Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can Generative AI Preserve Tail-Risk Networks? Evidence from Cryptocurrency Markets

Jiajing Sun
University of Chinese Academy of Sciences
Ruting Wang
rutiwang@cityu.edu.hk
City University of Hong Kong
Junda Wu
Beijing Institute of Mathematical Sciences and Applications

Research Article

Keywords: Generative AI, Synthetic financial data, Tail-risk networks, Quantile LASSO, Cryptocurrency, Expected shortfall

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11498528/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Can Generative AI Preserve Tail-Risk Networks? Evidence from Cryptocurrency Markets

Abstract

Generated financial scenarios inform risk estimation and portfolio decisions, but agreement with standard distributional diagnostics does not show that a generator preserves the structure a decision depends on. We evaluate generated scenarios in three layers: conventional distributional and risk diagnostics; preservation of a conditional lower-tail network, judged against the sampling uncertainty of the empirical network; and realised outcomes of an allocation rule that uses the network's node rankings. Using hourly data for ten cryptocurrencies over 2020–2025, we compare a stationary bootstrap, a GARCH-t copula model and a denoising diffusion probabilistic model (DDPM). The layers yield different conclusions about these generators. The discrepancy between the aggregated GARCH-t network and the observed network exceeds the bootstrap reference band at all 16 evaluation refits, yet overall distributional discrepancy stays within its band in 96 of the 157 GARCH-t panels that fall outside both the conventional-diagnostic and network bands. DDPM meets the joint conventional-diagnostic criterion in no panel but meets the single-panel network criterion in 120 of 160. Network differences change allocation weights and volatility exposure. After ex ante risk matching, differences in daily expected shortfall relative to the historical-network rule range from -0.09 to -0.02 percentage points, and every paired 95% interval contains zero. Controlled simulations show when ranking errors are costly: if tail hubs differ mainly in asymmetric downside dependence, a correlation-matching copula reverses the hub ranking and raises population expected shortfall by 3.1% relative to the true-network benchmark. Conventional diagnostics, network preservation and allocation outcomes give distinct assessments of these generators.

Keywords: Generative AI · Synthetic financial data · Tail-risk networks · Quantile LASSO · Cryptocurrency · Expected shortfall

JEL classification: C45 · C58 · G11 · G32

1 Introduction

Financial institutions use generated scenarios to estimate risk exposures, test allocation rules and develop models. Whether a generator is adequate depends on what the scenarios are used for. Validation usually targets marginal distributions, volatility dynamics, correlations and the tail losses of benchmark portfolios. Network-based risk monitoring relies on a different object: the conditional lower-tail relationships that identify which assets are central to joint downside risk. A generator can reproduce the usual targets and still misstate these relationships, or miss some of the usual targets and still recover them. This paper asks what each kind of evidence establishes when generated scenarios are used to estimate such a network and to allocate with it.

To make this question concrete, we follow the tail-event-driven network and Financial Risk Meter (FRM) literature (Härdle et al., 2016; Mihoci et al., 2020; Wang et al., 2024) and estimate each asset's 5% conditional return quantile from the contemporaneous returns of the other assets and from predetermined state variables, using an ℓ_1-penalised quantile regression. The coefficients form a directed network of conditional lower-tail associations, not estimates of causal transmission. Node scores rank assets by how strongly their returns enter the lower conditional quantiles of the other assets. Rankings of this kind are used in risk monitoring and can enter allocation rules directly, which makes them a concrete target for assessing generated scenarios.

Assessing preservation of this ranking involves two sources of difficulty. The observed network is itself an estimate: its coefficients and node rankings vary with the sample and with estimation conventions, so a discrepancy between a generated and an observed network has to be judged against the sampling variability of the observed estimate. In addition, a network discrepancy can change allocation weights without changing realised tail losses by a detectable amount, because its effect runs through the decision rule and the exposures that the rule creates.

We therefore evaluate generated scenarios in three layers that share one estimation and decision design. The first layer applies conventional diagnostics to marginal quantiles, serial dependence, correlations and the value-at-risk (VaR) and expected shortfall (ES) of benchmark portfolios. The second asks whether a generator preserves the network's coefficients and node rankings, using the bootstrap variability of the empirical reference network as the scale. The third traces network discrepancies through a fixed allocation rule to realised outcomes, with and without matching ex ante volatility. At the population level, equal joint distributions imply equal values of a uniquely defined network functional. The question is therefore a finite-sample one: what a finite set of diagnostics, an estimated network and a realised decision each reveal about a generator, and whether they agree.

We apply this design to an hourly panel of ten cryptocurrencies traded against USDT on Binance from January 2020 to December 2025. A liquidity rule that uses only pre-test information selects the assets and does not require them to survive to the end of the sample. We compare a joint stationary bootstrap, GJR-GARCH-t marginals with a Student-t copula, and a denoising diffusion probabilistic model (DDPM; Ho et al., 2020). Each generator is refitted quarterly on a 24-month window and evaluated over 2022–2025, and every network estimated from generated data uses the same sample length, covariates, scaling and penalty rule as the observed network. The primary network endpoint is the top-three overlap between aggregated generated networks and the observed network. The allocation rule halves the weights of the three highest-ranked assets, and the primary decision endpoint is the difference in realised ES between the rule applied to a generated network and to the historical network. The main comparison follows a time-stamped protocol, and analyses specified afterwards are reported as exploratory. Kraken US-dollar prices, a second asset panel and controlled simulations with a known network functional complement the main evidence.

The results first show why uncertainty in the reference matters. In the 24-month hourly windows it is nearly dense, so its information lies in connection strengths and rankings, and the observed top-three set appears in only 45% of bootstrap samples. Measured against this reference, the generators differ in ways that conventional diagnostics do not reveal. The stationary bootstrap reproduces both the diagnostics and the network. The discrepancy between the aggregated GARCH-t network and the observed network exceeds its reference band at all 16 refits, yet overall distributional discrepancy remains within its band in 96 of the 157 GARCH-t panels that fall outside both the conventional-diagnostic and the network bands.

DDPM shows the opposite pattern: no DDPM panel meets the joint conventional-diagnostic criterion, but 120 of 160 meet the single-panel network criterion. Coefficient-level and ranking-level agreement also diverge. On the primary endpoint, the mean top-three overlap is 2.69 for the stationary bootstrap, 2.19 for GARCH-t and 1.88 for DDPM, against 2.39 for an aggregated bootstrap network, so DDPM stays closer than GARCH-t to the observed coefficients but further from the observed ranking. The FRM index, an aggregate summary of penalty intensity, lies within its reference interval in 94–100% of refits for all three generators and does not distinguish them.

These ranking discrepancies also change allocation decisions. Relative to the historical network, the top-three set used for allocation is identical in 11 of 16 quarters for the stationary bootstrap, 5 for GARCH-t and 2 for DDPM. Because the most central assets in this panel, ETH and BTC, are also the least volatile, the rule moves weight towards more volatile assets, and part of the difference in realised tail loss reflects volatility exposure. After ex ante risk matching, synthetic-minus-historical differences in daily ES are -0.02, -0.09 and -0.09 percentage points for the stationary bootstrap, GARCH-t and DDPM, and all paired 95% intervals contain zero. Relative to the historical-network ES of 9.08%, the intervals are compatible with effects from about 2.3% lower to 0.9% higher tail loss. Controlled simulations show where ranking errors are costly. When tail hubs differ from the other assets mainly through asymmetric downside dependence, the correlation-matching GARCH-t copula reverses the hub ranking, and its allocations have a population ES 3.1% above the true-network benchmark.

Building on these findings, the paper contributes to the evaluation of synthetic financial data by linking the validation of an estimated network functional to its downstream use. It makes the sampling uncertainty of the empirical reference explicit and separates it from variation across generated panels and across training seeds. It measures how conventional and overall distributional diagnostics relate to network identification. It then follows the network into a fixed decision rule and reports realised effect sizes, their uncertainty and a controlled mechanism through which ranking errors become costly. Tail-risk-preserving generation, cross-asset tail dependence in generated data and decision-oriented evaluation are established in prior work (Cont et al., 2026; Huk et al., 2026; Cetingoz & Lehalle, 2025); the contribution here is the link between diagnostics, network identification under estimation uncertainty and realised decisions.

The remainder of the paper is organised as follows. Section 2 reviews related work, Sections 3 and 4 describe the data and the evaluation design, Section 5 reports the three layers of evidence, Section 6 presents controlled simulations and robustness checks, Section 7 discusses interpretation and scope, and Section 8 concludes.

2 Related literature

The study brings together research on financial scenario generation, the evaluation of synthetic data and conditional tail networks. The first two strands motivate the comparison of generators; the third defines the network that the comparison seeks to preserve.

2.1 Generative models for financial scenarios

Generative adversarial networks (Goodfellow et al., 2014) were adapted early to financial returns. They were used to reproduce univariate stylised facts (Takahashi et al., 2019; Wiese et al., 2020) and to calibrate trading strategies (Koshiyama et al., 2021). Later work generated multivariate and conditional paths, for

example with signature-based Wasserstein GANs (Ni et al., 2021; Liao et al., 2024), autoencoder and flow models for high-dimensional equity returns (Tepelyan & Gopal, 2023), and GANs that generate correlation matrices (Marti, 2020). Diffusion models (Ho et al., 2020; Song et al., 2021) have since become a second major family for time series (Tashiro et al., 2021; Rasul et al., 2021; Yuan & Qiao, 2024). Their financial applications are mostly recent preprints, which judge outputs by stylised facts, portfolio-level distributions or factor structure (Takahashi & Mizuno, 2024; Lesniewski & Trigila, 2024; Chen et al., 2025).

Within this literature, two approaches make tail dependence especially relevant to our question. Tail-GAN (Cont et al., 2026) makes tail risk the training target: its discriminator maps the sorted profit-and-loss outcomes of a set of benchmark strategies, through a differentiable sorting relaxation (Grover et al., 2019), to estimates of their value-at-risk and expected shortfall, and it is trained with a strictly consistent joint scoring function (Fissler & Ziegel, 2016). Its authors also report that historical simulation remains competitive on their own tail-risk metrics, an observation relevant to our comparison. Huk et al. (2026) separate marginal models from a diffusion-based copula, apply the method to cryptocurrency markets and report better forecasts of joint extremes. Generated data that preserve tail co-movement are also studied by Huh et al. (2026) and, from an extreme-value perspective, by Lhaut et al. (2025), and Asadi and Li (2026) train conditional generators so that the risk induced by downstream policies matches the data. We carry the tail-risk focus of this work to a different object: an estimated conditional lower-tail network and the allocation that uses its node rankings.

2.2 Evaluating synthetic financial data

The choice of generator also raises the question of how its output is evaluated. Surveys document that synthetic time series are usually evaluated with visual checks, summary statistics, distances between distributions, classifier-based discriminative scores and "train on synthetic, test on real" predictive scores (Esteban et al., 2017; Yoon et al., 2019; Assefa et al., 2020; Stenger et al., 2024). Kernel two-sample statistics (Gretton et al., 2012) and related energy distances (Székely & Rizzo, 2013) provide overall comparisons of distributions. A recent working paper (reference withheld for anonymous review) develops a formal finite-horizon specification test for generators of economic time series. The test compares the joint laws of observed and generated blocks with a characteristic-kernel discrepancy, and its calibration accounts for dependence within the observed series and within generated paths. By construction it targets the joint distribution, including marginal, serial, cross-series and tail features.

Alongside these distributional comparisons, a decision-oriented strand asks whether synthetic data are good enough for a specific use. Cetingoz and Lehalle (2025) argue that fidelity for portfolio construction must be judged against the intended use and the information content of the training sample. Aghapour et al. (2025) relate generator error to the stability of portfolio decisions, and Hounwanou and Gaba (2025) evaluate generators through portfolio and risk applications. Our paper belongs to this decision-oriented strand but differs in the downstream object: a conditional lower-tail network estimated from the generated data, and an allocation that uses the identity of its central nodes.

2.3 Conditional tail networks

For the network-based use considered here, quantile regression (Koenker & Bassett, 1978) underlies several measures of tail dependence between institutions or assets. CoVaR (Adrian & Brunnermeier, 2016) conditions one entity's quantile on another's distress, and CAViaR and multivariate regression-quantile

models describe the dynamics of conditional quantiles (Engle & Manganelli, 2004; White et al., 2015). Tail-event-driven networks (TENET) estimate penalised quantile regressions of each institution's return on those of all others and read the coefficients as a directed network (Härdle et al., 2016; Chen et al., 2019). The Financial Risk Meter averages the penalty parameters selected in these regressions to give an index of tail interconnectedness (Mihoci et al., 2020; Yu et al., 2023). FRM-based measures have been applied to emerging markets (Ben Amor et al., 2022) and to cryptocurrencies, where they are related to pricing-kernel volatility (Wang et al., 2024). The statistical foundations include ℓ_1-penalised quantile regression (Li & Zhu, 2008; Belloni & Chernozhukov, 2011), penalty selection by generalised approximate cross-validation (Yuan, 2006) or information criteria (Lee et al., 2014), and post-selection inference (Belloni et al., 2019). Connectedness measures based on variance decompositions (Diebold & Yılmaz, 2014) and quantile connectedness (Ando et al., 2022) have been widely applied to cryptocurrencies (Yi et al., 2018; Ji et al., 2019; Bouri et al., 2021).

These networks are estimated objects. Selected edges and node rankings depend on the sample, the quantile level and the penalty, and LASSO selection is a non-smooth function of the data. Whether a generator "preserves" such a network can therefore be judged only against a reference that reflects this estimation uncertainty, so that a discrepancy from the observed estimate can be compared with the discrepancies that resampling alone produces.

2.4 Contribution

Combining these strands, we evaluate generated scenarios by the conditional lower-tail network they produce and the decisions that use it. The quantile-LASSO network is re-estimated on generated panels that match the observed data in length, covariates, scaling and penalty rule, and the resulting discrepancies are measured against block-bootstrap uncertainty in the empirical reference network. Generator error is kept separate from variation across generated panels, from training sensitivity and from changes in the market state. The discrepancies are then traced through a prespecified allocation rule to realised out-of-sample losses, with effect sizes and the range of effects the data can exclude. Because the relationship between conventional diagnostics, network preservation and decision outcomes is the object of study, no layer is assumed to stand in for another.

This design makes it possible to ask whether agreement on standard diagnostics carries over to the network and, in turn, to the losses realised by the allocation rule.

3 Data

Implementing this comparison requires a common asset panel and a chronology that separates model fitting from evaluation. We construct the primary panel from Binance records and use Kraken prices to check sensitivity to the exchange and quotation currency.

