FinBench: A Benchmarking Framework for Stock Market Prediction and Portfolio Allocation

Sara Pederzoli
University of Modena and Reggio Emilia
Modena, Italy
sara.pederzoli@unimore.it

Marta Santacroce
University of Modena and Reggio Emilia
Modena, Italy
marta.santacroce@unimore.it

Francesco Guerra
University of Modena and Reggio Emilia
Modena, Italy
francesco.guerra@unimore.it

Marco Bergianti
Cristail
Carpi, Italy
marco.bergianti@cristail.com

Francesco Del Buono
Cristail
Carpi, Italy
francesco.delbuono@cristail.com

Abstract

Research on stock market prediction relies on public datasets to develop and compare learning-based models. However, existing datasets are often limited to price-based information, cover a restricted set of assets or time periods, or provide only a subset of the heterogeneous signals required by modern prediction approaches. Moreover, differences in data preprocessing, task formulation, and evaluation protocols make experimental results difficult to reproduce and hard to compare across studies. To address these challenges, we present FinBench, a benchmarking framework designed to support reproducible evaluation of stock market prediction models using heterogeneous financial data. FinBench integrates historical prices, inter-company relations, macroeconomic indicators, and financial news within a unified data construction pipeline, and standardizes feature extraction and evaluation across classification, regression, and ranking tasks. Model outputs are evaluated both at the task prediction level and through portfolio-based strategies under consistent experimental settings. Using FinBench, we benchmark a range of existing models on European and US equity markets, analyzing their behavior across different prediction tasks, portfolio constructions, and investment universes with diverse liquidity and structural characteristics. FinBench provides a practical reference for systematic and reproducible comparison of stock market prediction methods.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **General and reference** → *Evaluation.*

Keywords

Portfolio Allocation; Stock prediction; Computational finance

ACM Reference Format:

Sara Pederzoli, Marta Santacroce, Francesco Guerra, Marco Bergianti, and Francesco Del Buono. 2026. FinBench: A Benchmarking Framework for Stock Market Prediction and Portfolio Allocation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3817595

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3817595

1 Introduction

The stock market is a complex and dynamic system in which accurate forecasting is crucial for investment decisions and portfolio management. Much of the literature relies on quantitative financial datasets composed of historical prices, returns, volumes, and derived technical indicators. However, financial time series primarily encode market outcomes rather than the mechanisms driving price formation. Prices aggregate trading activity but do not explicitly represent the structural, economic, and informational factors underlying price movements. As a result, important sources of market variation, such as inter-company relationships, sectoral dependencies, macroeconomic regimes, and information flows from financial news, are largely absent from purely quantitative datasets. This limitation constrains the expressive power of models trained solely on price-based features, making it difficult to capture complex dependencies and regime shifts and often leading to unstable predictions and overfitting in non-stationary market conditions.

Recent work [16, 31, 37] has explored machine learning models that integrate heterogeneous information sources, such as company relations, macroeconomic indicators, and textual financial news. Despite their potential, existing approaches remain relatively scarce and exhibit several critical limitations: (L1) many models rely on strong simplifying assumptions [31], including static asset relationships, limited temporal context, or overly simplified preprocessing pipelines, which hinder adaptation to changing market conditions and limit real-world generalizability; (L2) reproducibility remains a major challenge, as dataset construction pipelines are often opaque, poorly documented, or tightly coupled to private data sources, and critical preprocessing, feature extraction, and data alignment choices are frequently underspecified despite their impact on downstream performance; (L3) prediction tasks and evaluation protocols lack standardization, with studies alternating between regression, classification, and ranking formulations [9, 24, 30], adopting heterogeneous portfolio construction strategies, and reporting inconsistent evaluation metrics [1, 21]; (L4) finally, identifying the true state of the art is difficult, as experimental evaluations are often conducted against weak or incomplete baselines [1, 24, 42], such as LSTM [14], XGBoost [2], or simple MLP architectures, without meaningful comparison to more advanced methods. Together, these issues hinder fair benchmarking, slow scientific progress, and limit the practical adoption of machine learning techniques for stock market prediction.

To address these challenges, we propose FinBench, an end-to-end framework designed as a unifying and reproducible infrastructure for financial prediction and portfolio evaluation. FinBench serves as a systematic collector and integrator of existing approaches, enabling the evaluation of diverse modeling assumptions within a shared experimental setting (L1, L4). The framework supports dataset construction by retrieving, cleaning, and aligning heterogeneous data sources, including historical prices, company relational information, and financial news, and provides a transparent and standardized preprocessing pipeline that makes data preparation explicit and reproducible (L2). On top of the resulting datasets, FinBench makes available a wide range of models from the existing literature and evaluates them under consistent experimental settings. The framework supports multiple prediction paradigms, i.e., regression, classification, and ranking, and explicitly connects model outputs to portfolio construction strategies that are evaluated using standardized financial metrics (L3). By unifying data construction, task formulation, and evaluation protocols within a single framework, FinBench enables fair benchmarking, improves comparability across studies, and facilitates a clearer identification of the state of the art approaches.

We apply FinBench to benchmark stock market prediction models across classification, regression, and ranking tasks, evaluating predictive and portfolio behavior across investment universes with diverse liquidity and structural characteristics. This setting enables a systematic assessment of model behavior across tasks, portfolio constructions, and market conditions, supporting a comprehensive and realistic evaluation of existing approaches.

Summarizing, the main contributions of this paper are: (1) a unified and reproducible framework for stock market prediction and portfolio evaluation that supports the construction of datasets from heterogeneous data sources; (2) a standardized experimental pipeline that enables the evaluation of models across multiple prediction paradigms; (3) a task- and portfolio-level evaluation protocol that connects predictive performance to downstream investment strategies; (4) a benchmark, analyzing model behavior across investment universes with different liquidity profiles and market structures.

2 Related Work

Stock movement prediction is commonly studied via classification, regression, and ranking tasks. This section reviews representative models and datasets.

Classification models. The goal is to predict the future direction of stock price movements, such as upward or downward trends. Deep learning models leverage temporal modeling, attention mechanisms, and graph neural networks. THGNN [36] models dynamic and heterogeneous relations among stocks with time-varying graphs and transformer-based graph attention. DTML [42] captures asymmetric inter-stock correlations using transformers with multi-level market contexts. MAN-SF [31] and HGTAN [5] exploit attention mechanisms for complex spatial–temporal dependencies. HATS [20] selectively aggregates relational information for both individual stock and market index prediction. DANSMP [45] uses a market knowledge graph to capture momentum spillover signals and inter-company relations. Sequence- and adversarial-based methods include Adv-ALSTM [9], which enhances LSTM predictions with adversarial learning. CNNPred [15] and DGDNN [43] adopt convolutional and graph-based architectures for local and relational patterns, respectively. Text- and social-driven methods include HAN [18], StockNet [39] and PEN [23], which integrate news and price data using attention or generative models. SLOT [32] learns stock-tweet embeddings with self-supervised learning, improving predictions for sparsely mentioned stocks. CH-RNN [35] applies cross-modal attention to model stock-text interactions. SEP [22] fine-tunes large language models to generate explainable stock predictions from social content. NumHTML [40] processes numeric and structured data in earnings calls using hierarchical transformers to enhance return and risk predictions.

Regression models. The aim is to forecast continuous quantities such as future prices, returns, or log-returns over a given horizon. Generative and sequential approaches model stochasticity and temporal dependencies. D-Va [21] combines a hierarchical variational autoencoder with diffusion modeling for multi-step price prediction, while ESTIMATE [19] captures multi-order correlations and stock dynamics using hypergraph attention and temporal generative filters based on LSTMs [14]. Latent-factor methods focus on extracting underlying predictive signals. HIST [37] decomposes temporal data into predictable components, and DiscoverPLF [16] extends this approach to improve factor disentanglement. FactorVAE [6] integrates a dynamic factor model with a variational autoencoder to predict returns and estimate variances. Simpler architectures explicitly mix temporal, stock, and indicator information. StockMixer [7] and DYCOR [4] employ an MLP-based design to efficiently capture stock-to-market and market-to-stock interactions. Transformer-based models address complex temporal and cross-asset dependencies: MASTER [24] models momentary and cross-time correlations, MATCC [1] decomposes time series into trend and fluctuation components [29], and Finformer [48] adopts a sparse static–dynamic Transformer with adaptive spatio-temporal fusion. More recent approaches include SAMBA [28], which combines bidirectional Mamba [13] with adaptive graph convolution, TRA [25], which introduces temporal routing with multiple predictors, and HD^{2}AT [26], which models heterogeneous causal and lead–lag interactions on stock graphs. Finally, recent work explores large language models for regression tasks. StockTime [34] treats stock prices as sequential tokens and integrates textual signals with temporal patterns for forecasting. Event-driven approaches have also been explored: REST [38] introduces a relational framework that models stock-specific contexts and propagates event effects across related stocks via a stock graph for trend forecasting.

