

MUSED-FM: A BENCHMARK FOR EVALUATING MULTIVARIATE TIME SERIES FOUNDATION MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

Multivariate Time Series Foundation Models (TSFMs) aim to identify patterns in multiple contexts to make meaningful predictions of the future. At their core is multivariate capability; models make use of information from multiple sources rather than relying on a single signal with limited information. Learning multivariate models requires meaningful evaluations, but current benchmarks are limited in two key ways: quantity and quality. There are a limited number of multivariate time series datasets, with existing ones lacking size and diversity across domains. Furthermore, although some collections of time series might be marketed as multivariate, it is not proven that they contain meaningful information in multiple contexts. This work takes a major step in both directions, providing a Multivariate Time Series Evaluation Dataset for Foundation Models (MuSED-FM). MuSED-FM spans 16 multivariate time series domains and introduces novel synthetic data techniques, comprising 67 billion data points and 2.6 million time series. To improve and prove the quality of multivariate data, we provide a powerful suite of benchmarking tools focused on identifying the multivariate predictability of a time series and introduce novel multivariate predictability aggregate metrics based on classical methods. Finally, we evaluate current state-of-the-art TSFMs for both univariate and multivariate capability, finding that despite multivariate predictability identifying correlation, univariate prediction often matches or outperforms multivariate prediction across models.

1 INTRODUCTION

Accurate time series forecasting is valuable in a wide variety of domains, such as scientific advancement [96], finance [88], operations analysis [89], and weather [55]. Catalyzed by recent advances in foundation models for vision and language [38; 63], Time Series Foundation Models (TSFMs) [30] achieve state-of-the-art forecasting accuracy by identifying patterns in-context. However, the history of a *single* time series, that is, a single sequence of numbers, often lacks sufficient information to forecast all but the simplest trends. Incorporating information about the patterns in other, covarying time series’ is essential to improve forecasting accuracy. Thus, it is vital that TSFMs take advantage of multivariate data.

Due to the necessity of multivariate TSFMs, it is imperative that time series forecasting evaluations reflect multivariate forecasting performance. However, evaluating the multivariate capabilities of TSFMs remains a challenging problem. Existing open multivariate time series datasets suffer from two key issues: quantity and quality. First, the quantity and diversity of data is insufficient. Existing TSFMs are trained on limited domains of multivariate data, leading to poor generalization to new multivariate domains. Second, the quality of existing multivariate evals is poor. Although existing multivariate datasets provide multiple contexts for time series forecasting, they fail to demonstrate that the additional contexts contain valuable information for forecasting.

Motivated by the challenges of multivariate time series evaluation, we introduce a **Multivariate Time Series Evaluation Dataset (MuSED-FM)** consisting of a significant number of multivariate datasets from a wide range of sources. We aggregate data first from open source multivariate time series benchmarks on a wide range of domains and frequencies, from climate information to web data, scientific domains, or government statistics. We then take univariate data from similar domains, such as finance or weather, and construct multivariate time series data by combining series with similar

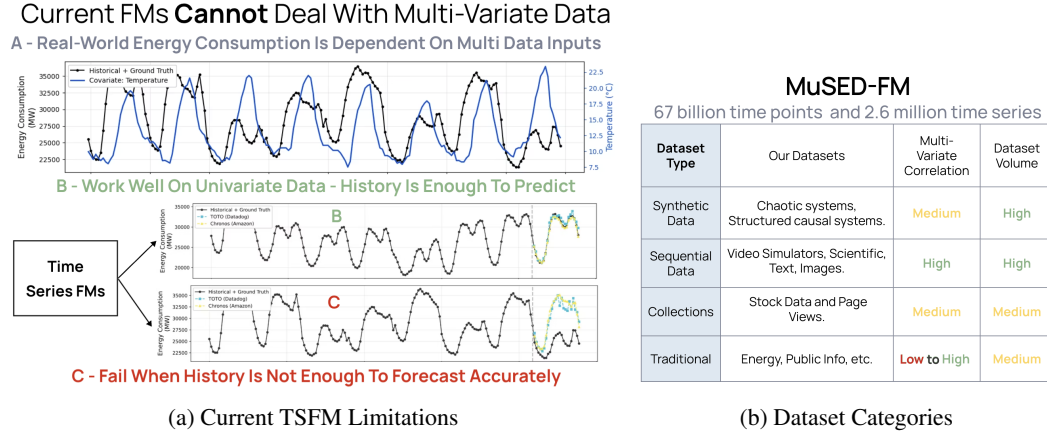


Figure 1: **Limitations of TSFMs and our proposed MuSED-FM** Left: Real data highlight a core limitation of current TSFMs: they perform well when the past alone suffices (B) but fail when prediction depends on exogenous, multivariate signals (C). In (A), energy demand depends on temperature; in (B), Chronos [5] (yellow) and Toto [29] (blue) succeed on a univariate case; in (C), the same models largely ignore covariates and default to seasonal extrapolation. **Right:** MuSED-FM is a curated multivariate benchmark—spanning sequential, synthetic, real-world, and combined datasets—designed to rigorously evaluate TSFMs’ ability to use cross-variable information.

descriptions. Then, we augment this data by representing various sequential data, including video, controls, and robotics data, which can be transformed into a usable range for multivariate foundation models. Finally, we include synthetic data generated according to either a distribution of multivariate ordinary differential equations [68], or using a structural causal model prior [56]. In Figure 1, we provide a categorization of the kinds of information present in the dataset and the predictive properties of the datasets by domain. In particular, this evaluation benchmark can answer: 1) How well does a multivariate TSFM perform on a wide range of realistic time series data? 2) How does a model’s multivariate performance compare with univariate performance? 3) How can one assess whether a novel multivariate dataset actually exhibits univariate *and* multivariate predictability? This work provides comprehensive foundation model evaluation on several existing foundation models to answer the first two questions, and a comprehensive univariate and multivariate predictability analysis to answer the last. The evaluation and analysis in this work illuminates the core premise: *Current multivariate time series foundation models ubiquitously fail to utilize multivariate information, and MuSED-FM is the key evaluation necessary for multivariate capability.*

In aggregate, the contributions of this paper are as follows:

1. **We propose MuSED-FM, the largest collection of multivariate time series datasets**, including open-source, univariate datasets with correlated information, multivariate series derived from multimodal data (including scientific, simulator, video, vision, and language data), and multivariate synthetic time series generated via multiple novel techniques. This dataset includes 67 billion time points, 2.6 million series, and an average of 26 variates across 45 datasets and 16 domains.
2. **We provide a novel metric** reflecting the univariate and multivariate “predictability” of a time-series, utilizing classic multivariate relational tools: transfer entropy [109], Granger Causality [52], Convergent Cross Mapping [115], lagged cross correlation [10] and DLinear [142].
3. **We evaluate current state of the art TSFMs**, illustrating that current state of the art multivariate models fail to exceed univariate performance across multivariate data (Figure 2).

2 RELATED WORK

TSFMs have emerged recently, with pioneering efforts [103; 44; 105] applying principles from language [38] and vision [63] to the field of time series, sometimes explicitly [53]. Previous datasets [84; 85], [7], [31], [47], served primarily to access domain-specific training data rather than as an evaluation of a single model operating across multiple domains. However, the subsequent capacity of foundation models [73; 91; 127] as tools that offer versatility and even performance over classical methods [90; 140] such as ARIMA or VARMAX, has inspired recent investigation into time

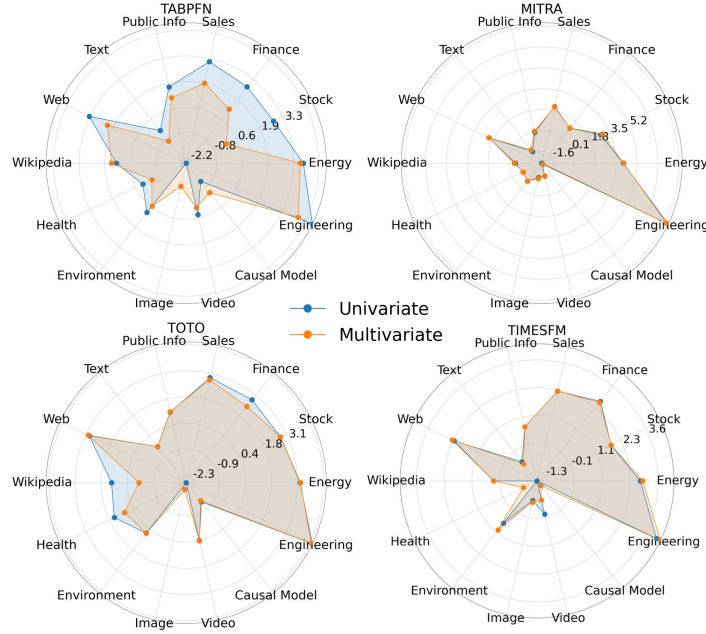


Figure 2: **Existing State-Of-The-Art (SOTA) Multivariate TSFMs fail to effectively use the available correlated information.** We showcase the comparison of the univariate and multivariate performance of 4 SOTA models on 14 diverse domains. Specifically, we plot the logarithm of the inverse of the Mean Average Percentage Error (MAPE) metric (higher is better). Surprisingly, providing additional correlated information (multivariate case) leads to similar or worsening of the model performance (TabPFN), with marginal improvements observed in very few cases.

series foundation models [133; 42; 139; 131; 43; 5; 36; 50; 74; 51; 113]. Nonetheless, initial efforts [105; 5] focused on univariate time series’, which are both simpler to process and more readily available. However, univariate time series’, by definition, do not encode contextual information, which limits their predictive power. Thus, recent work has begun to consolidate on the question of Multivariate Time Series Forecasting [139; 16; 29; 6; 105; 57], though in this work we demonstrate that these results, while promising, suggest that multivariate forecasting remains an open question. Works in foundation modeling often utilize large, privately owned datasets to train models, such as [29; 81; 105]. This work contributes to the effort for a multivariate time series foundation model by 1) providing the largest (number of data points, rows and correlates) and most diverse (spanning multiple domains and classes of data) multivariate time series evaluation benchmark and 2) providing aggregated metrics and benchmarking for *identifying multivariate predictability* of both models and data.

Meaningful evaluation and training data is the cornerstone for a multivariate time series foundation model, and while main benchmarks remain domain specific, focusing on classic fields such as energy [151; 128], weather and climate [134; 55], traffic [17; 79; 92], finance [88; 58], air quality [147], healthcare and brain signals [32; 62; 126; 96; 136] but also branching out into compute or network related domains [117; 130; 87; 75], cryptocurrency [4], biology [97], industrial devices [125], icicle formation [19], aerospace [144], action identification [49], environment [54]. Alternatively, many benchmarks focus on related problems such as anomaly detection [94; 59; 78; 135; 101], classification [98], robustness [146], etc. The substantial diversity of domains offers both a tantalizing promise and an obvious challenge for time series foundation models: the promise of widespread applicability and the challenge of myriad diversity. Nonetheless, recent work foundation model benchmarks include [84; 119; 85], but subsequently with [3; 47; 9; 1; 71], as well as benchmarks focused on balanced or fair assessment [99; 112], horizon [145], multimodality [76], irregular sampling [15], or dynamical systems [68; 120]. Other work focuses on benchmarking with different foundation models, including [70; 111; 77; 100], while also contributing some compilations of data. While these methods often include multivariate series, they often do not focus on the multivariate problem. First, this can lead to a limited volume of multivariate data, and datasets such as those used in [47], [1], and [70] are either entirely or heavily univariate. Furthermore, the analysis of the mul-

	Domain	Counts	Number of Timesteps	Average Number of variates	Dataset Count	Number of Time Series
Traditional	Energy	114,941,077	4,801,760	19.32	34	315
	Public Info	4,625,412	219,946	17.00	10	270
	Health	202,815	41,756	8.00	2	9
	Sales	280,348	29,275	16.83	6	15
	Climate & Env.	496,270,478	10,993,235	24.86	7	54
	Engineering	1,975,188	183,200	8.50	2	20
	Finance	99,666	6,328	15.75	8	8
	Web	15,169	2,167	7.00	1	1
Collections	Stock	8,852,800	354,112	25.00	1	352
	Wikipedia	18,084,600	602,820	30	1	1,530
Sequential	Images	126,720,000	11,520,000	11.00	1	20,000
	Text	190,308,864	10,016,256	19.00	1	19,563
	Video	9,403,332	154	10	1	6114
	Scientific	1,373,811,912	26,441,444	54.00	5	48,812
Synthetic	Dynamic	5,142,085,632	214,253,568	25	1	209,232
	Causal Model	57,404,227,584	1,808,007,168	25.4	5	2,354,176
Combined		64,791,804,797	2,077,472,139	25.7	85	2,680,671

Table 1: **High-Level Dataset Summary by Domain.** We provide a breakdown of our benchmark, **MuSED-FM**, including the average number of variates, time series, and datasets in each domain.

tivariate datasets in [1] is limited to mostly a small subset. Of existing time series foundation model datasets, [71] shows the closest resemblance to MuSED-FM, containing multiple real-world time series datasets. Our work subsumes this contemporary work by first including a greater volume of datasets spanning several more domains and time series patterns. Second, including novel datasets by combining related univariate series, transforming sequential data into time series and generating synthetic data to create a more comprehensive datasets, which we discuss further in Section 4. Third, the analysis of variates in [71] uses pearson correlation metrics, which only measure if two series look similar, regardless of how useful the variates are for prediction. By contrast this work introduces a novel multivariate *predictability* score discussed in Section 5. Overall, MuSED-FM exposes the limitations of existing multivariate models, provides meaningful correlation analysis tools and provides a much greater volume and scope of multivariate data. Extended discussion of related work can be found in Appendix A.

