OPEN-WORLD TOKENIZED EQUITIES: A DATASET LINKING BLOCKCHAINS AND EQUITY MARKETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tokenized equities are an emerging class of real-world assets that connect public blockchains with traditional equity markets. Yet this rapidly growing market remains underexplored. Existing AI-for-finance datasets model prices, text, and inter-stock relations, while blockchain datasets capture wallets and transactions; neither jointly observes public on-chain activity and the corresponding listed-equity information. We introduce, to our knowledge, the first large-scale ticker-aligned dataset bridging these domains, covering 890 tokenized equities, 249.5M transfers, and 31.3M swaps, together with exchange prices, news, and fundamentals. This alignment uncovers new empirical findings. Tokenized equities remain active when U.S. exchanges are closed, and fresh, liquid on-chain quotes can predict subsequent exchange price movements. A small set of liquid tokenized assets also carries information about the broader equity market, while account-level co-holding is more strongly associated with shared news attention than with return correlation. These findings motivate two complementary learning tasks enabled by the linked dataset: *dark-window equity forecasting*, which augments off-chain market data with evolving on-chain signals to predict the direction of the next opening price, and *open-world next-asset prediction*, which predicts participation in newly tokenized equities from account histories and asset information. Together, the dataset, findings, and tasks introduce a new learning setting spanning on-chain–equity forecasting and temporal asset recommendation. The dataset and code are available at `https://anonymous.4open.science/r/RWA-dataset-repo-CADC/`.

1 INTRODUCTION

Tokenized equities are an emerging class of blockchain-based assets that represent traditional equities and ETFs. Their economic exposure is designed to track an underlying equity, whose ticker links the token to conventional prices, news, and fundamentals (Malinova & Park, 2023; RWA.xyz, n.d.). At the same time, tokenization adds a distinct data layer: public, 24/7 account-level records of transfers, swaps, and participation. As illustrated in Figure 1, an asset such as TSLA can be observed simultaneously as an exchange-traded security and as an on-chain asset. This dual view is becoming increasingly relevant as tokenized equities expand from niche real-world-asset experiments into a broader market, supported by recent regulatory and market developments (U.S. Securities and Exchange Commission, 2026; Reuters, 2026). Figure 2 illustrates this rapid emergence.

However, existing datasets capture only one side of this market. Financial datasets provide prices, news, and fundamentals but generally lack public account-level activity (Xu & Cohen, 2018; Dong et al., 2024; Li et al., 2024; Fan & Shen, 2024; Zhang et al., 2024). Blockchain datasets, in contrast, capture wallets and transactions but typically focus on crypto-native assets that are not aligned with exchange-traded equities (Hu et al., 2020; Huang et al., 2023; Yu et al., 2023; Shamsi et al., 2022). As a result, there is no public large-scale dataset that jointly observes on-chain account activity and the corresponding equity-market information. We address this gap with a large-scale ticker-aligned dataset linking public blockchain activity to traditional equity markets. The dataset spans four chains and 890 tokenized equities, with 249.5 M transfers, 31.3 M swaps, and 140.1 M one-minute exchange price bars, together with ticker-aligned news and fundamentals, through August 20, 2026. This alignment enables direct study of how continuously evolving on-chain activity interacts with conventional equity markets.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

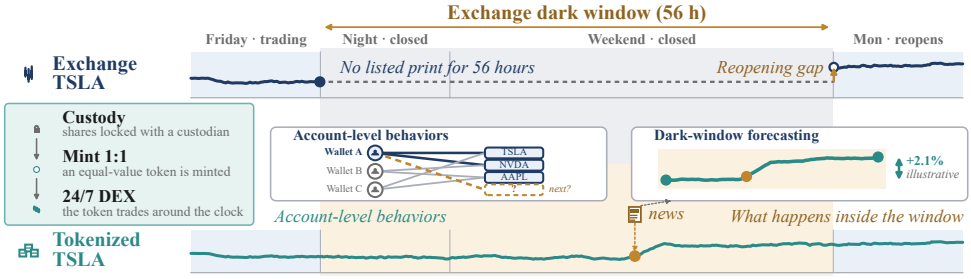


Figure 1: Listed and tokenized views of Tesla (TSLA).

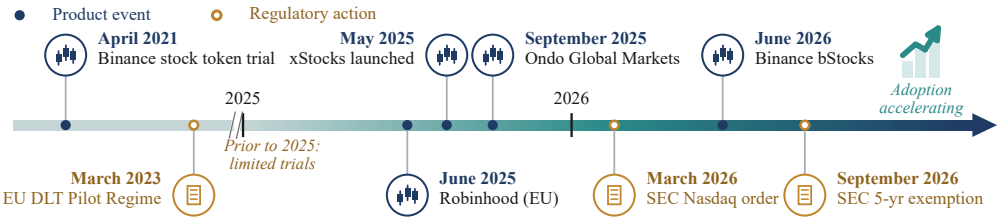


Figure 2: Evolution of the tokenized-equity market.

Our dataset reveals new empirical findings that are not observable from conventional equity or crypto-native data alone. Existing equity forecasting work largely studies regular-session prediction from prices, text, and inter-stock relations, showing that temporal patterns and multimodal fusion can improve ranking and movement prediction (Li et al., 2024; Fan & Shen, 2024; Sawhney et al., 2020; Zong et al., 2025). These settings, however, do not observe a public 24/7 token market or account-level holder dynamics on the same tickers. By jointly observing these signals, we identify four key patterns. First, tokenized equities remain highly active while U.S. exchanges are closed: 63.1% of swap dollar volume and 78.5% of transfers occur outside regular trading hours. Second, dark-window direction becomes more informative with liquidity and quote freshness, with directional accuracy exceeding 80% when a window contains at least 300 trades and the latest quote is no more than one hour old. Third, a small subset of liquid tokenized assets carries predictive information about a broader equity universe at market reopening. Finally, account-level co-holding is more strongly associated with shared news attention ($r = 0.37$–0.40) than with conventional return correlation ($r = 0.12$–0.16). These results show that tokenized equities provide both a cross-venue price signal and a public behavioral view of equity participation.