Ranking Models. In ranking-based stock prediction, the goal is to order stocks according to expected returns or relative performance, rather than predicting exact prices or directional movements. STHAN-SR [30] uses a temporal Hawkes attention hypergraph, RT-GCN [46] employs a relation-temporal graph convolutional network, and RSR [10] introduces temporal graph convolutions to jointly model stock relations and their temporal evolution. StockODE [12] leverages Ordinary Differential Equation Networks with a hierarchical hypergraph to model continuous-time volatility and domain-aware dependencies. RSSL [11] builds a probabilistic state-space model with attention-based higher-order relations and uncertainty estimates to improve return-risk trade-offs. FinGAT [17] learns latent stock and sector interactions via graph attention networks with hierarchical sequential modeling. SVAT [3] applies adversarial training on risky stocks to improve risk sensitivity and produce safer, risk-adjusted rankings.

Datasets and Benchmarks. Existing datasets often exhibit limitations in terms of asset coverage, temporal span, or the availability of heterogeneous information sources, which hinders comprehensive evaluation across different financial modeling scenarios. KDD17 [44] focuses on historical price series for a limited universe of 50 stocks. StockNet [39], BigData22 [32], and CIKM2018 [35] combine stock prices with textual sources such as tweets or news, but are constrained by limited temporal coverage and a restricted number of assets. RSR [10] provides price series and inter-stock relational information, including sector and corporate links, but covers only five years and lacks textual data. ECON [33] includes stock prices, macroeconomic indicators, and social media data; however, it does not provide explicit inter-stock relations, limiting its applicability to relational models. EDT [47], CMIN [27] and MarketNews [8] exclusively target news and social media data and do not include price series, making it unsuitable for price-driven prediction tasks. Finally, Qlib [41] offers a flexible data infrastructure and a rich collection of predefined alpha factors derived from market data, but does not include textual information or explicit inter-stock relational structures within a unified benchmarking setting. Taken together, these limitations motivate the need for a benchmark that integrates price series, textual information, and explicit inter-stock relations, while providing standardized preprocessing and evaluation protocols for comprehensive comparison of stock prediction models.

3 The FinBench Architecture

Figure 1 summarizes the architecture of FinBench, which is organized into three modular components corresponding to the main stages of a typical stock market prediction pipeline: FinBench-DATA, FinBench-MODEL, and FinBench-EVALUATION, responsible for data construction, model execution, and prediction and portfolio evaluation, respectively. The execution pipeline is provided in the accompanying GitHub repository[1].

3.1 FinBench-DATA: Data Engineering

Existing approaches to stock market prediction exploit a wide range of indicators that analyze market dynamics from complementary perspectives, including price microstructure, macroeconomic conditions, inter-company relations, and information flows. In FinBench-DATA, we collect a set of such indicators within a data construction pipeline. In particular, we organize features into two broad categories: *financial* and *contextual–semantic* indicators.

3.1.1 Financial Indicators. Financial indicators are reorganized into three groups: market microstructure, macroeconomic and global context, and market-driven relational dependencies.

Market microstructure indicators. Market microstructure indicators describe asset-specific behavior derived from price and volume data, capturing short- and medium-term trends, volatility patterns, and liquidity conditions. FinBench builds these indicators from daily *OHLCV* (Open, High, Low, Close, Volume) data[2] and organizes them as summarized in Table 1. We include raw OHLCV features and standard *Technical Indicators* based on established formulations[3] (e.g., momentum, volatility, and trend measures). To capture intra-day dynamics, Open, High, and Low prices are expressed in normalized form relative to the daily Close (*Adj. Prices*). We also extract rolling-window transformations, such as volatility and volume ratios, to quantify market stress and liquidity variations. Finally, we incorporate two widely used alpha factor representations, *Alpha158* and *Alpha360* [41]. These include price–volume interactions, candlestick patterns (e.g., shadow lengths), long-term trend descriptors, and rolling windows of recent *OHLCV* and *VWAP* (Volume Weighted Average Price) observations normalized by the current closing price. The definition of all microstructure-level features used in FinBench is reported in Appendix A.

Macroeconomic and global market context indicators. Macroeconomic and global market context indicators capture broad economic regimes and market-wide conditions that influence asset returns at a systemic level. FinBench organizes these indicators into *primitive* and *derived* feature groups, as summarized in Table 1. The primitive indicators include yield curve and credit spread measures (e.g., term spread and default spread), as well as returns on global commodities and currencies, corresponding to the *Commodities* category in Table 1. We also include indicators computed on major market indices (e.g., S&P 500, Euro Stoxx 50), corresponding to the *Market Ind.* category, to capture broad market movements and systemic risk. Derived *time-based macro features* are obtained via standardized temporal transformations, such as rolling statistics and windowed technical operators. All indicators are constructed from global data sources retrieved via the FRED API[4] and financial market data providers[5]. The list of primitive and derived macroeconomic features are reported in Appendix A.

Relational financial indicators. Relational financial indicators encode market-driven dependencies among assets based on observed price dynamics. These include correlation networks constructed from rolling correlation matrices of asset returns over fixed windows. By thresholding positive and negative correlations, we obtain dynamic adjacency matrices that capture co-movement patterns and hedging relationships between assets.

[1] https://github.com/softlab-unimore/finbench

[2] https://eodhd.com/

[3] https://ta-lib.org/

[4] https://fred.stlouisfed.org/docs/api/fred/

[5] https://finance.yahoo.com/

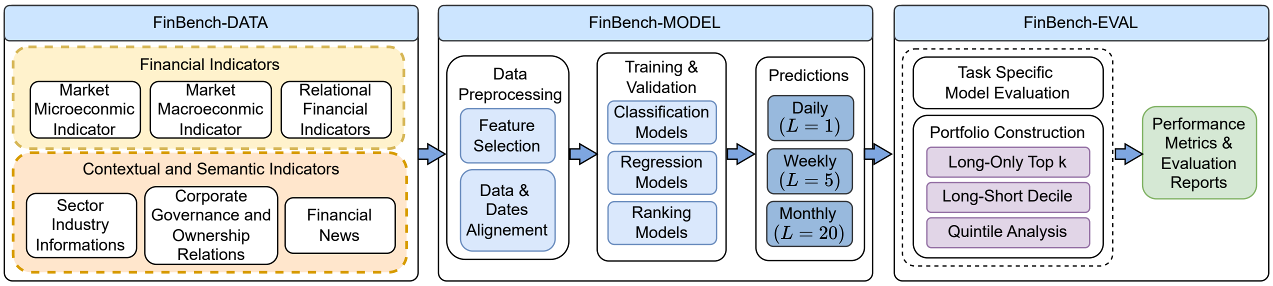

Figure 1: The FinBench architecture

3.1.2 Contextual and Semantic Indicators. Contextual and semantic indicators capture information that is not directly observable from numerical price time series, but that often explains the underlying drivers of market movements. These signals encode firm-level context, structural relationships, and information flow, and complement financial indicators by providing explanatory and semantic views of the market.

Sector and Industry Information. These indicators capture categorical relationships among firms, modeling thematic exposure and common economic drivers. Stocks belonging to the same sector or industry often exhibit co-movement due to shared sensitivities to macroeconomic conditions, regulatory changes, or input costs. In FinBench, these relationships are derived from metadata associated with each stock, obtained from the same data provider used for OHLCV data. We represent such relationships through categorical assignments and graph-based structures that explicitly encode sector- and industry-level groupings.

Corporate Governance and Ownership Relations. Governance- and ownership-related indicators capture long-term organizational and structural dependencies between firms. These include relationships such as ownership links, subsidiaries, and other corporate control structures, which may influence strategic decisions and risk propagation across companies. Such relations are extracted from external knowledge bases, including Wikidata[6], and are modeled as graph-structured data. Temporal qualifiers are used to enforce point-in-time validity, ensuring that evolving corporate structures do not introduce look-ahead bias.

Financial News. Textual indicators are derived from financial news associated with individual assets. News data can provide insight into why price movements occur. We aggregate and temporally align news articles with market data to model information arrival, firm-specific events, and shifts in market sentiment and risk perception, including periods of heightened uncertainty and market panic, which are not directly observable from price dynamics alone. However, FINBENCH does not impose a unified preprocessing or embedding pipeline for textual data. Instead, for models that use financial news, FinBench follow their original text encoding and aggregation strategies to ensure fair reproduction while preserving model-specific NLP designs.

3.2 FinBench-MODEL: Baseline Models and Training Methodology

The literature on stock market prediction includes a large number of models with heterogeneous assumptions, inputs, and learning objectives. Rather than providing an exhaustive survey, FinBench selects a representative set of methods to enable systematic and interpretable benchmarking under a unified experimental protocol. Model selection is guided by explicit criteria that reflect how prediction problems are formulated in practice. In particular, we consider the prediction operation adopted by each method (classification, regression, or ranking), and the categories of financial and contextual–semantic indicators explicitly exploited. These choices directly affect how model outputs are evaluated and how predictions are translated into downstream portfolio construction. An overview of all selected models, together with their input features, is reported in Table 1.

Data preprocessing and temporal alignment strictly follow model-specific requirements. When models are trained independently per asset, alignment is applied only at the ticker level. For models operating on the full asset universe, alignment depends on whether the architecture supports a varying number of assets per date; when a fixed set of tickers is required, missing values are forward-filled to ensure consistency across time. Feature normalization is performed according to the original implementations. For instance, Robust Z-score normalization for models based on alpha factors.