3 MULTIVARIATE TIME SERIES FOUNDATION MODELS

The multivariate time series forecasting problem involves the correlate variates time series $\mathbf{x} \in \mathbb{R}^{T \times d_x}$ and target variate $\mathbf{y} \in \mathbb{R}^{T \times d_y}$, where d_x is the number of correlate variates, d_y is the dimension of the target variate and T is the number of timesteps. The time series is then separated into the context (history) component $0 < h < T$ and the prediction component $p = T - h$, where we represent $\mathbf{x}^{:h}, \mathbf{y}^{:h}$ as the context components of correlate and target variates respectively, and $\mathbf{y}^{-p:}$ as the predict component, and $X^{:h}, Y^{:h}, Y^{-p:}$ as the random variables representing the distribution of multivariate time series correspondingly. The objective of a multivariate time series forecaster is to learn a model of the conditional distribution: $f(\mathbf{x}^{:h}, \mathbf{y}^{:h}) \rightarrow \mathbb{P}(Y^{-p:} | X^{:h} = \mathbf{x}^{:h}, Y^{:h} = \mathbf{y}^{:h})$. Foundation models for time series train the model on the space of all multivariate series to $(\mathcal{X}, \mathcal{Y})$ to perform the best estimate of this model. By contrast, a univariate time series forecaster simply models $f_{\text{uni}}(\mathbf{y}^{:h}) \rightarrow \mathbb{P}(Y^{-p:} | XY^{:h} = \mathbf{y}^{:h})$, assuming that $d_y = 1$.

To properly situate the datasets in this work, we provide univariate analysis, multivariate analysis and benchmarking with existing foundation models. Univariate analysis involves tools such as Entropy, Fourier analysis, etc., as described in Section 5, and aim to characterize the distribution of $\mathcal{Y}$, not only for particular domains, but also for synthetic data. By contrast multivariate analysis aims to assess the amount of relative information in the history of the variates $X^{:h}$ with the future of the target $Y^{-p:}$, which can be formalized as the time-based conditional mutual information $I(X^{:h}; Y^{-p:} | Y^{:h})$ [110]. While the distributions of all three random variables are unknown in practice, our analysis offers a suite of approximation

techniques for estimating this quantity. Finally, benchmarking assess the performance of the models $\mathbb{E}_{\mathbf{x}^{:h}, \mathbf{y}^{:h}, \mathbf{y}^{-p:} \sim \mathcal{D}_{\text{eval}}} [\|\mu(f_{\text{model}}(\mathbf{x}^{:h}, \mathbf{y}^{:h})) - \mathbf{y}^{-p:}\|]$, where $\mathcal{D}_{\text{eval}}$ is the evaluation dataset stratified by domain. We also compare the performance of univariate and multivariate forecasting $\mathbb{E}_{\mathbf{x}^{:h}, \mathbf{y}^{:h}, \mathbf{y}^{-p:} \sim \mathcal{D}_{\text{eval}}} [\|\mu(f_{\text{model}}(\mathbf{x}^{:h}, \mathbf{y}^{:h})) - \mathbf{y}^{-p:}\| - \|\mu(f_{\text{uni}}^*(\mathbf{y}^{:h})) - \mathbf{y}^{-p:}\|]$, where f_{uni}^* is the performance of the best univariate model. Because none of the existing state-of-the-art models show little performance difference between univariate and multivariate forecasting, we also introduce the variate counterfactual scenario, which also performs prediction also using $\mathbf{x}^{-p:}$, or the future values of the correlate variates. Formally, this is represented with: $\mathbb{E}_{\mathbf{x}^{:h}, \mathbf{x}^{-p:}, \mathbf{y}^{:h}, \mathbf{y}^{-p:} \sim \mathcal{D}_{\text{eval}}} [\|\mu(f_{\text{model}}(\mathbf{x}^{:h}, \mathbf{x}^{-p:}, \mathbf{y}^{:h})) - \mathbf{y}^{-p:}\|]$. This setting is successfully used by [56] and gives partial evidence of multivariate capabilities.

4 DATASET

To provide context for the data that makes up MuSED-FM, we focus on four primary classes of data, with a brief description of the constituent datasets that comprise these datasets. With 45 different datasets aggregated to form MuSED-FM specific dataset details are reserved for Appendix B, with a summary of dataset information in Table 1. The four categories of data are: Traditional multivariate time series datasets (traditional), related univariate time series combined as time series (combined), sequential data transformed into time series (sequence), and synthetic time series (synthetic). These four categories allow us to distinguish correlate selection as by intuition, heuristics, domain and construction, properties we discuss further below.

4.1 TRADITIONAL MULTIVARIATE TIME SERIES

We first form the core of our dataset with an aggregation of **37 different datasets** from a wide range of domains including energy, retail, finance, weather, healthcare, public services, healthcare and web traffic. The obvious power of these datasets is primarily that they are readings from a real world process, but also that they have been selected by domain experts, so that the variates provide useful information for forecasting. However, it is also possible that for some of these datasets, even though the variates are intuitively related, they provide no additional forecasting benefit. Thus, we provide analysis in Section 5 to provide initial evidence that these variates should help in forecasting. We were able to collect 650 time series comprising 0.5 Billion time points.

4.2 COLLECTIONS OF UNIVARIATE TIME SERIES

In addition to hand-curated multivariate time series, correlated univariate series, especially in domains such as web and finance, offers another rich data source. MuSED-FM includes **two domains**: stocks and wikipedia pageview data, constructing correlated variates using domain specific heuristics (see Appendix C). The degree of multivariate capability is decided by the efficacy of the heuristic, though we also provide metrics that suggest that the variates are useful for forecasting in Section 5. These collections add 1800 time series and 29 million data points, as seen in Table 1.

4.3 SEQUENTIAL DATA TRANSFORMED INTO TIME SERIES

Time series are a ubiquitous form of data, represented by any signal taken at timestamped intervals. However, most datasets focus on “classic” time series domains such as climate, web, finance, etc. However, sequential data can often be extracted from sources such as scientific data, sequential decision making data, images, language, and video, which we leverage in MuSED-FM, which we leverage to add **five additional** data-rich and correlated domains. With each of these domains, we apply a three step process to convert them into time series: **1)** transform the data into high dimensional tokens, **2)** Apply dimensionality reduction to a fixed set of variates. **3)** Apply smoothing and/or noising to better match the distribution of traditional time series. Specific details for this process are included in Appendix D. The two advantages over other multivariate time series datasets come through the volume of data, as well as high confidence of correlation in the variates, since they are generated according to the same underlying generative process.

Group	Statistic	Description
Univariate Basic	Outlier percentage	Percentage of data points > 3 st. dev. from mean
	Kurtosis	Fourth order moment, measures tailedness
	Skewness	Third order moment, measures asymmetry
	Entropy [41]	Measures noise as $-\sum_i p_i \log p_i$
	Hurst coefficient [60]	Measure of persistent features
	Number of mean crossing points [83]	Number of times the mean is crossed (variance)
	Unitroot KPSS statistic [66]	Kwiatkowski-Phillips-Schmidt-Shin measure of stationarity
	Unitroot Phillips-Perron statistic [95]	Phillips-Perron test statistic for stationarity.
Univariate Shape	Autocorrelation at lag 10 [10]	Measure of the correlation between a data point and another point that occurred 10 time periods earlier
	Differenced autocorrelation at lag 1 [10]	Degree to which autocorrelation explains the sequence
	Sum of squares of the first five autocorrelation coefficients [10]	Magnitude of simple autocorrelation
	Longest flat sequence [83]	length of flat periods, flatter is sometimes easier to predict.
	Normalized number of periods [41]	More periodic signals can be easier to predict.
	Nonlinearity via surrogate data [123]	measure of linearity in the signal.
Univariate Window	Stability [14]	variance of the means in the windows
	Lumpiness [14]	variance of the variances in the windows
	Max change per window [14]	largest delta inside a window (measures spikes)
	Max variance per window [14]	Highest window variance (measures noisiness)
Univariate Decomp.	10-feature Fourier decomposition residual [10]	Residual after decomposing the top 10 frequencies
	10-feature SINDy decomposition residual [13]	Residual after dynamic system decomposition with 10 parameters.
	10-feature Spectral decomposition residual [41]	Residual after decomposition using spectral Principle Component Analysis (linear).
	Trend [27]	Measure of trend of STL relative to residual.
	Seasonal strength [27]	Measure of seasonal of STL relative to residual.
	Spike [83]	Measure of variance of residual.
	Linearity [83]	linear consistency of the trend component.
	Curvature [83]	Measures curvature of trend component
	Peak [83]	measures the peaks of the seasonal component
	Trough [83]	Measures the troughs of the seasonal component.
	ST decomposition residual [27]	Measures the residual from just the trend and seasonal components.
Multivariate	Lagged cross correlation [10]	Convolves the shape of a lagged historical window of a correlate with the future
	Granger causality [52]	Measures the degree to which linear model of each variate improves over univariate autoregression.
	Transfer entropy [109]	Measures the relative information added by adding a variate.
	Convergent cross mapping [115]	Measures the degree of coupling of the variates to the target forecast.
	DLinear [142]	Learns a model of the STL decomposition of variates to forecast the target.

Table 2: **Description of univariate and multivariate predictability features.** Our novel univariate and multivariate predictability metrics are obtained by aggregating these individual metrics.

4.4 SYNTHETIC MULTIVARIATE TIME SERIES DATA

Training tabular and time series with synthetic data has seen recent success [102; 56; 6; 5], suggesting that synthetic data provides a substantial breadth of multivariate data that can overlap the space of real time series. Furthermore, synthetic datasets are multivariate by design, since the generative process dictates correlation in time and across variates. MuSED-FM leverages **two distinct lines of work** to generate multivariate time series. First, building on the success of TabPFN [56], especially when applied to time series, we provide an SCM-based implementation building on TabICL [102], but adapted to generate multivariate time series by replacing noise with sequentially correlated inputs, and shifting observations to ensure multivariate predictability. Second, building on the exciting work of Panda [68], we augment this method to construct multivariate dynamical systems with multiple components by composing these systems into arbitrary graph structures, to produce relationships between the graph components that form the multivariate relationships that a foundation model can identify. This allows us to generate 2.5 million time series with 6.2 billion data points, a vast and diverse set of multivariate data with correlations by construction. We provide additional details about these data formats in Appendix E and F.

	Traditional								Collections		Sequential				Synthetic		Combined
Predictability Type	Energy	Public Info	Health	Sales	Climate & Env.	Eng.	Finance	Web	Stock	Wiki-pedia	Images	Text	Video	Scientific	Dynamics	Causal Model	
Univariate Predictability	-0.233	-0.300	-0.328	-0.314	-0.363	-0.379	-0.162	-0.168	-0.198	-0.409	-0.422	-0.399	-0.427	-0.445	-0.311	-0.364	-0.285
Multivariate Predictability	0.384	0.373	0.331	0.408	0.258	0.359	0.440	0.494	0.336	0.286	0.335	0.342	0.439	0.351	0.281	0.266	0.367

Table 4: **Univariate and Multivariate predictability metrics per domain.** Here, we showcase both predictability metrics for 16 different domains in the MuSED-FM benchmark. Higher univariate predictability (lower negative values) indicates the availability of more information in the target time series history to forecast effectively. Similarly, higher multivariate predictability indicates the availability of more information in the variates to forecast effectively. **Key takeaway:** across many domains, covariates carry substantial predictive signal that good TSFMs should be able to exploit—beyond simple extrapolation from “single context”, i.e., just using single variate.

5 ANALYSIS OF CHARACTERISTIC FEATURES

The vital property of a dataset used for time series foundation models is to characterize the degree to which the dataset can be forecasted from historical data. This work introduces two “predictability” metrics: **univariate predictability** and **multivariate predictability**. The former measures the degree of difficulty of univariate forecasting, whereas the latter characterizes the degree of simple correlations in the data. Though neither is a perfect metric, this analysis is invaluable for aggregating a large corpus of datasets across a wide variety of different domains, since it allows for a meaningful comparison of foundation model performance in Section 6 across different approximations of difficulty. Both metrics are an averaged sum of different statistics described in Table 2, where the weighting ensures that each statistic can take on a value between [0,1] or [-1,0]. Thus, the metrics take on values between [-1,1] for univariate predictability, and [0,1] for multivariate predictability.

