CTBench: Cryptocurrency Time Series Generation Benchmark

Yihao Ang[1] **Qiang Wang**[1] **Qiang Huang**[2,*] **Yifan Bao**[1] **Xinyu Xi**[1]
Anthony K. H. Tung[1] **Chen Jin**[1] **Zhiyong Huang**[1]
[1]Department of Computer Science, National University of Singapore
[2]School of Intelligence Science and Engineering, Harbin Institute of Technology (Shenzhen)
{yihao_ang, yifan_bao, atung, huangzy}@comp.nus.edu.sg
{qwang, xinyu_xi}@u.nus.edu, disjinc@nus.edu.sg, huangqiang@hit.edu.cn

Abstract

Synthetic time series are essential tools for data augmentation, stress testing, and algorithmic prototyping in quantitative finance. However, in cryptocurrency markets, characterized by 24/7 trading, extreme volatility, and rapid regime shifts, existing Time Series Generation (TSG) methods and benchmarks often fall short, jeopardizing practical utility. Most prior work (1) targets non-financial or traditional financial domains, (2) focuses narrowly on classification and forecasting while neglecting crypto-specific complexities, and (3) lacks critical financial evaluations, particularly for trading applications. To address these gaps, we introduce **CTBench**, the first **C**rypto-centric **T**ime series generation **Bench**mark, comprising (1) an open dataset of 452 cryptocurrencies, (2) a dual-task evaluation on Predictive Utility and Statistical Arbitrage, and (3) 13 metrics across six dimensions (i.e., error, rank, trading performance, risk assessment, efficiency, and visualization). We systematically benchmark 8 representative TSG models from five families, uncovering trade-offs between statistical fidelity and real-world profitability. Notably, CTBench offers model ranking analysis and actionable guidance for selecting and deploying TSG models in crypto analytics and strategy development.

1 Introduction

Time Series Generation (TSG) underpins a wide range of applications, including data augmentation [5, 22], anomaly detection [4, 32], privacy [16, 27], and domain adaptation [8, 18], by capturing temporal dependencies and cross-asset correlations. Yet, despite progress from generalized benchmarks such as TSGBench [3, 2], they largely neglect the unique microstructure of cryptocurrencies, an asset class valued at about $4 trillion (May 2025) and characterized by 24/7 trading, speculative dynamics, extreme volatility, and irregular liquidity [23], warranting a crypto-specific evaluation standard.

Existing financial time series benchmarks, such as FinTSB [15] and FinTSBridge [33], have advanced evaluation practices, but they remain anchored in traditional markets and forecasting tasks. When applied to cryptocurrency, they face three key limitations: (1) limited generality, being stock-centric with fixed trading hours and lower volatility; (2) narrow task scope, prioritizing forecasting over crypto-relevant generation, arbitrage, and strategy evaluation [34]; and (3) lack of crypto-specific evaluation metrics that link fidelity with economic utility and tail-risk under 24/7 trading.

We introduce **CTBench**, the first open benchmark tailored to cryptocurrency markets. It (1) curates a high-volatility, hourly dataset with standardized preprocessing; (2) bridges TSG with practice through dual tasks: **Predictive Utility** (train on synthetic, trade on real) and **Statistical Arbitrage**