These findings naturally lead to two complementary learning tasks enabled by the linked data. (1) *Dark-window equity forecasting* predicts the next exchange opening from off-chain information and on-chain dark-window signals. (2) *Open-world next-asset prediction* ranks the next tokenized equity from public account histories. Experiments show that the dark-window state raises RankIC of the best model by $1.9\times$, while next-asset prediction is challenging because of cold start on newly listed tokens. Together, these tasks provide a testbed for learning across financial time series, temporal transaction graphs, and real-world asset markets.

Contributions. The key contributions of this paper are as follows:

- We introduce a large-scale ticker-aligned dataset linking tokenized equities across four public blockchains with exchange prices, news, and fundamentals.
- We identify patterns enabled by ticker-aligned on-chain and equity-market data: dark-window activity, on-chain informativeness, and account-level co-holding.
- We formulate two corresponding learning problems: dark-window equity forecasting with an on-chain channel, and open-world next-asset ranking from public account histories.

2 Related Work

Financial prediction datasets and cross-source stock models. Machine learning on equity prices and text has been studied for decades, from classical factor and sentiment models to large aligned

Table 1: Comparison with financial and blockchain datasets. TG denotes temporal-graph evaluation.

Resource	Price	News	On-chain	Ticker-aligned	TG protocol
StockNet (Xu & Cohen, 2018)	✓	✓		✓	
FNSPID (Dong et al., 2024)	✓	✓		✓	
FinAgent (Zhang et al., 2024)	✓	✓		✓	
Chartalist (Shamsi et al., 2022)			✓		
Live Graph Lab (Zhang et al., 2023)			✓		✓
EX-Graph (Wang et al., 2024)			✓		
Coin-Meme (Long et al., 2025)	✓	✓	✓		
TGB (Huang et al., 2023; Gastinger et al., 2024)			✓		✓
Ours	✓	✓	✓	✓	✓

corpora and cross-source forecasters. Representative datasets such as StockNet and FNSPID (Xu & Cohen, 2018; Dong et al., 2024) couple exchange equity prices with social or news text on a closed ticker universe. Existing stock forecasting models establish the value of several complementary information sources. MASTER uses market-guided feature selection and dynamic inter-stock modeling (Li et al., 2024), while StockMixer achieves strong performance through lightweight indicator, temporal, and stock mixing (Fan & Shen, 2024). MAN-SF demonstrates gains from combining prices, text, and company relations (Sawhney et al., 2020), and MSGCA shows that effective fusion matters because additional modalities can also degrade prediction (Zong et al., 2025). These studies evaluate conventional equity prediction settings. Financial studies further document off-hours pricing differences in tokenized assets (Scharnowski, 2026) and differing evidence on their ability to anticipate equity reopenings (Cong et al., 2025; Dombrowski & Arshadi, 2026). Our benchmark extends the evaluation to token-market observations during exchange closures and provides aligned data for quote reliability, incremental forecasting value, and public account-level participation.

Blockchain and Web3 datasets. Public ledgers have been studied as graphs for transaction forensics and token analytics. Labeled blockchain corpora and analysis platforms (Shamsi et al., 2022; Kalodner et al., 2020; Somin et al., 2025; Li et al., 2025) scale UTXO and account-based graphs. Live Graph Lab provides a dynamically updated Ethereum NFT transaction graph for temporal graph analysis (Zhang et al., 2023), while EX-Graph links Ethereum transactions with verified X accounts and social relations (Wang et al., 2024). Coin-Meme further adds images, text, and market metadata to memecoin records (Long et al., 2025). These resources expand the scale and context of blockchain graph analysis, but their entities and prediction targets remain crypto-native rather than ticker-aligned with listed equities.

Temporal graph benchmarks. Graph benchmarks standardize datasets, splits, evaluation metrics, and model comparison. Common tasks include node, link, and graph property prediction (Hu et al., 2020); dynamic link and node property prediction under chronological splits (Poursafaei et al., 2022; Huang et al., 2023); link prediction on temporal knowledge graphs and heterogeneous graphs (Gastinger et al., 2024); and next-destination prediction under complex sequential dynamics (Yi et al., 2025). These benchmarks commonly use ranking metrics such as MRR against sampled negatives, while unified libraries support reproducible comparison across temporal graph models (Yu et al., 2023). Typical nodes include users and items in interaction graphs, addresses in transaction graphs, and entities in temporal knowledge graphs.

Table 1 summarizes the most related state-of-the-art datasets and benchmarks. No prior resource has exchange equity prices and a public holder graph on the same tickers.