Formally, to construct both the univariate and multivariate “predictability” statistic, we combine these metrics by first normalizing unnormalized values between 0, 1 using the range observed across all datasets. Next, we take a weighted sum of the metrics, where the weights are -1 for metrics that indicate more challenging forecasting when lower, and $+1$ for the opposite (for multivariate, this is when the metric indicates greater information transfer). Finally, we divide by the total number of components to identify the final predictability scores. Thus, for a set of n positive normalized metric values $\{\gamma_1^+, \dots, \gamma_n^+\}$ and m negative normalized metric values $\{\gamma_1^-, \dots, \gamma_m^-\}$, the aggregated metric is computed as:

$$\frac{1}{n} \cdot \left(\sum_{i=1}^n \gamma_i^+ + \sum_{j=1}^m \gamma_j^- \right). \quad (1)$$

We illustrate these scores across different domains in Table 4.

We visualize higher univariate predictability values indicating greater predictability and thus lower average MAPE across

models through a loose correlation as seen in Figure 3a. While multivariate predictability should give a measure of the benefit of adding correlated information, we have observed that most existing multivariate time series foundation models have the same or worse performance as univariate, thus we do not observe that higher multivariate prediction score actually correlate with greater difference in univariate-multivariate in Figure 3b. However, to demonstrate that this is a limitation of the models and not the metric, we construct multivariate time series where we either sample series from a random distribution of Fourier and ARIMA samples, or from unrelated time series, and show that these collections of time series have lower average multivariate predictability than the datasets sampled from MuSED-FM, as seen in Table 3.

	Random	Other	Correlated
Multivariate Predictability	0.163	0.135	0.367

Table 3: **Multivariate predictability captures cross-variable signal.** We ablate the metric by replacing true covariates with (i) **Random**: randomly generated Fourier/ARIMA series, and (ii) **Other**: covariates randomly drawn from different series. In both cases, the score drops substantially relative to using the true covariates, validating that the metric assigns higher values when covariates are genuinely correlated and lower values when they are uninformative.

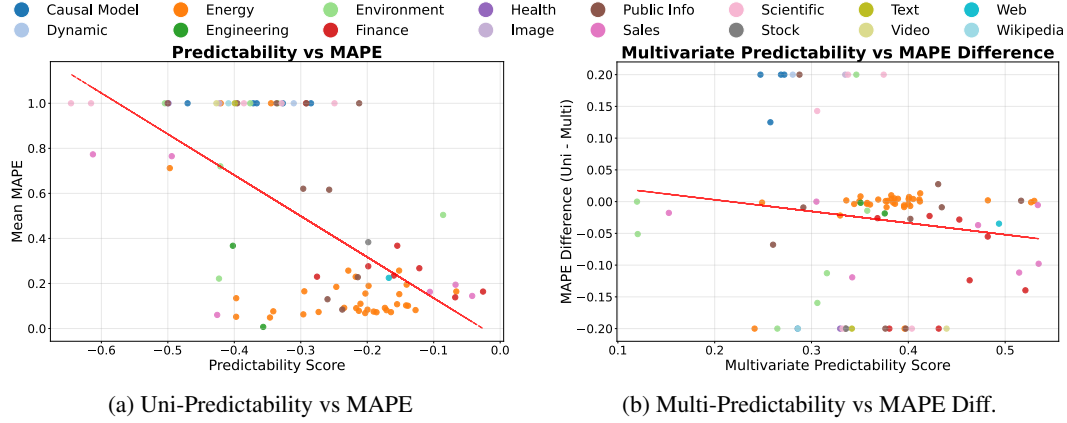


Figure 3: Points are datasets. **Left:** As predictability increases, MAPE decreases (better prediction from FMs). **Right:** Plot univariate MAPE - multivariate MAPE averaged across multivariate models. **Key takeaway :** As multivariate predictability increases, it does not necessarily indicate better performance of TSFMs, indicating their limited multivariate capability. MAPE values are clipped at 1.0/0.2, respectively, to prevent outlier skew.

Group	Time Series Model	Description
Classic	Mean [10]	Trivial Baseline Model that forecasts the mean value.
	Prophet [122]	Additive linear regression mixture model using STL.
	SARIMAX [10]	Seasonal ARIMA with exogenous regressors; classic statistical model with exogenous elements included as lagged regressors.
	Google Causal Impact [12]	Bayesian structural time series model for inferring the causal effects (interventions, counterfactuals, etc.).
Univariate	Chronos [5]	Transformer-based large-scale pretrained univariate Time Series Foundation Model.
Multivariate	TabPFN [56]	Tabular Transformer-based Model with row-column attention adapted for time series by including timestamps as features trained on synthetic tabular data. Univariate uses only these features, while multivariate uses the historical variates as context.
	Mitra [6]	Tabular Transformer-based Model inspired by TabPFN, which uses a similar structure but different implementation and improved performance. Univariate and Multivariate implementations are similar.
	Toto [29]	Time-series Optimal Transport Operator; Multivariate time series model trained on proprietary data, with native univariate and multivariate support by changing the number of input columns.
	TimesFM [105]	Current state of the art model for univariate time series foundation models. Implements multivariate prediction for the target by first predicting other variates followed by fitting a linear regression model for the target prediction.

Table 5: **Representative time-series models used in our experiments**, grouped as classic statistical baselines, univariate TSFMs, and multivariate/tabular approaches. The list emphasizes how multivariate modeling is handled: some methods natively learn cross-variable structure (e.g., Toto), while others treat covariates as features or apply post-hoc regression (e.g., TimesFM).

6 FOUNDATION MODEL EVALUATION

Identifying multivariate correlations in real-world time series remains a holy grail of TSFMs, as it allows the model to maximally leverage the most general form of information—other time series. As such, several state-of-the-art foundation models support multivariate capabilities. However, identifying the degree to which these models are capable of multivariate forecasting remains an open problem. MuSED-FM offers a comprehensive variety of data for identifying multivariate time series foundation model capabilities.

The core challenge behind evaluating a TSFM’s multivariate capability is the entanglement of a dataset’s multivariate information and a model’s ability to make use of multivariate infor-

		CLASSIC				UNIVARIATE					MULTIVARIATE				AGGREGATED
	DOMAIN	MEAN	PROPHET	SARIMAX	CAUSAL IMPACT	CHRONOS	TABPFN	MITRA	TOTO	TIMESFM	TABPFN	MITRA	TOTO	TIMESFM	
TRADITIONAL	ENERGY	N/A	3.836	3.784	13.313	0.186	0.126	0.178	11.745	1.837	0.191	0.175	16.385	0.288	4.337
	PUBLIC INFO	2.274	1.505	2.596	8.643	1.499	1.460	3.800	1.745	1.744	1.465	3.056	1.593	1.765	2.550
	HEALTH	5.553	4.719	1.533	2.266	0.434	2.528	3.956	0.453	3.804	5.278	4.118	0.768	2.230	2.895
	SALES	0.487	0.524	0.539	0.355	0.242	0.247	0.455	0.240	0.217	0.476	0.464	0.265	0.213	0.363
	CLIMATE & ENVIRONMENT	2.263	1.505	1.085	0.806	2.450	1.314	2.624	1.092	0.667	1.523	2.740	1.101	0.487	1.598
	ENGINEERING	0.217	0.216	0.009	0.008	0.209	0.147	0.006	0.217	0.132	0.214	0.006	0.216	0.132	0.133
	FINANCE	2.512	0.523	0.465	0.136	0.586	0.399	2.437	0.438	0.330	2.314	2.425	0.691	0.386	1.049
COLLECTIONS	WEB	0.288	0.165	0.038	0.034	0.167	0.133	0.289	0.156	0.141	0.290	0.287	0.149	0.130	0.174
	STOCK	1.132	0.268	0.026	0.027	0.251	0.274	0.136	0.202	0.200	0.885	0.129	0.204	0.199	0.303
SEQUENTIAL	WIKIPEDIA	2.029	5.747	N/A	N/A	0.462	0.584	1.348	0.551	0.802	0.481	1.282	1.583	0.802	1.425
	IMAGES	3.590	4.245	3.346	8.042	2.114	8.958	2.373	9.727	1.871	3.560	2.259	7.447	1.752	4.560
	TEXT	2.641	3.035	1.989	1.583	1.346	1.756	2.354	1.671	1.609	2.970	2.121	1.664	1.776	2.040
	VIDEO	1.404	0.910	2.358	1.649	0.982	1.141	2.538	1.010	1.133	1.508	2.554	0.998	1.897	1.545
	SCIENTIFIC	N/A	6.950	N/A	N/A	5.522	4.816	N/A	7.279	3.551	15.200	N/A	5.969	3.551	6.605
SYNTHETIC	DYNAMIC	3.023	2.796	2.792	2.887	3.325	3.081	2.204	2.675	4.393	1.094	2.219	2.298	1.610	2.646
	CAUSAL MODEL	1.570	16.030	10.329	N/A	3.753	3.922	6.883	4.434	5.496	2.113	6.430	4.408	6.489	5.988
COMBINED	COMBINED	2.055	3.358	2.236	3.079	1.368	2.001	2.018	2.823	1.772	2.419	1.937	2.961	1.520	2.273

Table 6: Domain groups vs forecasting method evaluations mean performance. Green indicates that multivariate performance is more than 0.1 better than the univariate performance of the same model. Note that the best multivariate exceeds the best univariate on only two domains.

mation. Poor performance in an evaluation could be attributable to either a dataset’s lack of useful information in multivariate signals or a model’s inability to take advantage of existing multivariate information. While in an ideal world, we might be confident with a multivariate dataset using a diversity of realistic, meaningful correlates, this volume of data simply does not exist in an open-source format. The core advantage of MuSED-FM is to aggressively leverage a diversity of data as described in Section 4.

MuSED-FM disentangles the limitations of a model from the limitations of a dataset by categorizing its forms of data as “synthetic”, “combined”, or “traditional and sequential”. In particular, “synthetic” datasets have multiple multivariate patterns by construction—there will *always* be additional information provided by the variates not present in the univariate time series. Thus, if the model is unable to capture these relationships, then that would serve as a strong indicator of the limitations of the model. On the other hand, “combined” datasets of hand-collated univariate time series will probably contain a good deal of redundant information, so if a model is able to improve performance by adding variates, this suggests generalized ability to ignore certain variates and keep others. Finally, “traditional and sequential” data is likely to contain positive transfer in the correlates, either because of strong human intuitions or because the underlying system from which the sequential information is derived is entangled. Consequently, these domains offer a meaningful blend of being in-distribution of “real” time series data, while in aggregate offering strong evidence. Thus, MuSED-FM is a powerful tool for fine-grained analysis of multivariate time series foundation model capabilities.

Through MuSED-FM we identify a key limitation of existing TSFMs— across most domains, these models show **little to no improvement** when presented with multivariate context. For the four models we assessed with multivariate capability (TabPFN, Mitra, Toto, and TimesFM), all of them showed little to no improvement when comparing univariate to multivariate forecasting performance, as seen in Figure 2. Assessing a model univariately involves only showing it the target variate, while multivariate assessment gives the model all the variates. On all domains, the best performance for any model was univariate rather than multivariate, as seen in Table 6. Thus even when a multivariate version of a model is able to improve over the corresponding univariate version, it may not be leveraging the multivariate information to do so.

7 CONCLUSION

This work introduces MuSED-FM, a benchmark that assesses time series foundation models multivariate capabilities across a wide range of domains and settings. Focusing on multivariate predictability, this work introduces both univariate and multivariate predictability metrics to estimate useful information from history, as well as from other variables. Combined, this dataset and analysis illustrate a core limitation of existing multivariate time series foundation models: when applied zero-shot on a variety of domains, *the multivariate performance does not exceed univariate performance, even on synthetic datasets which have multivariate correlation by construction*. This demonstrates the need not only for more extensive multivariate analysis, but also the importance of MuSED-FM as a key step for delivering on the promise of multivariate time series foundation models.

8 REPRODUCIBILITY STATEMENT

We intend for MuSED-FM as a competition on Huggingface to assess SOTA Multivariate time series foundation models, and eventually fully open source. The analysis computations and metrics leverage open source repositories, and we intend to open source our code for computing the univariate and multivariate predictability metrics.