3.1 Sources and acquisition

The primary data are hourly spot candlesticks (klines) from Binance Public Data, the exchange's official archive of monthly ZIP files (Binance, n.d.). We downloaded the monthly hourly archives for January 2020 to December 2025 and checked every file against its official SHA-256 checksum; all 938 archives

used in the study match. The raw files are stored unchanged. Timestamps are normalised before files are joined, because archives from January 2025 onward record open times in microseconds instead of milliseconds. A data manifest (Online Resource 1) records for every file its source, symbol, frequency, month, URL and retrieval time, the computed and official checksums, the row count, duplicate timestamps and missing hours.

The Kraken robustness sample uses hourly OHLCVT records for matched US-dollar pairs from the exchange's historical archive (Kraken, n.d.). Extracted members are verified by CRC-32 checks. Online Resource 1 documents the acquisition and explains how these member-level checks differ from verification of whole archive parts by SHA-256.

3.2 Asset eligibility and survivorship

Within the Binance archive, we select the primary universe using information dated on or before 31 December 2021. The Binance archive inventory contains 757 USDT-quoted symbols, including pairs that were later delisted. Excluding stablecoins, fiat-, gold- and other pegged tokens, and leveraged tokens leaves 679 symbols, and requiring an hourly archive under a single symbol for every month from January 2020 to December 2021 leaves 80. We rank these pairs by total USDT quote volume over 2020–2021 and retain the ten largest: BTC, ETH, BNB, DOGE, XRP, ADA, MATIC, LTC, LINK and EOS. An earlier development panel (BTC, ETH, BNB, XRP, LTC, ADA, TRX, LINK, XLM and ETC) serves as a check on panel composition. It was assembled for pipeline development rather than selected by the pre-test liquidity rule, and it shares seven assets with the primary panel.

Because survival to the end of the sample is not an eligibility condition, the rule admits assets whose trading changed later. Two selected tokens were migrated during the test period. Polygon's MATIC was converted into POL (last MATICUSDT bar 10 September 2024, 02:00 UTC; first POLUSDT bar 13 September 2024, 10:00 UTC), and EOS was rebranded as A (last EOSUSDT bar 26 May 2025, 02:00 UTC; first AUSDT bar 28 May 2025, 08:00 UTC). Binance's announcements state that both were one-for-one conversions, applied automatically to user balances, with spot trading suspended between the two dates (Binance, 2024, 2025). In the main analysis each asset stays in the universe through its successor at a one-for-one conversion. For portfolio accounting, the price change from the last predecessor close to the first successor close is booked when trading resumes; for estimation, no return is computed across the conversion gap. A sensitivity analysis instead liquidates the position at the last predecessor close, holds the proceeds in USDT and removes the asset from later allocations.

3.3 Cleaning and returns

After accounting for these migrations, we align the selected assets on a grid of 52,608 UTC hours. No archive contains duplicate timestamps or invalid OHLC records. Thirty-one hours are missing for every pair, reflecting exchange-wide interruptions (Online Resource 1). Hourly log returns are computed only from consecutive valid closes: a return is missing if either adjacent close is missing, missing intervals are never filled with zero returns, and no return is computed across a gap. An hour is *complete* if all ten assets have a valid return; 52,427 of the 52,608 hours in the primary panel are complete, and lagged quantities never cross an incomplete hour. Extreme returns are kept. The largest absolute hourly log return is 0.53 (DOGE, 29 January 2021), and several absolute returns above 0.29 occur in the 19 May 2021 sell-off.

The primary panel is quoted in USDT, so its returns combine changes in each asset's US-dollar price with changes in USDT's dollar value. The matched Kraken US-dollar subset measures sensitivity to the exchange and the quotation currency.

3.4 Chronology

To separate choices made before evaluation from subsequent performance, the protocol divides the sample into development (January 2020 to June 2021), validation (July to December 2021) and out-of-sample evaluation (January 2022 to December 2025). Models are refitted at every quarter start from 1 January 2022 to 1 October 2025, giving 16 test refits, each using the 24 months of complete hours before the refit date. The two validation refits (1 July and 1 October 2021) use the 18 and 21 months then available. Data after a refit date are never used for fitting, scaling, tuning or allocation at that date. The main comparison follows an analysis protocol written and time-stamped before test-period evaluation (Online Resource 2), which also records validation-stage choices and later protocol changes. Analyses specified after test-period results were available are identified as exploratory where they are introduced (Section 4.7), and the timing of the Tail-GAN audit is stated in Section 4.3.

4 Methodology

At each refit, the observed fitting window supplies both the data used to train the generators and the reference network. We first define that network and the generated panels, then describe how the three evaluation layers compare their distributional features, network structure and allocation outcomes.

4.1 Conditional lower-tail network

Let $r_{i,t}$ denote the hourly log return of asset i, and let $x_{i,t} = r_{i,t}/s_i$ be the return scaled by its standard deviation s_i in the fitting window. The same window-specific s_i scales the observed panel and every panel generated from a model fitted to that window. Following the quantile-LASSO approach behind tail-event-driven networks and the Financial Risk Meter (FRM) (Härdle et al., 2016; Mihoci et al., 2020; Wang et al., 2024), the conditional τ-quantile of each asset is modelled as

$$Q_\tau(x_{i,t} \mid x_{-i,t}, z_{i,t-1}) = \alpha_i + \sum_{j \neq i} \beta_{ij} x_{j,t} + \gamma_i' z_{i,t-1}, \tag{1}$$

where $x_{-i,t}$ contains the contemporaneous scaled returns of the other nine assets and $z_{i,t-1} = (x_{i,t-1}, m_{t-1}, v_{t-1})'$ contains predetermined state variables built from the same panel: the asset's own lagged return, the lagged equal-weighted market return m_{t-1} and the root-mean-square market return v_{t-1} over hours $t-24$ to $t-1$. The main quantile level is $\tau = 0.05$. The coefficients minimise

$$\frac{1}{n} \sum_t \rho_\tau \left(x_{i,t} - \alpha_i - \sum_{j \neq i} \beta_{ij} x_{j,t} - \gamma_i' z_{i,t-1} \right) + \lambda_i \left(\sum_{j \neq i} |\beta_{ij}| + \|\gamma_i\|_1 \right), \tag{2}$$

where $\rho_\tau(u) = u(\tau - 1\{u < 0\})$ is the check function. The intercept is not penalised. As in the FRM literature, the penalty λ_i is chosen by generalised approximate cross-validation (GACV; Yuan, 2006) over

20 log-spaced values between $\lambda_{\max}$ and $10^{-3}\lambda_{\max}$, where $\lambda_{\max}$ is the smallest penalty at which all slopes are zero.

We compute exact solutions with the Frisch–Newton interior-point algorithm (Portnoy & Koenker, 1997), representing the ℓ_1 penalty exactly by pseudo-observations through $\rho_\tau(u) + \rho_\tau(-u) = |u|$. On validation data the solutions agree with a general-purpose linear-programming solver to within 10^{-11} in every coefficient, and the computation is about 100 times faster, which makes the several thousand network estimates in the study feasible.

These fitted coefficients define the network as the 10×10 matrix W with entries $W_{ij} = \beta_{ij}$ for $i \neq j$. Rows index the response asset i and columns the conditioning asset j, so each entry is a directed edge $j \to i$: the conditional association between the contemporaneous return of asset j and the lower conditional quantile of asset i. These coefficients describe conditional dependence within the asset panel. Positive and negative entries are recorded separately. The primary node score is positive out-strength,

$$S_j = \sum_{i \neq j} \max(\beta_{ij}, 0), \tag{3}$$

which measures how strongly the moves of asset j enter the lower conditional quantiles of the other assets with a positive sign. Absolute out-strength, $\sum_{i \neq j} |\beta_{ij}|$, serves as a sensitivity score. Nodes are ranked by S_j, with ties broken by absolute out-strength and then by the pre-test liquidity order, and the *top-three set* consists of the first three nodes in this ranking.

To construct the state variables consistently across observed and generated data, each observation requires 24 hours of history within the same contiguous segment. No lagged quantity therefore crosses a missing hour, a bootstrap block boundary or the boundary between generated paths, and the first 24 hours of every segment or path serve only as history.

4.2 FRM index

The FRM index summarises the intensity of the selected penalty, whereas the coefficient network describes conditional connectivity. In the validation pilot, the GACV criterion was flat below about $10^{-3}\lambda_{\max}$ in long hourly windows ($n \approx 13{,}000$–$17{,}000$ with 12 regressors), so the selected coefficients are numerically close to unpenalised quantile regression and a mean selected penalty would describe the grid rather than the data. We therefore estimate the coefficient network on 24-month windows and the FRM index on short sub-windows in which the penalty is active. Each sub-window joins two 168-hour paths (2×144 usable hours) and is formed identically from observed blocks, bootstrap panels and generated panels. GACV searches 20 log-spaced values in $[10^{-4}\lambda_{\max}, \lambda_{\max}]$ on 20 evenly spaced sub-windows per panel, and the FRM index is the mean selected penalty across assets and sub-windows. In the pilot, 5% of selected penalties fell on a grid boundary. The sub-window specification was fixed before any test-period computation and is recorded in the protocol.

4.3 Scenario generators

To compare ways of reproducing these network features, all generators use the same information set: the scaled complete hours of the 24-month window before each refit date. A refit is a quarterly re-estimation using the data then available; a generated panel is one synthetic sample used to estimate a network. Each

produces independent 168-hour paths of the ten assets, and every path keeps its identifier.

The joint stationary bootstrap (SB) resamples rows of the ten-asset panel jointly, in blocks with geometrically distributed lengths (Politis & Romano, 1994). The mean block length follows the Politis–White rule with the Patton–Politis–White correction (Politis & White, 2004; Patton et al., 2009), applied to the squared equal-weighted market return, and blocks never span a missing hour.

The second generator (GARCH-t) combines GJR-GARCH-t marginals with a Student-t copula. Each asset follows a constant-mean GJR-GARCH(1,1) model with standardised Student-t innovations (Bollerslev, 1986; Glosten et al., 1993). Dependence between the probability-integral transforms of the standardised residuals is a constant-correlation Student-t copula (Demarta & McNeil, 2005). Its correlation matrix is obtained by inverting Kendall's τ and projecting to the nearest positive-definite matrix, and its degrees of freedom by profile likelihood. Each path starts from the fitted conditional state at a randomly drawn hour of the window.

The third generator is a denoising diffusion probabilistic model (DDPM; Ho et al., 2020) trained on all overlapping 168-hour windows. It uses a cosine noise schedule with 200 steps and a dilated one-dimensional convolutional residual denoiser with 12 blocks and 96 channels, about 1.09 million parameters. The model is trained to predict added Gaussian noise; this prediction is converted to a clean path for sampling, with the clean-path prediction clipped at 1.5 times the largest absolute training value. Optimisation and normalisation details are given in Online Resource 3, Section 3.4.

To keep network preservation separate from model selection, the neural configuration was chosen on the two validation refits by a prespecified rule that ranked candidates on conventional diagnostics only (marginal quantiles, autocorrelations of returns and squared returns, correlations, and benchmark-portfolio value-at-risk and expected shortfall) and never on network metrics, so that network preservation remains an independent evaluation target. The selected DDPM configuration uses 6,000 training iterations, batches of 128 windows, a learning rate of 2×10^{-4} and no marginal transformation; it was preferred to an inverse-hyperbolic-sine transformation in 11 of 12 validation comparisons. Seed 1 is the primary training run. Seeds 2 and 3 measure training sensitivity on the four first-quarter test refits (January 2022, 2023, 2024 and 2025), a subset fixed in the protocol.

The protocol also included Tail-GAN (Cont et al., 2026). Our implementation did not meet the eligibility criteria of a bounded benchmark audit on the original authors' synthetic design and is excluded from the comparison. The audit began after initial test-period results of the first implementation had been inspected. Online Resource 3 reports the audit procedure, its independent evaluation samples, the stopping rule, the version history and all results. The audit evaluates our implementation under the stated computing budget.

4.4 Synthetic panels and reference networks

The generated paths are assembled into panels that match the effective length of the observed sample. For refit date d, let n_d be the effective sample size of the observed network regression, that is, the number of complete hours with 24 hours of within-segment history. Each synthetic panel consists of $K_d = \lceil n_d/144 \rceil$ independent generated paths, truncated to exactly n_d usable rows, so the synthetic and observed regressions have the same length, covariates, scaling and tuning rule. Each generator produces $R = 10$ panels per refit. The *aggregated synthetic network* is the element-wise mean of the ten coefficient matrices; a *single-panel network* describes what a user would obtain from one generated history.

The *in-sample reference* for refit d is the network estimated on the observed fitting window. Because every generator is trained on that window, the in-sample reference is the natural target for asking whether a generator preserves the network of the data it learned from. The network estimated on the following quarter's observed hours serves only for ex post comparison and never enters an allocation decision.

4.5 Separating three sources of uncertainty

Discrepancies from this reference can reflect uncertainty in the reference itself as well as variation in generated data. We separate three sources of variation. The first is uncertainty in the observed reference. We resample 168-hour blocks of complete hours with a moving-block bootstrap (Künsch, 1989), drawing $B = 100$ bootstrap panels per refit. Each block is treated as a separate segment, so bootstrap panels have exactly the structure of generated panels, and the network is re-estimated with the same rule. From these networks we report edge-selection frequencies, sign stability, the distribution of top-three sets and *bootstrap reference bands*. The single-panel band is the distribution of a discrepancy between two independent bootstrap networks; the aggregated band compares the mean of ten bootstrap networks with one further bootstrap network. A generated network lies *within the band* for a discrepancy metric if its discrepancy from the observed reference is below the 95th percentile of the corresponding band, and for an agreement measure such as the Spearman correlation if it is above the 5th percentile. The bands are descriptive benchmarks for the size of a discrepancy. LASSO selection and node rankings are non-smooth functionals of the data, so the bands are not calibrated edge-level or rank-level tests (cf. Belloni et al., 2019), and lying within a band does not establish equivalence.

The second source is simulation variability given a fitted generator, measured by the dispersion across the ten panels drawn from one trained model. The third is training sensitivity, measured by differences between independently initialised training runs of DDPM. Generated paths and training seeds are never treated as additional independent market histories. Online Resource 3, Section 3.1, gives the bootstrap pairing counts and the construction of the reference thresholds.