3 Preliminaries: The Tokenized Equity Market

Blockchain and DEX. Public blockchains provide transparent, persistent, and independently auditable records of on-chain activity. On account-based blockchains, such as Ethereum, BNB Chain, and Solana, state is organized around addresses, and timestamped token transfers form temporal multigraphs of value flows (Kalodner et al., 2020; Shamsi et al., 2022). Decentralized exchanges (DEXs) execute trades through smart contracts, commonly using automated-market-maker pools whose reserves determine executable prices. DEX swaps are particularly important because they

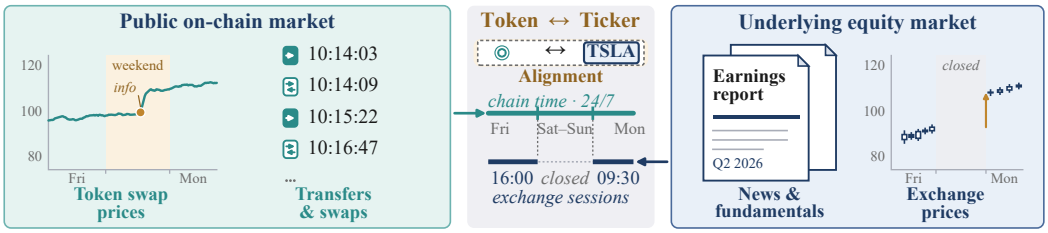


Figure 3: Overview of the ticker-aligned dataset.

Table 2: On-chain event coverage across blockchains.

	Transfers				Swaps			
Chain	Tickers	Records	Addresses	Start	Tickers	Records	Addresses	Start
BSC	880	113,233,730	732,597	2025-08-13	83	7,740,325	1,412	2025-10-30
Ethereum	874	4,572,453	69,719	2025-07-17	37	139,614	728	2025-09-05
Solana	802	42,063,312	492,885	2025-06-10	132	20,825,076	320,921	2025-06-27
RH-chain	96	89,654,643	135,453	2026-06-22	95	2,588,304	883	2026-07-02
Total	**890**	**249,524,138**	1,399,760	2025-06-10	216	**31,293,319**	323,881	2025-06-27

reveal transaction prices, whereas transfers, mints, and burns record movements or supply changes without necessarily revealing a traded price.

Tokenized-equity platforms. *Tokenized equities* are fungible tokens designed to track the economic exposure of listed equities or ETFs. Different platforms implement this linkage with different legal wrappers and venues. xStocks are 1:1-collateralized tracker certificates issued by a dedicated SPV, with the corresponding shares held in segregated custody, and trade mainly on Solana (xStocks, n.d.; 2025). Ondo Stocks are structured notes issued by a bankruptcy-remote SPV and backed by the underlying securities plus a collateral buffer; they appear on Ethereum, BNB Chain, and Solana (Ondo Finance, n.d.; 2024). Binance bStocks are 1:1-backed tokenized securities with 24/7 secondary trading on BNB Chain (Binance, 2026). Robinhood issues stock tokens on Robinhood Chain (Robinhood Markets, 2025). In all cases, token holders receive economic exposure rather than direct shareholder rights.

Distinctive properties of tokenized stocks. Three properties motivate our dataset and tasks. **(1) Public holder observability:** the ledger exposes fractional, address-level transfers and holdings that brokerage-only equity panels omit. **(2) A 24/7 secondary clock:** tokens remain transferable while U.S. exchanges are closed, making otherwise unobserved dark-window activity available on chain (xStocks, 2025). **(3) Multiple coexisting prices:** exchange prices, DEX pool prices, and issuer quotes may diverge when primary markets are closed or token liquidity is thin (Scharnowski, 2026). Additional market background is provided in Appendix A.

4 Dataset

We build a ticker-aligned multi-source dataset of tokenized equities in which transfers, swaps, exchange equity prices, news, and fundamentals share a common ticker identifier and clock. Each asset is covered from its on-chain launch through August 20, 2026, 23:59:59 UTC. The dataset is *continuously updated* as new on-chain activity and market information become available.

On-chain events. The on-chain module contains transfers and swaps. **Transfers** record token movements, including timestamped sender, receiver, and amount information for mints, burns, custody movements, and wallet-to-wallet transfers. **Swaps** record DEX trades, including token and quote amounts, execution prices, and trader addresses when observable. Table 2 summarizes the cleaned transfers and ticker coverage on the four chains with nonzero swap observations.

Prices. Every on-chain asset is matched to minute-level exchange equity prices from Polygon.io (Polygon.io, n.d.). The feed covers consolidated U.S. equity trades from 04:00 to 20:00 Eastern time, including premarket, regular-session, and after-hours activity. The series spans from each ticker's on-chain launch through the common study end of 2026-08-20. Across 890 matched tickers, the dataset contains 156.7 M minute bars, of which 84.4% fall in regular trading hours.

News and fundamentals. We collect ticker-linked news and fundamentals from two sources, Polygon.io (Polygon.io, n.d.) and Finnhub (Finnhub, n.d.). They aggregate articles from reliable publishers, together with company news and press-release wires. News articles retain the original publisher, publication time, headline, and ticker mentions. An article may name several tickers, which form a news co-mention. Fundamentals consist of ticker profiles, financial statements, dividends, and split events, keyed by ticker and reporting or effective date. The dataset contains 1,193,614 news articles and 24,927 fundamental records.

Alignment. All data sources are mapped to a canonical underlying ticker and retain their original timestamps. For each ticker, the usable window begins at its first on-chain observation and ends at the common study cutoff. Each analysis then applies only the source and activity requirements needed for that question rather than imposing one global completeness filter. Figure 3 presents the overview of the aligned dataset.

5 Empirical Analysis of Linked On-Chain and Equity Data

In this section, we examine how tokenized equities behave across on-chain and traditional equity markets. By aligning public on-chain activity with exchange prices, news, and fundamentals, we uncover patterns that neither conventional equity nor crypto-native datasets reveal alone. We summarize the evidence in four findings that motivate the learning tasks in Section 6.

5.1 Overall alignment with traditional equities

Finding 1. Tokenized equities remain active while U.S. exchanges are closed, and the on-chain–exchange price gap widens as listed quotes age.