REFERENCES

- [1] Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.
- [2] Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, et al. Gluonts: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020.
- [3] Yihao Ang, Qiang Huang, Yifan Bao, Anthony KH Tung, and Zhiyong Huang. Tsgbench: Time series generation benchmark. *arXiv preprint arXiv:2309.03755*, 2023.
- [4] Yihao Ang, Qiang Wang, Qiang Huang, Yifan Bao, Xinyu Xi, Anthony KH Tung, Chen Jin, and Zhiyong Huang. Ctbench: Cryptocurrency time series generation benchmark. *arXiv preprint arXiv:2508.02758*, 2025.
- [5] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.
- [6] AutoGluon Contributors. Mitra regressor. <https://huggingface.co/autogluon/mitra-regressor>, 2025. Accessed: 2025-09-17.
- [7] Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.
- [8] Lingfeng Bao, Jing Li, Zhenchang Xing, Xinyu Wang, Xin Xia, and Bo Zhou. Extracting and analyzing time-series hci data from screen-captured task videos. *Empirical Software Engineering*, 22(1):134–174, 2017.
- [9] André Bauer, Marwin Züfle, Simon Eismann, Johannes Grohmann, Nikolas Herbst, and Samuel Kounev. Libra: A benchmark for time series forecasting methods. In *Proceedings of the ACM/SPEC International Conference on Performance Engineering*, pp. 189–200, 2021.
- [10] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970.
- [11] Lorenzo Brigato, Rafael Morand, Knut Strømmen, Maria Panagiotou, Markus Schmidt, and Stavroula Mougiakakou. Position: There are no champions in long-term time series forecasting. *arXiv preprint arXiv:2502.14045*, 2025.
- [12] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247–274, 2015.
- [13] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *PNAS*, 113(15):3932–3937, 2016.
- [14] Mathieu Candel. tsfeatures: Time series feature extraction for r, 2020. URL <https://github.com/Nixtla/tsfeatures>. Accessed: 2025-09-17.

- [15] Ching Chang, Jeehyun Hwang, Yidan Shi, Haixin Wang, Wen-Chih Peng, Tien-Fu Chen, and Wei Wang. Time-imm: A dataset and benchmark for irregular multimodal multivariate time series. *arXiv preprint arXiv:2506.10412*, 2025.
- [16] Sameep Chattopadhyay, Pulkit Paliwal, Sai Shankar Narasimhan, Shubhankar Agarwal, and Sandeep P. Chinchali. Context matters: Leveraging contextual features for time series forecasting, 2025. URL <https://arxiv.org/abs/2410.12672>.
- [17] Chao Chen, Karl Petty, Alexander Skabardonis, Pravin Varaiya, and Zhanfeng Jia. Free-way performance measurement system: mining loop detector data. *Transportation research record*, 1748(1):96–102, 2001.
- [18] Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. Visions: Visual masked autoencoders are free-lunch zero-shot time series forecasters. *arXiv preprint arXiv:2408.17253*, 2024.
- [19] Xu Cheng, Fan Shi, Meng Zhao, Guoyuan Li, Houxiang Zhang, and Shengyong Chen. Temporal attention convolutional neural network for estimation of icing probability on wind turbine blades. *IEEE Transactions on Industrial Electronics*, 69(6):6371–6380, 2021.
- [20] Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery. *arXiv preprint arXiv:2310.01753*, 2023.
- [21] Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11525–11533, 2024.
- [22] Caleb Chuck, Supawit Chockchawat, and Scott Niekum. Hypothesis-driven skill discovery for hierarchical deep reinforcement learning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5572–5579. IEEE, 2020.
- [23] Caleb Chuck, Kevin Black, Aditya Arjun, Yuke Zhu, and Scott Niekum. Granger-causal hierarchical skill discovery. *arXiv e-prints*, pp. arXiv–2306, 2023.
- [24] Caleb Chuck, Carl Qi, Michael J Munje, Shuoze Li, Max Rudolph, Chang Shi, Siddhant Agarwal, Harshit Sikchi, Abhinav Peri, Sarthak Dayal, et al. Robot air hockey: A manipulation testbed for robot learning with reinforcement learning. *arXiv preprint arXiv:2405.03113*, 2024.
- [25] Caleb Chuck, Sankaran Vaidyanathan, Stephen Giguere, Amy Zhang, David Jensen, and Scott Niekum. Automated discovery of functional actual causes in complex environments. *arXiv preprint arXiv:2404.10883*, 2024.
- [26] Caleb Chuck, Fan Feng, Carl Qi, Chang Shi, Siddhant Agarwal, Amy Zhang, and Scott Niekum. Null counterfactual factor interactions for goal-conditioned reinforcement learning. *arXiv preprint arXiv:2505.03172*, 2025.
- [27] Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1): 3–73, 1990.
- [28] Oliver M Cliff, Annie G Bryant, Joseph T Lizier, Naotsugu Tsuchiya, and Ben D Fulcher. Unifying pairwise interactions in complex dynamics. *Nature Computational Science*, 3(10): 883–893, 2023.
- [29] Ben Cohen, Emaad Khwaja, Kan Wang, Charles Masson, Elise Ramé, Youssef Doubli, and Othmane Abou-Amal. Toto: Time series optimized transformer for observability. *arXiv preprint arXiv:2407.07874*, 2024.
- [30] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

- [31] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.
- [32] Nicolas Decat, Jasmine Walter, Zhao H Koh, Piengkwan Sribanditmongkol, Ben D Fulcher, Jennifer M Windt, Thomas Andriillon, and Naotsugu Tsuchiya. Beyond traditional sleep scoring: Massive feature extraction and data-driven clustering of sleep time series. *Sleep Medicine*, 98:39–52, 2022.
- [33] Edward J Delp, Rangasami L Kashyap, and O Robert Mitcheli. Image data compression using autoregressive time series models. *Pattern Recognition*, 11(5-6):313–323, 1979.
- [34] David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431, 1979. ISSN 01621459, 1537274X. URL <http://www.jstor.org/stable/2286348>.
- [35] Hossein Ebrahimi Dinaki, Shervin Shirmohammadi, Emil Janulewicz, and David Côté. Forecasting video qoe with deep learning from multivariate time-series. *IEEE Open Journal of Signal Processing*, 2:512–521, 2021.
- [36] Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam Nguyen, Wesley M Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. *Advances in Neural Information Processing Systems*, 37:74147–74181, 2024.
- [37] Ian Falconer, Georg A Gottwald, Ian Melbourne, and Kjetil Wormnes. Application of the 0-1 test for chaos to experimental data. *SIAM Journal on Applied Dynamical Systems*, 6(2):395–402, 2007.
- [38] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and machines*, 30(4):681–694, 2020.
- [39] John R Freeman. Granger causality and the times series analysis of political relationships. *American Journal of Political Science*, pp. 327–358, 1983.
- [40] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.
- [41] Ben D. Fulcher and Nick S. Jones. hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell Systems*, 5(5):527–531, 2017.
- [42] Ge Gao, Qitong Gao, Xi Yang, Miroslav Pajic, and Min Chi. A reinforcement learning-informed pattern mining framework for multivariate time series classification. In *31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.
- [43] Shanghua Gao, Teddy Koker, Owen Queen, Tom Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. Units: A unified multi-task time series model. *Advances in Neural Information Processing Systems*, 37:140589–140631, 2024.
- [44] Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. Timegpt-1. *arXiv preprint arXiv:2310.03589*, 2023.
- [45] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.
- [46] William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/ec5decca5ed3d6b8079e2e7e7bacc9f2-Paper-round2.pdf.

- [47] Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643*, 2021.
- [48] Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus. <http://SkyLion007.github.io/OpenWebTextCorpus>, 2019.
- [49] Dian Gong, Gerard Medioni, and Xuemei Zhao. Structured time series analysis for human action segmentation and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1414–1427, 2013.
- [50] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.
- [51] Lars Graf, Thomas Ortner, Stanisław Wołszniak, Angeliki Pantazi, et al. Flowstate: Sampling rate invariant time series forecasting. *arXiv preprint arXiv:2508.05287*, 2025.
- [52] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.
- [53] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36: 19622–19635, 2023.
- [54] Chenjuan Guo, Bin Yang, Ove Andersen, Christian S Jensen, and Kristian Torp. Ecomark 2.0: empowering eco-routing with vehicular environmental models and actual vehicle fuel consumption data. *GeoInformatica*, 19(3):567–599, 2015.
- [55] Tao Han, Song Guo, Zhenghao Chen, Wanghan Xu, and Lei Bai. How far are today’s time-series models from real-world weather forecasting applications? *arXiv preprint arXiv:2406.14399*, 2024.
- [56] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.
- [57] Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: How tabpfn-v2 outperforms specialized time series forecasting models. *arXiv preprint arXiv:2501.02945*, 2025.
- [58] Yifan Hu, Yuante Li, Peiyuan Liu, Yuxia Zhu, Naiqi Li, Tao Dai, Shu-tao Xia, Dawei Cheng, and Changjun Jiang. Fintsb: A comprehensive and practical benchmark for financial time series forecasting. *arXiv preprint arXiv:2502.18834*, 2025.
- [59] Xuanwen Huang, Yang Yang, Yang Wang, Chunping Wang, Zhisheng Zhang, Jiarong Xu, Lei Chen, and Michalis Vazirgiannis. Dgraph: A large-scale financial dataset for graph anomaly detection. *Advances in Neural Information Processing Systems*, 35:22765–22777, 2022.
- [60] H. E. Hurst. Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116:770–799, 1951.
- [61] Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. Position: What can large language models tell us about time series analysis. In *41st International Conference on Machine Learning*. MLResearchPress, 2024.
- [62] Maksims Kazijevs and Manar D Samad. Deep imputation of missing values in time series health data: A review with benchmarking. *Journal of biomedical informatics*, 144:104440, 2023.
- [63] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

- [64] Kosmas Kosmidis, Alkiviadis Kalampokis, and Panos Argyrakis. Language time series analysis. *Physica A: Statistical Mechanics and its Applications*, 370(2):808–816, 2006.
- [65] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.
- [66] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54(1-3):159–178, 1992.
- [67] Baptiste Lafabregue, Jonathan Weber, Pierre Gançarski, and Germain Forestier. End-to-end deep representation learning for time series clustering: a comparative study. *Data mining and knowledge discovery*, 36(1):29–81, 2022.
- [68] Jeffrey Lai, Anthony Bao, and William Gilpin. Panda: A pretrained forecast model for universal representation of chaotic dynamics. *arXiv preprint arXiv:2505.13755*, 2025.
- [69] Zekun Li, Shiyang Li, and Xifeng Yan. Time series as images: Vision transformer for irregularly sampled time series. *Advances in Neural Information Processing Systems*, 36:49187–49204, 2023.
- [70] Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Qingsong Wen, et al. Foundts: Comprehensive and unified benchmarking of foundation models for time series forecasting. 2024.
- [71] Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Christian S Jensen, et al. Tsfm-bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5595–5606, 2025.
- [72] Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Zhengmao Zhu, Guangyi Chen, and Kun Zhang. On the identification of temporally causal representation with instantaneous dependence. *arXiv preprint arXiv:2405.15325*, 2024.
- [73] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6555–6565, 2024.
- [74] Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. *Advances in Neural Information Processing Systems*, 37:106315–106345, 2024.
- [75] Shengsheng Lin, Weiwei Lin, Feiyu Zhao, and Haojun Chen. Benchmarking and revisiting time series forecasting methods in cloud workload prediction. *Cluster Computing*, 28(1):71, 2025.
- [76] Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Prabhakar Karthi, Aditya Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: Multi-domain multimodal dataset for time series analysis. *Advances in Neural Information Processing Systems*, 37:77888–77933, 2024.
- [77] Haoxin Liu, Zhiyuan Zhao, Shiduo Li, and B Aditya Prakash. Evaluating system 1 vs. 2 reasoning approaches for zero-shot time series forecasting: A benchmark and insights. *arXiv preprint arXiv:2503.01895*, 2025.
- [78] Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 37:108231–108261, 2024.
- [79] Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhen-guang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. *Advances in Neural Information Processing Systems*, 36:75354–75371, 2023.

- [80] Xu Liu, Taha Aksu, Juncheng Liu, Qingsong Wen, Yuxuan Liang, Caiming Xiong, Silvio Savarese, Doyen Sahoo, Junnan Li, and Chenghao Liu. Empowering time series analysis with synthetic data: A survey and outlook in the era of foundation models. *arXiv preprint arXiv:2503.11411*, 2025.
- [81] Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025.
- [82] Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pp. 509–525. PMLR, 2022.
- [83] Carl H. Lubba, Simo Sethi, Paul Knaute, Sebastian R. Schultz, Ben D. Fulcher, and Nick S. Jones. catch22: Canonical time-series characteristics. *Data Mining and Knowledge Discovery*, 33:1821–1852, 2019.
- [84] Spyros Makridakis and Michele Hibon. The m3-competition: results, conclusions and implications. *International journal of forecasting*, 16(4):451–476, 2000.
- [85] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: Results, findings, conclusion and way forward. *International Journal of forecasting*, 34(4): 802–808, 2018.
- [86] Claudia Malzer and Marcus Baum. A hybrid approach to hierarchical density-based cluster selection. *arXiv preprint arXiv:1911.02282*, 2019.
- [87] Ameen Maniparambath. Optimizing and benchmarking mamba for network traffic prediction on edge devices. Master’s thesis, A. Maniparambath, 2025.
- [88] Michael W McCracken and Serena Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.
- [89] Nijat Mehdiyev, Johannes Lahann, Andreas Emrich, David Enke, Peter Fettke, and Peter Loos. Time series classification using deep learning for process planning: A case from the process industry. *Procedia Computer Science*, 114:242–249, 2017.
- [90] Guy Mélard and J-M Pasteels. Automatic arima modeling including interventions, using time series expert software. *International Journal of Forecasting*, 16(4):497–508, 2000.
- [91] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.
- [92] Zheyi Pan, Yuxuan Liang, Weifeng Wang, Yong Yu, Yu Zheng, and Junbo Zhang. Urban traffic prediction from spatio-temporal data using deep meta learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1720–1730, 2019.
- [93] John Paparrizos, Chunwei Liu, Aaron J Elmore, and Michael J Franklin. Debunking four long-standing misconceptions of time-series distance measures. In *Proceedings of the 2020 ACM SIGMOD international conference on management of data*, pp. 1887–1905, 2020.
- [94] John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S Tsay, Themis Palpanas, and Michael J Franklin. Tsb-uad: an end-to-end benchmark suite for univariate time-series anomaly detection. *Proceedings of the VLDB Endowment*, 15(8):1697–1711, 2022.
- [95] Peter CB Phillips and Pierre Perron. Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346, 1988.
- [96] Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. *arXiv preprint arXiv:2410.20542*, 2024.