4.6 Conventional diagnostics

The first evaluation layer asks which distributional and risk features the generated data reproduce. Conventional diagnostics are computed on path panels with a common structure, with the observed data cut into consecutive non-overlapping 168-hour blocks. Six discrepancies cover the features that standard validation targets: the mean absolute error of marginal quantiles at 1%, 5%, 95% and 99%, in scaled units; the mean absolute errors of within-path autocorrelations of returns and of squared returns at lags 1–24; the mean absolute error of the pairwise correlations; and the relative errors of the value-at-risk and of the expected shortfall of 22 fixed benchmark portfolios (equal weight, a BTC–ETH portfolio and 20 random long-only portfolios) at horizons of 1 hour (95% and 99%), 24 hours (95% and 99%) and 168 hours (95%). Each discrepancy is compared with its own bootstrap band, constructed exactly as for the network metrics, and a panel meets the *joint conventional-diagnostic criterion* when all six lie within their bands. We measure overall distributional discrepancy using the energy distance between the joint distributions of 24-hour, ten-asset blocks (Székely & Rizzo, 2013), and compare it with its bootstrap reference band separately. This descriptive measure summarises joint-distribution discrepancies over 24-hour blocks; smaller values indicate closer distributions. The averaging conventions, portfolio weights and distance calculation are specified in Online Resource 3, Section 3.1.

4.7 Network-comparison metrics

The second layer asks whether those data reproduce the network itself. A generated network W^s is compared with the observed reference W^o through five metrics: the relative Frobenius error $\|W^s - W^o\|_F / \|W^o\|_F$ over off-diagonal entries; three-way signed-edge agreement (negative, zero or positive); the Spearman correlation between node scores; the number of common assets in the two top-three sets (0–3); and the difference in edge density. A single panel meets the *single-panel network criterion* when both its Frobenius error and its node-score Spearman correlation lie within their single-panel bands. The prespecified primary network endpoint is the top-three overlap of the aggregated synthetic network with the in-sample reference, averaged over the 16 test refits. In addition, each generator's top-three set is located within the bootstrap distribution of the observed top-three set, through its bootstrap frequency and the bootstrap inclusion probabilities of its assets. These frequencies are descriptive, not calibrated probabilities.

The results are organised in three layers: conventional diagnostics (Layer 1), network and ranking preservation against reference uncertainty (Layer 2), and consequences under the allocation rule (Layer 3). Two analyses were specified after the main results were available and use saved outputs only; they are reported as exploratory. The first recomputes node rankings from coefficients expressed in raw return units, $\beta_{ij} s_i / s_j$, which is exact for unpenalised quantile regression (the penalty is inactive here), and from coefficients with $|\beta_{ij}| < 0.05$ set to zero, and measures each generator's agreement with the observed ranking under the same convention. The second compares the observed network of each refit with the network estimated on the following quarter (ex post) and with the observed network of the next refit, relative to the bootstrap band.

4.8 Allocation rule and evaluation

The third layer links network rankings to an implementable decision. At each test refit date the allocation starts from equal weights $1/N$, halves the weights of the three highest-ranked nodes and redistributes the released $1.5/N$ equally across the other seven assets. With ten assets, the selected three each receive 5% and the remaining seven each receive approximately 12.14%. The portfolio is long-only and fully invested, rankings use information up to the refit date, and trades execute at the close of the first hour after that date, with a 24-hour lag as a sensitivity analysis. Between rebalancing dates, weights drift with observed prices. The rule is a transparent way to act on a network ranking: it reduces exposure to the assets whose returns are most strongly associated with the lower conditional quantiles of the others. It is fixed in advance and not optimised for portfolio performance, so differences between allocations that use different networks arise only from the ranking each network supplies.

The historical-network rule (HN) uses the in-sample observed network, and a synthetic-network rule (SN) for each generator uses its aggregated synthetic network; for each SN rule we also report the distribution of outcomes across its single-panel networks. Two benchmarks use no network: equal weights (EW), a demanding benchmark for portfolio rules (DeMiguel et al., 2009), and a historical-volatility rule (HV) that halves the three assets with the highest trailing 24-month volatility. A future-network rule (FN) uses the following quarter's realised network. It is unavailable at the allocation date and serves only as an ex post diagnostic, so HN is the implementable comparator.

Outcomes are daily (00:00 UTC) portfolio log losses from January 2022 to December 2025, summarised by the 95% expected shortfall (the mean of losses at or above the empirical 95% quantile), mean return, volatility, maximum drawdown and turnover. Transaction costs of 0, 10 and 25 basis points per unit of

one-way turnover are applied at each rebalance. The primary decision endpoint is the difference in realised 95% expected shortfall between each SN rule and HN. It is assessed with a paired stationary bootstrap of days (mean block length 20 days, 2,000 draws) that applies the same resampled days to every strategy. The resulting intervals are conditional on the implemented fitting procedures and exclude variation from retraining the generators. No equivalence margin was prespecified, so the intervals describe the uncertainty of effect sizes and do not establish equivalence.

Further prespecified sensitivity analyses address exposure and implementation. In an ex ante risk-matched version, each portfolio is scaled to the ex ante volatility of HN, estimated from the trailing 90 days. Scaling below one leaves cash; scaling above one uses borrowing, both at zero interest. Online Resource 3, Section 3.2, gives the scaling formula, trading-cost convention and interval construction. In an exit-rule version, migrated tokens are liquidated instead of converted. Quarters are classified as high-volatility if realised equal-weight volatility exceeds the 75th percentile of quarterly volatility in 2020–2021. Five event windows were fixed before the evaluation: the Terra/LUNA collapse, the FTX failure, the USDC de-peg, the yen-carry unwind of August 2024 and the October 2025 liquidation cascade. Because a few days do not support tail averages, the windows are described by cumulative return and drawdown.

4.9 Controlled simulations

To assess these comparisons against a known target, the controlled simulations use a ten-asset process with heavy tails, volatility clustering and a downside-only common shock:

$$\begin{aligned} x_{i,t} &= \sigma_{i,t}\varepsilon_{i,t}, \\ \varepsilon_{i,t} &= a_i f_t + c_i g_t + (1 - a_i^2 - c_i^2)^{1/2}\eta_{i,t}. \end{aligned} \tag{4}$$

Here f_t and $\eta_{i,t}$ are unit-variance Student-t(4) variables, g_t is the centred, unit-variance version of a shock that is negative with probability 0.05 and zero otherwise, and $\sigma_{i,t}^2$ follows a GARCH(1,1) recursion with $\alpha = 0.08$ and $\beta = 0.90$. Assets with larger c_i co-move more strongly in crashes and act as tail hubs. In design A the hubs differ from the other assets in both ordinary correlation and downside dependence. In design B, the tail-only design, the hubs' factor loading is lowered so that their innovation correlation with non-hubs equals the innovation correlation among non-hubs, leaving the hubs distinguished mainly by downside dependence. Online Resource 3, Section 3.3, specifies the shock distribution, loadings and regime changes.

Generators are trained on samples of 8,760 or 17,520 hours from the pre-break regime. Their networks and decisions are evaluated against the same regime, which isolates generator and estimation error, and against post-break regimes with stronger common downside shocks or a change in hub identity, which measure transfer across regimes. The structural loadings in (4) are not network coefficients. The reference network functional for each regime is the unpenalised τ-quantile regression of equation (1), the $\lambda \to 0$ limit of the estimator, computed on an independent sample of 500,000 hours. The population 95% expected shortfall of daily portfolio losses for each allocation is approximated from the same sample, using fixed-weight sums of simulated daily returns (Online Resource 3, Section 3.3).

4.10 Implementation

The code is written in Python 3.14 with NumPy, SciPy, pandas, statsmodels, arch and PyTorch 2.14, and the neural models were trained on an Apple M2 Max GPU (Metal Performance Shaders). Seeds, software versions, computing times and all protocol changes are logged in the replication package. Claude (Anthropic) assisted with analysis-code generation and execution and manuscript drafting. Codex (OpenAI) assisted with manuscript editing and document preparation.

5 Results

Unless stated otherwise, results refer to the 16 out-of-sample refits from January 2022 to October 2025, the configurations fixed in the protocol and training seed 1; the two validation refits of 2021 were used only to choose configurations. Table 1 describes the asset panel and Table 2 the generators and their computing budgets. Because the precision of the reference network sets the scale for every later comparison, the section begins with the reference and then reports the three layers in turn.

Table 1 Primary asset panel and hourly return characteristics, January 2020 to December 2025

Asset	Pre-test rank	Quote volume 2020–21 (bn USDT)	Test-period status	Hourly returns (n)	Ann. vol. (%)	Skewness	Excess kurtosis	Min (%)	Max (%)
BTC	1	1440.8	Continuous	52561	62.8	-0.95	54.0	-20.1	16.0
ETH	2	873.7	Continuous	52561	81.0	-0.93	28.9	-23.4	14.0
BNB	3	355.4	Continuous	52561	83.1	-1.13	42.8	-22.6	13.8
DOGE	4	336.8	Continuous	52561	129.1	2.21	110.7	-29.3	53.1
XRP	5	295.8	Continuous	52561	103.8	-0.05	42.7	-22.5	24.5
ADA	6	229.0	Continuous	52561	104.6	-0.18	30.8	-24.3	29.0
MATIC	7	120.9	Migrated to POL (Sep 2024)	52481	125.0	0.17	29.2	-29.3	30.4
LTC	8	108.9	Continuous	52561	95.3	-1.07	43.3	-30.0	22.3
LINK	9	108.2	Continuous	52561	110.1	-0.65	20.5	-26.2	16.3
EOS	10	83.9	Migrated to A (May 2025)	52507	102.9	-1.52	46.2	-30.0	14.7

Note: Pre-test rank and quote volume use 2020–2021 Binance spot data only. Statistics refer to hourly log returns computed from consecutive valid closes (USDT quotation); MATIC and EOS are continued by POL and A after the one-for-one migrations. Ann. vol. = hourly standard deviation $\times \sqrt{(24 \times 365)}$.

Table 2 Scenario generators, configurations and computing budgets (16 test refits, seed 1)

Generator	Configuration	Refits	Mean fit time (s)	Mean sampling time (s)	Device
Stationary bootstrap	mean block length 165 h (PW rule)	16	13	0.0	CPU
GJR-GARCH-t + t-copula	GJR-GARCH(1,1)-t margins; t-copula	16	4	1.9	CPU
Diffusion (DDPM)	6000 iterations, batch 128, T = 200; 1.09M parameters	16	1236	80.1	Apple GPU (MPS)

Note: Fit and sampling times are wall-clock seconds per refit on a shared Apple M2 Max (CPU or Metal GPU); sampling produces 10 panels of about 120 paths of 168 hours. Tail-GAN is not listed because it was excluded after the reproduction audit (Section 4.3; Online Resource 3).

5.1 The reference network and its uncertainty

In the 24-month hourly windows, which contain 17,139 to 17,494 effective observations, the estimator selects almost every edge. Network density averages 0.997, and an observed network contains 0.44 negative edges on average. The GACV-selected penalty lies in the flat region of the criterion, and the penalised and unpenalised networks give the same top-three set at all 16 refits. At this frequency and window length the quantile-LASSO network is in effect a dense matrix of conditional quantile coefficients, and its information lies in connection strengths and node rankings rather than in which edges are selected.

Fig. 1(a) shows the average coefficient matrix. Almost all associations are positive. The strongest associations link BTC and ETH in both directions (mean coefficients 0.47 and 0.41); ETH is the main conditioning asset for LINK, MATIC and BNB, and EOS loads mainly on LTC. Ranked by positive out-strength averaged over the test refits, ETH (1.52) leads BTC (1.15), LINK (1.09) and ADA (0.99), and DOGE (0.54) is the least central asset.

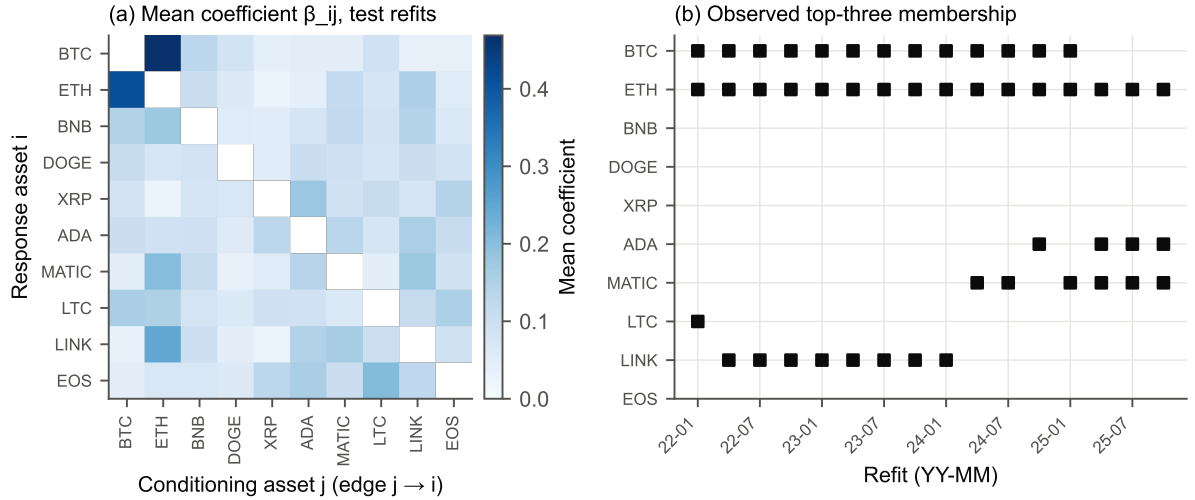


Fig. 1 Observed conditional lower-tail network. (a) Mean standardised coefficient β_{ij} across 16 quarterly refits in 2022–2025: columns are conditioning assets j, rows are response assets i, and darker cells indicate larger coefficients for edges $j \to i$. Blank diagonal cells exclude self-links. (b) Each square marks membership in the observed top-three set, ranked by positive out-strength $S_j = \sum_{i \neq j} \max(\beta_{ij}, 0)$. Dates denote refits (YY-MM).

The average ranking, however, conceals considerable estimation uncertainty. Over the test refits, ETH is in the observed top three 16 times, BTC 13 times, LINK 8 times, MATIC 6 times, ADA 4 times and LTC once (Fig. 1(b)). The 100 bootstrap networks of a refit produce on average 9.9 distinct top-three sets; the most frequent set occurs in 47% of replications on average (range 22–86%), and the observed set in 45%. Fig. 2 shows that ETH is almost always included, whereas the second and third places are contested among BTC, LINK, MATIC and ADA, and the contest changes over time. Two independent bootstrap networks differ by a relative Frobenius distance of 0.25, have a node-score Spearman correlation of 0.79 and share 2.24 of their top-three assets on average. The mean of ten bootstrap networks compared with a further one gives 0.19, 0.87 and 2.39 (Table 3). These are the reference levels for the generated networks: a generated network that reaches them differs from the observed network by about as much as a network estimated from a resampled history.

