AEQUITAS: A Toolkit for Stock Market Prediction and Portfolio Evaluation

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

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

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

Abstract

Stock market prediction has produced a large number of models: comparing them in practice remains difficult. Existing approaches are often released as isolated implementations, use different data construction and preprocessing pipelines, and report heterogeneous machine learning and portfolio metrics. As a result, analysts cannot easily determine whether observed performance differences are due to the models themselves or to differences in data preparation, experimental design, and evaluation protocols. To address these challenges, we present AEQUITAS, **A**n **E**valuation toolkit for **QU**ant**I**tative and **T**rading **A**naly**S**is. It supports the construction, storage, and reuse of heterogeneous financial datasets, integrates classification, regression, and ranking models under consistent settings, and links model outputs to standardized portfolio strategies. The demonstration showcases how users can configure stock prediction experiments, compare pipelines across horizons, investment universes, and portfolio settings, and interpret predictive and portfolio-level results through an interactive dashboard with LLM-based support.

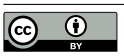
CCS Concepts

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

Keywords

Portfolio Allocation; Stock prediction; Computational finance

ACM Reference Format:
Sara Pederzoli, Marta Santacroce, Francesco Del Buono, Marco Bergianti, and Francesco Guerra. 2026. AEQUITAS: A Toolkit for Stock Market Prediction and Portfolio Evaluation. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3799682.3840240



This work is licensed under a Creative Commons Attribution 4.0 International License.
CIKM '26, Rome, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3840240

1 Introduction

Stock market prediction is a central problem in quantitative finance, supporting investment decisions and portfolio management. It is commonly addressed through classification (e.g., predicting price direction), regression (e.g., forecasting returns), and ranking (e.g., ordering assets by expected performance). These task formulations differ in their outputs but share the same objective: translating predictive signals into portfolio decisions.

Existing approaches are usually released as independent repositories, each tied to specific datasets, preprocessing steps, task formulations, and evaluation scripts. This fragmentation makes fair comparison difficult. Datasets are often proprietary, partially accessible [7], or reconstructed through undocumented pipelines [17, 25], limiting reuse across models. At the same time, evaluation protocols vary widely: some works report only task-level metrics, while others use different subsets of financial indicators, making it unclear whether a model has practical investment value. Moreover, many pipelines are evaluated under narrow settings and simplifying assumptions [16], with limited stress testing across horizons, universes, and market regimes. Comparisons are also frequently based on weak or outdated baselines [1, 22], or on incompatible experimental settings, which makes it difficult for analysts to identify which prediction pipeline should actually be implemented.

To address these limitations, we developed FinBench [14], a benchmark that defines a unified protocol for evaluating financial prediction models. It standardizes data construction, task formulation, and portfolio-oriented evaluation across heterogeneous financial signals, enabling fair comparison among classification, regression, and ranking approaches. While it provides the protocol and implementation needed for reproducible benchmarking, FinBench does not offer an interactive workflow for analysts to configure experiments, monitor executions, and inspect portfolio results through a graphical interface.

Building on this foundation, we propose AEQUITAS: **A**n **E**valuation toolkit for **QU**ant**I**tative and **T**rading **A**naly**S**is, designed for financial prediction and portfolio evaluation. AEQUITAS makes FinBench accessible through a configurable backend and a graphical interface that helps users configure experiments, monitor executions, visualize results, and compare model behavior across horizons, investment universes, and market regimes. The toolkit

is organized into three layers: data layer, model layer, and evaluation layer. The first supports the construction of reusable financial datasets by collecting, cleaning, aligning, and storing heterogeneous data from authoritative sources, including historical prices, company relations, macroeconomic indicators, and financial news. On top of this data layer, the model layer integrates a broad collection of stock prediction models from the literature, covering classification, regression, and ranking tasks. Finally, the evaluation layer connects model outputs to portfolio construction strategies and assesses them through standardized financial metrics. This layered design allows users to compare state-of-the-art models along two complementary dimensions: task-level predictive quality and downstream portfolio utility. As a result, financial analysts can select the models to evaluate and, accordingly, identify the data required to run them, while avoiding comparisons driven by incompatible datasets, preprocessing choices, temporal splits, prediction horizons, or metric families.