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

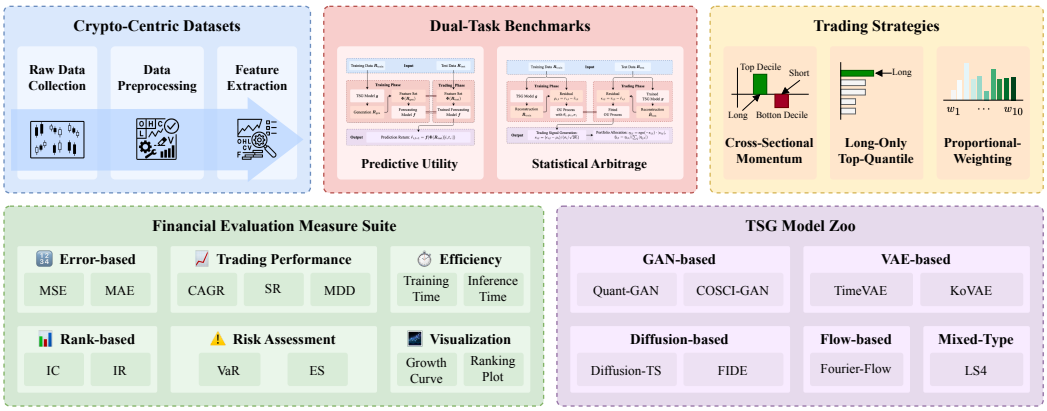


Figure 1: Overall architecture of CTBench with its dual-task benchmarks.

(reconstruct residuals for mean-reversion); (3) stress-tests models using diverse trading strategies and a holistic financial metric suite covering error, rank, trading performance, risk assessment, efficiency, and visualization; and (4) benchmarks 8 representative models across 5 families, providing systematic analyses that expose the trade-offs between fidelity, tradability, and robustness.

2 Preliminaries

Let $\boldsymbol{R} \in \mathbb{R}^{n \times l}$ denote the log-return matrix of n tradable crypto-assets over l hourly observations, where each vector $\boldsymbol{r}_t = [r_{1,t}, \cdots, r_{n,t}] \in \mathbb{R}^n$ contains hourly log-returns $r_{i,t} = \log(p_{i,t}/p_{i,t-1})$ with $p_{i,t}$ the price of asset i at hour t. For walk-forward evaluation, we use a rolling window with training size w and step s, split offsets $\tau \in \{w, w+s, \cdots, w+(k-1)s\}$ with $k = \lfloor \frac{l-w}{s} \rfloor$ yield:

$$\boldsymbol{R}_{\text{train}}^{(\tau)} = [\boldsymbol{r}_{\tau-w+1}, \cdots, \boldsymbol{r}_{\tau}], \ \boldsymbol{R}_{\text{test}}^{(\tau)} = [\boldsymbol{r}_{\tau+1}, \cdots, \boldsymbol{r}_{\tau+s}].$$

At each split, a TSG model $\boldsymbol{g}^{(\tau)}$ is trained on $\boldsymbol{R}_{\text{train}}^{(\tau)}$ and evaluated in two modes: (1) **Generation mode**, sampling synthetic sequences $\boldsymbol{R}_{\text{gen}} = \boldsymbol{g}^{(\tau)}(\boldsymbol{z})$, with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$; (2) **Reconstruction mode**, reconstructing real data as $\hat{\boldsymbol{R}}_{\text{train}} = \boldsymbol{g}^{(\tau)}(\boldsymbol{R}_{\text{train}}^{(\tau)})$ and $\hat{\boldsymbol{R}}_{\text{test}} = \boldsymbol{g}^{(\tau)}(\boldsymbol{R}_{\text{test}}^{(\tau)})$.

To link generation with financial evaluation, we simulate portfolio dynamics. Starting from initial capital $V_0 > 0$, the strategy allocates weights $\boldsymbol{\eta}_t = [\eta_{1,t}, \cdots, \eta_{n,t}] \in \mathbb{R}^n$ at each hour t, and the portfolio value evolves as $V_t = V_{t-1} \cdot (\boldsymbol{\eta}_t^\top \boldsymbol{r}_t)$ with profit-and-loss $\Delta V_t = V_t - V_{t-1}$.

3 CTBench

We present CTBench, a comprehensive benchmark for evaluating TSG models in cryptocurrency markets. As illustrated in Figure 1, it comprises crypto-centric datasets, dual-task evaluation, diverse trading strategies, a comprehensive financial metric suite, and a model zoo across five families.