The ranking target also depends on estimation conventions, which the exploratory comparisons in Table 4 quantify. Coefficients estimated on window-standardised returns rank nodes differently from the

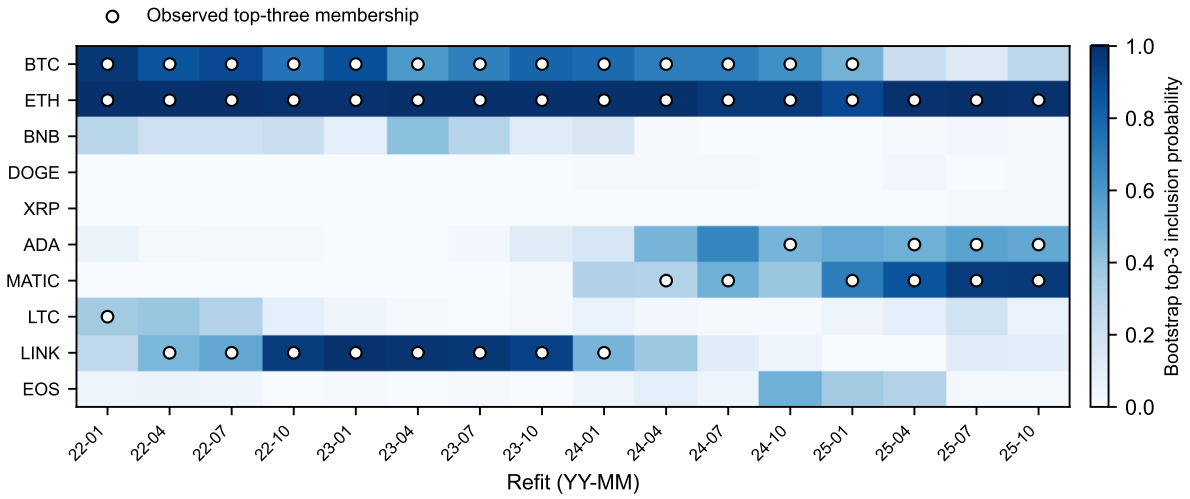


Fig. 2 Uncertainty of the observed node ranking. Each cell gives the fraction of 100 moving-block bootstrap networks (168-hour blocks) in which an asset enters the top-three set at a quarterly refit. Darker shading indicates a larger fraction; the colour bar runs from 0 to 1. Black-outlined, white-filled circles mark the three assets selected by the observed fitting-window network, ranked by positive out-strength, and do not denote significance. Each column therefore contains exactly three circles. Dates denote refits (YY-MM).

same coefficients expressed in raw return units, $\beta_{ij} s_i / s_j$, and the observed top-three set is the same under both conventions at only 6 of 16 refits. Setting coefficients below 0.05 in absolute value to zero leaves the top-three set unchanged at 11 of 16 refits. Because the identity of the most connected assets depends on these conventions, generated and observed networks are always compared under the same convention.

Beyond sampling uncertainty and estimation conventions, the target may change as market conditions evolve. The network estimated on the quarter after each refit differs from the 24-month fitting-window network by a relative Frobenius distance of 0.67, with a node-score Spearman correlation of 0.28 and a top-three overlap of 1.50 (Table 5). This distance exceeds the 95% reference band at all 16 refits and exceeds the within-window generator discrepancies reported below. The quarterly networks are estimated on 2,080–2,184 hours, however, so the comparison combines changes in market conditions with a shorter window and lower estimation precision. Consecutive 24-month refits, which share 21 months of data, differ by a distance of 0.13 and have a top-three overlap of 2.67. The controlled simulations in Section 6.1 isolate the effect of a change in market regime.

5.2 Layer 1: conventional distributional and risk diagnostics

With the reference variability established, we turn to the generated panels. Table 6 reports the conventional diagnostics, each measured against a bootstrap band computed in the same way. The stationary bootstrap meets the joint conventional-diagnostic criterion in 98% of panels, so by these diagnostics its panels are close to resampled observed data. GARCH-t and DDPM meet the criterion in no panel, but they fail in different ways. GARCH-t has the larger marginal-quantile discrepancy (mean error 0.25, against 0.22 for DDPM and 0.09 for the reference) and the larger error in benchmark-portfolio ES (0.16). DDPM is closer on serial dependence and on portfolio ES (relative ES error 0.10, against 0.09 for the reference) but has the larger correlation error (0.17, against 0.10 for GARCH-t).

The overall distributional discrepancy on 24-hour blocks discriminates less sharply. It lies within its

Table 3 Network preservation against the in-sample observed network (means over 16 test refits)

Network	Refits	Rel. Frobenius error	Sign agreement	Spearman (node scores)	Top-3 overlap (0–3)	Top-3 identical (share)	Within band: Frobenius	Within band: Spearman
Reference band: two bootstrap networks	16	0.251	0.969	0.790	2.244	0.337		
Reference band: mean of 10 vs one bootstrap network	16	0.188	0.979	0.866	2.385	0.431		
Stationary bootstrap — single panel	16	0.177	0.978	0.883	2.462	0.494	1.000	1.000
Stationary bootstrap — aggregated	16	0.068	0.995	0.971	2.688	0.688	1.000	1.000
GJR-GARCH-t + t-copula — single panel	16	0.434	0.946	0.621	1.994	0.169	0.019	0.625
GJR-GARCH-t + t-copula — aggregated	16	0.277	0.986	0.757	2.188	0.312	0.000	0.750
Diffusion (DDPM) — single panel	16	0.280	0.968	0.682	1.844	0.125	0.881	0.794
Diffusion (DDPM) — aggregated	16	0.244	0.980	0.730	1.875	0.125	0.438	0.625

Note: Single panel: average over the 10 generated panels; aggregated: element-wise mean of the 10 panel networks. Reference bands from 100 moving-block bootstrap (168-hour block) networks of the observed window. Within band: share of refits (aggregated) or panels (single) whose discrepancy lies inside the 95% band (Frobenius) or above its 5% quantile (Spearman).

Table 4 Sensitivity of node-ranking agreement to modelling conventions (aggregated networks, 16 test refits)

Generator	Top-3 overlap with HN: standardised	Spearman with HN: standardised	Top-3 overlap with HN: raw units	Spearman with HN: raw units	Top-3 overlap with HN: threshold $\|\beta\|<0.05 \to 0$	Spearman with HN: threshold $\|\beta\|<0.05 \to 0$
Stationary bootstrap	2.69	0.97	2.88	0.98	2.44	0.96
GJR-GARCH-t + t-copula	2.19	0.76	2.56	0.89	1.94	0.71
Diffusion (DDPM)	1.88	0.73	2.44	0.88	1.88	0.71

Note: Standardised: main convention (coefficients estimated on window-standardised returns). Raw units: $\beta_{ij}\ s_i/s_j$. Threshold: coefficients with $|\beta| < 0.05$ set to zero (exploratory). Agreement with the observed network (HN) is measured under the same convention. Exploratory sensitivity analyses on saved outputs.

Table 5 Historical and subsequent networks: distance between the observed network and later networks

Comparison	Refits	Rel. Frobenius distance	Spearman	Top-3 overlap	Share above bootstrap band (95%)
Following-quarter network vs HN (ex post; 2,080–2,184 h)	16	0.665	0.281	1.500	1.0
HN at next refit vs HN (21 of 24 months shared)	15	0.125		2.667	
Two bootstrap networks of the same window (reference)	16	0.251	0.790	2.244	

Note: The following-quarter network is estimated on the quarter after each refit (ex post; never used for decisions). Share above bootstrap band: refits at which the Frobenius distance exceeds the 95th percentile of the distance between two bootstrap networks of the same window.

band in 62% of GARCH-t panels and 65% of DDPM panels, although no panel of either generator meets the joint criterion. A tail-risk training target of the Tail-GAN type, the relative error of the VaR and ES of 60 benchmark trading strategies, is 0.04 for the reference, 0.06 for the stationary bootstrap, 0.18 for DDPM and 0.30 for GARCH-t. It separates the stationary bootstrap from the two model-based generators and places DDPM closer to the data than GARCH-t.

Table 6 Conventional diagnostics, overall distributional discrepancy and network fidelity (means over 160 panels per generator)

Generator	Quantiles	ACF of squares	Corr.	Portfolio ES	Distrib. distance	All conventional within band	Distrib. within band	Network within band	Strategy VaR/ES error
Bootstrap reference (MBB-168)	0.090	0.019	0.021	0.087	0.076	0.968	0.969	0.995	0.040
Stationary bootstrap	0.089	0.019	0.018	0.082	0.076	0.975	0.975	1.000	0.055
GJR-GARCH-t + t-copula	0.245	0.066	0.101	0.155	0.095	0.000	0.619	0.019	0.300
Diffusion (DDPM)	0.217	0.064	0.165	0.097	0.092	0.000	0.650	0.750	0.182

Note: Discrepancies: mean absolute error of marginal quantiles (1, 5, 95, 99%), of autocorrelations of returns and squared returns (lags 1–24) and of pairwise correlations; mean relative error of benchmark-portfolio VaR and ES (22 portfolios, 1-, 24- and 168-hour horizons); overall distributional discrepancy between 24-hour ten-asset blocks. Bands: 95th percentile of the discrepancy between two independent bootstrap panels. Strategy VaR/ES error: mean relative error of 5% VaR and ES of 60 benchmark trading strategies (results for an author-aligned strategy set and for return ACF and portfolio VaR are in the replication outputs). Within-band columns are shares of panels.

5.3 Layer 2: preservation of the network and of node rankings

Whether these diagnostic differences carry over to the network is the next question. Table 3 reports network discrepancies against the observed reference, Fig. 3 shows them refit by refit, and Table 7 places each generator's top-three set within the bootstrap distribution of the observed set.

The stationary bootstrap preserves the network. Its aggregated network has a relative Frobenius error of 0.07, a Spearman correlation of 0.97 and a top-three overlap of 2.69, and it reproduces the observed top-three set in 11 of 16 refits. Because the bootstrap resamples the rows of the fitting window, this closeness partly reflects direct reuse of the observed data; the bootstrap is therefore a benchmark for what preservation looks like rather than a model of the market.

Relative to this resampling benchmark, GARCH-t distorts the network by more than the reference uncertainty. The Frobenius error of its aggregated network (0.28) exceeds the 95% band at every test refit. Its Spearman correlation is 0.76, and its top-three overlap averages 2.19, with identical sets in 5 of 16 refits. In 6 of 16 refits it selects a top-three set that the bootstrap never produced, and in 5 refits a set containing an asset whose bootstrap inclusion probability is below 0.10.

DDPM's discrepancies are of a different kind. Its aggregated Frobenius error of 0.24 lies within the band in 7 of 16 refits, so its coefficient discrepancy is often of the size that resampling produces. Its ranking departs from the observed one more often: the top-three overlap is 1.88, the sets are identical in only 2 of 16 refits, and the network systematically promotes BNB and demotes BTC. The DDPM top-three set nonetheless occurs among the bootstrap sets of the observed window in 14 of 16 refits, with a mean bootstrap frequency of 0.19. DDPM thus tends to select sets that a resampled history could also produce,

but rarely the set that the observed history produced.

Neither the penalty nor panel-to-panel noise explains these discrepancies. Re-estimating every synthetic panel with the observed window's penalties, or without any penalty, leaves all metrics unchanged to the second decimal. Aggregation over ten panels reduces panel-to-panel variability, but GARCH-t and DDPM retain their coefficient and ranking discrepancies after aggregation, with smaller reductions in Frobenius error than the bootstrap and little change in top-three overlap. On the prespecified primary endpoint, the mean top-three overlap of the aggregated network, the ordering is stationary bootstrap 2.69, GARCH-t 2.19 and DDPM 1.88, against a bootstrap reference level of 2.39. GARCH-t therefore agrees more closely than DDPM with the observed ranking, whereas DDPM agrees more closely with the observed coefficients relative to their sampling variability. Coefficient-level and ranking-level preservation are distinct properties, and the allocation rule acts on the ranking.

The unit convention affects the size of the ranking discrepancies (Table 4). With coefficients in raw return units, the top-three overlap with the observed network rises to 2.56 for GARCH-t and 2.44 for DDPM, and their Spearman correlations rise to 0.89 and 0.88. With small coefficients set to zero, the overlaps are 1.94 and 1.88. The qualitative ordering, with the bootstrap closest and GARCH-t and DDPM further away, holds under every convention, but the magnitude of a ranking discrepancy is not a convention-free property of a generator. Matching the convention is part of defining the preservation target.

Table 7 Generated top-three sets against the bootstrap distribution of the observed network's top-three set (16 test refits)

Network	Refits	Bootstrap frequency of the top-3 set (mean)	Refits with set never produced by bootstrap	Mean inclusion probability of chosen assets	Refits with an asset below 0.10 inclusion
Observed network (HN)	16	0.453	0	0.801	0
Stationary bootstrap	16	0.439	0	0.797	0
GJR-GARCH-t + t-copula	16	0.233	6	0.667	5
Diffusion (DDPM)	16	0.186	2	0.665	2

Note: For each refit, 100 moving-block bootstrap networks of the observed window give the frequency of each top-three set and each asset's top-three inclusion probability. Bootstrap frequency: share of bootstrap networks with exactly the same set (0 = never produced). Descriptive only: bootstrap frequencies of non-smooth ranking functionals are not calibrated probabilities and are centred on the observed estimate.

The FRM index, computed on short sub-windows in which the penalty is active, summarises the network more coarsely (Table 8). For all three generators, the mean index of the generated panels lies within the 95% interval of the structure-matched bootstrap reference in 94–100% of refits. Agreement in aggregate penalty intensity can therefore coexist with discrepancies in coefficient structure and node rankings.

5.4 How Layers 1 and 2 relate

The preceding comparisons suggest that diagnostic performance and network preservation need not move together. Table 9 and Fig. 4 examine this correspondence by classifying each generated panel under both layers. For the stationary bootstrap and GARCH-t the two layers agree: the bootstrap meets both criteria in 156 of 160 panels, and GARCH-t falls outside both in 157. The overall distributional discrepancy does not share this verdict. It stays within its band in 96 of those 157 GARCH-t panels, so a favourable overall distributional discrepancy can coexist with a network discrepancy that exceeds reference uncertainty.