The demonstration presents AEQUITAS as an interactive environment for stock prediction and portfolio evaluation. It shows how analysts can move from reproducible dataset construction to model execution and portfolio-level inspection, supporting two common analytical needs: selecting prediction pipelines for a given market setting and assessing the robustness of their behavior across different universes and portfolio segments. We illustrate these needs through two scenarios: cross-universe model selection and quintile-based ranking stability analysis. The source code is available online[1], together with an explanatory video of the tool[2].

2 Related Work

Stock movement prediction is commonly studied via classification [3, 8, 18], regression [1, 4, 11, 12], and ranking [2, 15, 24] tasks. These models span sequential, attention-based, graph neural network, and signal decomposition approaches [13], often incorporating heterogeneous signals such as news and social media, reflecting both the stochastic nature of financial markets and the importance of inter-stock dependencies. However, existing datasets and evaluation protocols are fragmented: some focus only on price data [21], others include textual [20, 23] or relational information [6] but remain limited in scale, temporal coverage, or modality integration. Evaluation practices are often inconsistent, as models are typically compared using standard machine learning metrics (e.g., accuracy or MSE), while economic performance measures—such as returns or risk-adjusted metrics—are not systematically considered [9]. FinTSB [10] attempts to provide a unified benchmarking framework, but is limited to purely numerical time series data and standard forecasting models, without supporting heterogeneous information (e.g., relational or textual data) or finance-specific architectures. FinBench [14] is the only existing framework that jointly supports heterogeneous financial data, multiple model families, and standardized portfolio-oriented evaluation. However, it remains a benchmark rather than an interactive environment for experiment configuration, execution monitoring, and result analysis, motivating the development of AEQUITAS.

[1]https://github.com/softlab-unimore/aequitas
[2]https://youtu.be/kzshWMwvhjM

3 The AEQUITAS Architecture

Figure 1 summarizes the architecture of AEQUITAS, which builds on the FinBench benchmark and turns its data, model, and evaluation components into an operational toolkit for stock prediction and portfolio evaluation. The Data layer retrieves and prepares the input datasets. The Model layer manages training and testing with the chosen hyperparameters, while the Evaluation layer analyzes predictive and portfolio-level results. On top of these components, a graphical interface exposes the workflow to users, supporting experiment configuration, execution monitoring, result visualization, and LLM-based results explanation. The selected baselines include classification models such as Adv-ALSTM [5], THGNN [18], and MAN-SF [16]; regression models such as MASTER [12], MATCC [1], HIST [19], DiscoverPLF [9]; and ranking models such as STHAN-SR [15], SVAT [2], and RT-GCN [24]. The complete list of implemented models and features is reported in the GitHub repository.

3.1 Data Layer

The Data layer is the entry point through which AEQUITAS turns the data specification of FinBench into an executable and reusable data construction workflow. Starting from the selected model and investment universe, AEQUITAS derives the set of required inputs, checks their availability in the local storage, and exposes this status through the graphical interface. Users can therefore verify which data are already available, identify missing inputs, and trigger a targeted retrieval of only the required sources.

After retrieval, the layer cleans, temporally aligns, and stores the resulting signals, making them reusable across subsequent experiments. This process is important because different models rely on different combinations of inputs, including price-based features, macroeconomic variables, relational structures, textual signals, and company metadata. By making these inputs explicit and reusable, AEQUITAS enables users to run and compare heterogeneous models under the same experimental conditions.

The signals made available by AEQUITAS follow FinBench and support complementary modeling needs. Market microstructure indicators, including *OHLCV* data, adjusted prices, volumes, liquidity measures, technical indicators, and alpha factors [21], describe asset-specific trading behavior. Macroeconomic and global market indicators, such as indices and commodities, capture broader market conditions. Relational financial indicators, such as correlation-based asset networks, represent co-movement and hedging relations among stocks. Contextual–semantic indicators, including sector and industry metadata, corporate governance and ownership relations, and financial news [16], provide information about common exposures, structural dependencies, and event-driven signals.