- [97] Rafael Poyatos, Víctor Granda, Roberto Molowny-Horas, Maurizio Mencuccini, Kathy Steppe, and Jordi Martínez-Vilalta. Sapfluxnet: towards a global database of sap flow measurements, 2016.
- [98] Zhongzheng Qiao, Quang Pham, Zhen Cao, Hoang H Le, Ponnuthurai N Suganthan, Xudong Jiang, and Savitha Ramasamy. Class-incremental learning for time series: Benchmark and evaluation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5613–5624, 2024.
- [99] Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S Jensen, Zhenli Sheng, et al. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. *arXiv preprint arXiv:2403.20150*, 2024.
- [100] Xiangfei Qiu, Xiuwen Li, Ruiyang Pang, Zhicheng Pan, Xingjian Wu, Liu Yang, Jilin Hu, Yang Shu, Xuesong Lu, Chengcheng Yang, et al. Easytime: Time series forecasting made easy. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pp. 4564–4567. IEEE, 2025.
- [101] Xiangfei Qiu, Zhe Li, Wanghui Qiu, Shiyan Hu, Lekui Zhou, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Aoying Zhou, Zhenli Sheng, et al. Tab: Unified benchmarking of time series anomaly detection methods. *arXiv preprint arXiv:2506.18046*, 2025.
- [102] Jingang QU, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=0VvDlPmNzM>.
- [103] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.
- [104] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [105] Google Research. Timesfm-2.5-200m: A decoder-only foundation model for time-series forecasting. <https://github.com/DataArcTech/timesfm-pytorch>, 2025. PyTorch implementation. To appear in ICML 2025.
- [106] Thomas Robiglio, Matteo Neri, Davide Coppes, Cosimo Agostinelli, Federico Battiston, Maxime Lucas, and Giovanni Petri. Synergistic signatures of group mechanisms in higher-order systems. *Physical review letters*, 134(13):137401, 2025.
- [107] Michael T. Rosenstein, James J. Collins, and Carlo J. De Luca. A practical method for calculating largest lyapunov exponents from small data sets. *Phys. D*, 65(1–2):117–134, May 1993. ISSN 0167-2789. doi: 10.1016/0167-2789(93)90009-P. URL [https://doi.org/10.1016/0167-2789\(93\)90009-P](https://doi.org/10.1016/0167-2789(93)90009-P).
- [108] Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.
- [109] Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000.
- [110] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- [111] Zezhi Shao, Fei Wang, Yongjun Xu, Wei Wei, Chengqing Yu, Zhao Zhang, Di Yao, Tao Sun, Guangyin Jin, Xin Cao, et al. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

- [112] Zezhi Shao, Yujie Li, Fei Wang, Chengqing Yu, Yisong Fu, Tangwen Qian, Bin Xu, Boyu Diao, Yongjun Xu, and Xueqi Cheng. Blast: Balanced sampling time series corpus for universal forecasting models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2502–2513, 2025.
- [113] Hao-Nan Shi, Ting-Ji Huang, Lu Han, De-Chuan Zhan, and Han-Jia Ye. One-embedding-fits-all: Efficient zero-shot time series forecasting by a model zoo. *arXiv preprint arXiv:2509.04208*, 2025.
- [114] Rolf Simoes, Gilberto Camara, Gilberto Queiroz, Felipe Souza, Pedro R Andrade, Lorena Santos, Alexandre Carvalho, and Karine Ferreira. Satellite image time series analysis for big earth observation data. *Remote Sensing*, 13(13):2428, 2021.
- [115] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *Science*, 338(6106):496–500, 2012.
- [116] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *Science*, 338(6106):496–500, 2012. doi: 10.1126/science.1227079. URL <https://www.science.org/doi/abs/10.1126/science.1227079>.
- [117] Luming Sun, Shijin Gong, Tieying Zhang, Fuxin Jiang, Zhibing Zhao, Jianjun Chen, and Xinyu Zhang. Sufs: A generic storage usage forecasting service through adaptive ensemble learning. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pp. 3168–3181. IEEE, 2023.
- [118] Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. Nts-notears: Learning non-parametric dbns with prior knowledge. *arXiv preprint arXiv:2109.04286*, 2021.
- [119] Souhaib Ben Taieb, Gianluca Bontempi, Amir F Atiya, and Antti Sorjamaa. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. *Expert systems with applications*, 39(8):7067–7083, 2012.
- [120] Yuki Takemoto. Scaling law for large-scale pre-training using chaotic time series and predictability in financial time series. *arXiv preprint arXiv:2509.04921*, 2025.
- [121] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.
- [122] Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1): 37–45, 2018.
- [123] James Theiler, Stephen Eubank, André Longtin, Brian Galdrikian, and J. Doyne Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, 58(1-4): 77–94, 1992.
- [124] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.
- [125] Hao Wang, Zhichao Chen, Zhaoran Liu, Licheng Pan, Hu Xu, Yilin Liao, Haozhe Li, and Xinggao Liu. Spot-i: Similarity preserved optimal transport for industrial iot data imputation. *IEEE Transactions on Industrial Informatics*, 2024.
- [126] Pengfei Wang, Huanran Zheng, Silong Dai, Yiqiao Wang, Xiaotian Gu, Yuanbin Wu, and Xiaoling Wang. A comprehensive survey on emerging techniques and technologies in spatio-temporal eeg data analysis. *Chinese Journal of Information Fusion*, 1(3):183–211, 2024.
- [127] Yuxuan Wang, Haixu Wu, Jiayang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*, 2024.

- [128] Zhixian Wang, Qingsong Wen, Chaoli Zhang, Liang Sun, Leandro Von Krannichfeldt, Shirui Pan, and Yi Wang. Benchmarks and custom package for electrical load forecasting. 2023.
- [129] Zizhao Wang, Jiaheng Hu, Caleb Chuck, Stephen Chen, Roberto Martín-Martín, Amy Zhang, Scott Niekum, and Peter Stone. Skild: Unsupervised skill discovery guided by factor interactions. *Advances in Neural Information Processing Systems*, 37:77748–77776, 2024.
- [130] Gerald Woo, Chenghao Liu, Akshat Kumar, and Doyen Sahoo. Pushing the limits of pre-training for time series forecasting in the cloudops domain. *arXiv preprint arXiv:2310.05063*, 2023.
- [131] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. 2024.
- [132] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.
- [133] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.
- [134] Haixu Wu, Hang Zhou, Mingsheng Long, and Jianmin Wang. Interpretable weather forecasting for worldwide stations with a unified deep model. *Nature Machine Intelligence*, 5(6): 602–611, 2023.
- [135] Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. Catch: Channel-aware multivariate time series anomaly detection via frequency patching. *arXiv preprint arXiv:2410.12261*, 2024.
- [136] Ke Yang, Xin Wang, Quan Zhang, Guangchao Geng, and Quanyuan Jiang. Dynamics enhanced quasi-steady-state model of lcc-hvdc systems based on neural network. *IEEE Transactions on Power Delivery*, 2025.
- [137] Qingren Yao, Chao-Han Huck Yang, Renhe Jiang, Yuxuan Liang, Ming Jin, and Shirui Pan. Towards neural scaling laws for time series foundation models. *arXiv preprint arXiv:2410.12360*, 2024.
- [138] Hao Ye, Ethan R Deyle, Luis J Gilarranz, and George Sugihara. Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific reports*, 5(1):14750, 2015.
- [139] Chengqing Yu, Fei Wang, Zezhi Shao, Tao Sun, Lin Wu, and Yongjun Xu. Dsformer: A double sampling transformer for multivariate time series long-term prediction. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 3062–3072, 2023.
- [140] Umi Kalsom Yusof, Mohd Nor Akmal Khalid, Abir Hussain, and Haziqah Shamsudin. Financial time series forecasting using prophet. In *International Conference of Reliable Information and Communication Technology*, pp. 485–495. Springer, 2020.
- [141] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.
- [142] Ailing Zeng, Muxi Chen, Lei Zhang, Qiang Xu, et al. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. DLinear formulation (linear channel-independent residual).
- [143] Zhen Zeng, Tucker Balch, and Manuela Veloso. Deep video prediction for time series forecasting. In *Proceedings of the Second ACM International Conference on AI in Finance*, pp. 1–7, 2021.

- [144] Da Zhang, Junyu Gao, and Xuelong Li. Learning long-range relationships for temporal aircraft anomaly detection. *IEEE Transactions on Aerospace and Electronic Systems*, 60(5): 6385–6395, 2024.
- [145] Jiawen Zhang, Xumeng Wen, Zhenwei Zhang, Shun Zheng, Jia Li, and Jiang Bian. Propts: Benchmarking point and distributional forecasting across diverse prediction horizons. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 48045–48082. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/55f2a27b1ac39dbfdd0fc83742dc87d7-Paper-Datasets_and_Benchmarks_Track.pdf.
- [146] Jiawen Zhang, Zhenwei Zhang, Shun Zheng, Xumeng Wen, Jia Li, and Jiang Bian. Are time-series foundation models deployment-ready? a systematic study of adversarial robustness across domains. *arXiv preprint arXiv:2505.19397*, 2025.
- [147] Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, 2017.
- [148] Xiangxiang Zhang, Wenkai Hu, Fan Yang, Weihua Cao, and Min Wu. A new transfer entropy approach based on information granulation and clustering for root cause analysis. *Control Engineering Practice*, 140:105669, 2023.
- [149] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*, 2020.
- [150] Zhenwei Zhang, Jiawen Zhang, Shun Zheng, Yuantao Gu, and Jiang Bian. Does cross-domain pre-training truly help time-series foundation models? In *ICLR 2025 Workshop on Foundation Models in the Wild*.
- [151] Xiangtian Zheng, Nan Xu, Loc Trinh, Dongqi Wu, Tong Huang, S Sivaranjani, Yan Liu, and Le Xie. A multi-scale time-series dataset with benchmark for machine learning in decarbonized energy grids. *Scientific Data*, 9(1):359, 2022.

Appendix

A EXTENDED RELATED WORK

Creating meaningful evaluation and training datasets underlies the push for a multivariate time series foundation model, and while main benchmarks remain domain specific, focusing on classic fields such as energy [151; 128], weather and climate [134; 55], traffic [17; 79; 92], finance [88; 58], air quality [147], healthcare and brain signals [32; 62; 126; 96; 136] but also branching out into compute or network related domains [117; 130; 87; 75], cryptocurrency [4], biology [97], industrial devices [125], icicle formation [19], aerospace [144], action identification [49], environment [54]. Alternatively, many benchmarks focus on related problems such as anomaly detection [94; 59; 78; 135; 101], classification [98], robustness [146], etc. The substantial diversity of domains offers both a tantalizing promise and an obvious challenge for time series foundation models: the promise of widespread applicability and the challenge of myriad diversity. Nonetheless, recent work has sought to aggregate and stratify datasets into meaningful benchmarks to assess general time series modeling, including early work such as [84; 119; 85], but subsequently with [3; 47; 9; 1; 71], as well as benchmarks focused on balanced or fair assessment [99; 112], horizon [145], multimodality [76], irregular sampling [15], or dynamical systems [68; 120]. Other work focuses on benchmarking with different foundation models, including [70; 111; 77; 100], while also contributing some compilations of data. While these methods often include multivariate series, they often do not focus on the multivariate problem. First, this can lead to limited volume of multivariate data, and datasets such as those used in [47], [1] and [70] are either entirely or heavily univariate. Furthermore, the analysis of the multivariate datasets in [1] is limited to mostly a small subset.