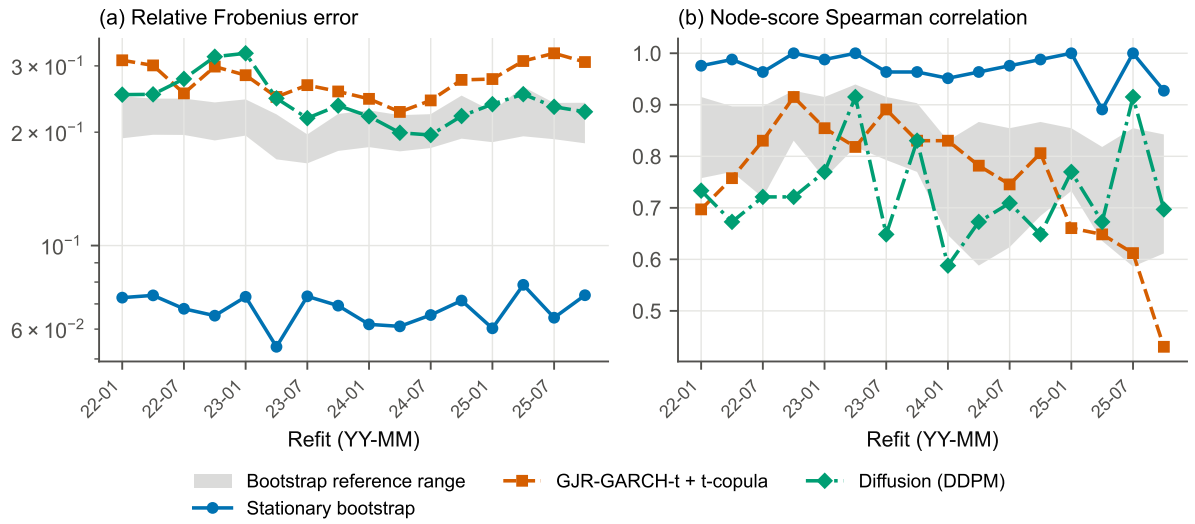


Fig. 3 Aggregated-network discrepancies by quarterly refit (2022–2025). Each generated network is the mean of ten panel networks. Circles with solid lines: stationary bootstrap; squares with dashed lines: GARCH-t; diamonds with dash-dot lines: denoising diffusion probabilistic model (DDPM). Left: relative Frobenius error (log scale; smaller is closer). Right: Spearman correlation of node scores (larger is closer). Grey shading spans the bootstrap median to the 95th percentile on the left and the 5th percentile to the median on the right, comparing the mean of ten bootstrap networks with one further bootstrap network. Dates denote refits (YY-MM).

Table 8 FRM index on 336-hour sub-windows (means over 16 test refits)

Panels	**Mean FRM** $\times 10^4$	**Ratio to bootstrap mean**	**Within bootstrap 95% interval (share)**
Observed (consecutive weeks)	8.366	1.055	0.750
Bootstrap reference (MBB-168)	7.942	1.000	
Stationary bootstrap	8.249	1.039	1.000
GJR-GARCH-t + t-copula	7.604	0.959	0.938
Diffusion (DDPM)	7.387	0.930	0.938

Note: FRM = mean GACV-selected penalty over assets and 20 sub-windows per panel. The bootstrap reference shares the generated panels' structure (independent 168-hour paths); the observed index pairs consecutive weeks and is shown for completeness.

DDPM separates the layers in the opposite direction. It falls outside the joint conventional-diagnostic criterion in every panel, yet 120 of its 160 panels meet the single-panel network criterion. Benchmark portfolio risk does not stand in for the network either: DDPM matches benchmark ES more closely than GARCH-t but reproduces the observed ranking less often under the main convention.

To check that the network comparison responds to dependence rather than just marginal features, we use a circular-shift negative control (Table 10). It shows that the network layer responds to cross-asset structure. Shifting each asset's series by a different multiple of 672 hours desynchronises the assets while leaving marginal quantiles and autocorrelations nearly unchanged (quantile discrepancy 0.12, ACF discrepancy 0.009). The network changes sharply, with a relative Frobenius error of 0.97, a Spearman correlation of 0.14 and a top-three overlap of 0.88. The correlations and the overall distributional discrepancy also detect the change.

Table 9 Cross-classification of generated panels by conventional diagnostics and network fidelity

Generator	Conventional diagnostics	Network	Panels (of 10 × refits)	Distributional discrepancy within band among these
Stationary bootstrap	within band	outside band	0	0
Stationary bootstrap	outside band	within band	4	4
Stationary bootstrap	within band	within band	156	152
Stationary bootstrap	outside band	outside band	0	0
GJR-GARCH-t + t-copula	within band	outside band	0	0
GJR-GARCH-t + t-copula	outside band	within band	3	3
GJR-GARCH-t + t-copula	within band	within band	0	0
GJR-GARCH-t + t-copula	outside band	outside band	157	96
Diffusion (DDPM)	within band	outside band	0	0
Diffusion (DDPM)	outside band	within band	120	73
Diffusion (DDPM)	within band	within band	0	0
Diffusion (DDPM)	outside band	outside band	40	31

Note: Conventional diagnostics within band: all six conventional discrepancies inside their bootstrap bands (joint conventional-diagnostic criterion). Network within band: Frobenius error and Spearman correlation inside their single-panel bands (single-panel network criterion). Last column: number of these panels whose overall distributional discrepancy is also within its band.

Table 10 Circular-shift negative control: each asset's series shifted by a different multiple of 672 hours (16 test refits)

Refits	Rel. Frobenius error	Spearman	Top-3 overlap	Quantile discrepancy	ACF discrepancy	ACF^2 discrepancy	Correlation discrepancy	Distrib. distance
16	0.971	0.141	0.875	0.12	0.009	0.022	0.665	0.156

Note: Network metrics compare the shifted-data network with the observed network; diagnostics compare shifted and unshifted observed path panels. For comparison, bootstrap-band means are about 0.09 (quantiles), 0.01 (ACF), 0.02 (ACF of squares), 0.02 (correlations) and 0.08 (overall distributional discrepancy).

5.5 Layer 3: consequences under the prespecified allocation rule

To examine the economic consequences of these differences, the allocation rule converts the top-three ranking into portfolio weights, so differences between networks become differences in exposure. In this panel the most central assets, ETH and BTC, are also the least volatile (Table 1). Downweighting them shifts capital towards more volatile altcoins, and a network that ranks them lower leaves the portfolio less volatile. The resulting changes in decisions are visible in both asset selection and turnover. Relative to the historical-network rule (HN), the top-three set used for allocation is identical in 11 of 16 quarters for the stationary bootstrap, in 5 for GARCH-t and in 2 for DDPM, and average turnover rises from 0.20 per rebalance for HN to 0.24 for DDPM. HN halves BTC in 13 quarters, the GARCH-t network in 4 and the DDPM network in 1.

Table 11 reports realised daily outcomes from January 2022 to December 2025 (1,461 days). HN has a daily 95% ES of 9.08%, higher than equal weights (8.91%; paired difference −0.17 pp, 95% interval [−0.30, −0.03]) and the historical-volatility rule (8.67%; −0.41 pp [−0.60, −0.23]). The stationary-bootstrap network reproduces the HN outcome (−0.01 pp [−0.10, 0.08]). The GARCH-t and DDPM networks give lower realised ES than HN, with differences of −0.24 pp [−0.38, −0.11] and −0.11 pp [−0.25, 0.02], in line with their lower volatility exposure. The infeasible rule based on the following quarter's realised network gives 9.02% (−0.06 pp [−0.19, 0.09]), and its paired interval includes both lower and higher tail

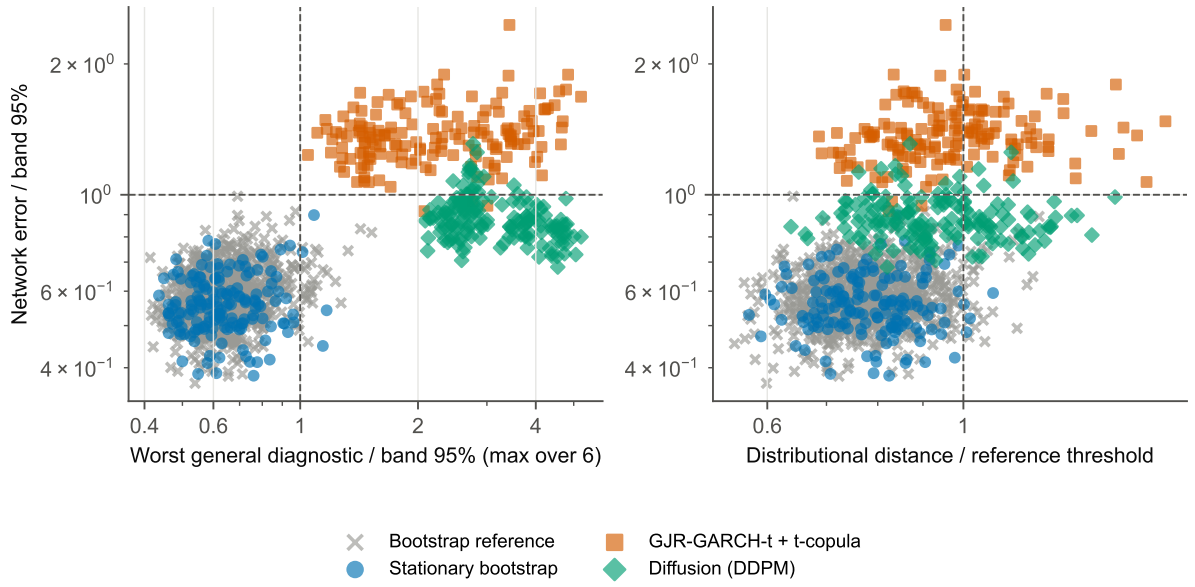


Fig. 4 Conventional diagnostics and network discrepancy for individual panels. Circles: stationary bootstrap; squares: GARCH-t; diamonds: denoising diffusion probabilistic model (DDPM); grey crosses: moving-block bootstrap reference. Each generator contributes ten panels at each of 16 refits. The overall distributional discrepancy is measured by energy distance. The vertical axis is the relative Frobenius network error divided by its bootstrap 95th percentile. The horizontal axis shows the maximum of the six similarly normalised conventional discrepancies (left) and normalised 24-hour-block overall distributional discrepancy (right). Both axes use logarithmic scales. Dashed lines at 1 mark the respective thresholds; larger values exceed them. The vertical axis represents the coefficient-error component of the network criterion; the full criterion also uses node-score Spearman correlation.

loss than HN.

Table 11 Realised daily outcomes of the allocation rules, January 2022 to December 2025 (1,461 days, no transaction costs)

Strategy	ES95 daily (%)	VaR95 daily (%)	Mean daily return (%)	Ann. volatility (%)	Max drawdown (%)	Avg turnover	ΔES vs HN (pp)	95% interval
Equal weight	8.915	5.991	-0.026	67.138	72.052	0.165	-0.165	[-0.297, -0.033]
Historical volatility	8.666	5.738	-0.021	64.810	70.824	0.187	-0.413	[-0.595, -0.230]
Historical network	9.080	5.845	-0.026	68.205	72.282	0.201	0.000	
Synthetic network: stationary bootstrap	9.065	5.972	-0.028	68.205	72.178	0.207	-0.015	[-0.104, 0.081]
Synthetic network: GARCH-t	8.837	5.845	-0.018	66.632	71.604	0.204	-0.243	[-0.378, -0.106]
Synthetic network: DDPM	8.974	5.845	-0.022	67.592	72.634	0.237	-0.106	[-0.245, 0.022]
Future network (ex post, infeasible)	9.019	5.864	-0.008	68.135	72.316	0.409	-0.061	[-0.190, 0.086]

Note: Daily losses are negative log changes of portfolio value at 00:00 UTC. ΔES: paired difference in 95% ES relative to the historical-network rule (HN), percentage points; intervals from 2,000 paired stationary-bootstrap draws of days (mean block 20 days). The future-network rule uses the following quarter's realised network and is infeasible (ex post diagnostic).

Ex ante risk matching assesses how much of this comparison is explained by differences in volatility exposure. After each portfolio is scaled ex ante to HN's trailing 90-day volatility, the synthetic-minus-historical point estimates are −0.02 pp for the stationary bootstrap, −0.09 pp for GARCH-t and −0.09 pp for DDPM, and every paired 95% interval contains zero: [−0.11, 0.08], [−0.21, 0.04] and [−0.21, 0.03]

(Table 12). Relative to HN's ES of 9.08%, the interval endpoints span approximately 2.3% lower to 0.9% higher tail loss. The network discrepancies thus produce observable changes in allocation weights, whereas the realised-loss comparison, once volatility exposure is matched, cannot distinguish effects within this range from zero. The intervals give the range of effect sizes compatible with this rule and sample.

The implementation checks give a similar picture: the unadjusted differences are essentially unchanged with 25-bp transaction costs and with a 24-hour execution lag. Under the exit rule for migrated tokens the point estimates are slightly more negative, and the DDPM interval just excludes zero ($[-0.30, -0.01]$). Across the ten single-panel networks of each generator, realised ES varies over less than 0.2 pp (Table 13), which shows how much the outcome depends on the particular panel drawn from a fitted generator.

Table 12 Differences in realised 95% ES relative to the historical-network rule under alternative implementations

Variant	Strategy	ΔES vs HN (pp)	Lower	Upper
Main	Equal weight	-0.165	-0.297	-0.033
Main	Historical volatility	-0.413	-0.595	-0.230
Main	Synthetic network: stationary bootstrap	-0.015	-0.104	0.081
Main	Synthetic network: GARCH-t	-0.243	-0.378	-0.106
Main	Synthetic network: DDPM	-0.106	-0.245	0.022
25 bp costs	Equal weight	-0.165	-0.297	-0.038
25 bp costs	Historical volatility	-0.413	-0.595	-0.225
25 bp costs	Synthetic network: stationary bootstrap	-0.015	-0.104	0.083
25 bp costs	Synthetic network: GARCH-t	-0.243	-0.372	-0.105
25 bp costs	Synthetic network: DDPM	-0.106	-0.246	0.022
24-h lag	Equal weight	-0.164	-0.291	-0.038
24-h lag	Historical volatility	-0.412	-0.592	-0.222
24-h lag	Synthetic network: stationary bootstrap	-0.014	-0.101	0.083
24-h lag	Synthetic network: GARCH-t	-0.241	-0.370	-0.104
24-h lag	Synthetic network: DDPM	-0.108	-0.255	0.024
Risk-matched	Equal weight	-0.037	-0.172	0.092
Risk-matched	Historical volatility	0.016	-0.168	0.197
Risk-matched	Synthetic network: stationary bootstrap	-0.021	-0.109	0.082
Risk-matched	Synthetic network: GARCH-t	-0.090	-0.206	0.041
Risk-matched	Synthetic network: DDPM	-0.085	-0.208	0.031
Exit rule	Equal weight	-0.199	-0.330	-0.070
Exit rule	Historical volatility	-0.489	-0.688	-0.279
Exit rule	Synthetic network: stationary bootstrap	-0.054	-0.175	0.053
Exit rule	Synthetic network: GARCH-t	-0.297	-0.422	-0.161
Exit rule	Synthetic network: DDPM	-0.148	-0.295	-0.010

Note: Main: one-hour execution lag, no costs; 25 bp costs: 25 bp per unit of one-way turnover; 24-h lag: 24-hour execution lag; Risk-matched: weights scaled ex ante to the HN portfolio's trailing 90-day volatility (zero-return cash or borrowing); Exit rule: migrated tokens liquidated at the last predecessor close. Lower/Upper: paired bootstrap 95% interval (percentage points).