Crypto-Centric Datasets. CTBench curates a crypto-focused dataset from Binance hourly spot data (Jan 2020–Dec 2024) [6], restricted to USDT pairs and filtered for continuity, resulting in 452 tradable series across diverse regimes. For each asset i at hour t, we record OHLC data and compute close-to-close log-returns $r_{i,t} = \log(C_{i,t}/C_{i,t-1})$, forming $\boldsymbol{R} \in \mathbb{R}^{n \times l}$. To assess whether generators preserve tradable structure, we apply the same feature pipeline to real and synthetic data, combining Alpha101 factors [17] with standard technical indicators (e.g., Bollinger Bands, RSI, moving averages), widely used in quantitative finance [26, 42, 40, 29]. The dataset reveals strong cross-sectional dispersion and cap-dependent volatility, motivating crypto-specific evaluation that tests predictive dependencies, cross-asset relations, and financial viability beyond marginal similarity.

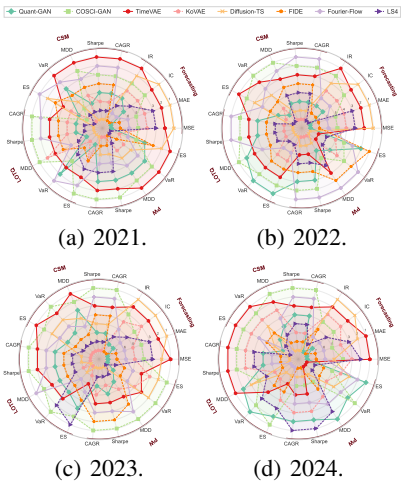
Dual-Task Benchmarks. To connect TSG with financial utility, CTBench introduces two complementary tasks. **Predictive Utility** evaluates whether synthetic data can train forecasters that trade profitably on real markets. Specifically, a TSG model $\boldsymbol{g}$ samples synthetic returns $\boldsymbol{R}_{\text{gen}}$, features $\boldsymbol{\Phi}(\cdot)$ are extracted from both real and synthetic data, and a lightweight forecaster (e.g., XGBoost)

[9, 31, 19, 38] is trained on $\mathbf{\Phi}(\boldsymbol{R}_{\text{gen}})$ to predict next-hour returns. Out-of-sample on real data, predicted returns are ranked each hour to form a dollar-neutral long-short portfolio, with performance scored by CTBench's financial suite. **Statistical Arbitrage** instead the TSG as a denoiser: reconstructed returns $\hat{\boldsymbol{R}}$ yield residuals $\epsilon_{i,t} = r_{i,t} - \hat{r}_{i,t}$, modeled via Ornstein–Uhlenbeck processes [30] (or extensions such as Lévy-type dynamics [12] or neural SDEs [21]). Standardized s-scores generate mean-reverting signals, with trades opened when deviations exceed thresholds and portfolios rebalanced hourly. Together, these tasks test fidelity in prediction and recovery of tradable signals.

Trading Strategies. To remain strategy-agnostic, CTBench evaluates each signal under three canonical styles: (1) Cross-Sectional Momentum (**CSM**), long top-decile vs. short bottom-decile; (2) Long-Only Top-Quantile (**LOTQ**), equal-weight long top 20%; and (3) Proportional Weighting (**PW**), allocating weights proportional to predicted returns. This design stresses models under diverse market paradigms, while allowing plug-and-play integration of proprietary strategies.

Financial Evaluation Measure Suite. CTBench consolidates 13 measures into six categories aligned with practitioner needs: (1) **Error**: MSE and MAE for bias vs. noise; (2) **Rank**: Information Coefficient (IC) and its stability (IR) for cross-sectional order preservation [28, 24]; (3) **Trading Performance**: CAGR, Sharpe Ratio (SR), Maximum Drawdown (MDD) quantify annualized returns and risk-adjusted efficiency; (4) **Risk Assessment**: 95% Value-at-Risk (VaR), and Expected Shortfall (ES) for tail and path-dependent losses; (5) **Efficiency**: training and inference time for deployability; and (6) **Visualization**: Simulated growth curves of a $10,000 investment and ranking plots on diverse market regimes. These capture fidelity, profitability, risk sensitivity, and operational feasibility.