Of existing time series foundation model datasets, [71] shows the closest resemblance to MuSED-FM, containing multiple real-world time series datasets. Our work subsumes this contemporary work by first including a greater volume of datasets spanning several more domains and time series patterns. Second, including novel datasets by combining related univariate series, transforming sequential data into time series and generating synthetic data to create a more comprehensive datasets, which we discuss further in Section 4. Third, the analysis of variates in [71] uses pearson correlation metrics, which only measure if two series look similar, regardless of how useful the variates are for prediction. By contrast this work introduces a novel multivariate forecasting score that combines Granger Causality [39], transfer entropy [108], convergent cross mapping [138; 116], lagged cross correlation and decomposition-based linear prediction [141], all measures of how the variate can be used for prediction. These tools builds on a rich tradition of time series analysis tools, including collections of metrics such as [83; 14; 2] to modern adaptations [93; 67] to more recent analysis of these tools from a large model perspective [137; 61; 11]. Recent work has investigated other learning metrics for multivariate predictability through dynamics [28; 106], Granger causality [22; 121; 23; 129; 26; 25; 24], causal graphs [82; 20; 21], dynamic bayes networks [118] and transfer entropy [148]. While this work just applies concepts from this field, the analysis applies these concepts on a much larger scale multivariate time series dataset than previously done.

In this work, we add a substantial volume of data in the form of synthetic data and also data derived from classically non-time-series sources: images, video and language. This method for deriving time series follows from understanding time series as timestamped, noisy measurements of a complex, real, dynamic system. From that perspective, time series have been derived from video in Human computer interface [8], video QoE [35], video prediction [143], Satellite images [114], image compression [33], masked autoencoding [18], time series as images [69], Language [64], Time-series derived from video [72]. Despite strong historic ties—especially considering that many classical methods for analyzing images and video such as FFT and wavelet analysis are derived from time series, this work is the first that we know of applying deep learning transformations to generate multivariate time series from these data sources for time series foundation models, a notion which has seen some recent traction [150]. Furthermore, synthetic data is also a classic tool for the training and evaluation of time series [80], with recent support for synthetic data as an effective training tool [56; 6; 68]. This work includes novel implementations built on [56] and [68] to generate a substantial synthetic evaluation dataset which allows the assessment of a wide range of multivariate relationships which are known to be present by construction.

B SUMMARY OF DATA PER DATASET

This section provides per-dataset visualizations of time series, evaluation performance and model forecasts. Additional per-data data summaries and example data from each dataset can be found in the supplementary attachment.

Domain	Dataset	CLASSIC				CAUSAL IMPACT		CHRONOS		UNIVARIATE		TIMESFM				MULTIVARIATE			
		MEAN	PROPHET	SARIMAX	CAUSAL IMPACT	CHRONOS	TaPFN	UNIVARIATE	TOTO	TIMEFM	TaPFN	MIRA	TOTO	TIMEFM	TaPFN	MIRA	TOTO	TIMEFM	
ENERGY	EUROPE ELECTRICITY	0.117 ± 0.046	0.069 ± 0.035	0.108 ± 0.087	0.078 ± 0.033	0.064 ± 0.040	0.058 ± 0.029	0.059 ± 0.032	0.076 ± 0.049	0.057 ± 0.031	0.082 ± 0.037	0.060 ± 0.034	0.073 ± 0.046	0.110 ± 0.087	0.082 ± 0.037	0.060 ± 0.034	0.073 ± 0.046	0.110 ± 0.087	
	DAILY ELECTRICITY	0.134 ± 0.064	0.154 ± 0.082	0.122 ± 0.068	0.078 ± 0.041	0.078 ± 0.045	0.083 ± 0.052	0.097 ± 0.045	0.074 ± 0.040	0.082 ± 0.050	0.082 ± 0.049	0.091 ± 0.039	0.074 ± 0.040	0.074 ± 0.040	0.082 ± 0.049	0.091 ± 0.039	0.074 ± 0.040	0.074 ± 0.040	
	HOURLY ELECTRICITY	0.129 ± 0.054	0.159 ± 0.136	0.213 ± 0.153	0.087 ± 0.059	0.097 ± 0.067	0.107 ± 0.052	0.091 ± 0.053	0.083 ± 0.046	0.093 ± 0.049	0.106 ± 0.052	0.089 ± 0.049	0.084 ± 0.049	0.084 ± 0.049	0.106 ± 0.052	0.089 ± 0.049	0.084 ± 0.049	0.084 ± 0.049	
	MON MICROGRID	7708.626 ± 489.123	124.391 ± 14.456	113.663 ± 21.789	3.099 ± 0.987	0.650 ± 0.104	395.502 ± 47.321	553.344 ± 58.210	57.008 ± 9.476	0.713 ± 0.089	0.870 ± 0.115	553.344 ± 58.210	5.809 ± 1.054	5.809 ± 1.054	0.870 ± 0.115	553.344 ± 58.210	5.809 ± 1.054	5.809 ± 1.054	0.870 ± 0.115
	EUROPE LOAD	0.125 ± 0.021	0.088 ± 0.015	0.182 ± 0.025	0.069 ± 0.010	0.078 ± 0.012	0.078 ± 0.013	0.081 ± 0.014	0.081 ± 0.014	0.117 ± 0.018	0.119 ± 0.019	0.076 ± 0.012	0.077 ± 0.011	0.077 ± 0.008	0.119 ± 0.019	0.076 ± 0.012	0.077 ± 0.011	0.077 ± 0.011	
	SOLAR ENERGY	2.722 ± 0.283	0.752 ± 0.145	6.337 ± 1.026	0.144 ± 0.035	0.302 ± 0.062	0.307 ± 0.061	1.253 ± 0.219	1.253 ± 0.222	2.771 ± 0.558	2.770 ± 0.562	0.282 ± 0.086	1.343 ± 0.276	1.343 ± 0.276	2.771 ± 0.558	0.282 ± 0.086	1.343 ± 0.276	1.343 ± 0.276	
	AUSTRALIA ENERGY	0.225 ± 0.035	0.247 ± 0.023	0.139 ± 0.108	0.110 ± 0.029	0.133 ± 0.031	0.171 ± 0.036	0.139 ± 0.030	0.141 ± 0.027	0.178 ± 0.033	0.146 ± 0.027	0.149 ± 0.029	0.136 ± 0.025	0.136 ± 0.024	0.146 ± 0.027	0.149 ± 0.029	0.136 ± 0.025	0.136 ± 0.024	
	SPAN ENERGY	0.153 ± 0.025	0.085 ± 0.018	0.100 ± 0.015	0.070 ± 0.012	0.052 ± 0.010	0.069 ± 0.011	0.082 ± 0.014	0.025 ± 0.007	0.061 ± 0.011	0.068 ± 0.012	0.079 ± 0.014	0.012 ± 0.004	0.012 ± 0.001	0.068 ± 0.012	0.079 ± 0.014	0.012 ± 0.004	0.012 ± 0.001	
	CASUAL RIVERS	2.263 ± 0.530	1.505 ± 0.220	1.085 ± 0.170	0.806 ± 0.120	2.450 ± 0.380	1.114 ± 0.210	2.424 ± 0.400	1.092 ± 0.180	0.667 ± 0.110	1.522 ± 0.250	2.740 ± 0.760	1.101 ± 0.170	0.487 ± 0.046	1.522 ± 0.250	2.740 ± 0.760	1.101 ± 0.170	0.487 ± 0.046	
CLIMATE & ENVIRONMENT	CHOKAL WEATHER	0.567 ± 0.045	0.197 ± 0.030	0.422 ± 0.051	0.219 ± 0.033	0.215 ± 0.034	0.316 ± 0.048	0.214 ± 0.032	0.147 ± 0.027	0.318 ± 0.049	0.318 ± 0.049	0.318 ± 0.049	0.318 ± 0.049	0.318 ± 0.049	0.318 ± 0.049	0.318 ± 0.049	0.318 ± 0.049	0.318 ± 0.049	
	BEHIND EMBASSY	2.152 ± 0.432	2.235 ± 0.410	2.757 ± 0.498	3.833 ± 0.475	1.877 ± 0.176	2.022 ± 0.487	1.768 ± 0.354	0.772 ± 0.153	3.109 ± 0.512	2.955 ± 0.491	1.864 ± 0.361	0.456 ± 0.100	0.456 ± 0.100	2.955 ± 0.491	1.864 ± 0.361	0.456 ± 0.100	0.456 ± 0.100	
	BEHIND AQ	3.173 ± 0.512	2.083 ± 0.402	2.812 ± 0.534	6.472 ± 0.645	2.873 ± 0.457	3.322 ± 0.498	1.854 ± 0.376	1.078 ± 0.185	2.174 ± 0.410	3.320 ± 0.500	1.990 ± 0.388	0.613 ± 0.131	0.613 ± 0.131	3.320 ± 0.500	1.990 ± 0.388	0.613 ± 0.131	0.613 ± 0.131	
	OPEN AQ	4.797 ± 1.025	0.202 ± 0.050	2.718 ± 0.563	1.261 ± 0.291	1.162 ± 0.228	0.006 ± 1.083	1.054 ± 0.211	1.058 ± 0.212	1.234 ± 0.241	0.890 ± 0.202	0.890 ± 0.202	0.890 ± 0.202	0.890 ± 0.202	0.890 ± 0.202	0.890 ± 0.202	0.890 ± 0.202	0.890 ± 0.202	
	GAS SENSOR	0.541 ± 0.085	0.324 ± 0.050	0.748 ± 0.090	0.443 ± 0.085	0.444 ± 0.066	0.553 ± 0.075	0.491 ± 0.070	0.541 ± 0.085	0.780 ± 0.090	0.495 ± 0.070	0.484 ± 0.070	0.483 ± 0.062	0.483 ± 0.062	0.495 ± 0.070	0.484 ± 0.070	0.483 ± 0.062	0.483 ± 0.062	
ENGINEERING	RTV SENSOR	0.417 ± 0.065	0.426 ± 0.050	0.412 ± 0.078	0.413 ± 0.060	0.280 ± 0.045	0.429 ± 0.065	0.429 ± 0.065	0.260 ± 0.042	0.417 ± 0.063	0.260 ± 0.042	0.417 ± 0.063	0.260 ± 0.042	0.260 ± 0.042	0.417 ± 0.063	0.260 ± 0.042	0.417 ± 0.063	0.260 ± 0.042	
	VOIP	0.018 ± 0.007	0.007 ± 0.003	0.007 ± 0.005	0.006 ± 0.004	0.005 ± 0.002	0.006 ± 0.003	0.005 ± 0.003	0.004 ± 0.002	0.011 ± 0.004	0.006 ± 0.003	0.005 ± 0.003	0.004 ± 0.002	0.004 ± 0.002	0.006 ± 0.003	0.005 ± 0.003	0.004 ± 0.002	0.004 ± 0.002	
FINANCE	FRED-MD	2.512 ± 0.500	0.523 ± 0.080	0.645 ± 0.077	0.586 ± 0.088	0.599 ± 0.055	2.437 ± 0.312	0.438 ± 0.058	0.130 ± 0.052	2.314 ± 0.445	2.425 ± 0.507	0.601 ± 0.079	0.386 ± 0.063	0.386 ± 0.063	2.425 ± 0.507	0.601 ± 0.079	0.386 ± 0.063	0.386 ± 0.063	
	BITCOIN PRICES	0.873 ± 0.150	1.590 ± 0.280	1.997 ± 0.320	0.540 ± 0.090	0.604 ± 0.092	0.571 ± 0.084	0.441 ± 0.075	0.443 ± 0.060	0.944 ± 0.090	0.606 ± 0.085	0.601 ± 0.085	0.425 ± 0.070	0.425 ± 0.043	0.606 ± 0.085	0.601 ± 0.085	0.425 ± 0.070	0.425 ± 0.043	
HEALTH	COM	0.394 ± 0.151	1.106 ± 0.783	0.571 ± 0.280	0.423 ± 0.142	0.443 ± 0.177	0.398 ± 0.162	0.415 ± 0.181	0.446 ± 0.132	0.394 ± 0.152	0.399 ± 0.160	0.419 ± 0.136	0.446 ± 0.132	0.473 ± 0.121	0.399 ± 0.160	0.419 ± 0.136	0.446 ± 0.132	0.473 ± 0.121	
	SHAP-LAB	0.425 ± 2.149	8.331 ± 38.380	2.496 ± 10.067	4.109 ± 21.217	0.425 ± 2.149	4.662 ± 29.700	7.821 ± 36.395	0.442 ± 2.156	2.230 ± 12.807	0.610 ± 30.020	7.492 ± 35.212	1.091 ± 5.127	10.673 ± 40.426	0.610 ± 30.020	7.492 ± 35.212	1.091 ± 5.127	10.673 ± 40.426	
PUBLIC INFO	TAC	3.590 ± 3.211	9.566 ± 13.125	10.030 ± 17.200	67.261 ± 122.854	0.071 ± 8.225	3.990 ± 3.000	12.507 ± 16.448	7.365 ± 10.528	9.662 ± 9.482	3.900 ± 3.000	19.301 ± 28.149	9.153 ± 15.729	12.000 ± 10.000	3.900 ± 3.000	19.301 ± 28.149	9.153 ± 15.729	12.000 ± 10.000	
	TAC	1.438 ± 12.436	3.472 ± 0.000	3.677 ± 17.904	11.859 ± 20.000	3.146 ± 13.223	1.475 ± 12.888	1.805 ± 15.727	3.693 ± 15.715	3.472 ± 15.618	1.476 ± 12.906	1.831 ± 15.855	3.551 ± 15.567	4.200 ± 18.000	1.476 ± 12.906	1.831 ± 15.855	3.551 ± 15.567	4.200 ± 18.000	
	MTA RIDEBOOK	0.266 ± 0.049	0.213 ± 0.036	0.360 ± 0.114	0.194 ± 0.297	0.175 ± 0.040	0.330 ± 0.130	0.268 ± 0.030	0.178 ± 0.035	0.177 ± 0.068	0.216 ± 0.132	0.227 ± 0.038	0.175 ± 0.036	0.180 ± 0.070	0.216 ± 0.132	0.227 ± 0.038	0.175 ± 0.036	0.180 ± 0.070	
	PARKS MOBILITY	0.172 ± 0.017	0.062 ± 0.023	0.141 ± 0.058	0.110 ± 0.010	0.059 ± 0.021	0.086 ± 0.013	0.105 ± 0.009	0.048 ± 0.012	0.045 ± 0.018	0.046 ± 0.019	0.046 ± 0.019	0.046 ± 0.019	0.046 ± 0.019	0.046 ± 0.019	0.046 ± 0.019	0.046 ± 0.019	0.046 ± 0.019	
	REDEARTH	0.023 ± 0.015	0.033 ± 0.016	1.281 ± 0.006	0.033 ± 0.016	0.988 ± 0.074	0.222 ± 0.119	0.401 ± 0.134	0.981 ± 0.013	0.570 ± 0.131	0.255 ± 0.145	0.401 ± 0.134	0.989 ± 0.014	0.570 ± 0.131	0.255 ± 0.145	0.401 ± 0.134	0.989 ± 0.014	0.570 ± 0.131	
	BLUE RIKES	0.119 ± 0.040	0.434 ± 0.270	0.461 ± 0.731	0.297 ± 0.861	0.434 ± 0.270	0.245 ± 0.924	0.773 ± 3.423	0.688 ± 0.055	0.510 ± 0.138	0.479 ± 0.236	0.766 ± 3.757	0.682 ± 0.032	0.522 ± 0.297	0.479 ± 0.236	0.766 ± 3.757	0.682 ± 0.032	0.522 ± 0.297	
	ASTIN WATER	0.076 ± 0.014	0.095 ± 0.055	0.233 ± 0.04	0.135 ± 0.088	0.119 ± 0.036	0.091 ± 0.001	0.094 ± 0.001	0.087 ± 0.011	0.089 ± 0.022	0.087 ± 0.011	0.089 ± 0.022	0.087 ± 0.011	0.089 ± 0.022	0.087 ± 0.011	0.089 ± 0.022	0.087 ± 0.011	0.089 ± 0.022	
	TRATIC	0.184 ± 0.114	0.167 ± 0.180	0.379 ± 0.202	0.325 ± 0.195	0.379 ± 0.202	0.325 ± 0.195	0.379 ± 0.202	0.325 ± 0.195	0.379 ± 0.202	0.325 ± 0.195	0.379 ± 0.202	0.325 ± 0.195	0.379 ± 0.202	0.325 ± 0.195	0.379 ± 0.202	0.325 ± 0.195	0.379 ± 0.202	
	CURRENT WATER	0.069 ± 0.139	1.478 ± 0.200	0.666 ± 0.215	2.156 ± 0.250	2.156 ± 0.250	2.156 ± 0.250	2.156 ± 0.250	2.156 ± 0.250	2.156 ± 0.250	2.156 ± 0.250	2.156 ± 0.250	2.156 ± 0.250	2.156 ± 0.250	2.156 ± 0.250	2.156 ± 0.250	2.156 ± 0.250	2.156 ± 0.250	
SALES	WALMART SALES	0.041 ± 0.020	0.041 ± 0.021	0.045 ± 0.021	0.043 ± 0.021	0.043 ± 0.021	0.049 ± 0.021	0.050 ± 0.027	0.050 ± 0.027	0.057 ± 0.027	0.064 ± 0.027	0.064 ± 0.026	0.065 ± 0.028	0.074 ± 0.035	0.064 ± 0.026	0.065 ± 0.028	0.074 ± 0.035	0.074 ± 0.035	
	ALCOHOL SALES	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	
	PAYROLL SALES	0.527 ± 0.710	0.531 ± 0.514	0.539 ± 0.174	0.577 ± 0.181	0.608 ± 0.167	0.609 ± 0.340	0.613 ± 0.341	0.613 ± 0.145	0.713 ± 0.609	0.687 ± 1.263	0.693 ± 0.746	1.023 ± 1.438	0.900 ± 0.775	0.687 ± 1.263	0.693 ± 0.746	1.023 ± 1.438	0.900 ± 0.775	
	RETAIL SALES	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	0.007 ± 0.004	
WEB	GOLD PRICES	0.589 ± 0.064	0.122 ± 0.077	0.074 ± 0.027	0.080 ± 0.058	0.079 ± 0.034	0.073 ± 0.031	0.150 ± 0.092	0.082 ± 0.024	0.074 ± 0.025	0.043 ± 0.063	0.183 ± 0.079	0.091 ± 0.025	0.074 ± 0.025	0.043 ± 0.063	0.183 ± 0.079	0.091 ± 0.025	0.074 ± 0.025	
	WEB VISITORS	0.288 ± 0.049	0.163 ± 0.033	0.040 ± 0.162	0.031 ± 0.011	0.167 ± 0.047	0.133 ± 0.048	0.289 ± 0.062	0.156 ± 0.033	0.141 ± 0.038	0.289 ± 0.057	0.287 ± 0.058	0.149 ± 0.012	0.130 ± 0.048	0.289 ± 0.057	0.287 ± 0.058	0.149 ± 0.012	0.130 ± 0.048	
SCIENTIFIC	MICROSD	1.346 ± 0.824	1.695 ± 1.269	6.115 ± 6.839	456.120 ± 1.94	8.931 ± 14.442	0.472 ± 2.28	2.116 ± 2.522	5.197 ± 7.899	3.434 ± 3.657	4.310 ± 3.194	2.080 ± 2.819	5.552 ± 7.924	3.434 ± 3.657	4.310 ± 3.194	2.080 ± 2.819	5.552 ± 7.924	3.434 ± 3.657	
	STREETSIDE	1.687 ± 0.514	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	2.430 ± 0.947	
IMAGES	CIF100	0.290 ± 1.124	0.424 ± 0.066	1.361 ± 0.513	0.842 ± 0.145	2.141 ± 0.987	8.958 ± 0.122	2.973 ± 0.744	1.877 ± 2.355	1.871 ± 1.210	2.259 ± 0.814	2.259 ± 0.814	2.259 ± 0.814	2.259 ± 0.814	2.259 ± 0.814	2.259 ± 0.814	2.259 ± 0.814	2.259 ± 0.814	
	TEXT OPENRETEXT	2.641 ± 0.590	3.035 ± 0.102	1.541 ± 2.806	1.551 ± 0.875	1.346 ± 0.988	1.756 ± 0.912	2.534 ± 0.879	1.671 ± 0.865	1.699 ± 3.599	2.970 ± 1.029	2.121 ± 0.895	1.664 ± 0.900	1.997 ± 0.683	2.970 ± 1.029	2.121 ± 0.895	1.664 ± 0.900	1.997 ± 0.683	