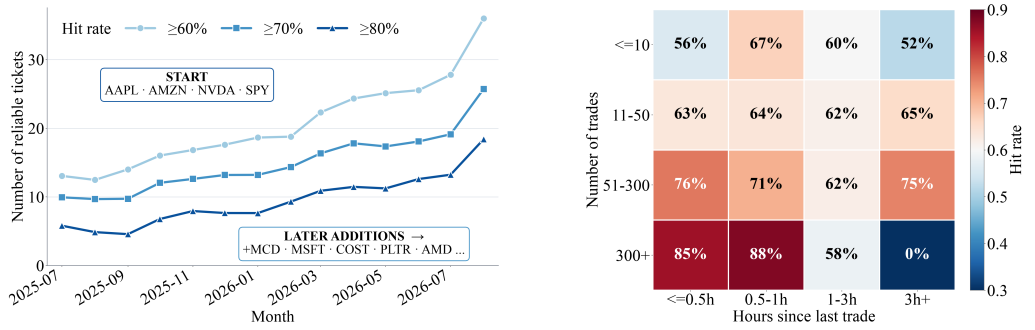
Hours and price gaps. Tokenized names remain active after the listed close: of 23.4 M matched swaps on 168 tickers, only 36.9% of dollar volume falls in regular hours, while 41.5% is weekday off-hours and 21.6% is weekends and holidays. Transfers are even more off-hours (21.4%, 53.4%, and 25.1%). Meanwhile, the on-chain–exchange price gap then widens as the listed quote ages. The median absolute gap rises from 0.19% in regular hours to 0.32% in weekday off-hours, 0.48% on holidays, and 0.62% on weekends, and the share of gaps above 1% rises from 6.5% to 15.8–31.6%. Within weekday off-hours the jump is overnight: the median stays near 0.21–0.22% while premarket and after-hours prints still exist, then reaches 0.51% once the listed tape is closed. This is consistent with wider off-hours pricing differences (Scharnowski, 2026), here separated by premarket, after-hours, overnight, weekend, and holiday.

Volume co-movement. We next ask whether listed and on-chain activity still move together when both markets are open. On the 78 tickers with enough on-chain history, daily exchange volume has median Spearman correlations of 0.45 with swap volume and 0.40 with transfer counts, and the on-chain series rise by $2.84\times$ and $1.71\times$ on the busiest listed days. However, these associations are only moderate, which indicates that on-chain activity has other drivers.

5.2 Trading beyond exchange hours

U.S. equities trade only during fixed market hours, while tokenized equities remain tradable overnight and on weekends. This creates a *dark window* in which the token can provide a continuously updating public price before the exchange reopens. We conduct the analysis below to tickers with a valid swap price in that window.

Finding 2. On-chain dark-window returns help predict equity reopening direction, and this signal is strongest when trading is active and the latest on-chain quote is fresh.



(a) Monthly coverage and correct directions.

(b) Hit rate by activity and freshness.

Figure 4: When dark-window token returns indicate equity reopening direction.

Liquidity and freshness. On-chain prices over nights and weekends give a new public observation of each equity while the exchange is closed, so we can test whether the token's dark-window direction matches the subsequent reopening. However, on-chain quotes also contain noise, so we stratify these windows by trading activity and quote age. Results show that the hit rate rises sharply with on-chain activity and quote freshness. Specifically, the hit rate exceeds 80% when a window has at least 300 trades and the last quote is at most one hour old, and exceeds 70% at 51–300 trades with the same freshness. With at most 50 trades it stays near 52–67%, and even 300+ windows fall to 58% once the last trade is 1–3 hours old (Figure 4(b)). Furthermore, as the on-chain market develops, more tickers become useful (Figure 4(a)): the reliable set starts with AAPL, AMZN, NVDA, and SPY, then gradually adds names such as MCD, MSFT, and AMD, with an increasing trend.

Finding 3. Reliable tokenized quotes provide an additional public observation of dark-window market direction for equities beyond the tokenized set itself.

Cross-equity proxies. The analysis above shows that a small set of liquid tickers carry usable dark-window information on-chain, while most equities still lack a reliable token quote of their own. We therefore examine whether this signal can be transmitted to related equities over nights and weekends, not only for the tokenized names themselves. We find two pathways. First, a liquid index token can transmit market-wide direction to individual equities that lack a usable quote of their own. For example, only 12% of NASDAQ-100 constituents open higher after weekends in which tokenized QQQ falls by more than 1%, compared with 80% after weekends in which it rises by more than 1%. Second, a liquid sector token still tracks related names after the common market factor is removed. For example, the residual weekend returns of the gold and silver tokens (GLD and SLV) correlate at 0.74 with those of the corresponding miner ETFs (GDX and SIL). These pathways motivate the liquid-token proxy used in the dark-window state $s_{i,t}$ of Section 6.1.

5.3 Account-level behavior across tokenized assets

Public blockchain records provide an account-level view of tokenized-equity participation that is largely absent from traditional equity datasets. In particular, they reveal holding preferences: which tokenized equities an address holds, and which names it holds together. Although an address does not necessarily correspond to a unique investor, we exclude non-human addresses, mainly contract addresses and large exchange addresses, and use the remaining accounts for the analysis below.

Platform-specific co-holding. Table 3 summarizes co-holding patterns within each chain. As shown, co-holding is common, but its breadth and timing vary substantially across platforms. Addresses spanning multiple issuers hold more equities on average than those confined to a single platform: 2.19 tickers on BSC and 1.83 on Solana. Cross-platform acquisition is also faster on BSC, with a median gap of 5.4 days, compared with 27.1 days on Solana. It is notable that Robinhood is a clear outlier, with 4.59 tickers per address and a median acquisition gap of zero days, consistent with more batch-like acquisition. This shows that holding preferences are platform-specific rather than a single pattern across venues.