TSG Model Zoo. CTBench benchmarks 8 representative models across five methodological families: (1) GANs (Quant-GAN [35], COSCI-GAN [25]), where GANs are used only for forecasting, as they do not natively support reconstruction [14, 11]; (2) VAEs (TimeVAE [10], KoVAE [20]); (3) Diffusion models (Diffusion-TS [37], FIDE [13]); (4) Flow-based models (Fourier-Flow [1]); and (5) Mixed-type models (LS4 [41]). This diverse model zoo supports architecture-agnostic comparison and highlights trade-offs between fidelity, tradability, and robustness.



(a) 2021. (b) 2022. (c) 2023. (d) 2024.

Figure 2: Rankings of TSG models on the Predictive Utility task.

4 Experiments

Setup. We evaluate CTBench on 452 USDT trading pairs using a walk-forward scheme: 500 days for training, followed by 30 days for the Predictive Utility task and 15 days for the Statistical Arbitrage task, with retraining at each cycle. To isolate generator quality, transaction fees are set to zero by default; for Statistical Arbitrage, we additionally report results under a 0.03% fee, reflecting typical exchange costs [39, 36, 7]. We benchmark eight TSG models across five architectural families, applying recommended or stably tuned hyperparameters, and score them with the evaluation suite in §3.

Predictive Utility Task. Ranking analysis (Figure 2) reveals distinct trade-offs. Diffusion models often achieve the lowest forecasting errors but yield weak trading profits. In contrast, TimeVAE excels

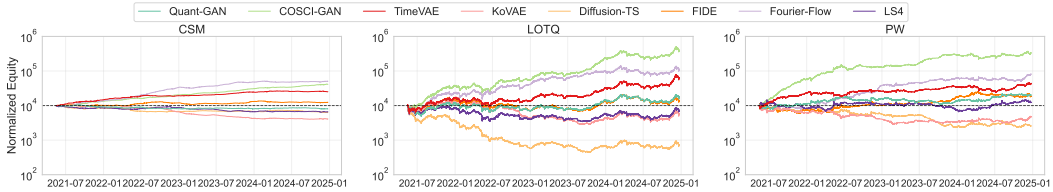


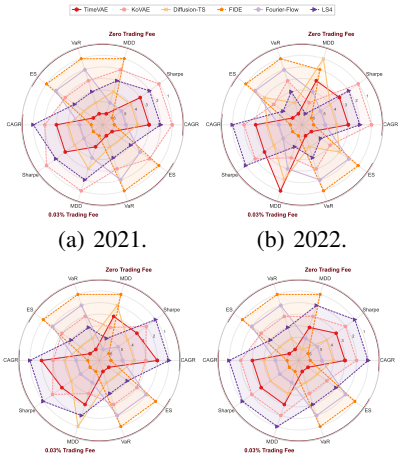
Figure 3: Simulated growth curves of $10,000 over four years under three trading strategies.

Scenario	Recommended TSG Models	Rationale
Trend-following / Directional Markets	COSCI-GAN, KoVAE	COSCI-GAN amplifies trend and dispersion; KoVAE offers alpha with higher drawdowns
Mean-reverting / Range-bound Regimes	TimeVAE, Fourier-Flow, Diffusion-TS	TimeVAE/Fourier-Flow provide balance; Diffusion-TS preserves rank order
Fee-sensitive / Low-turnover Settings	TimeVAE, Diffusion-TS	Smooth residuals, stable Sharpe under transaction costs
Risk Tolerance / Portfolio Design	KoVAE, LS4, TimeVAE, Diffusion-TS, FIDE	KoVAE/LS4 maximize returns with risk; TimeVAE/Diffusion-TS balance Sharpe and drawdown; FIDE is defensive
Deployment Efficiency	TimeVAE, LS4	Fast retraining and low-latency inference; diffusion models better suited for offline use