The selected baselines span classification, regression, and ranking paradigms and cover both price-based and relational approaches, reflecting the main ways in which stock market prediction problems are formulated in the literature. This diversity allows FinBench to evaluate models that target short-term directional movements, continuous return estimation, and cross-sectional asset ordering within a unified experimental setting. Classification models include CNNPred [15], Adv-ALSTM [9], and DGDNN [43], as well as relational architectures such as THGNN [36], MAN-SF [31], and HGTAN [5]. Regression models include D-Va [21], ESTIMATE [19], StockMixer [7], MASTER [24], MATCC [1], HIST [37], Discover-PLF [16], FinFormer [48] and FactorVAE [6]. Ranking-based approaches include STHAN-SR [30], SVAT [3], and RT-GCN [46].

[6] https://www.wikidata.org/wiki/Wikidata:Main_Page

Table 1: Overview of the indicators used by each model in FinBench. For each method, we report the categories of *financial indicators* and *contextual and semantic indicators* it explicitly exploits.

Type	Model	Financial Indicators								Contextual & Sematics Indicators		
		Microeconomic Indicators						Macroeconomic Indicators		Sector Industry Relation	Corporate & Governance Relations	Financial News
		OHLCV (Prim.)	OHLCV (Der.)					Commodities (Prim.-Der.)	Market Index (Prim.-Der.)			
			Adj. Prices	Alpha		Tech. Indic.	Rel. Fin. Indicators					
				158	360							
Classification	THGNN	✔	–	–	–	–	✔	–	–	–	–	–
	MAN-SF	✔	–	–	–	–	–	–	–	–	✔	✔
	Adv-ALSTM	–	–	–	–	✔	–	–	–	–	–	–
	HGTAN	✔	–	–	–	–	–	–	–	✔	–	–
	CNNPred (2D/3D)	–	–	–	–	✔	–	✔	–	–	–	–
	DGDNN	✔	–	–	–	–	–	–	–	–	–	–
Regression	D-Va	–	✔	–	–	–	–	–	–	–	–	–
	ESTIMATE	–	✔	–	–	–	–	–	–	✔	–	–
	StockMixer	–	✔	–	–	–	–	–	–	–	–	–
	MASTER	–	–	✔	–	–	–	–	✔	–	–	–
	MATCC	–	–	✔	–	–	–	–	✔	–	–	–
	HIST	–	–	–	✔	✔	–	–	–	–	✔	–
	DiscoverPLF	–	–	–	✔	✔	–	–	–	–	✔	–
	FactorVAE	–	–	✔	–	–	–	–	–	–	–	–
	FinFormer	–	–	–	✔	–	–	–	–	–	–	–
Ranking	STHAN-SR	✔ *(close)*	–	–	–	–	–	–	–	–	✔	–
	SVAT	✔ *(close)*	–	–	–	–	–	–	–	–	✔	–
	RT-GCN	✔ *(close)*	–	–	–	–	–	–	–	✔	✔	–

3.3 FinBench-EVALUATION: Prediction Tasks and Portfolio Evaluation

The evaluation component defines a unified protocol that assesses models at both the task level and the portfolio level, where predictions are evaluated in terms of downstream financial performance.

3.3.1 Task-Level Evaluation. FinBench evaluates models under a unified task formulation. For classification, the target is binary price movement (up/down); for regression, models predict future returns over the selected horizon; for ranking, the objective is to predict the relative ordering of assets. Predictions are generated at discrete decision times t using only information available up to t, and evaluated at horizon L. Labels are thus comparable across models within each task, with differences limited to task-specific preprocessing, such as normalization in regression settings. For models using Alpha indicators as input features, daily z-score normalization is applied to standardize signals across assets.

3.3.2 Portfolio-Level Evaluation. Portfolios are constructed at decision times t using a ranking score $S_{i,t}$ produced by each model for stock i. For a given forecast horizon L (denoted as N trading days), the portfolio formed at time t is held until $t + L$, after which it is rebalanced using updated predictions. To support a unified assessment across different prediction tasks, model outputs are mapped to a common ranking score. For regression models, the ranking score corresponds to the predicted continuous output, $S_{i,t} = \hat{y}_{i,t}$. For classification models, it is defined as the predicted probability of a positive outcome, $S_{i,t} = p^{+}_{i,t}$. For ranking models, the model output directly defines the ranking score, $S_{i,t} = s_{i,t}$. As a preprocessing step, model predictions are considered valid only for dates in which at least 10% of the assets in the universe are available. This filtering mitigates the impact of irregular trading days, such as market holidays or partial trading sessions, and ensures that portfolio decisions are based on a sufficiently representative cross-section of assets. We evaluate multiple portfolio construction strategies to assess complementary aspects of predictive quality.

Long-only Top-K Strategy. This strategy evaluates a model's ability to identify the highest-performing assets for capital appreciation. At each rebalancing step, we select the set $\mathcal{L}_t$ containing the K stocks with the highest ranking scores $S_{i,t}$. The portfolio is constructed as an equally weighted long-only portfolio, where $w_{i,t}$ denotes the portfolio weight assigned to stock i at time t: $w_{i,t} = \frac{1}{K}\mathbf{1}\{i \in \mathcal{L}_t\}$. The main metric for this strategy is the Compound Annual Growth Rate (CAGR%). The objective is to assess whether the model-selected portfolio achieves a higher CAGR% than the corresponding universe benchmark. The benchmark is defined as the index representing the tested asset universe, and this comparison evaluates the model's ability to generate excess returns in a long-only setting.

Long–Short Top-K Strategy. To evaluate market-neutral performance, we construct a long–short portfolio by taking a long position in the top K stocks ($\mathcal{L}_t$) and a short position in the bottom K stocks ($\mathcal{S}_t$) according to the ranking scores. Let $R_t(\mathcal{A})$ denote the realized return of an equally weighted portfolio formed by the set of stocks $\mathcal{A}$ over the holding period starting at time t. The long–short return is defined as $R^{\mathrm{L/S}}_t = R_t(\mathcal{L}_t) - R_t(\mathcal{S}_t)$. This strategy isolates the quality of the ranking signal from general market movements by neutralizing market exposure. Performance is evaluated using the Sharpe Ratio and the Sortino Ratio to assess whether the model delivers superior risk-adjusted returns compared to the corresponding universe benchmark.

Quintile Analysis. We assess signal stability across the full cross-section by partitioning the ranked universe into quintiles, from Q_1 (highest scores) to Q_5 (lowest scores). For each quintile, we compute the annualized return. This analysis evaluates *monotonicity*: a robust predictive model should exhibit an ordered return profile such that $\text{CAGR}(Q_1) > \text{CAGR}(Q_2) > \cdots > \text{CAGR}(Q_5)$. This behavior indicates that predictive power is distributed consistently across the asset universe, and is not driven by extreme-ranked assets.

4 Benchmarking Stock Prediction Models

We evaluate FinBench in a unified benchmarking setting that assesses both predictive performance and downstream portfolio behavior across heterogeneous market conditions. Experiments are conducted on multiple investment universes addressing the following research questions:

RQ1 Which model performs best for each prediction task? (Section 4.3)
RQ2 Do Top-K and Long–Short strategies outperform the market in terms of returns and risk-adjusted metrics? (Section 4.4)
RQ3 Do models produce robust ranking signals with monotonic returns across quintiles? (Section 4.5)
RQ4 How robust are models during periods of market stress? (Section 4.6)

4.1 Setup

Experiments are conducted on a workstation equipped with 4 × NVIDIA Ampere A100 GPUs (64GB VRAM), 256GB RAM, and a dual AMD EPYC 9254 24-Core CPU. We adopt a rolling evaluation protocol with five consecutive phases, each consisting of four years for training, one year for validation, and one year for testing. The initial split covers 2015–2018 for training, 2019 for validation, and 2020 for testing, with windows shifted forward by one year at each phase, yielding a final test period in 2024. This rolling setup also enables evaluation across different market regimes, including stressed and crisis periods, thereby providing a more robust assessment of model performance. Models are evaluated under multiple configurations of lookback window T and prediction horizon L, namely $(T = 5, L = 1)$, $(T = 20, L = 5)$, and $(T = 60, L = 20)$, corresponding to daily, weekly, and monthly investment horizons, respectively. For models based on Alpha360 features, the effective lookback is fixed to $T = 1$, as these indicators already embed information from the previous 60 trading days. All models are trained independently for each split and configuration using grid search following the original implementations. To ensure robustness, experiments are repeated with three random seeds, $S = \{0, 5, 42\}$. Data normalization follows the procedures specified by each model in their original works.

4.2 Investment Universes

To evaluate model robustness across different liquidity profiles and market environments, we conduct experiments on five investment universes, grouped into three size-based categories, to assess the impact of market scale and liquidity on portfolio performance. All universes are constructed in a survivorship-bias-free manner: at each prediction time, only stocks that were actually part of the index at that time are included.

Small (Blue-Chip) Universes. This category includes the Dow Jones Industrial Average (DJI) and the EURO STOXX 50 (SX5E). These universes consist of highly liquid, large-cap stocks operating in mature and heavily arbitraged markets, characterized by relatively low volatility and strong informational efficiency.

Medium (Growth/Technology) Universes. This category includes the NASDAQ 100, dominated by growth- and technology-oriented firms. Relative to blue-chip universes, it exhibits higher volatility, stronger momentum effects, and greater sensitivity to news and market sentiment.