Table 3: Co-holding within the same chain: same-platform versus cross-platform spell overlap.

Chain	Platform(s)	Cross	Addresses	Avg. tickers	Med. days	Most common pair
BNB Chain	Ondo	No	234,493	1.74	5.6	GOOGL / QQQ
BNB Chain	Binance	No	282,302	1.96	1.9	QQQ / SPX
BNB Chain	Ondo + Binance	Yes	438,826	2.19	5.4	SPCX / FXI
Solana	Ondo	No	19,786	1.72	4.0	AAPL / NVDA
Solana	xStocks	No	366,088	1.81	1.1	NVDA / TSLA
Solana	Ondo + xStocks	Yes	379,918	1.83	27.1	MRNA / SPY
Ethereum	Ondo	No	49,088	1.80	5.5	NVDA / TSLA
Robinhood Chain	Robinhood	No	103,392	4.59	0.0	AAPL / NVDA

Finding 4. On-chain tokenized-equity co-holding is associated with off-chain information, and more strongly with shared news attention than with return co-movement.

News versus returns. To characterize the factors associated with pairwise co-holding, we compare ticker-pair co-holding with news co-mention, on-chain activity, and off-chain return correlation, reporting the Pearson correlation r in each case. News co-mention shows the strongest association with co-holding ($r = 0.37$–0.40), followed by on-chain activity ($r = 0.28$–0.36). In contrast, return correlation is substantially weaker ($r = 0.12$–0.16). These results indicate that tokenized-equity co-holding is more closely associated with shared attention and on-chain participation than with conventional return co-movement. These associations motivate the open-world next-asset experiments in Section 6.2.

6 Tasks and Experimental Protocol

The empirical patterns enabled by this linked dataset naturally motivate two learning tasks: *(1) dark-window equity forecasting* uses off-chain market information together with evolving on-chain signals to predict the next exchange opening price; *(2) open-world next-asset prediction* predicts participation in newly tokenized equities from public account histories and asset information.

6.1 Dark-window equity forecasting

Task. Findings 2 and 3 show that liquid, fresh dark-window token returns inform the next equity print, and that reliable tokens can proxy a broader universe. We formalize this observation as overnight/weekend gap prediction on the point-in-time S&P 500 universe (500 constituents after excluding equities added after the study cutoff). For equity i and session date t, the prediction target is $y_{i,t} = \frac{\text{open}_i(t+1)}{\text{close}_i(t)} - 1$. Each instance is associated with a scalar dark-window state $s_{i,t}$. If equity i has a reliable tokenized quote in that window under the activity and freshness conditions of Section 5.2, $s_{i,t}$ is the token's own dark-window return. Otherwise, $s_{i,t} = \beta_i \bar{r}_t^{\text{tok}} + \delta_{i,t}$, where $\bar{r}_t^{\text{tok}}$ is the cross-sectional mean return of equities with reliable token quotes on day t, β_i is a historical exposure of i to that cross-section, and $\delta_{i,t}$ is a theme adjustment when a sector proxy is available. All components of $s_{i,t}$ are frozen before listed premarket prices become available, so the covariate uses only information that is public before any next-session exchange print.

Baselines. We evaluate three families of published equity forecasters under a common input protocol. The sequential / adaptive baselines are MASTER (Li et al., 2024), StockMixer (Fan & Shen, 2024), and DoubleAdapt (Zhao et al., 2023). The graph-relational baselines are CI-STHPAN (Xia et al., 2024), and MAN-SF (Sawhney et al., 2020). The LLM-assisted or LLM-based baselines are MSGCA (Zong et al., 2025) and GS-FUSE (Zhang et al., 2026). Model details are provided in Appendix C.1. Every completed baseline is run in two settings. In the off-chain setting, the published model consumes exchange prices and news, together with its own equity graph when applicable. In the on-chain setting, we keep that off-chain forecast $\hat{y}_{i,t}^{(0)}$ and train a residual head $\hat{y}_{i,t} = \hat{y}_{i,t}^{(0)} + f_\phi(s_{i,t})$, so the dark-window state enters as an additive correction rather than as a raw concatenated feature.

Table 4: Experiment results of overnight gap prediction across seeds. Best performance in each column is in bold.

Model	No on-chain		With on-chain	
	Market correlation	RankIC	Market correlation	RankIC
Sequential / adaptive				
MASTER (Li et al., 2024)	0.069±0.052	0.015±0.012	0.186±0.061	0.047±0.020
DoubleAdapt (Zhao et al., 2023)	0.119±0.021	0.043±0.008	0.196±0.017	0.110±0.009
StockMixer (Fan & Shen, 2024)	0.274±0.027	0.057±0.002	0.309±0.024	0.126±0.005
Graph-relational				
CI-STHPAN (Xia et al., 2024)	0.092±0.038	0.015±0.009	0.173±0.054	0.038±0.024
MAN-SF (Sawhney et al., 2020)	0.025±0.061	0.026±0.006	0.172±0.000	0.048±0.015
LLM-assisted or LLM-based				
GS-FUSE (Zhang et al., 2026)	0.255±0.007	0.045±0.001	**0.314**±0.006	0.070±0.002
MSGCA (Zong et al., 2025)	**0.289**±0.259	**0.082**±0.046	0.201±0.047	**0.158**±0.033