Table 1: Scenario-based recommendations for selecting TSG models in cryptocurrency markets.

in stable or mean-reverting regimes, COSCI-GAN excels in volatile, directional markets, and Fourier-Flow maintains robust mid-to-high "all-weather" rankings. Equity curves (Figure 3) corroborate this. Under **CSM**, COSCI-GAN and TimeVAE compound steadily while Diffusion-TS and FIDE decay. Under **LOTQ**, COSCI-GAN captures right-tail gains, whereas TimeVAE and Fourier-Flow grow modestly and Diffusion-TS misses rare spikes. Under **PW**, COSCI-GAN again dominates, with TimeVAE and Fourier-Flow compounding smoothly, and LS4 remaining flat. The core lesson is that **low prediction error does not necessarily guarantee tradability**. Over-regularized generators (e.g., Diffusion-TS, LS4) might wash out alpha-bearing variance, whereas models that retain structural noise and tails (e.g., TimeVAE, COSCI-GAN) provide higher economic utility.

Statistical Arbitrage Task. Radar plots (Figure 4) show further trade-offs. KoVAE and LS4 score high on returns (CAGR, Sharpe) but collapse on risk in turbulent regimes, while FIDE shows the opposite: tight risk control but weak returns. TimeVAE and Diffusion-TS maintain balanced, regime-agnostic profiles. Introducing fees compresses rank distances: high-turnover KoVAE loses Sharpe positions, while smoother TimeVAE and Diffusion-TS degrade less. Year-over-year sensitivities also emerge, such as LS4's sharp CAGR expansion in 2023 but large drawdowns in 2022, or KoVAE's peaks in volatile regimes but underperformance in calmer markets. Equity curves (Figure 5) with a 0.03% fee (starting at 10,000) reveal distinct dynamics: LS4 compounds almost monotonically with pronounced surges in mid-2022 and early-2023; KoVAE grows convexly but loses momentum by 2024; TimeVAE advances steadily before plateauing; Diffusion-TS delivers the smallest drawdowns yet ends with the lowest terminal value; FIDE collapses prematurely; and Fourier-Flow exhibits a slow, persistent capital bleed. Overall, **robust deployment requires balancing fidelity and dispersion with regime adaptability and fee resilience**, favoring models that yield smoother, lower-turnover residual signals rather than those optimized for any single metric.



(a) 2021. (b) 2022. (c) 2023. (d) 2024.

Figure 4: Rankings of TSG models on the Statistical Arbitrage task.

Recommendations. Our analysis uncovers a four-way trade-off across TSG families: (1) VAEs deliver fast, stable reconstruction but may under-react to rapid regime shifts. (2) GANs capture trend alpha yet suffer from volatility. (3) Diffusion models handle clustering and fat tails but weaken in low-signal regimes. (4) Flow-based models prioritize likelihood but show limited utility, while mixed-type models are efficient yet inconsistent in risk-return. From these patterns, Table 1 distills actionable guidance: practitioners should (1) diagnose their target regime, alpha source, and operational constraints; (2) select inductive biases that amplify the desired tradable structure; and (3) evaluate using task-metric pairs aligned with production goals. CTBench provides this decision surface through its dual-task design and comprehensive financial evaluation suite.

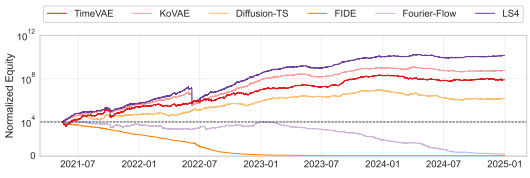


Figure 5: Simulated growth curves of $10,000 for the Statistical Arbitrage task (with 0.03% fee).

5 Conclusion and Future Work

In this work, we introduce CTBench, the first crypto-specific TSG benchmark, unifying (1) a curated high-frequency dataset, (2) a dual-task evaluation framework, and (3) a comprehensive suite of financial metrics that assess both statistical fidelity and practical viability. Extensive experiments surface actionable trade-offs and provide deployment guidance under diverse regimes, positioning it as a collaborative resource to standardize evaluation and spur innovation. We plan to extend CTBench with additional tokens and cross-exchange coverage, ensemble and regime-aware switching, and automated evaluation tuning to further improve robustness and usability.

