skewness (3F model)	Bali, Engle and Murray	2015	Book	Monthly
81	RevenueSurprise	Revenue Surprise	Jegadeesh and Livnat	2006	JAE	Quarterly

Table D.1: Descriptions of firm-level predictors from Chen and Zimmermann (2022) (cont'd).

No.	Acronym	Firm-level Predictor	Authors	Year	Journal	Frequency
82	roaq	Return on assets (qtrly)	Balakrishnan, Bartov and Faurel	2010	JAE	Quarterly
83	ShareIss1Y	Share issuance (1 year)	Pontiff and Woodgate	2008	JF	Annual
84	ShareVol	Share Volume	Datar, Naik and Radcliffe	1998	JFM	Monthly
85	Size	Size	Banz	1981	JFE	Monthly
86	STreversal	Short term reversal	Jegadeesh	1990	JF	Monthly
87	Tax	Taxable income to income	Lev and Nissim	2004	AR	Annual
88	TotalAccruals	Total accruals	Richardson et al.	2005	JAE	Annual
89	TrendFactor	Trend Factor	Han, Zhou, Zhu	2016	JFE	Monthly
90	VolSD	Volume Variance	Chordia, Subra, Anshuman	2001	JFE	Monthly
91	XFIN	Net external financing	Bradshaw, Richardson, Sloan	2006	JAE	Annual
92	zerotrade	Days with zero trades	Liu	2006	JFE	Monthly
93	zerotradeAlt1	Days with zero trades	Liu	2006	JFE	Monthly
94	zerotradeAlt12	Days with zero trades	Liu	2006	JFE	Monthly
95	Beta	CAPM beta	Fama and MacBeth	1973	JPE	Monthly
96	BetaFP	Frazzini-Pedersen Beta	Frazzini and Pedersen	2014	JFE	Monthly
97	BidAskSpread	Bid-ask spread	Amihud and Mendelsohn	1986	JFE	Monthly
98	Coskewness	Coskewness	Harvey and Siddique	2000	JF	Monthly
99	DebtIssuance	Debt Issuance	Spiess and Affleck-Graves	1999	JFE	Annual
100	FirmAge	Firm age based on CRSP	Barry and Brown	1984	JFE	Monthly
101	GrLTNOA*	Growth in long term operating assets	Fairfield, Whisenant and Yohn	2003	AR	Annual
102	HerfAsset	Industry concentration (assets)	Hou and Robinson	2006	JF	Monthly

Table D.1: Descriptions of firm-level predictors from Chen and Zimmermann (2022) (cont'd).

No.	Acronym	Firm-level Predictor	Authors	Year	Journal	Frequency
103	High52	52 week high	George and Hwang	2004	JF	Monthly
104	MRreversal	Medium-run reversal	De Bondt and Thaler	1985	JF	Monthly
105	NumEarnIncrease	Earnings streak length	Loh and Warachka	2012	MS	Quarterly
106	PriceDelaySlope	Price delay coeff	Hou and Moskowitz	2005	RFS	Annual
107	PriceDelayTstat	Price delay SE adjusted	Hou and Moskowitz	2005	RFS	Biennial
108	RoE	net income / book equity	Haugen and Baker	1996	JFE	Annual
109	ShareRepurchase	Share repurchases	Ikenberry, Lakonishok, Vermaelen	1995	JFE	Annual
110	SP	Sales-to-price	Barbee, Mukherji and Raines	1996	FAJ	Monthly
111	Spinoff	Spinoffs	Cusatis, Miles and Woolridge	1993	JFE	Monthly
112	VarCF	Cash-flow to price variance	Haugen and Baker	1996	JFE	Monthly
113	VolMkt	Volume to market equity	Haugen and Baker	1996	JFE	Monthly
114	VolumeTrend	Volume Trend	Haugen and Baker	1996	JFE	Monthly

Note: Predictors marked with * are inherently defined as change or growth rates.

D.2 Macroeconomic predictors

Table D.2: Descriptions of macroeconomic predictors from Welch and Goyal (2008).