Large (Broad-Market) Universes. This category includes the S&P 500 and the STOXX Europe 600 (SXXP), covering a broad range of sectors and market capitalizations. The presence of mid- and small-cap stocks increases cross-sectional heterogeneity poses a challenging setting for prediction and portfolio construction.

4.3 RQ1: Predictive Performance Across Tasks

We report task-level results for classification, regression, and ranking models. Due to space limitations, we show a subset of representative methods for each task based on overall performance in FinBench. Full results are available in the GitHub repository.

Classification results. Table 2 reports classification results. All models formulate the task as a binary prediction of price direction (up/down) and are evaluated under identical experimental conditions across multiple investment universes and prediction horizons.

Discussion. Results highlight the intrinsic difficulty of directional stock movement prediction across all investment universes. Classification performance remains close to random baselines, with Accuracy values typically around 0.50 and consistently low MCC scores. This behavior is observed across models, horizons, and markets, confirming that short-horizon classification remains challenging even under controlled experimental settings. Performance also varies substantially across investment universes. Models exhibit different behaviors on smaller and highly liquid markets (e.g., DJI, SX5E) compared to broader and more heterogeneous universes (e.g., S&P 500, SXXP), indicating that liquidity and cross-sectional complexity significantly affect classification difficulty. CNNPred2D does not report results for $L = 1$ because the model requires a fixed CNN kernel size of 8, which is incompatible with the adopted lookback window of 5.

Regression results. Table 3 reports regression results. The table distinguishes models that apply target normalization (i.e., Daily Z-score) from models using the original target scale.

Discussion. Among models using alpha indicators (MATCC, MASTER, HIST, and FinFormer), HIST achieves the lowest errors across most horizons and investment universes. However, when compared to StockMixer, which operates on the original target scale, alpha-based models generally exhibit substantially higher error magnitudes. This contrast highlights the strong impact of target normalization and preprocessing choices on regression performance, underscoring a key challenge in regression-based stock prediction: reported accuracy is often driven as much by preprocessing decisions as by modeling improvements. Regression performance also varies across investment universes. Errors are generally lower on smaller and more liquid markets, while broader and more heterogeneous universes such as the S&P 500 and SXXP exhibit larger errors, especially at longer prediction horizons, suggesting that cross-sectional complexity increases regression difficulty.

Ranking results. Table 4 reports results for ranking-based models, which directly optimize the cross-sectional ordering of assets.

Table 2: Classification performance under the considered investment universes and prediction horizons (L). Results are shown for a subset of classification models from the benchmark and reported in terms of Accuracy (Acc), F1-score (F1), and Matthews Correlation Coefficient (MCC) for multiple forecast lengths L. Results are aggregated across the three reported seeds and across the five testing years.

Model	L	DJI			SX5E			NASDAQ 100			S&P500			SXXP		
		Acc	F1	MCC	Acc	F1	MCC	Acc	F1	MCC	Acc	F1	MCC	Acc	F1	MCC
THGNN	1	0.513	0.348	-0.001	0.513	0.424	0.001	0.516	0.363	0.004	0.500	0.466	-0.001	0.502	0.499	0.005
	5	0.535	0.354	0.007	0.535	0.406	-0.003	0.530	0.356	-0.004	0.511	0.463	0.012	0.505	0.498	0.002
	20	0.553	0.354	-0.001	0.542	0.389	-0.021	0.552	0.370	0.009	0.507	0.498	0.013	0.497	0.485	0.001
Adv-ALSTM	1	0.580	0.363	0.000	0.637	0.389	0.000	0.608	0.378	0.000	0.382	0.277	-0.001	0.429	0.305	0.000
	5	0.467	0.318	0.000	0.485	0.326	0.000	0.480	0.324	0.000	0.478	0.323	0.000	0.461	0.318	-0.001
	20	0.521	0.342	0.000	0.519	0.341	0.000	0.532	0.346	0.000	0.530	0.346	0.003	0.504	0.335	-0.011
CNNPred2D	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	5	0.486	0.381	0.004	0.485	0.376	0.006	0.494	0.3177	0.000	0.470	0.233	-0.002	0.480	0.224	0.003
	20	0.480	0.356	-0.002	0.454	0.369	0.026	0.466	0.246	0.037	0.460	0.213	0.000	0.483	0.176	0.012

Table 3: Regression performance under multiple prediction horizons (L), reported in terms of MSE and MAE. Results are shown for a subset of regression models from the benchmark. Models are grouped by the target normalization strategy adopted in their original implementations. Results are aggregated across the three reported seeds and across the five testing years.

Model	L	DJI		SX5E		NASDAQ 100		S&P500		SXXP	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
StockMixer	1	0.018	0.092	0.000	0.015	0.632	0.032	0.022	0.099	0.632	0.038
	5	0.007	0.056	0.002	0.036	0.009	0.069	0.021	0.067	0.100	0.049
	20	0.018	0.099	0.015	0.095	0.027	0.124	0.025	0.110	2.727	0.773
MATCC	1	1.000	0.759	1.041	0.749	1.037	0.712	1.002	0.684	0.999	0.673
	5	1.130	0.818	1.495	0.941	1.713	0.981	1.271	0.824	1.047	0.752
	20	1.404	0.971	1.590	0.984	1.623	0.965	1.744	1.010	1.790	1.011
HIST	1	0.993	0.734	0.983	0.730	0.996	0.693	0.999	0.683	0.998	0.672
	5	0.997	0.747	1.113	0.796	1.030	0.721	1.005	0.711	1.008	0.693
	20	1.016	0.766	1.090	0.800	1.031	0.746	1.006	0.734	1.027	0.717
FinFormer	1	1.338	0.871	1.320	0.864	1.414	0.878	1.495	0.878	1.801	0.899
	5	1.657	0.982	1.430	0.922	1.458	0.867	1.677	0.950	1.544	0.889
	20	2.190	1.111	1.757	1.035	1.895	0.974	1.783	0.985	1.701	0.910

Discussion. Ranking-based models generally extract more stable predictive signals than classification or regression approaches, but signal strength varies substantially across models, horizons, and investment universes. While some methods (e.g., SVAT) achieve positive IC and Rank IC values, particularly at longer horizons, others frequently yield near-zero or negative correlations. Performance also depends on market structure: smaller markets, such as DJI and SX5E, tend to yield higher and more stable ranking correlations, whereas broader universes like the S&P 500 and SXXP introduce additional noise that reduces performance. These results indicate that ranking effectiveness is highly model- and context-dependent, reinforcing the role of architectural and relational design choices.

4.4 RQ2: Portfolio Performance

We evaluate whether model predictions translate into profitable portfolio strategies using a Long-only Top-K and a market-neutral Long–Short setup. For the Long-only setting, Tables 5 and 7 report CAGR% for equal-weighted Top-K portfolios, with $K = 10$ on the S&P500 and SXXP universes and $K = 5$ on the DJI, SX5E, and NASDAQ 100 universes, respectively. For the Long–Short setting, Tables 6 and 8 report Sharpe and Sortino ratios for market-neutral Top-K portfolios across the same groups of universes and holding periods. SVAT provides predictions only on $L = 1$ due to the preprocessing applied by FinBench-EVAL.

Discussion. Portfolio performance depends strongly on strategy type, universe, and prediction horizon. Long-only results are generally more robust, especially for Top-10 portfolios on the S&P500 (Table 5), where HIST and MATCC achieve high short-horizon CAGR%, while THGNN and FinFormer show stable gains. SXXP results are more moderate, with outperformance mainly at medium or long horizons. For Top-5 Long-only portfolios (Table 7, DJI is mostly below benchmark, SX5E is the most favorable universe, and NASDAQ100 is harder to beat due to its strong benchmark, except for selected cases such as Adv-ALSTM and STHAN-SR. Long–Short strategies are less robust. In the Top-10 setting (Table 6), only a few S&P500 configurations improve over the benchmark, while SXXP shows better market-neutral performance for StockMixer, RT-GCN, CNNPred2D, and THGNN. Top-5 Long–Short portfolios (Table 8) further amplify instability: DJI remains weak, SX5E is again the most favorable, and NASDAQ100 is mostly dominated by the benchmark. Overall, no model consistently dominates across universes, horizons, and portfolio constructions.

4.5 RQ3: Ranking Signal Monotonicity

We evaluate ranking signal quality through quintile analysis. Figure 2 reports the annualized returns (CAGR%) of portfolios formed from five ranking-based quintiles across prediction horizons. Due to space constraints, we report one model per category (classification, regression, ranking), selected as the best-performing method on average within each group.

Discussion. Figure 2 shows that monotonic return patterns emerge only in specific settings, typically for longer horizons and in smaller markets. Deviations from monotonicity are frequent in shorter horizons and broader universes, indicating that ranking signal quality depends jointly on the model, the horizon, and the market context. These results underscore the importance of evaluating ranking behavior across the full cross-section rather than relying on isolated performance indicators.

Table 4: Ranking performance for multiple prediction horizons (L) reported in terms of IC, Rank IC, ICIR, and Rank ICIR. Results are aggregated across the three reported seeds and across the five testing years.