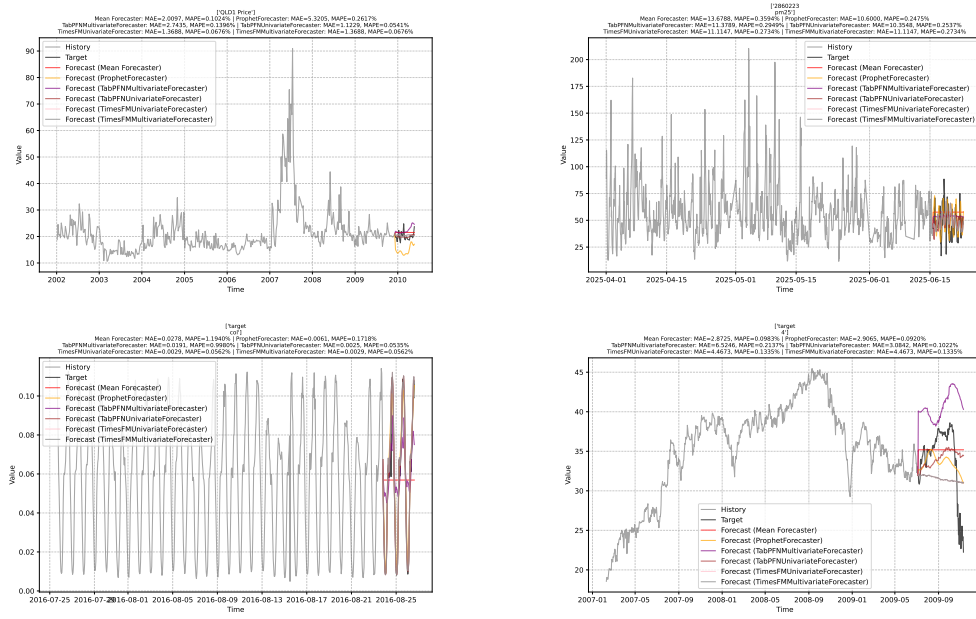


Figure 5: Cross-domain samples II: electricity, air quality, traffic, dynamics

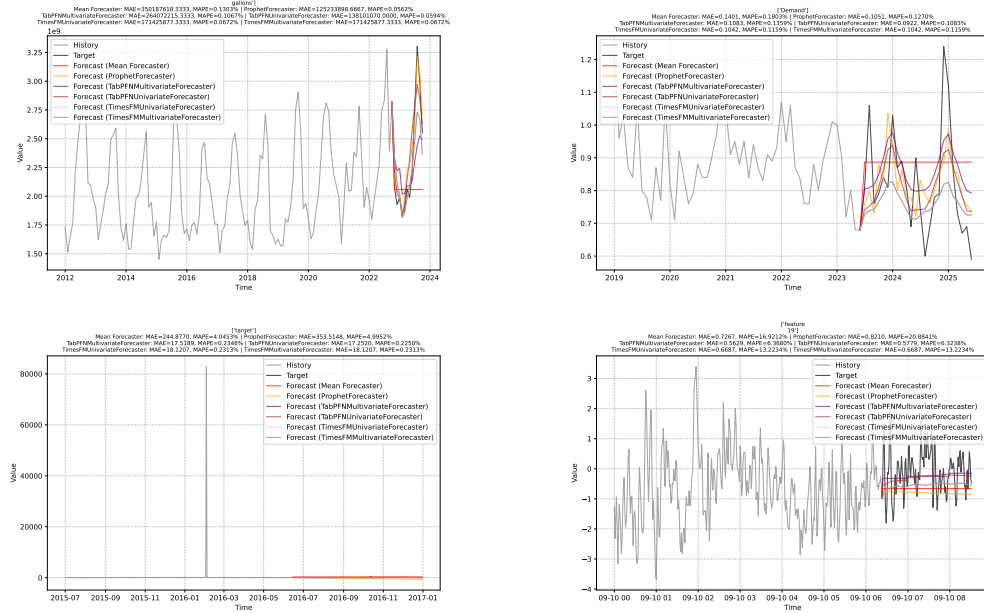


Figure 6: Cross-domain samples III: water, regional electricity, wikipedia, language

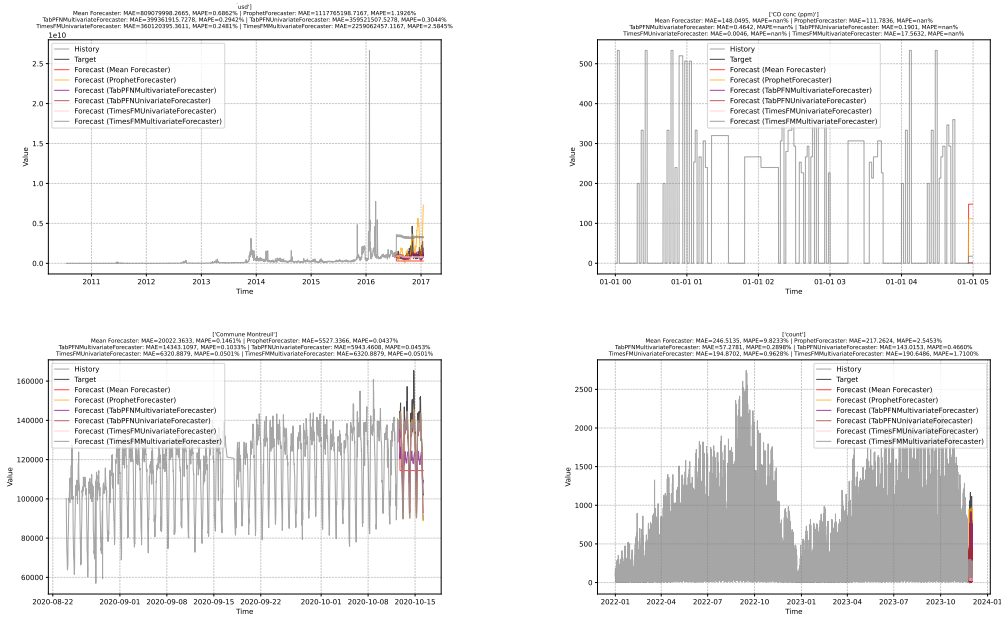


Figure 7: Cross-domain samples IV: Bitcoin price, gas sensors, mobility, bicycle sales