No.	Acronym	Macroeconomic Predictor	Description
1	dp	Dividend-price ratio	The difference between the log of dividends and the log of prices
2	ep	Earnings-price ratio	The difference between the log of earnings and the log of prices
3	bm	Book-to-market ratio	The ratio of book value to market value for the Dow Jones Industrial Average
4	ntis	Net equity expansion	The ratio of 12-month moving sums of net issues by NYSE listed stocks divided by the total end-of-year market capitalization of NYSE stocks
5	tbl	Treasury-bill rate	The 3-Month Treasury Bill: Secondary Market Rate
6	tms	Term spread	The difference between the long term yield on government bonds and the Treasury-bill
7	dfy	Default yield spread	The difference between BAA and AAA-rated corporate bond yields
8	svar	Stock variance	Sum of squared daily returns on the S&P 500

Table D.3: Descriptions of macroeconomic predictors from FRED-MD (McCracken and Ng, 2016).

No.	Group	Acronym	Macroeconomic Predictor
1	Output and Income	RPI	Real Personal Income
2	Output and Income	W875RX1	Real personal income ex transfer receipts
3	Output and Income	INDPRO	IP Index
4	Output and Income	IPFPNSS	IP: Final Products and Nonindustrial Supplies
5	Output and Income	IPFINAL	IP: Final Products (Market Group)
6	Output and Income	IPCONGD	IP: Consumer Goods
7	Output and Income	IPDCONGD	IP: Durable Consumer Goods
8	Output and Income	IPNCONGD	IP: Nondurable Consumer Goods
9	Output and Income	IPBUSEQ	IP: Business Equipment
10	Output and Income	IPMAT	IP: Materials
11	Output and Income	IPDMAT	IP: Durable Materials
12	Output and Income	IPNMAT	IP: Nondurable Materials
13	Output and Income	IPMANSICS	IP: Manufacturing (SIC)
14	Output and Income	IPFUELS	IP: Fuels
15	Output and Income	CUMFNS	Capacity Utilization: Manufacturing
16	Labor Market	HWI	Help-Wanted Index for United States
17	Labor Market	HWIURATIO	Ratio of Help Wanted/No. Unemployed
18	Labor Market	CLF16OV	Civilian Labor Force
19	Labor Market	CE16OV	Civilian Employment
20	Labor Market	UNRATE	Civilian Unemployment Rate
21	Labor Market	UEMPMEAN	Average Duration of Unemployment (Weeks)

Table D.3: Descriptions of macroeconomic predictors from FRED-MD (McCracken and Ng, 2016) (cont'd).

No.	Group	Acronym	Macroeconomic Predictor
22	Labor Market	UEMPLT5	Civilians Unemployed - Less Than 5 Weeks
23	Labor Market	UEMP5TO14	Civilians Unemployed for 5-14 Weeks
24	Labor Market	UEMP15OV	Civilians Unemployed - 15 Weeks and Over
25	Labor Market	UEMP15T26	Civilians Unemployed for 15-26 Weeks
26	Labor Market	UEMP27OV	Civilians Unemployed for 27 Weeks and Over
27	Labor Market	CLAIMSx	Initial Claims
28	Labor Market	PAYEMS	All Employees: Total nonfarm
29	Labor Market	USGOOD	All Employees: Goods-Producing Industries
30	Labor Market	CES1021000001	All Employees: Mining and Logging: Mining
31	Labor Market	USCONS	All Employees: Construction
32	Labor Market	MANEMP	All Employees: Manufacturing
33	Labor Market	DMANEMP	All Employees: Durable goods
34	Labor Market	NDMANEMP	All Employees: Nondurable goods
35	Labor Market	SRVPRD	All Employees: Service-Providing Industries
36	Labor Market	USTPU	All Employees: Trade, Transportation, and Utilities
37	Labor Market	USWTRADE	All Employees: Wholesale Trade
38	Labor Market	USTRADE	All Employees: Retail Trade
39	Labor Market	USFIRE	All Employees: Financial Activities
40	Labor Market	USGOVT	All Employees: Government
41	Labor Market	CES0600000007	Avg Weekly Hours : Goods-Producing
42	Labor Market	AWOTMAN	Avg Weekly Overtime Hours : Manufacturing

Table D.3: Descriptions of macroeconomic predictors from FRED-MD (McCracken and Ng, 2016) (cont'd).