Model	L	DJI				SX5E				NASDAQ 100				S&P500				SXXP			
		IC	RankIC	ICIR	RankICIR	IC	RankIC	ICIR	RankICIR	IC	RankIC	ICIR	RankICIR	IC	RankIC	ICIR	RankICIR	IC	RankIC	ICIR	RankICIR
STHAN-SR	1	-0.014	-0.016	-0.044	-0.060	-0.013	-0.017	-0.057	-0.076	-0.009	-0.014	-0.032	-0.054	-0.004	-0.014	-0.016	-0.054	0.000	-0.011	0.002	-0.057
	5	-0.014	-0.025	-0.038	-0.097	-0.017	-0.023	-0.064	-0.085	-0.007	-0.014	-0.024	-0.057	-0.002	-0.011	-0.011	-0.046	0.000	-0.016	0.007	-0.078
	20	-0.009	-0.025	-0.031	-0.149	-0.032	-0.038	-0.160	-0.180	-0.001	-0.010	0.028	-0.017	-0.001	-0.005	-0.036	-0.047	-0.003	-0.022	-0.006	-0.142
SVAT	1	0.009	0.016	0.040	0.068	0.013	0.012	0.071	0.065	0.003	0.011	0.009	0.055	0.000	0.010	-0.007	0.084	0.000	0.006	-0.005	0.063
	5	0.008	0.015	0.043	0.059	0.030	0.021	0.168	0.110	0.006	0.014	0.043	0.095	0.003	0.008	0.050	0.088	0.006	0.011	0.137	0.127
	20	0.030	0.045	0.143	0.210	0.069	0.040	0.422	0.245	0.010	0.030	0.036	0.195	-0.005	0.013	-0.081	0.161	0.003	-0.001	0.058	-0.031
RT-GCN	1	-0.027	-0.032	-0.096	-0.134	-0.020	-0.023	-0.083	-0.099	-0.024	-0.022	-0.104	-0.089	-0.018	-0.022	-0.083	-0.099	-0.018	-0.013	-0.142	-0.109
	5	-0.061	-0.065	-0.222	-0.277	-0.046	-0.046	-0.195	-0.199	-0.050	-0.056	-0.234	-0.243	-0.030	-0.046	-0.157	-0.225	-0.031	-0.029	-0.254	-0.268
	20	-0.094	-0.098	-0.365	-0.462	-0.082	-0.080	-0.350	-0.361	-0.077	-0.096	-0.347	-0.471	-0.054	-0.087	-0.323	-0.482	-0.085	-0.059	-0.792	-0.470

Table 5: Long-only Top-10 portfolio performance (CAGR%) across S&P 500 and SXXP investment universes and prediction horizons (L). The first row reports the market benchmark; subsequent rows report model-based portfolio metrics. Results are aggregated across the five testing years.

Model	S&P 500			SXXP		
	L=1	L=5	L=20	L=1	L=5	L=20
Benchmark	0.13	0.13	0.13	0.04	0.04	0.04
THGNN	0.38	0.30	0.40	-0.02	0.10	0.16
Adv-ALSTM	0.00	-0.02	0.11	-0.07	0.14	0.15
CNNPred2D	–	0.39	0.21	–	0.10	0.27
StockMixer	-0.04	0.15	0.08	0.09	0.19	0.06
MATCC	1.81	0.16	0.06	-0.03	-0.09	0.05
HIST	2.32	0.19	0.25	-0.05	0.01	0.09
FinFormer	0.47	0.08	0.42	0.03	-0.02	0.10
STHAN-SR	0.02	0.16	0.12	0.03	0.13	0.16
SVAT	0.06	0.18	0.14	-0.07	–	–
RT-GCN	0.02	0.13	0.12	0.03	0.15	0.06

Table 6: Long–Short Top-10 portfolio performance across S&P 500 and SXXP investment universes and prediction horizons (L), reported in terms of Sharpe and Sortino ratios. The first row reports the market benchmark; subsequent rows report model-based Long–Short portfolio metrics. Results are aggregated across the five testing years.

Model	S&P500						SXXP					
	Sharpe			Sortino			Sharpe			Sortino		
	L=1	L=5	L=20	L=1	L=5	L=20	L=1	L=5	L=20	L=1	L=5	L=20
Benchmark	0.67	0.67	0.67	0.94	0.94	0.94	0.31	0.31	0.31	0.42	0.42	0.42
THGNN	-0.08	0.40	0.70	0.48	0.61	1.07	-0.30	0.23	0.75	-0.29	0.24	1.00
Adv-ALSTM	-1.07	-1.52	-0.59	-0.75	-1.13	-0.50	-0.98	-0.71	0.27	-0.88	-0.64	0.27
CNNPred2D	–	0.76	0.20	–	1.57	0.33	–	0.17	0.89	–	0.21	1.02
StockMixer	0.41	0.37	0.04	0.44	0.39	0.04	1.13	0.20	-0.15	1.25	0.22	-0.15
MATCC	-0.07	-0.08	0.04	0.34	-0.03	0.07	-0.06	-0.44	0.06	-0.04	-0.40	0.08
FinFormer	0.35	0.39	0.68	3.98	0.63	0.86	0.44	-0.05	0.27	0.47	-0.05	0.30
HIST	0.54	0.22	0.70	13.48	0.42	0.80	0.25	-0.21	0.42	0.24	-0.20	0.46
STHAN-SR	0.30	0.22	0.29	0.32	0.24	0.32	0.76	0.11	-0.15	0.84	0.12	-0.16
SVAT	0.50	0.60	-0.05	0.51	0.62	-0.05	0.02	–	–	0.02	–	–
RT-GCN	0.30	0.31	0.13	0.31	0.34	0.14	0.54	0.93	0.38	0.57	1.20	0.47

Table 7: Long-only Top-5 portfolio performance (CAGR%) across DJI, SX5E and NASDAQ100 investment universes and prediction horizons (L). The first row reports the market benchmark for reference. Results are aggregated across the five testing years.

Model	DJI			SX5E			NASDAQ100		
	L=1	L=5	L=20	L=1	L=5	L=20	L=1	L=5	L=20
Benchmark	0.08	0.08	0.08	0.06	0.06	0.06	0.19	0.19	0.19
THGNN	-0.05	-0.03	0.00	0.01	0.06	0.08	0.03	0.04	0.07
Adv-ALSTM	-0.07	0.11	0.00	0.09	0.24	0.12	0.09	0.15	0.25
CNNPred2D	–	0.02	-0.02	–	0.15	0.11	–	0.14	0.09
StockMixer	-0.08	-0.04	0.01	-0.02	0.06	0.07	0.01	0.07	0.04
MATCC	0.01	0.02	0.02	-0.07	0.11	0.15	-0.05	0.05	0.06
FinFormer	0.02	0.05	0.06	-0.01	0.03	0.03	0.00	0.02	0.07
HIST	0.02	0.03	0.01	-0.01	0.10	0.13	-0.01	-0.07	0.01
STHAN-SR	-0.12	-0.07	0.01	0.09	0.09	0.07	0.09	0.21	0.06
SVAT	-0.01	-0.01	-0.06	0.01	0.03	0.09	0.05	0.13	0.09
RT-GCN	-0.01	0.01	0.08	0.07	0.12	0.11	0.02	0.12	0.05

4.6 RQ4: Robustness under Market Stress

We evaluate temporal robustness by comparing End-of-Year (EOY) returns of Long-only Top-10 strategies against the market benchmark for each year in the test period. Figure 3 illustrates portfolio performance across different market regimes. The models shown follow the same selection criterion adopted in the previous section, with one representative model reported for each category due to space constraints.

Discussion. Figure 3 highlights substantial temporal variability in portfolio performance. While model-based strategies can outperform the benchmark in certain years, these gains are not consistent over time and vary across universes and horizons. Performance differences tend to compress during stressed or volatile periods, indicating that robustness under market stress remains a challenging and model-dependent property.

4.7 Discussion

Across both task- and portfolio-level evaluations, results show that model performance is highly sensitive to market context, preprocessing choices, horizon length, and evaluation settings. No model consistently dominates across prediction tasks, investment universes, or market regimes, and conclusions drawn from isolated experimental configurations can therefore be misleading. Portfolio

Table 8: Long–Short Top-5 portfolio performance across DJI, SX5E and NASDAQ100 investment universes and prediction horizons (L), evaluated using Sharpe and Sortino ratios. The first row reports the market benchmark for reference. Results are aggregated across the five testing years.