Splits and metrics. (1) Splits. Discovery (September 2024–March 2026) and confirmation (March 2026–August 2026) are chronological, with three walk-forward folds in discovery. All reported numbers are confirmation-split means over n=106 days. **(2) Metrics.** We evaluate two metrics: market-level correlation and stock-level RankIC. Let $\hat{y}_{i,t}$ be the model score and $y_{i,t}$ the realized gap for equity i on day t, with daily means

$$\bar{\hat{y}}_t = |I_t|^{-1} \sum_{i \in I_t} \hat{y}_{i,t}, \qquad \bar{y}_t = |I_t|^{-1} \sum_{i \in I_t} y_{i,t}.$$

Market-level correlation is the Pearson correlation of these daily means: $\mathrm{Corr}_{\mathrm{mkt}} = \mathrm{Corr}_t\big(\bar{\hat{y}}_t, \bar{y}_t\big)$. Stock-level RankIC averages the daily Spearman correlation of demeaned scores and labels:

$$\mathrm{RankIC} = \frac{1}{|T|} \sum_{t \in T} \rho\big(\hat{y}_{\cdot,t} - \bar{\hat{y}}_t,\ y_{\cdot,t} - \bar{y}_t\big).$$

Results. Table 4 summarizes the results. (1) The residual dark-window correction improves stock-level RankIC for all models with completed paired results, with relative gains ranging from 56% to 213%. The gains span both sequential and graph-relational backbones, indicating incremental ranking value from the on-chain channel under our protocol. (2) Market-level effects are less uniform. For MSGCA, mean market correlation decreases by 30%, while its variation across seeds falls by 82%. We therefore distinguish improved cross-sectional ranking from improvements in market-level direction and stability. (3) StockMixer leads the sequential family, MAN-SF the graph-relational family, and MSGCA achieves the highest overall RankIC. For MSGCA, RankIC rises by 93% despite the lower market-level correlation, suggesting that liquid tokenized quotes act mainly as equity-specific ranking signals, consistent with Findings 2 and 3.

6.2 Open-world next-asset prediction

Task. Motivated by evidence that repeated interactions can dominate temporal-graph evaluation (Poursafaei et al., 2022; Yi et al., 2025), we focus on transitions from an account's existing asset to a different acquisition, using only pre-event history. Public account histories make cross-asset acquisitions directly observable, allowing us to formulate the task as temporal next-item ranking (Section 5.3). An item is a platform–ticker pair. Address filters and ticker coverage are in Appendix B. For each evaluation event $e = (u, y, t)$, account u has acquired exactly one item before time t and subsequently acquires a different item y. Given the pre-event history $\mathcal{H}_u(t)$, a model scores every eligible candidate j using $s_\theta(u, j, t \mid \mathcal{H}_u(t))$ and ranks the realized item y. The candidate set contains all items observed for at least one day in the same chain universe before t, excluding the account's existing holding.

Baselines. We select representative baselines from three model families to cover complementary modeling assumptions. ID-based sequential and collaborative methods include SASRec (Kang & McAuley, 2018), FDSA (Zhang et al., 2019), transductive UniSRec (Hou et al., 2022),

Table 5: Experiment results of next-asset acquisition ranking across available folds and seeds. BSC is split into Ondo and Binance. Best performance in each column is in bold.

	MRR					Recall@10				
Model	Ondo	Binance	Solana	Ethereum	RH-chain	Ondo	Binance	Solana	Ethereum	RH-chain
ID-embedding sequential / collaborative filtering										
EASE	0.270±0.101	0.068±0.042	0.312±0.081	0.186±0.000	0.229±0.000	0.562±0.222	0.128±0.074	0.627±0.113	0.347±0.000	0.460±0.000
SASRec	0.299±0.131	0.071±0.045	0.307±0.078	**0.232**±0.005	0.283±0.011	0.605±0.197	0.143±0.093	0.650±0.116	**0.534**±0.012	0.619±0.017
FDSA	0.317±0.138	0.068±0.048	0.313±0.071	0.230±0.003	**0.296**±0.004	**0.611**±0.198	0.132±0.096	0.668±0.116	0.463±0.044	**0.662**±0.007
UniSRec	**0.321**±0.117	0.090±0.049	**0.314**±0.088	0.227±0.005	0.288±0.004	0.606±0.171	0.156±0.088	0.663±0.124	0.519±0.014	0.646±0.028
Dynamic graph / temporal										
DyGFormer	0.286±0.132	**0.274**±0.153	0.233±0.061	0.071±0.012	0.139±0.034	0.607±0.172	0.376±0.136	0.509±0.146	0.171±0.049	0.373±0.215
CAWN	0.273±0.097	0.187±0.089	0.293±0.046	0.191±0.004	0.178±0.081	0.572±0.061	**0.461**±0.275	**0.683**±0.086	0.453±0.009	0.432±0.264
LLM-based										
LLMRank	0.185±0.020	0.085±0.046	0.202±0.002	0.167±0.002	0.204±0.001	0.562±0.109	0.216±0.126	0.560±0.003	0.451±0.006	0.567±0.004
LLM-ESR	0.288±0.125	0.252±0.168	0.289±0.010	0.165±0.031	0.225±0.012	0.569±0.219	0.402±0.224	0.622±0.011	0.329±0.043	0.503±0.111

and EASE (Steck, 2019). Temporal-graph methods include DyGFormer (Yu et al., 2023) and CAWN (Wang et al., 2021), which model time-ordered account–asset interactions. LLM-based methods include LLMRank (Hou et al., 2024) and LLM-ESR (Liu et al., 2024). Further model details are provided in Appendix C.1.