No.	Group	Acronym	Macroeconomic Predictor
43	Labor Market	AWHMAN	Avg Weekly Hours : Manufacturing
44	Labor Market	CES0600000008	Avg Hourly Earnings : Goods-Producing
45	Labor Market	CES2000000008	Avg Hourly Earnings : Construction
46	Labor Market	CES3000000008	Avg Hourly Earnings : Manufacturing
47	Housing	HOUST	Housing Starts: Total New Privately Owned
48	Housing	HOUSTNE	Housing Starts, Northeast
49	Housing	HOUSTMW	Housing Starts, Midwest
50	Housing	HOUSTS	Housing Starts, South
51	Housing	HOUSTW	Housing Starts, West
52	Consumption, Orders, and Inventories	DPCERA3M086SBEA	Real personal consumption expenditures
53	Consumption, Orders, and Inventories	CMRMTSPLx	Real Manu. and Trade Industries Sales
54	Consumption, Orders, and Inventories	RETAILx	Retail and Food Services Sales
55	Consumption, Orders, and Inventories	AMDMNOx	New Orders for Durable Goods
56	Consumption, Orders, and Inventories	AMDMUOx	Unfilled Orders for Durable Goods
57	Consumption, Orders, and Inventories	BUSINVx	Total Business Inventories
58	Consumption, Orders, and Inventories	ISRATIOx	Total Business: Inventories to Sales Ratio
59	Money and Credit	M1SL	M1 Money Stock
60	Money and Credit	M2SL	M2 Money Stock
61	Money and Credit	M2REAL	Real M2 Money Stock
62	Money and Credit	BOGMBASE	Monetary Base
63	Money and Credit	TOTRESNS	Total Reserves of Depository Institutions

Table D.3: Descriptions of macroeconomic predictors from FRED-MD (McCracken and Ng, 2016) (cont'd).

No.	Group	Acronym	Macroeconomic Predictor
64	Money and Credit	NONBORRES	Reserves Of Depository Institutions
65	Money and Credit	BUSLOANS	Commercial and Industrial Loans
66	Money and Credit	REALLN	Real Estate Loans at All Commercial Banks
67	Money and Credit	NONREVSL	Total Nonrevolving Credit
68	Money and Credit	CONSPI	Nonrevolving consumer credit to Personal Income
69	Money and Credit	DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding
70	Money and Credit	DTCTHFNM	Total Consumer Loans and Leases Outstanding
71	Money and Credit	INVEST	Securities in Bank Credit at All Commercial Banks
72	Interest and Exchange Rates	FEDFUNDS	Effective Federal Funds Rate
73	Interest and Exchange Rates	CP3Mx	3-Month AA Financial Commercial Paper Rate
74	Interest and Exchange Rates	TB3MS	3-Month Treasury Bill:
75	Interest and Exchange Rates	TB6MS	6-Month Treasury Bill:
76	Interest and Exchange Rates	GS1	1-Year Treasury Rate
77	Interest and Exchange Rates	GS5	5-Year Treasury Rate
78	Interest and Exchange Rates	GS10	10-Year Treasury Rate
79	Interest and Exchange Rates	AAA	Moody Seasoned Aaa Corporate Bond Yield
80	Interest and Exchange Rates	BAA	Moody Seasoned Baa Corporate Bond Yield
81	Interest and Exchange Rates	COMPAPFFx	3-Month Commercial Paper Minus FEDFUNDS
82	Interest and Exchange Rates	TB3SMFFM	3-Month Treasury C Minus FEDFUNDS
83	Interest and Exchange Rates	TB6SMFFM	6-Month Treasury C Minus FEDFUNDS
84	Interest and Exchange Rates	T1YFFM	1-Year Treasury C Minus FEDFUNDS

Table D.3: Descriptions of macroeconomic predictors from FRED-MD (McCracken and Ng, 2016) (cont'd).