Model	DJI						SX5E						NASDAQ100					
	Sharpe			Sortino			Sharpe			Sortino			Sharpe			Sortino		
	L=1	L=5	L=20	L=1	L=5	L=20	L=1	L=5	L=20	L=1	L=5	L=20	L=1	L=5	L=20	L=1	L=5	L=20
Benchmark	0.49	0.49	0.49	0.68	0.68	0.68	0.36	0.36	0.36	0.49	0.49	0.49	0.81	0.81	0.81	1.14	1.14	1.14
THGNN	-0.66	-0.79	-0.54	-0.63	-0.76	-0.52	-0.28	0.20	0.47	-0.28	0.20	0.49	0.08	-0.40	-0.18	0.08	-0.38	-0.16
Adv-ALSTM	-1.04	0.73	-0.19	-0.99	0.74	-0.18	0.40	0.48	-0.47	0.41	0.51	-0.46	-0.12	0.61	0.72	-0.15	0.62	0.74
CNNPred2D	–	-0.52	-0.77	–	-0.50	-0.75	–	0.04	-0.04	–	0.05	-0.03	–	0.11	0.12	–	0.12	0.13
StockMixer	-0.12	-0.13	-0.23	-0.12	-0.14	-0.24	0.27	0.16	-0.02	0.29	0.16	-0.01	0.60	0.10	-0.04	0.33	0.11	-0.04
MATCC	0.19	0.20	-0.16	0.19	0.21	-0.16	-12.23	0.26	0.31	-1.15	0.29	0.34	-0.06	0.18	0.12	-0.05	0.19	0.13
FinFormer	0.20	0.03	0.15	0.21	0.04	0.18	-0.34	-0.06	0.01	-0.33	-0.05	0.01	0.06	-0.36	0.34	0.05	-0.35	0.36
HIST	0.38	0.16	0.11	0.40	0.17	0.11	-0.26	0.04	0.69	-0.26	0.05	0.73	0.16	-0.49	-0.17	0.16	-0.48	-0.17
STHAN-SR	-0.45	-0.45	-0.16	-0.45	-0.47	-0.16	0.83	0.23	0.12	0.91	0.24	0.13	0.62	0.07	-0.19	0.67	0.07	-0.20
SVAT	-0.24	-0.12	-0.81	-0.23	-0.11	-0.77	0.22	-0.61	0.34	0.23	-0.58	0.35	0.77	0.01	1.27	0.87	0.01	1.36
RT-GCN	0.13	-0.16	0.17	0.13	-0.17	0.18	0.52	0.15	-0.04	0.55	0.16	-0.04	0.32	0.05	0.19	0.34	0.05	0.20

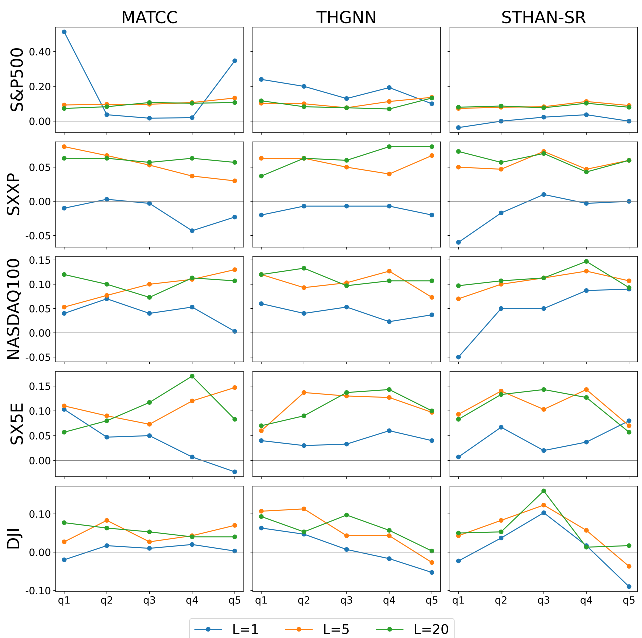

Figure 2: Quintile analysis for MATCC, THGNN, and STHAN-SR models. Panels reports the annualized return (CAGR%) of quintile portfolios for the prediction horizons $L = 1, 5, 20$.

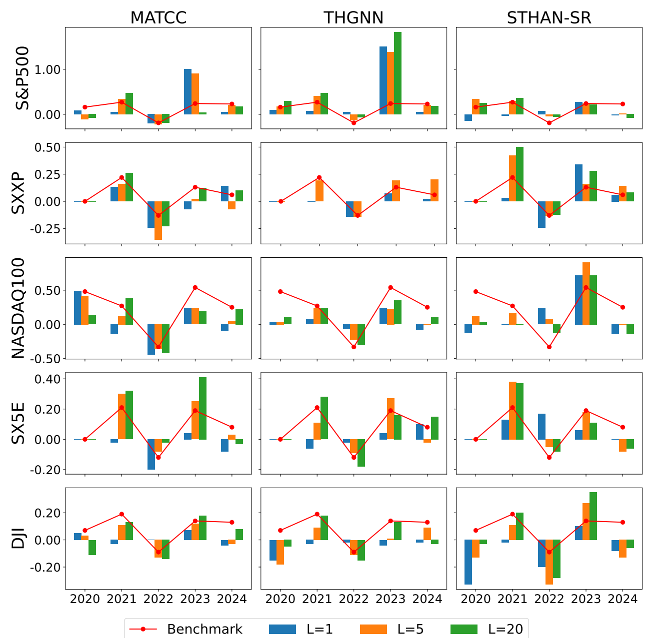

Figure 3: End-of-Year (EOY) returns of Long-only Top-K strategies across investment universes and prediction horizons for MATCC, THGNN and STHAN-SR, compared to the market benchmark.

gains, ranking signal quality, and temporal robustness vary with market structure and evaluation design, reinforcing the need for a unified benchmarking framework that evaluates models across tasks, portfolio constructions, and market conditions.

5 Conclusion

This work introduces FinBench and applies it to benchmark a wide range of stock market prediction models across multiple tasks, portfolio constructions, and market conditions. The empirical analysis shows that both predictive accuracy and economic performance vary substantially with evaluation choices, market structure, and temporal context, making conclusions drawn from isolated experimental settings unreliable. These findings demonstrate the necessity of a benchmarking framework such as FinBench for fair and systematic comparison of future methods.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support.

References

[1] Zhiyuan Cao, Jiayu Xu, Chengqi Dong, Peiwen Yu, and Tian Bai. 2024. MATCC: A Novel Approach for Robust Stock Price Prediction Incorporating Market Trends and Cross-time Correlations. In *CIKM*. ACM, 187–196.

[2] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *KDD*. ACM, 785–794.

[3] Jiezhu Cheng, Kaizhu Huang, and Zibin Zheng. 2024. Can Perturbations Help Reduce Investment Risks? Risk-aware Stock Recommendation via Split Variational Adversarial Training. *ACM Trans. Inf. Syst.* 42, 4 (2024), 101:1–101:28.

[4] Kangmin Choi, Geon Shin, Jungwoo Yang, and Hyunjoon Kim. 2025. DYCOR: Capturing Hidden Stock Relationships for Stock Trend Prediction. In *CIKM*. ACM, 458–467.

[5] Chaoran Cui, Xiaojie Li, Chunyun Zhang, Weili Guan, and Meng Wang. 2023. Temporal-Relational hypergraph tri-Attention networks for stock trend prediction. *Pattern Recognit.* 143 (2023), 109759.

[6] Yitong Duan, Lei Wang, Qizhong Zhang, and Jian Li. 2022. FactorVAE: A Probabilistic Dynamic Factor Model Based on Variational Autoencoder for Predicting Cross-Sectional Stock Returns. In *AAAI*. AAAI Press, 4468–4476.

[7] Jinyong Fan and Yanyan Shen. 2024. StockMixer: A Simple Yet Strong MLP-Based Architecture for Stock Price Forecasting. In *AAAI*. AAAI Press, 8389–8397.

[8] Saeede Anbaee Farimani, Majid Vafaei Jahan, Amin Milani Fard, and Seyed Reza Kamel Tabbakh. 2022. Investigating the informativeness of technical indicators and news sentiment in financial market price prediction. *Knowl. Based Syst.* 247 (2022), 108742.

[9] Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. 2019. Enhancing Stock Movement Prediction with Adversarial Training. In *IJCAI*. ijcai.org, 5843–5849.

[10] Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. 2019. Temporal Relational Ranking for Stock Prediction. *ACM Trans. Inf. Syst.* 37, 2 (2019), 27:1–27:30.

[11] Qiang Gao, Zhengxiang Liu, Li Huang, Kunpeng Zhang, Jun Wang, and Guisong Liu. 2025. Relational Stock Selection via Probabilistic State Space Learning. *IEEE Trans. Knowl. Data Eng.* 37, 2 (2025), 865–880.

[12] Qiang Gao, Xinzhu Zhou, Li Huang, Kunpeng Zhang, Siyuan Liu, and Fan Zhou. 2024. Relational Fusion-based Stock Selection with Neural Recursive Ordinary Differential Equation Networks. *Inf. Fusion* 110 (2024), 102468.

[13] Albert Gu and Tri Dao. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *CoRR* abs/2312.00752 (2023).

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), 1735–1780.

[15] Ehsan Hoseinzade and Saman Haratizadeh. 2019. CNNpred: CNN-based stock market prediction using a diverse set of variables. *Expert Syst. Appl.* 129 (2019), 273–285.

[16] Jingyi Hou, Zhen Dong, Jiayu Zhou, and Zhijie Liu. 2024. Discovering Predictable Latent Factors for Time Series Forecasting. *IEEE Trans. Knowl. Data Eng.* 36, 10 (2024), 5106–5119.

[17] Yi-Ling Hsu, Yu-Che Tsai, and Cheng-Te Li. 2023. FinGAT: Financial Graph Attention Networks for Recommending Top-KK Profitable Stocks. *IEEE Trans. Knowl. Data Eng.* 35, 1 (2023), 469–481.

[18] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to Chaotic Whispers: A Deep Learning Framework for News-oriented Stock Trend Prediction. In *WSDM*. ACM, 261–269.

[19] Thanh Trung Huynh, Minh Hieu Nguyen, Thanh Tam Nguyen, Phi Le Nguyen, Matthias Weidlich, Quoc Viet Hung Nguyen, and Karl Aberer. 2023. Efficient Integration of Multi-Order Dynamics and Internal Dynamics in Stock Movement Prediction. In *WSDM*. ACM, 850–858.