3.2 Model Layer

The Model layer turns model specifications into executable prediction pipelines. Through the graphical interface, users can choose among classification, regression, and ranking models, and configure the main experimental settings required for execution, including date selection, input sequence length, prediction horizon, and model-specific hyperparameters. Users can then launch training and testing procedures and monitor their execution through logs and configuration summaries exposed by the interface.

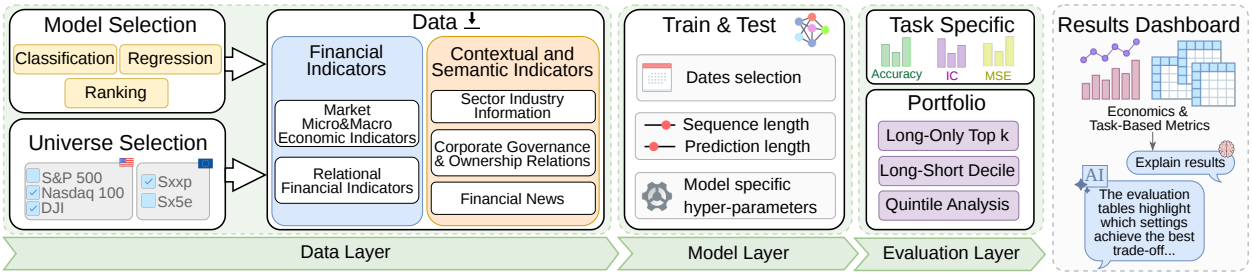


Figure 1: The AEQUITAS architecture. The framework combines data construction, model execution, task- and portfolio-level evaluation, and result exploration through a graphical interface with LLM-based support for interpreting reports.

AEQUITAS applies temporal alignment and preprocessing strategy for each model: models trained independently per asset are aligned at ticker level, while models operating on the full universe are aligned across dates; when a fixed set of tickers is required, missing values are forward-filled to ensure consistency across time. Feature normalization follows the original implementations. By exposing these choices through a unified execution layer, AEQUITAS allows users to run heterogeneous stock prediction models under consistent experimental settings, while preserving the model-specific assumptions required by the original implementations.

3.3 Evaluation Layer

The Evaluation layer provides an interactive result-analysis workflow. After training and testing, AEQUITAS collects model predictions and exposes them through a dashboard that jointly reports task-level and portfolio-level results. This allows users to verify whether predictive performance translates into economically meaningful investment outcomes, since machine learning metrics do not necessarily imply profitable portfolio decisions.

At the task level, the dashboard reports metrics tailored to the selected prediction paradigm, following those implemented in FinBench [14]: classification models are evaluated for price-movement prediction, regression models for return estimation, and ranking models for cross-sectional ordering.

At portfolio level, AEQUITAS makes heterogeneous model outputs comparable by mapping them to a common stock-level ranking score. This score is obtained from the predicted return for regression models, the probability of positive movement for classification models, and the ranking score for ranking models. Portfolios are then formed at decision time t, held for the selected horizon L, and rebalanced consistently across models. The dashboard supports multiple portfolio views. In the Long-only Top-K strategy, AEQUITAS selects the K stocks with the highest scores and builds an equally weighted portfolio. In the Long–Short Top-K strategy, it takes a long position in the top-K stocks and a short position in the bottom-K stocks, isolating ranking quality from market exposure. In the Quintile Analysis, the ranked universe is partitioned from Q_1 to Q_5, and AEQUITAS computes the Compound Annual Growth Rate (CAGR) of each quintile. A monotonic pattern, i.e., $\mathrm{CAGR}(Q_1) > \mathrm{CAGR}(Q_2) > \cdots > \mathrm{CAGR}(Q_5)$, indicates that predictive power is distributed across the asset universe rather than driven only by extreme-ranked assets.