No.	Group	Acronym	Macroeconomic Predictor
85	Interest and Exchange Rates	T5YFFM	5-Year Treasury C Minus FEDFUNDS
86	Interest and Exchange Rates	T10YFFM	10-Year Treasury C Minus FEDFUNDS
87	Interest and Exchange Rates	AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS
88	Interest and Exchange Rates	BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS
89	Interest and Exchange Rates	EXSZUSx	Switzerland / U.S. Foreign Exchange Rate
90	Interest and Exchange Rates	EXJPUSx	Japan / U.S. Foreign Exchange Rate
91	Interest and Exchange Rates	EXUSUKx	U.S. / U.K. Foreign Exchange Rate
92	Interest and Exchange Rates	EXCAUSx	Canada / U.S. Foreign Exchange Rate
93	Prices	WPSFD49207	PPI: Finished Goods
94	Prices	WPSFD49502	PPI: Finished Consumer Goods
95	Prices	WPSID61	PPI: Intermediate Materials
96	Prices	WPSID62	PPI: Crude Materials
97	Prices	OILPRICEx	Crude Oil, spliced WTI and Cushing
98	Prices	PPICMM	PPI: Metals and metal products:
99	Prices	CPIAUCSL	CPI : All Items
100	Prices	CPIAPPSL	CPI : Apparel
101	Prices	CPITRNSL	CPI : Transportation
102	Prices	CPIMEDSL	CPI : Medical Care
103	Prices	CUSR0000SAC	CPI : Commodities
104	Prices	CUSR0000SAD	CPI : Durables
105	Prices	CUSR0000SAS	CPI : Services

Table D.3: Descriptions of macroeconomic predictors from FRED-MD (McCracken and Ng, 2016) (cont'd).

No.	Group	Acronym	Macroeconomic Predictor
106	Prices	CPIULFSL	CPI : All Items Less Food
107	Prices	CUSR0000SA0L2	CPI : All items less shelter
108	Prices	CUSR0000SA0L5	CPI : All items less medical care
109	Prices	PCEPI	Personal Cons. Expend.: Chain Index
110	Prices	DDURRG3M086SBEA	Personal Cons. Exp: Durable goods
111	Prices	DNDGRG3M086SBEA	Personal Cons. Exp: Nondurable goods
112	Prices	DSERRG3M086SBEA	Personal Cons. Exp: Services
113	Stock Market	S&P 500	S&P500 Common Stock Price Index: Composite
114	Stock Market	S&P div yield	S&P500 Composite Common Stock: Dividend Yield
115	Stock Market	S&P PE ratio	S&P500 Composite Common Stock: Price-Earnings Ratio

D.3 Analysts' consensus variables

Table D.4: Descriptions of analysts' consensus variables from Chen and Zimmermann (2022).

No.	Acronym	Analyst Consensus	Authors	Year	Journal
1	AnalystRevision	EPS forecast revision	Hawkins, Chamberlin, Daniel	1984	FAJ
2	ChangeInRecommendation	Change in recommendation	Jegadeesh et al.	2004	JF
3	ChForecastAccrual	Change in Forecast and Accrual	Barth and Hutton	2004	RAS
4	EarningsForecastDisparity	Long-vs-short EPS forecasts	Da and Warachka	2011	JFE
5	FEPS	Analyst earnings per share	Cen, Wei, and Zhang	2006	WP
6	ForecastDispersion	EPS Forecast Dispersion	Diether, Malloy and Scherbina	2002	JF
7	REV6	Earnings forecast revisions	Chan, Jegadeesh and Lakonishok	1996	JF
8	AnalystValue	Analyst Value	Frankel and Lee	1998	JAE
9	AOP	Analyst Optimism	Frankel and Lee	1998	JAE