Splits and metrics. (1) Splits. BSC, Solana, Ethereum, and RobinhoodChain are evaluated as separate datasets. Within each dataset, contiguous walk-forward test folds use earlier events for training; BSC is additionally split at the Binance venue launch and reported as Ondo and Binance. The last four training weeks form the validation window. **(2) Metrics.** We evaluate two metrics: MRR and Recall@10. MRR is the mean reciprocal rank of the realized next item, $\frac{1}{|E|}\sum_{e\in E} 1/\operatorname{rank}_e$, and therefore rewards placing the correct acquisition near the top of the list. Recall@10 is the fraction of events for which the realized item appears among the ten highest-ranked candidates, measuring top-10 shortlist coverage.

Results. Table 5 summarizes the results across chains and model families. Main observations are: (1) No model dominates, and the ranking follows the market regime rather than the chain. Sequential recommenders lead on Ondo, Solana, Ethereum, and RH-chain (best MRR 0.23–0.32), but drop by 72% to 0.07–0.09 after the Binance venue launch on the same chain. DyGFormer and LLM-ESR hold up there (0.274 and 0.252), about $3\times$ the strongest ID score. On Ethereum and RH-chain the best ID MRR still exceeds the strongest graph method by 21% and 66%. This shows the challenge of learning across regimes. (2) The launch drop is a cold-start failure of ID embeddings. On the hardest Binance fold, SASRec, FDSA, and UniSRec perform poorly, while neighborhood and text models stay above it. Learned item IDs carry no information for tokens listed after training; models that score candidates from interaction neighborhoods or text do not depend on item identity and therefore transfer to newly listed tokens. Therefore, this dataset introduces new challenges for open-world next-asset ranking: models must handle non-stationarity and represent newly listed tokens, rather than relying on larger backbones. Furthermore, we tested adding off-chain news and price information. It improves ranking on hard and cold-start events but not on the full population, because on-chain purchase sequences already capture most of the same signal (Appendix D).

7 Conclusion

Tokenized equities make previously invisible parts of the equity market observable by placing exchange claims on public ledgers. Our dataset bridges two data worlds previously studied largely in isolation: traditional equity markets and public blockchain markets. This alignment reveals that tokenized equities remain active while exchanges are closed, that liquid and fresh on-chain quotes are strongly informative about the next equity print, that a small set of liquid tokens can inform a broader equity universe, and that account-level co-holding is more closely associated with shared attention than with conventional return correlation. These observations motivate two new learning problems: dark-window equity forecasting and open-world next-asset ranking. Together, the dataset and evaluation protocols support future research in AI for finance at the intersection of equity forecasting and temporal graph learning.

Ethics statement

This work uses publicly observable blockchain transfers and swaps, together with exchange prices, news, and fundamentals obtained through vendor APIs. We do not redistribute these vendor files and will provide data collection code. We do not attempt to deanonymize addresses beyond what is already public on chain, and we discourage using the release for surveillance of individuals.

The dataset and benchmarks are research artifacts, not investment advice. Predictive metrics (market-level correlation, RankIC, MRR, and Recall@10) must not be read as trading performance. We are not aware of human-subjects issues that would require IRB review. Potential dual-use concerns are inherent to any public market dataset; we mitigate them by stating limitations and by not framing cross-modal associations as causal trading rules.

Reproducibility statement

All baseline implementations are taken from the public repositories of the original papers. We release our code at https://anonymous.4open.science/r/RWA-dataset-repo-CADC/.

AI use statement

In this work, we used generative AI tools to help polish manuscript text from internal dataset-analysis notes. We have not used generative AI tools to generate synthetic datasets, to invent experimental results, or to fabricate citations. All AI-assisted prose was reviewed and revised by the authors; every empirical number is the authors' responsibility. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Binance. Introducing bstocks: Tokenized securities 1:1 backing with 24/7 trading. https://www.binance.com/en/support/announcement/detail/2c0c92ed15ac42d1b14bb1eac00d22bb, June 2026.

Lin William Cong, Wayne R. Landsman, Daniel Rabetti, Che Zhang, and Wenqi Zhao. Tokenized stocks. Working paper, available at SSRN, 2025.

Timothy Dombrowski and Nasser Arshadi. Tokenized stock markets: Tracking quality, arbitrage limits, and price discovery. Working paper, available at SSRN, 2026.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. FNSPID: A comprehensive financial news dataset in time series. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

Jinyong Fan and Yanyan Shen. StockMixer: A simple yet strong MLP-based architecture for stock price forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Finnhub. Company news API. https://finnhub.io/docs/api/company-news, n.d. Accessed September 26, 2026.

Julia Gastinger, Shenyang Huang, Mikhail Galkin, Erfan Loghmani, Ali Parviz, Farimah Poursafaei, Jacob Danovitch, Emanuele Rossi, Ioannis Koutis, Heiner Stuckenschmidt, Reihaneh Rabbany, and Guillaume Rabusseau. TGB 2.0: A benchmark for learning on temporal knowledge graphs and heterogeneous graphs. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-4450.

Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. Towards universal sequence representation learning for recommender systems. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022. doi: 10.1145/3534678.3539381.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval (ECIR)*, pp. 364–381, 2024. doi: 10.1007/978-3-031-56060-6_24.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. Temporal graph benchmark for machine learning on temporal graphs. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Harry Kalodner, Malte Möser, Kevin Lee, Steven Goldfeder, Martin Plattner, Alishah Chator, and Arvind Narayanan. BlockSci: Design and applications of a blockchain analysis platform. In *USENIX Security Symposium*, 2020.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018. doi: 10.1109/ICDM.2018.00035.