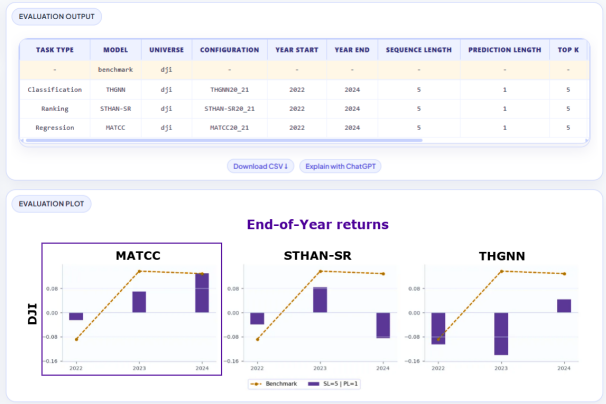
Finally, AEQUITAS includes an AI-assisted explanation module that helps users interpret evaluation reports. The module summarizes trends, highlights discrepancies between task- and portfolio-level behavior, and supports the analysis of model robustness across horizons, investment universes, and portfolio segments.

4 The Demonstration

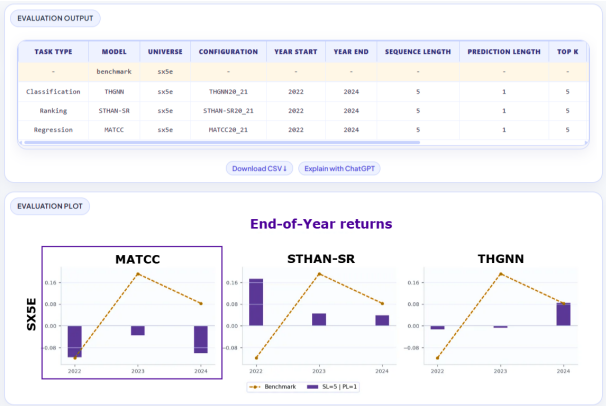
In portfolio evaluation tasks, the final ranking of stocks is crucial, as investment decisions directly follow the ordering induced by the model prediction. Ranking stability must also be assessed, since unstable rankings may lead to inconsistent portfolio returns. AEQUITAS supports this analysis across investment universes and portfolio segments. We demonstrate this through two use cases: an end-of-year analysis showing that the same model may perform well on one universe but not another, and a quintile analysis showing that models on the same universe may exhibit stable or unstable return patterns. In both cases, the dashboard serves as the operational entry point: analysts configure the universe, models, temporal split, and portfolio strategy, launch the evaluation, and inspect the resulting portfolio-level views.

4.1 Use case 1: End-of-Year Model Performance Comparison Across Universes

To perform the end-of-year evaluation, the analyst uses AEQUITAS to run a cross-market robustness check under controlled experimental conditions, assessing whether a prediction pipeline that is effective in one investment universe remains reliable in a different market. In this use case, we consider three representative models, each associated with a different prediction task: THGNN [18] for classification, MATCC [1] for regression, and STHAN-SR [15] for ranking. To ensure a fair comparison, all models are trained under the same temporal setting, using 2019–2020 for training and 2021 for validation. The trained models are then evaluated under the same portfolio configuration according to the Long–Short Top-5 strategy, over the 2022–2024 testing period, considering both the Dow Jones and Euro Stoxx 50 universes. These universes represent different geographical markets, with potentially different liquidity profiles, sector compositions, and market dynamics. In the evaluation layer, AEQUITAS supports the translation of model outputs into portfolio-oriented performance measures. This allows the analyst to compare their impact on investment decisions under consistent experimental and investment settings.



(a) End-Of-Year return on Dow Jones universe.



(b) End-Of-Year return on Euro Stoxx 50 universe.

Figure 2: End-of-year CAGR% return comparison on two selected investment universes. MATCC (highlighted) is the best-performing model on Dow Jones, but it is the worst-performing one on Euro Stoxx 50, showing that performance does not transfer across markets.

Figure 2 reports the resulting end-of-year returns. By comparing Figure 2a with Figure 2b, we observe a clear reversal in the relative behavior of the models. MATCC achieves the best performance on the Dow Jones universe, outperforming both the other considered models and the corresponding benchmark. Conversely, on Euro Stoxx 50, MATCC becomes the worst-performing model, also underperforming the benchmark. This result shows that strong performance on one market universe does not necessarily transfer to another. From an analyst's perspective, this is critical: the apparent superiority of a model in a single backtest cannot be assumed to generalize across markets. Rather than selecting a pipeline from one favorable evaluation setting, the analyst can use AEQUITAS to identify performance instability across investment universes and detect models whose behavior is strongly market-dependent. The LLM-assisted explanation module further supports this analysis by summarizing the reversal of MATCC across the two universes, helping the analyst recognize a market-dependent behavior that would be difficult to detect from a single-universe evaluation.