Table 13 Realised 95% ES of allocations based on single generated panels (10 panels per generator)

Generator	Single-panel ES95 min (%)	median	max
Stationary bootstrap	8.956	9.030	9.115
GJR-GARCH-t + t-copula	8.858	8.898	8.990
Diffusion (DDPM)	8.943	8.985	9.037

Note: Simulation variability conditional on the fitted generator, seed 1.

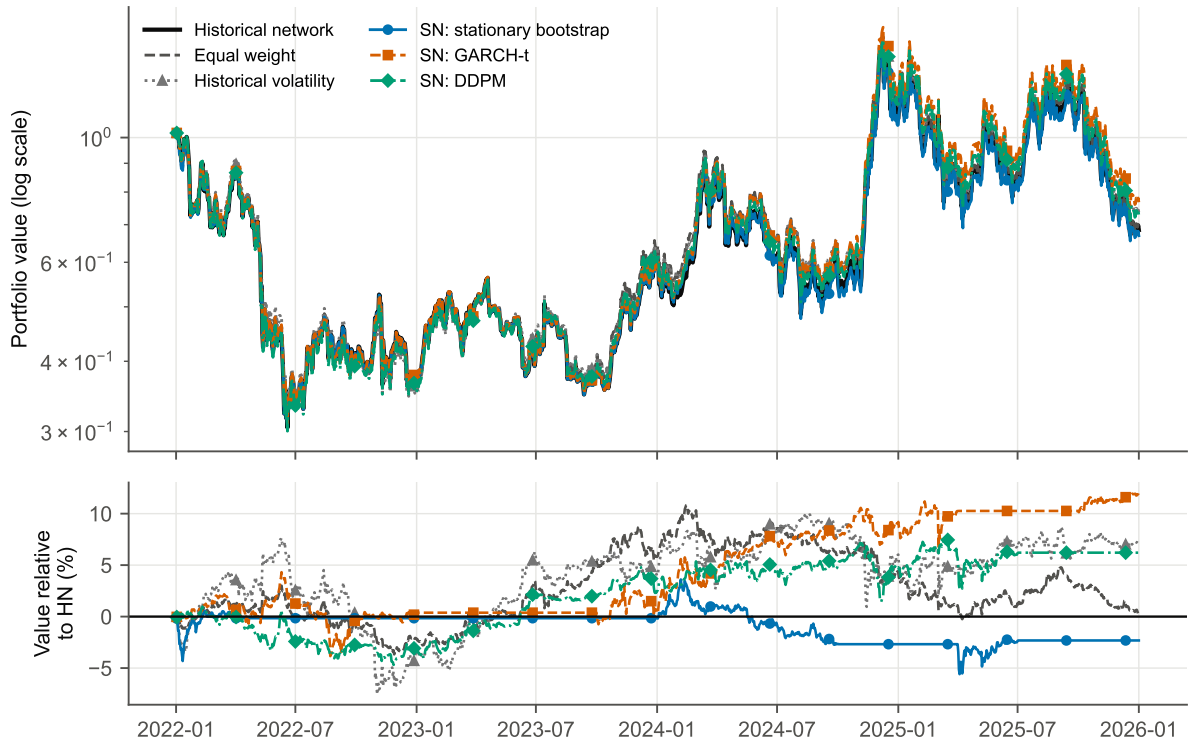


Fig. 5 Allocation paths, January 2022–December 2025, without transaction costs. Top: portfolio value on a logarithmic scale. Bottom: $100(V_t/V_t^{\mathrm{HN}} - 1)$, the percentage difference in value relative to the historical-network rule (HN); the horizontal zero line indicates equal value. HN is a solid black line; equal weight is dashed grey; historical volatility is dotted with triangle markers. Synthetic-network (SN) rules use circles with solid lines (stationary bootstrap), squares with dashed lines (GARCH-t), and diamonds with dash-dot lines (DDPM). Markers are spaced along the curves for identification. DDPM denotes the denoising diffusion probabilistic model.

The remaining question is whether these outcomes differ across market conditions. The prespecified comparison between volatility states has no high-volatility observations. No evaluation quarter reaches the threshold of 121.3%: quarterly volatility ranges from 44.7% (2023Q3) to 101.1% (2022Q2). A check based on daily returns gives the same classification (threshold 114.6%, test maximum 100.6%). In the five prespecified event windows (Table S1 in Online Resource 3), cumulative returns of the implementable rules differ by at most 2.1 percentage points, in the Terra/LUNA window. For example, cumulative returns are −31.4% for HN and −30.7% for GARCH-t during Terra/LUNA, and −13.2% and −12.7% during the October 2025 liquidation. Windows of a few days describe exposure to events but do not support estimates of tail averages.

6 Controlled simulations and robustness

The empirical results show clear changes in rankings and exposures, but their risk-matched loss effects remain imprecisely estimated. Controlled simulations help identify when ranking errors impose a cost; the subsequent checks examine how the empirical findings depend on training, network specification and the data used.

6.1 Simulations with a known network functional

The controlled simulations ask how the dependence structure and changes of regime affect network recovery and allocation cost when the true network functional is known. The reference functional is computed separately in each regime from an independent 500,000-hour sample. Table 14 and Fig. 6 summarise the results. The conventional generators are evaluated on 50 replications at each training-sample size; DDPM contributes four exploratory replications per design at 17,520 hours.

In design A, where the tail hubs differ in both correlation and downside dependence, the historical network estimated from 17,520 training hours of a stable process recovers the true top-three set in 98% of replications, and so does the aggregated stationary-bootstrap network. The aggregated GARCH-t network does so in 58%. Relative to the allocation chosen with the true network, population daily ES rises by 0.04% for the historical network and by 0.78% for GARCH-t.

Design B isolates asymmetric downside dependence: the hubs have the same correlation with the other assets as the non-hubs and differ mainly in downside co-movement. The historical network still recovers the true top-three set in 84% of replications and the stationary bootstrap in 58%. The GARCH-t network reverses the ranking. Its Spearman correlation with the true node scores is -0.55, and it never recovers the hubs. Its Student-t copula imposes the same dependence in the upper and lower tails, whereas in this design the hubs are distinguished mainly in the lower tail. The resulting allocations have a population ES 3.1% above the true-network benchmark and 1.8% above equal weights. A generator that matches correlations and marginal tails can therefore misidentify tail hubs systematically when tail dependence is asymmetric, and the error carries a decision cost. Halving the training sample to 8,760 hours lowers recovery rates (in design B, to 56% for the historical network and 34% for the bootstrap) but leaves the qualitative ranking of generators unchanged.

In its four exploratory replications per design, the aggregated DDPM network recovers the true top-three set in all 4 replications of design A, as the historical network and the stationary bootstrap do on the same replications; GARCH-t does so in 2. In design B it recovers the hubs in 2 of 4 replications (mean top-three overlap 2.50; Spearman correlation with the true node scores 0.56), against 4 of 4 for the historical network and 0 of 4 for GARCH-t (Spearman -0.67), and its allocations have a population ES 0.37% above the true-network benchmark, against 3.42% for GARCH-t. These four replications illustrate DDPM behaviour under the two dependence designs; comparisons among the conventional generators use 50 replications.

Changes of regime after training separate fidelity to the training network from transfer to a new dependence structure. Strengthening the common downside shocks while keeping the hubs fixed (regimes R1m, R1s and T1s) preserves node identification and reduces decision losses. Changing the identity of the hubs (regimes R2 and T2) raises the population ES of every history-based network allocation, including the historical-network rule, to 2.9–8.3% above the regime-specific true-network benchmark and above equal weights. These are failures to transfer to a changed market rather than generator errors: a generator that reproduced its training data perfectly would share them.

6.2 Training sensitivity

Returning to the empirical results, we first ask whether DDPM's network discrepancies depend on a particular training seed. Table 15 compares three independently initialised DDPM training runs at the four

Table 14 Controlled simulations: network recovery and population decision loss

Design	Training hours	Regime	Source	Reps	Spearman	Top-3 = 3 (share)	ES excess (%)
A	8760	R0	Historical network	50	0.61	1.00	0.00
A	8760	R0	Stationary bootstrap	50	0.63	1.00	0.00
A	8760	R0	GARCH-t	50	0.50	0.38	1.30
A	8760	R1s	Historical network	50	0.60	1.00	0.00
A	8760	R1s	Stationary bootstrap	50	0.62	1.00	0.00
A	8760	R1s	GARCH-t	50	0.49	0.38	0.60
A	8760	R2	Historical network	50	-0.27	0.00	8.35
A	8760	R2	Stationary bootstrap	50	-0.24	0.00	8.35
A	8760	R2	GARCH-t	50	-0.14	0.00	6.98
A	17520	R0	Historical network	50	0.62	0.98	0.04
A	17520	R0	Stationary bootstrap	50	0.63	0.98	0.04
A	17520	R0	GARCH-t	50	0.55	0.58	0.78
A	17520	R0	DDPM	4	0.68	1.00	0.00
A	17520	R1s	Historical network	50	0.63	0.98	0.02
A	17520	R1s	Stationary bootstrap	50	0.64	0.98	0.02
A	17520	R1s	GARCH-t	50	0.54	0.58	0.39
A	17520	R1s	DDPM	4	0.72	1.00	0.00
A	17520	R2	Historical network	50	-0.21	0.00	8.34
A	17520	R2	Stationary bootstrap	50	-0.18	0.00	8.34
A	17520	R2	GARCH-t	50	-0.26	0.00	7.78
A	17520	R2	DDPM	4	-0.22	0.00	8.35
B	8760	T0	Historical network	50	0.58	0.56	0.43
B	8760	T0	Stationary bootstrap	50	0.48	0.34	0.77
B	8760	T0	GARCH-t	50	-0.58	0.00	3.10
B	8760	T1s	Historical network	50	0.58	0.56	0.16
B	8760	T1s	Stationary bootstrap	50	0.48	0.34	0.30
B	8760	T1s	GARCH-t	50	-0.60	0.00	1.06
B	8760	T2	Historical network	50	-0.05	0.00	5.59
B	8760	T2	Stationary bootstrap	50	-0.05	0.00	5.21
B	8760	T2	GARCH-t	50	0.06	0.00	2.91
B	17520	T0	Historical network	50	0.61	0.84	0.12
B	17520	T0	Stationary bootstrap	50	0.57	0.58	0.43
B	17520	T0	GARCH-t	50	-0.55	0.00	3.08
B	17520	T0	DDPM	4	0.56	0.50	0.37
B	17520	T1s	Historical network	50	0.61	0.84	0.02
B	17520	T1s	Stationary bootstrap	50	0.57	0.58	0.15
B	17520	T1s	GARCH-t	50	-0.54	0.00	1.11
B	17520	T1s	DDPM	4	0.62	0.50	0.12
B	17520	T2	Historical network	50	-0.12	0.00	6.01
B	17520	T2	Stationary bootstrap	50	-0.09	0.00	5.69
B	17520	T2	GARCH-t	50	0.02	0.00	3.37
B	17520	T2	DDPM	4	-0.29	0.00	5.92

Note: Network functional: unpenalised 5% quantile-regression network on an independent 500,000-hour sample of each regime. Top-3 = 3: share of replications recovering the true top-three set. ES excess: population daily 95% ES of the allocation relative to the allocation based on the true network (percent). Regimes: R0/T0 training regime; R1m, R1s, T1s stronger common downside shocks; R2, T2 change of hub identity. Reps: replications; DDPM was run for 4 replications per design at 17,520 hours (exploratory).

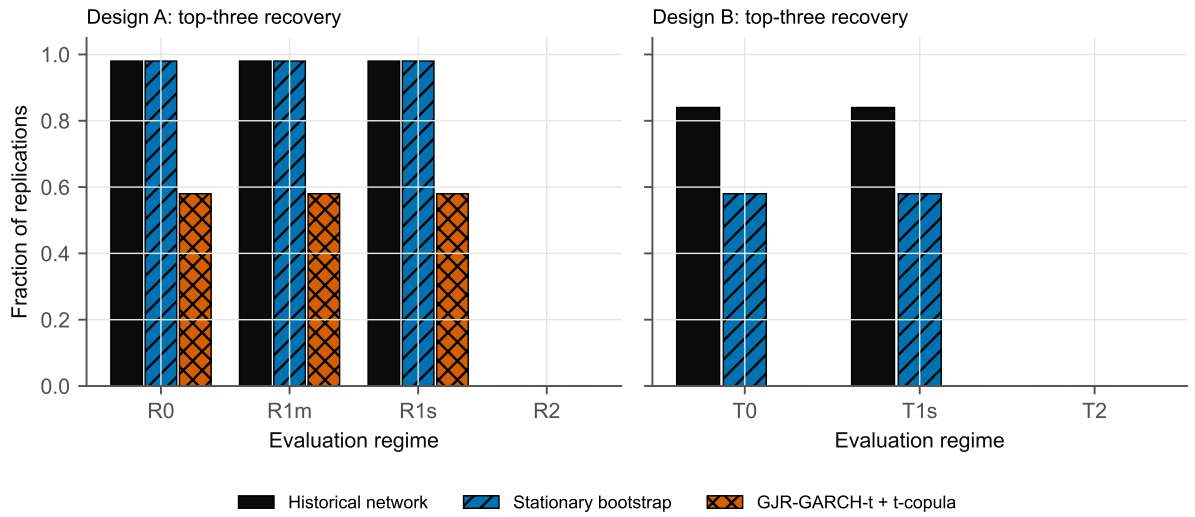


Fig. 6 Controlled simulations: fraction of 50 replications recovering the true top-three set with $n = 17{,}520$ training hours. Design A has hubs distinguished by correlation and downside dependence; design B has hubs distinguished mainly by asymmetric downside dependence. Solid bars: historical network; diagonal hatching: stationary bootstrap; cross-hatching: GARCH-t. Generated networks average five panels. R0/T0 denote the training regime; R1m and R1s denote moderately and strongly increased downside shocks, respectively, and T1s denotes strongly increased downside shocks; R2/T2 change the hub identity. DDPM results (four replications) appear in Table 14.

first-quarter test refits. The relative Frobenius error against the observed network averages 0.27, with a mean within-refit range of 0.03 across seeds. Networks from two seeds have a Spearman correlation of 0.90, and their top-three sets are identical in 75% of seed pairs. DDPM's departures from the observed network therefore recur across independently initialised fits at these refits. Section 5.3 compares these departures with the sampling uncertainty of the reference.