Haoyuan Li, Mengxiao Zhang, Maoyuan Li, Jianzheng Li, Junyi Yang, Shuangyan Deng, Zijian Zhang, and Jiamou Liu. Multi-source multi-level multi-token Ethereum dataset and benchmark platform. *arXiv preprint arXiv:2501.11906*, 2025.

Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. MASTER: Market-guided stock transformer for stock price forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Qidong Liu, Xian Wu, Yejing Wang, Zijian Zhang, Feng Tian, Yefeng Zheng, and Xiangyu Zhao. LLM-ESR: Large language models enhancement for long-tailed sequential recommendation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-0839.

Hou-Wan Long, Nga-Man Wong, and Wei Cai. Bridging culture and finance: A multimodal analysis of memecoins in the Web3 ecosystem. In *Companion Proceedings of the ACM on Web Conference (WWW)*, 2025.

Katya Malinova and Andreas Park. Tokenized stocks for trading and capital raising. Working paper, available at SSRN, 2023.

Ondo Finance. Introducing ondo global markets—a new paradigm for securities tokenization. https://ondo.finance/blog/introducing-ondo-global-markets, February 2024.

Ondo Finance. Ondo stocks: Legal and regulatory. https://docs.ondo.finance/ondo-stocks/legal-and-regulatory, n.d. Accessed September 26, 2026.

Polygon.io. Stocks API: Aggregates and news. https://polygon.io/, n.d. Accessed September 26, 2026.

Farimah Poursafaei, Shenyang Huang, Kellin Pelrine, and Reihaneh Rabbany. Towards better evaluation for dynamic link prediction. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Reuters. US securities regulator rolls out five-year exemption for tokenized stock trading. *Reuters*, September 2026.

Robinhood Markets. Building with stock tokens. Robinhood Chain documentation, https://docs.robinhood.com/chain/building-with-stock-tokens/, 2025.

RWA.xyz. Tokenized stocks. https://app.rwa.xyz/stocks, n.d. Accessed September 26, 2026.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. MANSF.

Stefan Scharnowski. Fractional and around the clock: Trading activity in tokenized financial assets. *Journal of International Financial Markets, Institutions and Money*, 110:102355, July 2026. doi: 10.1016/j.intfin.2026.102355.

Kiarash Shamsi, Friedhelm Victor, Murat Kantarcioglu, Yulia R. Gel, and Cuneyt G. Akcora. Chartalist: Labeled graph datasets for UTXO and account-based blockchains. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Shahar Somin, Yaniv Altshuler, and Alex Pentland. Crypto-asset trading on top of Ethereum blockchain comprehensive dataset. *Scientific Data*, 12(1), 2025. doi: 10.1038/s41597-025-05662-w.

Harald Steck. Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference (WWW)*, 2019. doi: 10.1145/3308558.3313710.

U.S. Securities and Exchange Commission. Self-regulatory organizations; the Nasdaq stock market LLC; order approving a proposed rule change, as modified by amendment no. 2, to enable the trading of securities on Nasdaq in tokenized form. https://www.sec.gov/files/rules/sro/nasdaq/2026/34-105047.pdf, 2026. Release No. 34-105047; File No. SR-NASDAQ-2025-072; March 18, 2026.

Qian Wang, Zhen Zhang, Zemin Liu, Shengliang Lu, Bingqiao Luo, and Bingsheng He. EX-Graph: A pioneering dataset bridging Ethereum and X. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=juE0rWGCJW.

Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. Inductive representation learning in temporal networks via causal anonymous walks. In *International Conference on Learning Representations (ICLR)*, 2021.

Hongjie Xia, Huijie Ao, Long Li, Yu Liu, Sen Liu, Guangnan Ye, and Hongfeng Chai. CI-STHPAN: Pre-trained attention network for stock selection with channel-independent spatio-temporal hypergraph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. doi: 10.1609/aaai.v38i8.28770.

xStocks. xstocks are going live: Tokenized stocks for the DeFi era. https://xstocks.fi/news/xstocks-are-going-live-tokenized-stocks-for-the-defi-era, June 2025.

xStocks. Product legal overview. https://docs.xstocks.fi/docs/product-legal-overview, n.d. Accessed September 26, 2026.

Yumo Xu and Shay B. Cohen. Stock movement prediction from tweets and historical prices. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Lu Yi, Jie Peng, Yanping Zheng, Fengran Mo, Zhewei Wei, Yuhang Ye, Zixuan Yue, and Zengfeng Huang. TGB-Seq benchmark: Challenging temporal GNNs with complex sequential dynamics. In *International Conference on Learning Representations (ICLR)*, 2025.

Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. DyGFormer.

Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S. Sheng, Jiajie Xu, Deqing Wang, Guanfeng Liu, and Xiaofang Zhou. Feature-level deeper self-attention network for sequential recommendation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4320–4326, 2019. doi: 10.24963/ijcai.2019/600.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024. doi: 10.1145/3637528.3671801.

Yang Zhang, En Chun, Ziyun Mao, Yulu Wu, and Jun Wang. GS-Fuse: Granger-supervised gated fusion and multi-granularity alignment for event-driven financial forecasting. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 6570–6581, 2026. doi: 10.1145/3770855.3817927.

Zhen Zhang, Bingqiao Luo, Shengliang Lu, and Bingsheng He. Live graph lab: Towards open, dynamic and real transaction graphs with NFT. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=zr1e15kczE.

Lifan Zhao, Shuming Kong, and Yanyan Shen. DoubleAdapt: A meta-learning approach to incremental learning for stock trend forecasting. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2023.

Chang Zong, Jian Wan, Lucia Cascone, and Hang Zhou. Stock movement prediction with multimodal stable fusion via gated cross-attention mechanism. *Complex & Intelligent Systems*, 11, 2025. MSGCA.