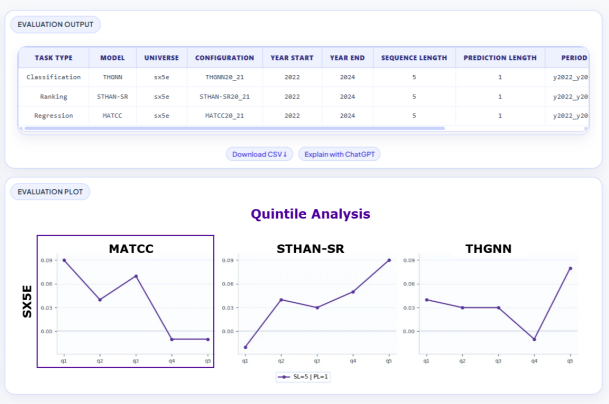


Figure 3: Quintile analysis on Euro Stock 50 universe. MATCC (highlighted) shows a monotonic decreasing trend across quintiles, indicating a more stable ranking behavior, whereas the other considered models do not preserve the same pattern and produce less consistent return distributions.

4.2 Use Case 2: Quintile Analysis on the Euro Stoxx 50 Universe

This use case is performed using the same experimental settings adopted in the previous section, including the selected models, temporal split, and portfolio setting, but we focus only on the Euro Stoxx 50 universe to assess whether the rankings induced by the models are stable across portfolio segments.

Figure 3 reports the CAGR% obtained by each ranking quintile. MATCC is the only model showing a monotonic decreasing trend across quintiles, suggesting that its predicted ranking is more consistently aligned with the realized returns. In contrast, THGNN and STHAN-SR exhibit irregular curves, with no evident decreasing monotonicity, indicating less stable ranking signals for portfolio construction. This analysis further highlights the relevance of AEQUITAS, as it enables financial analysts to go beyond aggregate portfolio performance and inspect the stability of model-induced rankings. In this way, AEQUITAS helps identify models that provide more reliable signals across different portfolio segments.

5 Conclusion

We demonstrate AEQUITAS, an interactive toolkit for stock market prediction and portfolio evaluation. AEQUITAS supports the complete evaluation workflow, from reusable data construction to model execution and portfolio-level result analysis, enabling users to compare classification, regression, and ranking pipelines under consistent settings. The reported use cases show how AEQUITAS can support practical model selection and robustness analysis by comparing model behavior across investment universes and inspecting the stability of model-induced rankings.

Acknowledgments

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

GenAI Usage Disclosure

ChatGPT was used to refine sentence structure and assist in evaluating the clarity and soundness of paragraph-level writing. All original content and ideas were created by the authors. Grammarly was also used for additional grammar and fluency checks.

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] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Yifan Hu, Yuante Li, Peiyuan Liu, Yuxia Zhu, Naiqi Li, Tao Dai, Shutao Xia, Dawei Cheng, and Changjun Jiang. 2025. FinTSB: A Comprehensive and Practical Benchmark for Financial Time Series Forecasting. *CoRR* abs/2502.18834 (2025).

[11] 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.

[12] 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.

[13] 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.

[14] 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 *KDD (2)*. ACM, 9661–9672. doi:10.1145/3770855.3817595

[15] 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.

[16] 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.

[17] Quanzhou Xiang, Zhan Chen, Qi Sun, and Rujun Jiang. 2024. RSAP-DFM: Regime-Shifting Adaptive Posterior Dynamic Factor Model for Stock Returns Prediction. In *IJCAI*. ijcai.org, 6116–6124.

[18] 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.

[19] 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).

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

[21] 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).

[22] 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.

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

[24] 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.

[25] Peng Zhu, Yuante Li, Yifan Hu, Qinyuan Liu, Dawei Cheng, and Yuqi Liang. 2024. LSR-IGRU: Stock Trend Prediction Based on Long Short-Term Relationships and Improved GRU. In *CIKM*. ACM, 5135–5142.