Table 15 Training-seed sensitivity of DDPM (seeds 1–3, four first-quarter test refits)

Generator	Refits × seeds	Rel. Frobenius error vs observed: mean	range across seeds (mean over refits)	Spearman vs observed: mean	Top-3 overlap with HN: mean	Seed pairs: Spearman	Seed pairs: top-3 overlap	Seed pairs with identical top-3 (share)
Diffusion (DDPM)	12	0.27	0.03	0.688	1.583	0.903	2.667	0.75

Note: Aggregated networks (10 panels per seed). Range across seeds: max minus min of the relative Frobenius error over the three seeds, averaged over refits. Seed pairs compare aggregated networks of two seeds at the same refit.

6.3 Alternative specifications

We next vary the tail probability and sampling frequency. Table 16 repeats the network comparison for $\tau = 0.10$, $\tau = 0.01$ and 4-hour returns, each against its own bootstrap band, using 5 generated panels per generator and 20 bootstrap panels. At $\tau = 0.10$ and with 4-hour returns the pattern of the main specification holds. The stationary bootstrap stays within the band at every refit, GARCH-t lies outside it at every refit (Frobenius errors 0.28 and 0.41), and DDPM lies within it in 19% and 44% of refits. At $\tau = 0.01$ the reference variability increases: the bootstrap Frobenius discrepancy is 0.34, against 0.15 at $\tau = 0.10$. GARCH-t then lies within the band in 31% of refits and DDPM in 81%, although their own Frobenius errors are larger than at $\tau = 0.10$ (Table 16). More frequent band membership in the extreme

tail thus reflects a wider band rather than closer preservation.

The allocation results do not depend on the choice of node score or on the penalty. Ranking nodes by absolute rather than positive out-strength, or using the unpenalised historical network, leaves the historical-network ES at 9.08%, and absolute out-strength changes the synthetic-network ES by at most 0.04 pp.

Table 16 Network preservation under alternative specifications (16 test refits)

Specification	Network	Refits	Rel. Frobenius error	Spearman	Top-3 overlap	Frobenius within band (share)
$\tau = 0.10$	Bootstrap band (mean of 5 vs one)	16	0.153	0.891	2.552	
$\tau = 0.10$	Stationary bootstrap (aggregated, 5 panels)	16	0.065	0.964	2.750	1.000
$\tau = 0.10$	GJR-GARCH-t + t-copula (aggregated, 5 panels)	16	0.280	0.709	2.125	0.000
$\tau = 0.10$	Diffusion (DDPM) (aggregated, 5 panels)	16	0.231	0.792	1.875	0.188
$\tau = 0.01$	Bootstrap band (mean of 5 vs one)	16	0.340	0.741	2.025	
$\tau = 0.01$	Stationary bootstrap (aggregated, 5 panels)	16	0.168	0.910	2.500	1.000
$\tau = 0.01$	GJR-GARCH-t + t-copula (aggregated, 5 panels)	16	0.471	0.509	1.812	0.312
$\tau = 0.01$	Diffusion (DDPM) (aggregated, 5 panels)	16	0.391	0.439	1.438	0.812
4-hour returns ($\tau = 0.05$)	Bootstrap band (mean of 5 vs one)	16	0.299	0.781	2.170	
4-hour returns ($\tau = 0.05$)	Stationary bootstrap (aggregated, 5 panels)	16	0.198	0.886	2.688	1.000
4-hour returns ($\tau = 0.05$)	GJR-GARCH-t + t-copula (aggregated, 5 panels)	16	0.412	0.570	1.750	0.000
4-hour returns ($\tau = 0.05$)	Diffusion (DDPM) (aggregated, 5 panels)	16	0.359	0.575	1.750	0.438

Note: Aggregated networks from the first 5 generated panels per generator (seed 1); bands from 20 moving-block bootstrap panels (mean of 5 vs one). Frobenius within band: share of refits inside the 95% band.

6.4 Exchange, quotation and asset panel

The final checks address the data source and asset composition. Kraken US-dollar prices give reference networks very similar to those from Binance USDT prices on the eight assets available on both exchanges throughout the sample. Over the 16 test refits, the Spearman correlation of node scores between exchanges ranges from 0.88 to 1.00; the top-three sets share at least two assets at every refit and are identical in 11. At the two validation refits the correlations are 0.93 and 0.86, with two common assets. Re-evaluating all allocation decisions on Kraken prices, with weights restricted and renormalised to the eight common assets, changes realised daily ES by at most 0.02 percentage points and leaves the ordering of strategies unchanged (Online Resource 3). For these eight assets, the choice of exchange and quotation currency has little effect on the reference networks or on the allocation results.

To examine asset composition separately, the development panel replaces DOGE, MATIC and EOS

with TRX, XLM and ETC, and the conventional generators were rerun on it (Table 17). Its network results match those of the primary panel. The stationary-bootstrap network lies within the reference band at every refit (Frobenius error 0.06; top-three overlap 2.75), and the GARCH-t network lies outside it at every refit (0.29; Spearman 0.64; overlap 1.94). GARCH-t repeatedly promotes XLM, which enters its top-three set in 11 of 16 quarters but never enters the observed network's. The allocation results also repeat. The realised ES of the GARCH-t-network rule is again lower than that of the historical-network rule (−0.15 pp, 95% interval [−0.26, −0.02]), and after ex ante risk matching the difference is +0.02 pp, with interval [−0.11, 0.13].

Table 17 Development panel (BTC, ETH, BNB, XRP, LTC, ADA, TRX, LINK, XLM, ETC): network preservation of the conventional generators (16 test refits)

Network	Refits	Rel. Frobenius error	Spearman	Top-3 overlap	Frobenius within band (share)
Stationary bootstrap (aggregated)	16	0.064	0.942	2.750	1.0
GJR-GARCH-t + t-copula (aggregated)	16	0.290	0.641	1.938	0.0
Bootstrap band (mean of 10 vs one; B = 30)	16	0.178	0.821	2.318	

Note: Same estimator and protocol as the primary panel; the reference band uses 30 moving-block bootstrap networks per refit, against 100 for the primary panel.

7 Discussion

Across the empirical and simulation comparisons, preserving a network and improving the allocation that uses it emerge as related but distinct objectives. This distinction explains both the disagreements among the evaluation layers and the conditions under which they matter economically.

7.1 What the three layers establish

Conventional diagnostics, network preservation and realised allocation outcomes answer different questions about generated scenarios, and in this study they lead to different conclusions about the same generators. All three layers are computed on the same data, estimator and decision rule, so the disagreements reflect what each layer measures rather than differences in design.

The first disagreement concerns what is measured by distributional summaries and network metrics. The GARCH-t network departs from the observed network by more than the reference uncertainty at every refit, yet the overall distributional discrepancy stays within its band in 96 of the 157 panels that fall outside both the conventional-diagnostic and the network bands. This is consistent with the difference in target: the overall distributional discrepancy on 24-hour blocks weighs the whole joint distribution, whereas the network depends on how assets co-move in the lower conditional tail. The FRM index shows the same limitation for network summaries. It lies within its reference interval in 94–100% of refits for all three generators, although their coefficient networks and top-three sets differ. A summary of overall discrepancy or of penalty intensity does not identify the central nodes.

DDPM reveals a further distinction between coefficient agreement and ranking agreement. No DDPM panel meets the joint conventional-diagnostic criterion, but 120 of 160 meet the single-panel network

criterion, and the aggregated DDPM network lies within the Frobenius band in 7 of 16 refits. Relative to GARCH-t, DDPM departs less in its coefficients and more in its ranking. Its top-three sets are ones that resampled histories also produce, in 14 of 16 refits, but they coincide with the observed set less often than those of GARCH-t, and the seed comparisons show that this pattern recurs across independent training runs. The allocation rule uses only the top-three set, so for this decision the ranking-level departures are the relevant ones.

The reference analysis sets the scale for these comparisons. The observed top-three set appears in only 45% of bootstrap samples, and an aggregated bootstrap network shares on average 2.39 of the observed top-three assets. Because the networks are nearly dense, connection strengths and rankings carry the structural information, and both depend on conventions such as coefficient units. Judging a generator against a single observed estimate, or under a convention different from that of the observed network, would misstate how far it departs from the data.

7.2 From network rankings to allocation outcomes

These ranking differences matter through the exposures chosen by the allocation rule. Under the fixed rule, downweighting the most central assets often reduces exposure to BTC and ETH, the least volatile assets in the panel, and raises exposure to more volatile ones. Equal weights and the historical-volatility rule, which do not shift weight in this way, have lower daily ES (8.91% and 8.67%) than the historical-network rule (9.08%). Closer reproduction of the observed network and lower realised tail loss therefore need not coincide under this rule.

Ex ante risk matching separates volatility exposure from the network input. After it, the paired 95% intervals for the synthetic-minus-historical ES difference are $[-0.11, 0.08]$, $[-0.21, 0.04]$ and $[-0.21, 0.03]$ percentage points for the stationary bootstrap, GARCH-t and DDPM. The sample resolves the changes in allocation weights but leaves a range of positive and negative tail-loss differences compatible with the data. These estimates quantify the uncertainty in the decision effect under the specified allocation rule and evaluation period.

The simulations show when a ranking error becomes costly. Where hubs are distinguished by asymmetric downside dependence, the copula model ranks them in reverse, with a node-score correlation of -0.55, and its allocations carry a population ES 3.1% above the true-network benchmark and 1.8% above equal weights. The empirical sample did not reveal a detectable cost of this kind after risk matching, while the simulation identifies asymmetric downside dependence as a setting in which ranking errors raise tail loss.

Fidelity to the historical network and transfer to a new one are separate requirements. Empirically, the network of the following quarter is further from the 24-month network than any generated network is, although that comparison also reflects the shorter quarterly window. In the simulations, a change of hub identity raises the ES of every history-based allocation to 2.9–8.3% above the regime-specific true-network benchmark. A generator trained on history inherits this limitation however faithfully it reproduces its training data.

7.3 What the complementary comparisons reveal

Taken together, the comparisons identify where disagreements arise between diagnostics, network estimates and allocation outcomes. Changing coefficient units changes the observed top-three set at ten of the sixteen refits, while resampling reveals uncertainty even under a fixed convention. Across generated panels and training seeds, the discrepancies persist. Their allocation effects then depend on which assets the ranking downweights and on the resulting volatility exposure.

The additional analyses distinguish several sources of this disagreement. The circular-shift control shows that the network metrics respond to cross-asset structure, the cross-exchange and development-panel checks show how the results depend on the empirical setting, and risk matching shows how portfolio exposure affects the reading of allocation outcomes. Formal distributional tests remain complementary: they address equality of joint laws, whereas the evaluation proposed here concerns one estimated functional and its use.

7.4 Scope and interpretation

The extent to which these mechanisms generalise depends on the market, estimator and generator. The empirical findings concern ten liquid cryptocurrencies over 2022–2025, with checks on Kraken US-dollar prices and on an alternative asset panel. In the 24-month hourly windows the conditional-quantile networks are nearly dense and the penalty is effectively inactive. Sparse high-dimensional networks, shorter windows or lower frequencies may behave differently, and the edges represent conditional associations, so questions of causal transmission require a different estimator and reference design. No test quarter reached the prespecified high-volatility threshold, so the evidence does not cover that state.

The neural evidence concerns one unconditional DDPM architecture and training configuration, and its simulation results rest on four replications per design, against 50 for the conventional generators. Conditional and copula-based diffusion models, such as those studied by Huk et al. (2026), may preserve the network differently. Tail-GAN is not part of the comparison for the reasons given in Section 4.3.

The in-sample reference measures fidelity to the fitting-window network. Joint resampling reuses that window directly, and the bootstrap reference shares the same empirical basis, which favours the stationary bootstrap by construction. The bands and ranking frequencies are descriptive benchmarks for non-smooth estimated functionals (Section 4.5). The distributional comparison also uses a descriptive reference band. The allocation intervals are conditional on the implemented fitting procedures, and the allocation rule is an evaluation device rather than an optimised strategy.

8 Conclusion

This study evaluates generated financial scenarios through three related outcomes: conventional diagnostics, preservation of the network relative to the sampling uncertainty of the empirical reference, and realised outcomes under a fixed allocation rule.

In hourly cryptocurrency data these kinds of evidence diverge. A GARCH-t copula model distorts the network by more than the reference uncertainty at all 16 refits, while the overall distributional discrepancy stays within its band for most of the affected panels. A diffusion model meets the joint conventional-diagnostic criterion in no panel but meets the single-panel network criterion in 120 of 160, and, relative

to GARCH-t, it departs less in its coefficients and more in its ranking. The FRM index, an aggregate summary, does not distinguish the generators. The empirical reference is itself uncertain, nearly dense and sensitive to estimation conventions, and the bootstrap comparisons quantify the corresponding reference variability.

The economic consequence of these discrepancies depends on how the ranking affects exposure. After ex ante risk matching, the realised differences in daily expected shortfall relative to the historical-network rule range from -0.09 to -0.02 percentage points, with paired intervals that contain zero, leaving the sign of the risk-matched effect unresolved. The controlled simulations show where a cost arises: when tail hubs differ mainly in asymmetric downside dependence, a correlation-matching copula reverses the hub ranking and raises population expected shortfall by 3.1% relative to the true-network benchmark.

For the generators and cryptocurrency panel studied here, agreement on conventional diagnostics, network coefficients and allocation losses does not coincide. The clearest decision cost occurs in the simulation with asymmetric downside dependence, where the copula model reverses the tail-hub ranking.

These findings suggest several directions for future research. Comparisons could examine larger asset universes and other markets to establish whether the same discrepancies arise in sparse networks. Comparisons of state-conditioned generators across calm and stressed periods could assess whether conditioning improves the preservation of changing network rankings. Testing alternative allocation rules, with transaction costs and leverage limits, would clarify when improved network preservation translates into lower realised losses. These comparisons could use loss tolerances specified before evaluation to distinguish economically material differences from estimation noise.

Statements and declarations

Data availability. All data are publicly available from Binance Public Data and from Kraken's downloadable OHLCVT archive (see References). Online Resource 4 contains acquisition scripts, a file-level manifest with SHA-256 or CRC-32 checks, cleaning code and saved analytical outputs. Raw and processed exchange price data are not redistributed.

Code availability. Acquisition, estimation, simulation and evaluation code, seeds, software versions and the time-stamped analysis protocol are supplied in the anonymised replication package (Online Resource 4).

Funding, competing interests, ethics, author contributions and the statement on the use of generative AI are provided on the separate title page, in line with the journal's double-blind procedure.

Online Resources. Online Resource 1: data manifest and cleaning report. Online Resource 2: analysis protocol with change log and freeze record. Online Resource 3: supplementary tables (event windows, allocation checks on Kraken prices and on the development panel), the Tail-GAN implementation-audit record, and computational details for diagnostics, portfolio accounting, simulations and DDPM optimisation. Online Resource 4: anonymised replication package.

References

Adrian, T. and M. K. Brunnermeier (2016). "CoVaR". In: *American Economic Review* 106.7, pp. 1705–1741. DOI: 10.1257/aer.20120555.