References

[1] Ahmed M. Alaa, Alex James Chan, and Mihaela van der Schaar. Generative time-series modeling with fourier flows. In *ICLR*, 2021.

[2] Yihao Ang, Yifan Bao, Qiang Huang, Anthony KH Tung, and Zhiyong Huang. Tsgassist: An interactive assistant harnessing llms and rag for time series generation recommendations and benchmarking. *Proceedings of the VLDB Endowment*, 17(12):4309–4312, 2024.

[3] Yihao Ang, Qiang Huang, Yifan Bao, Anthony KH Tung, and Zhiyong Huang. Tsgbench: Time series generation benchmark. *Proceedings of the VLDB Endowment*, 17(3):305–318, 2023.

[4] Yihao Ang, Qiang Huang, Anthony KH Tung, and Zhiyong Huang. A stitch in time saves nine: Enabling early anomaly detection with correlation analysis. In *ICDE*, pages 1832–1845, 2023.

[5] Yifan Bao, Yihao Ang, Qiang Huang, Anthony KH Tung, and Zhiyong Huang. Towards controllable time series generation. *arXiv preprint arXiv:2403.03698*, 2024.

[6] Binance Exchange. Binance exchange. `https://binance.com/`, 2025. Accessed: 1 March 2025.

[7] Binance Exchange. Trading fee schedule, 2025.

[8] Ruichu Cai, Jiawei Chen, Zijian Li, Wei Chen, Keli Zhang, Junjian Ye, Zhuozhang Li, Xiaoyan Yang, and Zhenjie Zhang. Time series domain adaptation via sparse associative structure alignment. In *AAAI*, pages 6859–6867, 2021.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *KDD*, pages 785–794, 2016.

[10] Abhyuday Desai, Cynthia Freeman, Zuhui Wang, and Ian Beaver. Timevae: A variational auto-encoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095*, 2021.

[11] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. In *ICLR*, 2017.

[12] Tsukasa Fujiwara and Hiroshi Kunita. Stochastic differential equations of jump type and lévy processes in diffeomorphisms group. *Journal of mathematics of Kyoto University*, 25(1):71–106, 1985.

[13] Asadullah Hill Galib, Pang-Ning Tan, and Lifeng Luo. Fide: Frequency-inflated conditional diffusion model for extreme-aware time series generation. In *NeurIPS*, 2024.

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

[15] Y. Hu et al. Fintsb: A comprehensive benchmark for financial time series forecasting. In *arXiv:2502.18834*, 2025.

[16] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *ICLR*, 2018.

[17] Zura Kakushadze. 101 formulaic alphas. *Wilmott Magazine*, (84):72–80, 2016. 22 pages; no changes (excepting this line); to appear; also available as arXiv:1601.00991v3 [q-fin.PM].

[18] Yuening Li, Zhengzhang Chen, Daochen Zha, Mengnan Du, Jingchao Ni, Denghui Zhang, Haifeng Chen, and Xia Hu. Towards learning disentangled representations for time series. In *KDD*, pages 3270–3278, 2022.

[19] Guang Liu, Yuzhao Mao, Qi Sun, Hailong Huang, Weiguo Gao, Xuan Li, Jianping Shen, Ruifan Li, and Xiaojie Wang. Multi-scale two-way deep neural network for stock trend prediction. In *IJCAI*, pages 4555–4561, 2021.

[20] Ilan Naiman, N Benjamin Erichson, Pu Ren, Michael W Mahoney, and Omri Azencot. Generative modeling of regular and irregular time series data via koopman vaes. In *ICLR*.

[21] YongKyung Oh, Dongyoung Lim, and Sungil Kim. Stable neural stochastic differential equations in analyzing irregular time series data. In *The Twelfth International Conference on Learning Representations*, 2024.