[20] Raehyun Kim, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, and Jaewoo Kang. 2019. HATS: A Hierarchical Graph Attention Network for Stock Movement Prediction. *CoRR* abs/1908.07999 (2019).

[21] Kelvin J. L. Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2023. Diffusion Variational Autoencoder for Tackling Stochasticity in Multi-Step Regression Stock Price Prediction. In *CIKM*. ACM, 1087–1096.

[22] Kelvin J. L. Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2024. Learning to Generate Explainable Stock Predictions using Self-Reflective Large Language Models. In *WWW*. ACM, 4304–4315.

[23] Shuqi Li, Weiheng Liao, Yuhan Chen, and Rui Yan. 2023. PEN: Prediction-Explanation Network to Forecast Stock Price Movement with Better Explainability. In *AAAI*. AAAI Press, 5187–5194.

[24] Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. 2024. MASTER: Market-Guided Stock Transformer for Stock Price Forecasting. In *AAAI*. AAAI Press, 162–170.

[25] Hengxu Lin, Dong Zhou, Weiqing Liu, and Jiang Bian. 2021. Learning Multiple Stock Trading Patterns with Temporal Routing Adaptor and Optimal Transport. In *KDD*. ACM, 1017–1026.

[26] Haotian Liu, Yadong Zhou, Yuxun Zhou, and Bowen Hu. 2024. Heterogeneous Dual-Dynamic Attention Network for Modeling Mutual Interplay of Stocks. *IEEE Trans. Artif. Intell.* 5, 7 (2024), 3595–3606.

[27] Di Luo, Weiheng Liao, Shuqi Li, Xin Cheng, and Rui Yan. 2023. Causality-Guided Multi-Memory Interaction Network for Multivariate Stock Price Movement Prediction. In *ACL (1)*. Association for Computational Linguistics, 12164–12176.

[28] Ali Mehrabian, Ehsan Hoseinzade, Mahdi Mazloum, and Xiaohong Chen. 2025. Mamba Meets Financial Markets: A Graph-Mamba Approach for Stock Price Prediction. In *ICASSP*. IEEE, 1–5.

[29] Sara Pederzoli, Francesco del Buono, Maurizio Vincini, and Francesco Guerra. 2026. Trend-Error Decomposition for Self-Supervised Time Series Learning in Multivariate Forecasting Task. *IEEE Access* 14 (2026), 8618–8631.

[30] Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, Tyler Derr, and Rajiv Ratn Shah. 2021. Stock Selection via Spatiotemporal Hypergraph Attention Network: A Learning to Rank Approach. In *AAAI*. AAAI Press, 497–504.

[31] Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020. Deep Attentive Learning for Stock Movement Prediction From Social Media Text and Company Correlations. In *EMNLP (1)*. Association for Computational Linguistics, 8415–8426.

[32] Yejun Soun, Jaemin Yoo, Minyong Cho, Jihyeong Jeon, and U Kang. 2022. Accurate Stock Movement Prediction with Self-supervised Learning from Sparse Noisy Tweets. In *IEEE Big Data*. IEEE, 1691–1700.

[33] Shengkun Wang, Yangxiao Bai, Taoran Ji, Kaiqun Fu, Linhan Wang, and Chang-Tien Lu. 2023. Stock Movement and Volatility Prediction from Tweets, Macroeconomic Factors and Historical Prices. In *IEEE Big Data*. IEEE, 1863–1872.

[34] Shengkun Wang, Taoran Ji, Linhan Wang, Yanshen Sun, Shang-Ching Liu, Amit Kumar, and Chang-Tien Lu. 2024. StockTime: A Time Series Specialized Large Language Model Architecture for Stock Price Prediction. *CoRR* abs/2409.08281 (2024).

[35] Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. 2018. Hybrid Deep Sequential Modeling for Social Text-Driven Stock Prediction. In *CIKM*. ACM, 1627–1630.

[36] Sheng Xiang, Dawei Cheng, Chencheng Shang, Ying Zhang, and Yuqi Liang. 2022. Temporal and Heterogeneous Graph Neural Network for Financial Time Series Prediction. In *CIKM*. ACM, 3584–3593.

[37] Wentao Xu, Weiqing Liu, Lewen Wang, Yingce Xia, Jiang Bian, Jian Yin, and Tie-Yan Liu. 2021. HIST: A Graph-based Framework for Stock Trend Forecasting via Mining Concept-Oriented Shared Information. *CoRR* abs/2110.13716 (2021).

[38] Wentao Xu, Weiqing Liu, Chang Xu, Jiang Bian, Jian Yin, and Tie-Yan Liu. 2021. REST: Relational Event-driven Stock Trend Forecasting. In *WWW*. ACM / IW3C2, 1–10.

[39] Yumo Xu and Shay B. Cohen. 2018. Stock Movement Prediction from Tweets and Historical Prices. In *ACL (1)*. Association for Computational Linguistics, 1970–1979.

[40] Linyi Yang, Jiazheng Li, Ruihai Dong, Yue Zhang, and Barry Smyth. 2022. NumHTML: Numeric-Oriented Hierarchical Transformer Model for Multi-Task Financial Forecasting. In *AAAI*. AAAI Press, 11604–11612.

[41] Xiao Yang, Weiqing Liu, Dong Zhou, Jiang Bian, and Tie-Yan Liu. 2020. Qlib: An AI-oriented Quantitative Investment Platform. *CoRR* abs/2009.11189 (2020).

[42] Jaemin Yoo, Yejun Soun, Yong-chan Park, and U Kang. 2021. Accurate Multivariate Stock Movement Prediction via Data-Axis Transformer with Multi-Level Contexts. In *KDD*. ACM, 2037–2045.

[43] Zinuo You, Zijian Shi, Hongbo Bo, John Cartlidge, Li Zhang, and Yan Ge. 2024. DGDNN: Decoupled Graph Diffusion Neural Network for Stock Movement Prediction. In *ICAART (2)*. SCITEPRESS, 431–442.

[44] Liheng Zhang, Charu C. Aggarwal, and Guo-Jun Qi. 2017. Stock Price Prediction via Discovering Multi-Frequency Trading Patterns. In *KDD*. ACM, 2141–2149.

[45] Yu Zhao, Huaming Du, Ying Liu, Shaopeng Wei, Xingyan Chen, Fuzhen Zhuang, Qing Li, and Gang Kou. 2023. Stock Movement Prediction Based on Bi-Typed Hybrid-Relational Market Knowledge Graph via Dual Attention Networks. *IEEE Trans. Knowl. Data Eng.* 35, 8 (2023), 8559–8571.

[46] Zetao Zheng, Jie Shao, Jia Zhu, and Heng Tao Shen. 2023. Relational Temporal Graph Convolutional Networks for Ranking-Based Stock Prediction. In *ICDE*. IEEE, 123–136.

[47] Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the Event: Corporate Events Detection for News-Based Event-Driven Trading. In *ACL/IJCNLP (Findings) (Findings of ACL, Vol. ACL/IJCNLP 2021)*. Association for Computational Linguistics, 2114–2124.

[48] Yi Zu, Jiacong Mi, Lingning Song, Shan Lu, and Jieyue He. 2023. Finformer: A Static-dynamic Spatiotemporal Framework for Stock Trend Prediction. In *IEEE Big Data*. IEEE, 1460–1469.

A Features

The primitive micro- and macro-structure features are reported directly in the GitHub repository, together with the corresponding data source from which each feature is collected. Tables 9 and 10 report only the derived features computed within FinBench.