Aghapour, A., E. Bayraktar, and F. Yuan (2025). *Dynamic data generation and dynamic portfolio selection: An application of a score-based diffusion model*. [Preprint]. arXiv. DOI: 10.48550/arXiv.2507.09916.

Ando, T., M. Greenwood-Nimmo, and Y. Shin (2022). "Quantile connectedness: Modeling tail behavior in the topology of financial networks". In: *Management Science* 68.4, pp. 2401–2431. DOI: 10.1287/mnsc.2021.3984.

Asadi, S. and J. Y.-M. Li (2026). *Generative adversarial regression (GAR): Learning conditional risk scenarios*. [Preprint]. arXiv. DOI: 10.48550/arXiv.2603.08553.

Assefa, S. A., D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, and M. Veloso (2020). "Generating synthetic data in finance". In: *Proceedings of the First ACM International Conference on AI in Finance (ICAIF '20)*, pp. 1–8. DOI: 10.1145/3383455.3422554.

Belloni, A. and V. Chernozhukov (2011). "ℓ1-penalized quantile regression in high-dimensional sparse models". In: *The Annals of Statistics* 39.1, pp. 82–130. DOI: 10.1214/10-AOS827.

Belloni, A., V. Chernozhukov, and K. Kato (2019). "Valid post-selection inference in high-dimensional approximately sparse quantile regression models". In: *Journal of the American Statistical Association* 114.526, pp. 749–758. DOI: 10.1080/01621459.2018.1442339.

Ben Amor, S., M. Althof, and W. K. Härdle (2022). "Financial risk meter for emerging markets". In: *Research in International Business and Finance* 60, p. 101594. DOI: 10.1016/j.ribaf.2021.101594.

Binance (n.d.). *Binance public data*. [Data set and documentation]. URL: https://data.binance.vision/ (visited on 10/14/2026). Documentation: https://github.com/binance/binance-public-data.

— (Aug. 28, 2024). *Binance will support the Polygon (MATIC) token swap to Polygon (POL)*. [Announcement]. URL: https://www.binance.com/en/support/announcement/detail/6a6de383727f4659a3050f7982e1620f.

— (May 14, 2025). *Binance will support the EOS (EOS) token swap and rebranding to Vaulta (A)*. [Announcement]. URL: https://www.binance.com/en/support/announcement/detail/1e89a9ca957c4b0ca7502e60b993e201.

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of Econometrics* 31.3, pp. 307–327. DOI: 10.1016/0304-4076(86)90063-1.

Bouri, E., T. Saeed, X. V. Vo, and D. Roubaud (2021). "Quantile connectedness in the cryptocurrency market". In: *Journal of International Financial Markets, Institutions and Money* 71, p. 101302. DOI: 10.1016/j.intfin.2021.101302.

Cetingoz, A. R. and C.-A. Lehalle (2025). *Synthetic data for portfolios: A throw of the dice will never abolish chance*. [Preprint]. arXiv. DOI: 10.48550/arXiv.2501.03993.

Chen, C. Y.-H., W. K. Härdle, and Y. Okhrin (2019). "Tail event driven networks of SIFIs". In: *Journal of Econometrics* 208.1, pp. 282–298. DOI: 10.1016/j.jeconom.2018.09.016.

Chen, M., R. Xu, Y. Xu, and R. Zhang (2025). *Diffusion factor models: Generating high-dimensional returns with factor structure*. [Preprint]. arXiv. DOI: 10.48550/arXiv.2504.06566.

Cont, R., M. Cucuringu, R. Xu, and C. Zhang (2026). "Tail-GAN: Learning to simulate tail risk scenarios". In: *Management Science* 72.4, pp. 2917–2936. DOI: 10.1287/mnsc.2023.00936.

Demarta, S. and A. J. McNeil (2005). "The t copula and related copulas". In: *International Statistical Review* 73.1, pp. 111–129. DOI: 10.1111/j.1751-5823.2005.tb00254.x.

DeMiguel, V., L. Garlappi, and R. Uppal (2009). "Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?" In: *The Review of Financial Studies* 22.5, pp. 1915–1953. DOI: 10.1093/rfs/hhm075.

Diebold, F. X. and K. Yılmaz (2014). "On the network topology of variance decompositions: Measuring the connectedness of financial firms". In: *Journal of Econometrics* 182.1, pp. 119–134. DOI: 10.1016/j.jeconom.2014.04.012.

Engle, R. F. and S. Manganelli (2004). "CAViaR: Conditional autoregressive value at risk by regression quantiles". In: *Journal of Business & Economic Statistics* 22.4, pp. 367–381. DOI: 10.1198/073500104000000370.

Esteban, C., S. L. Hyland, and G. Rätsch (2017). *Real-valued (medical) time series generation with recurrent conditional GANs*. [Preprint]. arXiv. DOI: 10.48550/arXiv.1706.02633.

Fissler, T. and J. F. Ziegel (2016). "Higher order elicitability and Osband's principle". In: *The Annals of Statistics* 44.4, pp. 1680–1707. DOI: 10.1214/16-AOS1439.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). "On the relation between the expected value and the volatility of the nominal excess return on stocks". In: *The Journal of Finance* 48.5, pp. 1779–1801. DOI: 10.1111/j.1540-6261.1993.tb05128.x.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). "Generative adversarial nets". In: *Advances in Neural Information Processing Systems*. Vol. 27. URL: https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). "A kernel two-sample test". In: *Journal of Machine Learning Research* 13.25, pp. 723–773. URL: https://www.jmlr.org/papers/v13/gretton12a.html.

Grover, A., E. Wang, A. Zweig, and S. Ermon (2019). "Stochastic optimization of sorting networks via continuous relaxations". In: *International Conference on Learning Representations (ICLR 2019)*. DOI: 10.48550/arXiv.1903.08850.

Härdle, W. K., W. Wang, and L. Yu (2016). "TENET: Tail-Event driven NETwork risk". In: *Journal of Econometrics* 192.2, pp. 499–513. DOI: 10.1016/j.jeconom.2016.02.013.

Ho, J., A. Jain, and P. Abbeel (2020). "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 6840–6851. URL: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Hounwanou, C. D. and Y. U. Gaba (2025). *Deep generative models for synthetic financial data: Applications to portfolio and risk modeling*. [Preprint]. arXiv. DOI: 10.48550/arXiv.2512.21798.

Huh, J., S. Jeong, H.-G. Kim, H. K. Koo, and B. H. Lim (2026). *MarketGANs: Multivariate financial time-series data augmentation using generative adversarial networks*. [Preprint]. arXiv. DOI: 10.48550/arXiv.2601.17773.

Huk, D., D. Wang, and M. Bresar (2026). *Probabilistic multivariate time series forecasting with diffusion copulas*. [Preprint]. arXiv. DOI: 10.48550/arXiv.2605.19685.

Ji, Q., E. Bouri, C. K. M. Lau, and D. Roubaud (2019). "Dynamic connectedness and integration in cryptocurrency markets". In: *International Review of Financial Analysis* 63, pp. 257–272. DOI: 10.1016/j.irfa.2018.12.002.

Koenker, R. and G. Bassett (1978). "Regression quantiles". In: *Econometrica* 46.1, pp. 33–50. DOI: 10.2307/1913643.

Koshiyama, A., N. Firoozye, and P. Treleaven (2021). "Generative adversarial networks for financial trading strategies fine-tuning and combination". In: *Quantitative Finance* 21.5, pp. 797–813. DOI: 10.1080/14697688.2020.1790635.

Kraken (n.d.). *Downloadable historical OHLCVT (open, high, low, close, volume, trades) data*. [Data set]. URL: https://support.kraken.com/articles/360047124832-downloadable-historical-ohlcvt-open-high-low-close-volume-trades-data (visited on 10/14/2026).

Künsch, H. R. (1989). "The jackknife and the bootstrap for general stationary observations". In: *The Annals of Statistics* 17.3, pp. 1217–1241. DOI: 10.1214/aos/1176347265.

Lee, E. R., H. Noh, and B. U. Park (2014). "Model selection via Bayesian information criterion for quantile regression models". In: *Journal of the American Statistical Association* 109.505, pp. 216–229. DOI: 10.1080/01621459.2013.836975.

Lesniewski, A. and G. Trigila (2024). *Beyond Monte Carlo: Harnessing diffusion models to simulate financial market dynamics*. [Preprint]. arXiv. DOI: 10.48550/arXiv.2412.00036.

Lhaut, S., H. Rootzén, and J. Segers (2025). *Simulation of multivariate extremes: A Wasserstein-Aitchison GAN approach*. [Preprint]. arXiv. DOI: 10.48550/arXiv.2504.21438.

Li, Y. and J. Zhu (2008). "L1-norm quantile regression". In: *Journal of Computational and Graphical Statistics* 17.1, pp. 163–185. DOI: 10.1198/106186008X289155.

Liao, S., H. Ni, M. Sabate-Vidales, L. Szpruch, M. Wiese, and B. Xiao (2024). "Sig-Wasserstein GANs for conditional time series generation". In: *Mathematical Finance* 34.2, pp. 622–670. DOI: 10.1111/mafi.12423.

Marti, G. (2020). "CorrGAN: Sampling realistic financial correlation matrices using generative adversarial networks". In: *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8459–8463. DOI: 10.1109/ICASSP40776.2020.9053276.

Mihoci, A., M. Althof, C. Y.-H. Chen, and W. K. Härdle (2020). "FRM financial risk meter". In: *The econometrics of networks*. Ed. by Á. de Paula, E. Tamer, and M.-C. Voia. Vol. 42. Advances in Econometrics. Emerald Publishing, pp. 335–368. DOI: 10.1108/S0731-905320200000042016.

Ni, H., L. Szpruch, M. Sabate-Vidales, B. Xiao, M. Wiese, and S. Liao (2021). "Sig-Wasserstein GANs for time series generation". In: *Proceedings of the Second ACM International Conference on AI in Finance (ICAIF '21)*, pp. 1–8. DOI: 10.1145/3490354.3494393.

Patton, A., D. N. Politis, and H. White (2009). "Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White". In: *Econometric Reviews* 28.4, pp. 372–375. DOI: 10.1080/07474930802459016.

Politis, D. N. and J. P. Romano (1994). "The stationary bootstrap". In: *Journal of the American Statistical Association* 89.428, pp. 1303–1313. DOI: 10.1080/01621459.1994.10476870.

Politis, D. N. and H. White (2004). "Automatic block-length selection for the dependent bootstrap". In: *Econometric Reviews* 23.1, pp. 53–70. DOI: 10.1081/ETC-120028836.

Portnoy, S. and R. Koenker (1997). "The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators". In: *Statistical Science* 12.4, pp. 279–300. DOI: 10.1214/ss/1030037960.

Rasul, K., C. Seward, I. Schuster, and R. Vollgraf (2021). "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting". In: *Proceedings of the 38th International Conference on Machine Learning (PMLR 139)*, pp. 8857–8868. URL: https://proceedings.mlr.press/v139/rasul21a.html.

Song, Y., J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole (2021). "Score-based generative modeling through stochastic differential equations". In: *International Conference on Learning Representations (ICLR 2021)*. DOI: 10.48550/arXiv.2011.13456.

Stenger, M., R. Leppich, I. Foster, S. Kounev, and A. Bauer (2024). "Evaluation is key: A survey on evaluation measures for synthetic time series". In: *Journal of Big Data* 11.1. DOI: 10.1186/s40537-024-00924-7.

Székely, G. J. and M. L. Rizzo (2013). "Energy statistics: A class of statistics based on distances". In: *Journal of Statistical Planning and Inference* 143.8, pp. 1249–1272. DOI: 10.1016/j.jspi.2013.03.018.

Takahashi, S., Y. Chen, and K. Tanaka-Ishii (2019). "Modeling financial time-series with generative adversarial networks". In: *Physica A: Statistical Mechanics and its Applications* 527, p. 121261. DOI: 10.1016/j.physa.2019.121261.

Takahashi, T. and T. Mizuno (2024). *Generation of synthetic financial time series by diffusion models*. [Preprint]. arXiv. DOI: 10.48550/arXiv.2410.18897.

Tashiro, Y., J. Song, Y. Song, and S. Ermon (2021). "CSDI: Conditional score-based diffusion models for probabilistic time series imputation". In: *Advances in Neural Information Processing Systems*. Vol. 34, pp. 24804–24816. URL: https://proceedings.neurips.cc/paper/2021/hash/cfe8504bda37b575c70ee1a8276f3486-Abstract.html.

Tepelyan, R. and A. Gopal (2023). "Generative machine learning for multivariate equity returns". In: *Proceedings of the 4th ACM International Conference on AI in Finance (ICAIF '23)*, pp. 159–166. DOI: 10.1145/3604237.3626884.

Wang, R., V. Potì, and W. K. Härdle (2024). "Assessing network risk with FRM: Links with pricing kernel volatility and application to cryptocurrencies". In: *Quantitative Finance* 24.7, pp. 975–992. DOI: 10.1080/14697688.2024.2370311.

White, H., T.-H. Kim, and S. Manganelli (2015). "VAR for VaR: Measuring tail dependence using multivariate regression quantiles". In: *Journal of Econometrics* 187.1, pp. 169–188. DOI: 10.1016/j.jeconom.2015.02.004.

Wiese, M., R. Knobloch, R. Korn, and P. Kretschmer (2020). "Quant GANs: Deep generation of financial time series". In: *Quantitative Finance* 20.9, pp. 1419–1440. DOI: 10.1080/14697688.2020.1730426.

Yi, S., Z. Xu, and G.-J. Wang (2018). "Volatility connectedness in the cryptocurrency market: Is Bitcoin a dominant cryptocurrency?" In: *International Review of Financial Analysis* 60, pp. 98–114. DOI: 10.1016/j.irfa.2018.08.012.

Yoon, J., D. Jarrett, and M. van der Schaar (2019). "Time-series generative adversarial networks". In: *Advances in Neural Information Processing Systems*. Vol. 32. URL: https://proceedings.neurips.cc/paper/2019/hash/c9efe5f26cd17ba6216bbe2a7d26d490-Abstract.html.

Yu, L., W. K. Härdle, L. Borke, and T. Benschop (2023). "An AI approach to measuring financial risk". In: *The Singapore Economic Review* 68.5, pp. 1529–1549. DOI: 10.1142/S0217590819500668.

Yuan, M. (2006). "GACV for quantile smoothing splines". In: *Computational Statistics & Data Analysis* 50.3, pp. 813–829. DOI: 10.1016/j.csda.2004.10.008.

Yuan, X. and Y. Qiao (2024). "Diffusion-TS: Interpretable diffusion for general time series generation". In: *International Conference on Learning Representations (ICLR 2024)*. DOI: 10.48550/arXiv.2403.01742.