[22] Giorgia Ramponi, Pavlos Protopapas, Marco Brambilla, and Ryan Janssen. T-CGAN: conditional generative adversarial network for data augmentation in noisy time series with irregular sampling. *arXiv preprint arXiv:1811.08295*, 2018.

[23] Reuters. Crypto sector breaches $4 trillion in market value during pivotal week. *Reuters*, July 18 2025.

[24] C Grinold Richard and Ronald Kahn. Active portfolio management: A quantitative approach for producing superior returns and controlling risk, 2000.

[25] Ali Seyfi, Jean-François Rajotte, and Raymond T. Ng. Generating multivariate time series with common source coordinated GAN (COSCI-GAN). In *NeurIPS*, pages 32777–32788, 2022.

[26] Shuo Sun, Rundong Wang, and Bo An. Reinforcement learning for quantitative trading. *ACM Transactions on Intelligent Systems and Technology*, 14(3):1–29, 2023.

[27] Muhang Tian, Bernie Chen, Allan Guo, Shiyi Jiang, and Anru R Zhang. Reliable generation of privacy-preserving synthetic electronic health record time series via diffusion models. *JAMIA*, 31(11):2529–2539, 2024.

[28] Jack L Treynor and Fischer Black. How to use security analysis to improve portfolio selection. *The journal of business*, 46(1):66–86, 1973.

[29] Chih-Fong Tsai and Yu-Chieh Hsiao. Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision support systems*, 50(1):258–269, 2010.

[30] George E Uhlenbeck and Leonard S Ornstein. On the theory of the brownian motion. *Physical review*, 36(5):823, 1930.

[31] László Vancsura, Tibor Tatay, and Tibor Bareith. Navigating ai-driven financial forecasting: A systematic review of current status and critical research gaps. *Forecasting*, 7(3):36, 2025.

[32] Chengyu Wang, Kui Wu, Tongqing Zhou, Guang Yu, and Zhiping Cai. Tsagen: synthetic time series generation for kpi anomaly detection. *IEEE Transactions on Network and Service Management*, 19(1):130–145, 2021.

[33] Yanlong Wang, Jian Xu, Tiantian Gao, Hongkang Zhang, Shao-Lun Huang, Danny Dongning Sun, and Xiao-Ping Zhang. Fintsbridge: A new evaluation suite for real-world financial prediction with advanced time series models. *arXiv preprint arXiv:2503.06928*, 2025.

[34] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*, 2024.

[35] Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. Quant gans: deep generation of financial time series. *Quantitative Finance*, 20(9):1419–1440, 2020.

[36] Julian Winkel and Wolfgang Karl Härdle. Pricing kernels and risk premia implied in bitcoin options. *Risks*, 11(5):85, 2023.

[37] Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. In *ICLR*, 2024.

[38] Kyung Keun Yun, Sang Won Yoon, and Daehan Won. Prediction of stock price direction using a hybrid ga-xgboost algorithm with a three-stage feature engineering process. *Expert Systems with Applications*, 186:115716, 2021.

[39] Chuheng Zhang, Yitong Duan, Xiaoyu Chen, Jianyu Chen, Jian Li, and Li Zhao. Towards generalizable reinforcement learning for trade execution. In *IJCAI*, pages 4975–4983, 2023.

[40] Chuheng Zhang, Yuanqi Li, Xi Chen, Yifei Jin, Pingzhong Tang, and Jian Li. Doubleensemble: A new ensemble method based on sample reweighting and feature selection for financial data analysis. In *ICDM*, pages 781–790, 2020.

[41] Linqi Zhou, Michael Poli, Winnie Xu, Stefano Massaroli, and Stefano Ermon. Deep latent state space models for time-series generation. In *ICML*, pages 42625–42643, 2023.

[42] Zhoufan Zhu and Ke Zhu. Alphaqcm: Alpha discovery in finance with distributional reinforcement learning. In *ICML*, 2025.