Table 9: Derivative Market Microstructure Indicators

#	Variable	Description	Calculation
1	**Alpha158**	Alpha Factors Indicators	See Qlib [41]
2	**Alpha360**	Alpha Factors Indicators	See Qlib [41]
3	$\mathbf{c}_p$	p price normalized by close price minus one, $p \in \{\text{Open, High, Low}\}$	$\mathrm{c}_p = \frac{p}{\text{Close}} - 1$
4	***p*_price_change**	Daily p price return, $p \in \{\text{Open, High, Low, Close}\}$	$p\text{\_price\_change} = \frac{p_t}{p_{t-1}} - 1$
5	**adj_*p*_price_change**	Daily adjusted p price return, $p \in \{\text{Open, High, Low, Close}\}$	$\text{adj\_}p\text{\_price\_change} = \frac{\text{Adj. } p_t}{\text{Adj. } p_{t-1}} - 1$
6	***p*_close_ratio_d1**	p price to previous close ratio, $p \in \{\text{Open, High, Low, Close}\}$	$p\text{\_close\_ratio\_d1} = \frac{p_t}{\text{Close}_{t-1}}$
7	**adj_*p*_close_ratio_d1**	Adjusted p to previous adjusted close ratio, $p \in \{\text{Open, High, Low, Close}\}$	$\text{adj\_}p\text{\_close\_ratio\_d1} = \frac{\text{Adj. } p_t}{\text{Adj. Close}_{t-1}}$
8	**adj_close_abs_return**	Absolute adjusted close price change	$\text{Adj. Close}_t - \text{Adj. Close}_{t-1}$
9	**close_abs_return**	Absolute close price change	$\text{adj\_close\_abs\_return} = \text{Close}_t - \text{Close}_{t-1}$
10	$\mathbf{z}_{dk}$	Normalized k-day adjusted close moving average, $k \in \{5, 10, 15, 20, 25, 30\}$	$\frac{\sum_{i=0}^{k-1} \text{Adj. Close}_{t-i}}{k \cdot \text{Adj. Close}_t} - 1$
11	$\mathbf{f}_{dk}$	Alternative normalized k-day adjusted close moving average, $k \in \{5, 10, 15, 20, 25, 30\}$	$\frac{\sum_{i=0}^{k-1} \text{Adj. Close}_{t-i}}{k(\text{Adj. Close}_t - 1)}$
12	**vol**	Daily volume return	$\frac{\text{Volume}_t}{\text{Volume}_{t-1}} - 1$
13	$\mathbf{mom}_k$	k-day adjusted close momentum, $k \in \{1, 2, 3\}$	$\text{Adj. Close}_t - \text{Adj. Close}_{t-k}$
14	$\mathbf{roc}_k$	k-day rate of change of adjusted close, $k \in \{5, 10, 15, 20\}$	$100 \cdot \frac{\text{Adj. Close}_t - \text{Adj. Close}_{t-k}}{\text{Adj. Close}_{t-k}}$
15	$\mathbf{ema}_k$	k-day exponential moving average of adjusted close, $k \in \{10, 20, 50, 200\}$	$\text{EMA}(\text{Adj. Close}, k)$
16	**Adj. *p***	p price adjusted using the adjustment factor, $p \in \{\text{Open, High, Low}\}$	$\text{Adj. } p = p \cdot \frac{\text{Adj. Close}}{\text{Close}}$
17	**TTM Dividends Paid**	Trailing twelve months of cash dividends paid	$\text{TTM Dividends Paid} = \sum_{i=0}^{3} \text{Dividends Paid}_{t-i}$
18	**TTM Net Income**	Trailing twelve months of net income	$\text{TTM Net Income} = \sum_{i=0}^{3} \text{Net Income}_{t-i}$
19	**TTM Total Revenue**	Trailing twelve months of total revenue	$\text{TTM Total Revenue} = \sum_{i=0}^{3} \text{Total Revenue}_{t-i}$
20	**TTM Gross Profit**	Trailing twelve months of gross profit	$\text{TTM Gross Profit} = \sum_{i=0}^{3} \text{Gross Profit}_{t-i}$
21	**Sales Growth**	Growth rate of total revenue over the last year	$\text{Sales Growth} = \frac{\text{Total Revenue}_t}{\text{Total Revenue}_{t-4}} - 1$
22	**Earnings Growth**	Growth rate of net income over the last year	$\text{Earnings Growth} = \frac{\text{Net Income}_t}{\text{Net Income}_{t-4}} - 1$
23	**Gross Profit to Assets**	Ratio of TTM gross profit to total assets	$\text{Gross Profit to Assets} = \frac{\text{TTM Gross Profit}}{\text{Total Assets}}$
24	**ROE**	Return on equity	$\text{ROE} = \frac{\text{TTM Net Income}}{\text{Total Equity}}$
25	**ROA**	Return on assets	$\text{ROA} = \frac{\text{TTM Net Income}}{\text{Total Assets}}$
26	**Market Leverage**	Total liabilities to total assets ratio	$\text{Market Leverage} = \frac{\text{Total Liabilities}}{\text{Total Assets}}$
27	**Market Cap**	Market capitalization	$\text{Market Cap} = \text{Adj. Close} \cdot \text{Shares Outstanding}$
28	**Dividend Yield**	TTM dividends over market cap	$\text{Dividend Yield} = \frac{\text{TTM Dividends Paid}}{\text{Market Cap}}$
29	**Earnings to Price**	TTM net income over market cap	$\text{Earnings to Price} = \frac{\text{TTM Net Income}}{\text{Market Cap}}$
30	**Sales to Price**	TTM total revenue over market cap	$\text{Sales to Price} = \frac{\text{TTM Total Revenue}}{\text{Market Cap}}$
31	**Book Value to Price**	Total equity over market cap	$\text{Book Value to Price} = \frac{\text{Total Equity}}{\text{Market Cap}}$
32	**Volatility**	Annualized rolling 252-day volatility of adjusted close	$\text{Volatility} = \text{Std}(\text{Adj. Close eturns}, 252)$
33	**Momentum**	21-day vs 252-day price momentum	$\text{Momentum} = \frac{\text{Adj. Close}_{t-21}}{\text{Adj. Close}_{t-252}} - 1$
34	**REV**	Recent 1-month return	$\text{REV} = \frac{\text{Adj. Close}_t}{\text{Adj. Close}_{t-21}} - 1$
35	**Share Turnover**	21-day volume over shares outstanding	$\text{Share Turnover} = \frac{\sum_{i=0}^{20} \text{Volume}_{t-i}}{\text{Shares Outstanding}}$

Table 10: Derivative Market Macrostructure Features

#	Variable	Description	Calculation
1	**price_change**	Daily adjusted close return of market index.	$\text{price\_change} = \frac{\text{Adj. Close}_t}{\text{Adj. Close}_{t-1}} - 1$
2	**mean_price_change**$_k$	Rolling mean of daily adjusted close returns, $k \in \{5, 10, 20, 30, 60\}$	$\text{mean\_price\_change}_k = \frac{1}{k}\sum_{i=0}^{k-1} \text{price\_change}_{t-i}$
3	**std_price_change**$_k$	Rolling standard deviation of daily adjusted close returns, $k \in \{5, 10, 20, 30, 60\}$	$\text{std\_price\_change}_k = \sqrt{\frac{1}{k}\sum_{i=0}^{k-1} \left(\text{price\_change}_{t-i} - \mu_k\right)^2}$
4	**mean_vol**$_k$	Rolling mean of trading volume normalized by current volume, $k \in \{5, 10, 20, 30, 60\}$	$\text{mean\_vol}_k = \frac{\frac{1}{k}\sum_{i=0}^{k-1} \text{Volume}_{t-i}}{\text{Volume}_t}$
5	**std_vol**$_k$	Rolling standard deviation of trading volume normalized by current volume, $k \in \{5, 10, 20, 30, 60\}$	$\text{std\_vol}_k = \frac{\sqrt{\frac{1}{k}\sum_{i=0}^{k-1} \left(\text{Volume}_{t-i} - \bar{V}_k\right)^2}}{\text{Volume}_t}$
6	**Commodity_Return**$_c$	Daily return of commodity futures, $c \in$ {HG, NG, GC, SI}	$\text{Commodity\_Return}_c = \frac{P_t^{(c)}}{P_{t-1}^{(c)}} - 1$
7	**World_Index_Return**$_i$	Daily return of global equity indices, $i \in$ {GSPC, IXIC, DJI, NYA, HSI, SSE, FCHI, GDAXI}	$\text{World\_Index\_Return}_i = \frac{P_t^{(i)}}{P_{t-1}^{(i)}} - 1$
8	**ETF_Return**$_f$	Daily return of equity ETFs, $f \in$ {IWM, ISF}	$\text{ETF\_Return}_f = \frac{P_t^{(f)}}{P_{t-1}^{(f)}} - 1$
9	**FX_Return**$_x$	Daily return of foreign exchange rates and metal FX pairs, $x \in$ { USDJPY, USDGBP, USDCAD, USDCNY,USDAUD, USDNZD, USDCHF, USDEUR, XAUUSD, XAGUSD}	$\text{FX\_Return}_x = \frac{P_t^{(x)}}{P_{t-1}^{(x)}} - 1$
10	**USD_Index_Return**	Daily return of the US Dollar Index (DXY)	$\text{USD\_Index\_Return} = \frac{P_t^{(\text{DXY})}}{P_{t-1}^{(\text{DXY})}} - 1$
11	**US_Stock_Return**$_s$	Daily return of selected US equities, $s \in$ {XOM, JPM, AAPL, MSFT, GE, JNJ, WFC, AMZN}	$\text{US\_Stock\_Return}_s = \frac{P_t^{(s)}}{P_{t-1}^{(s)}} - 1$
12	**TE1**	Term spread (10Y – 4W)	$\text{TE1} = \text{DGS10}_t - \text{DTB4WK}_t$
13	**TE2**	Term spread (10Y – 3M)	$\text{TE2} = \text{DGS10}_t - \text{DTB3}_t$
14	**TE3**	Term spread (10Y – 6M)	$\text{TE3} = \text{DGS10}_t - \text{DTB6}_t$
15	**TE5**	Term spread (3M – 4W)	$\text{TE5} = \text{DTB3}_t - \text{DTB4WK}_t$
16	**TE6**	Term spread (6M – 4W)	$\text{TE6} = \text{DTB6}_t - \text{DTB4WK}_t$
17	**DE1**	Default spread (Baa – Aaa)	$\text{DE1} = \text{DBAA}_t - \text{DAAA}_t$
18	**DE2**	Default spread (Baa – 10Y)	$\text{DE2} = \text{DBAA}_t - \text{DGS10}_t$
19	**DE4**	Default spread (Baa – 6M)	$\text{DE4} = \text{DBAA}_t - \text{DTB6}_t$
20	**DE5**	Default spread (Baa – 3M)	$\text{DE5} = \text{DBAA}_t - \text{DTB3}_t$
21	**DE6**	Default spread (Baa – 4W)	$\text{DE6} = \text{DBAA}_t - \text{DTB4WK}_t$