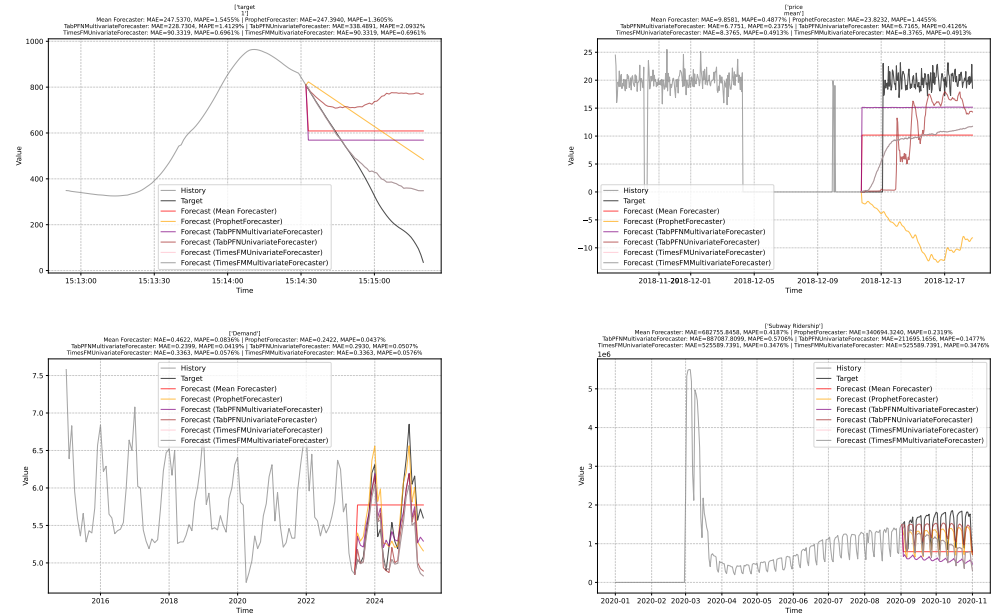


Figure 8: Cross-domain samples V: Video, rideshare, power, transit

B.2 FULL MULTIVARIATE SERIES

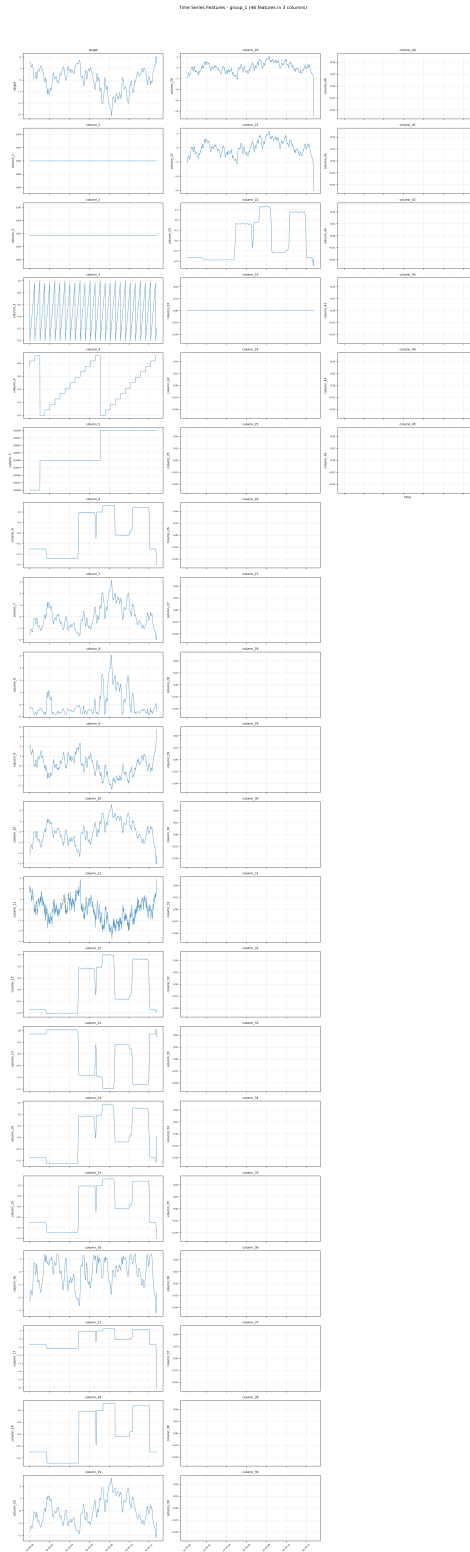


Figure 9: Synthetic Causal Data

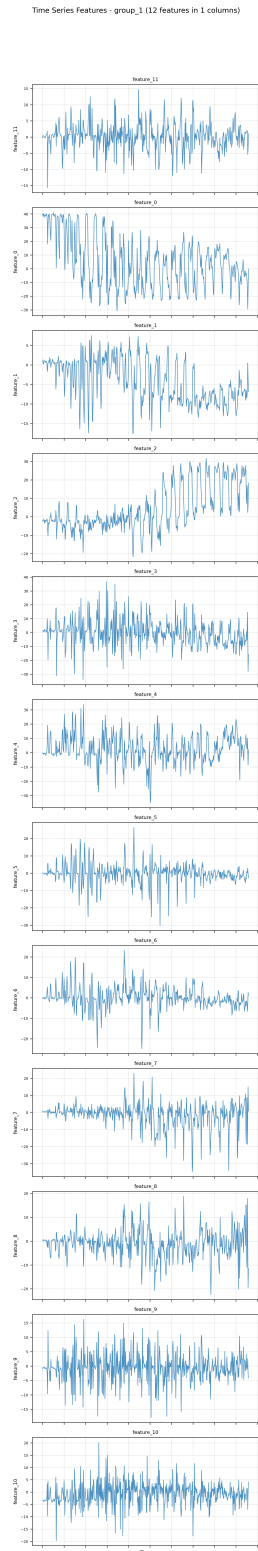


Figure 10: Sequential CIFAR-100 image data

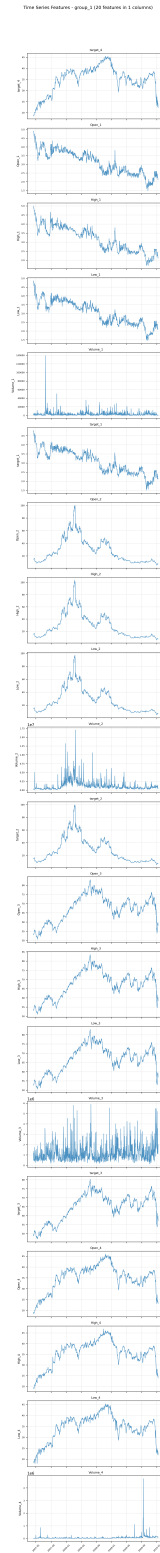


Figure 11: Collected Stock data

Time Series Features - group_1 (6 features in 1 columns)

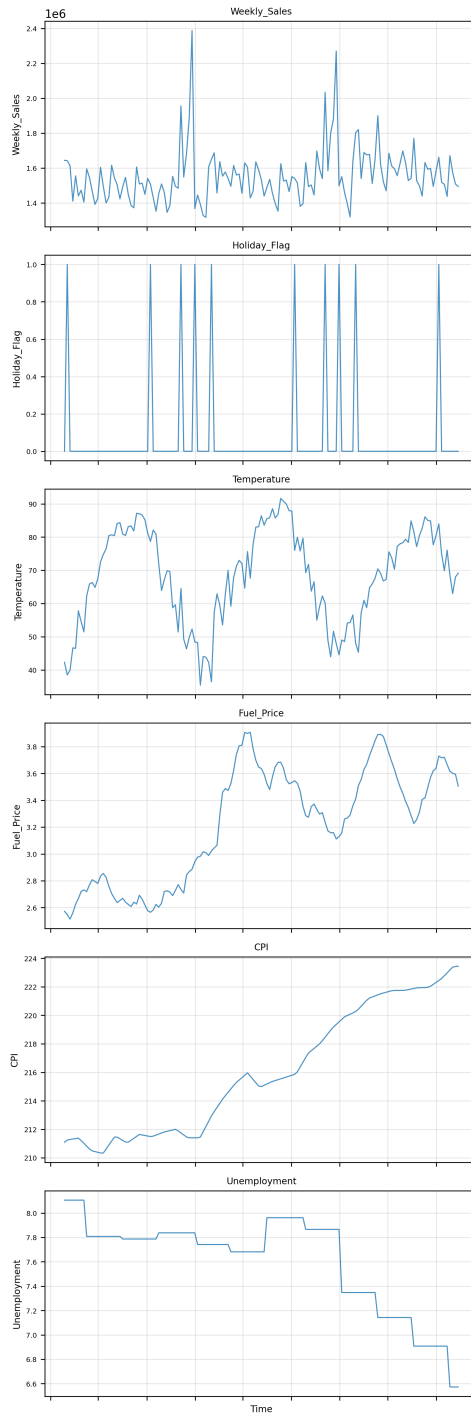


Figure 12: Walmart Sales Dataset

C COMBINING COLLECTIONS OF RELATED UNIVARIATE TIME SERIES

C.1 STOCK DATA

We collect historical daily stock price data for the tickers currently trading on the NASDAQ exchange. The list of active tickers is obtained from Nasdaq Trader, and the historical time series are downloaded from Yahoo Finance via the `yfinance` Python package. For each ticker, data are saved as CSV with the following fields

- Date: trading day.
- Open: opening price.
- High: highest price during the day.
- Low: lowest price during the day.
- Adj Close: closing price adjusted for both dividends and splits.
- Volume: the number of shares traded that day.

For our evaluation, we select the period 2016-04-05 $\rightarrow$ 2020-04-01 and include 2347 tickers. We group companies by type and construct sequences in which each sequence contains five stocks. The prediction target is the **Close** price for one of the stocks, randomly selected.

C.2 WIKIMEDIA DATA

We collect daily Wikipedia pageview data for 145,000 pages over the period 2015-07-01 $\rightarrow$ 2016-12-31. For each page, semantic categories are obtained from the Wikipedia API category query interface. Categories that appear in more than 5% of all pages are removed to avoid generic groupings. Each page is then represented as text by concatenating its title with the remaining categories.

We generate embeddings for these text representations using Sentence-BERT [104]. Clustering is performed with HDBSCAN [86] to group pages by semantic similarity.

From the clusters, we construct multivariate time series datasets. Clusters with more than 30 pages are split into groups of 30 pages each. Each group forms a multivariate series with 30 covariates, where each covariate is the daily pageview count for one page. This process yields 1,530 multivariate series in total.

D SEQUENTIAL DATA TRANSFORMATION PROCESS

Multivariate time series foundation models must learn to identify patterns from the context and leverage them to predict future values. Thus, when evaluating the capability of these models it can be useful to introduce series that are likely to contain varied correlational patterns from various sources. This is the motivation for transforming data from different sequential processes. In this section we describe the process that we apply to transform data from image, video, language, and simulator domains:

1. Transform into a meaningful sequential space using a tokenizer. This can vary for different domains.
 - (a) For images, apply spiral patching with a visual transformer [132] using images from CIFAR-100 [65] by resizing the images to 450×450 , then using a patch size of 16 with a stride of 8.
 - (b) For scientific data, we took two simulators, Spriteworld [26], which uses a box2d physics simulation backend where the number of objects and their initial positions and velocities are randomized, and Mujoco [124] data collected using policies of varying capability in D4RL [40]. Each episode lasts 1000 time steps, and since the state for these simulators is already low-dimensional, we do not need to apply additional processing
 - (c) For Text, we take the openwebtext dataset [48], and apply an open sourced language model [149] to convert the words into vectors.
 - (d) For video data, we apply keypoint detection to the KITTI dataset [45], using the keypoint positions as the states over the course of a time series.

2. Apply dimensionality reduction in the form of PCA to the sequencized data to reduce it to a meaningful form. Given the set of tokens, this takes a substantial random sample to apply SVD, then uses the learned matrix for dimensionality reduction. The number of dimensions taken by PCA is tuned to allow for 80-90% of the variance to be explained by the features. For images, this is 12 dimensions, and for text, this is 20 dimensions. Video and Simulator data require no additional dimensionality reduction.
3. Apply smoothing and/or noise, where smoothing is applied in the form of a Gaussian filter, where the window size is tuned based on the noisiness of the data. Noise is re-added from a normal distribution. For images and text, the Gaussian filter used a window size of 5, whereas for D4RL data, we used a window size of 20, though for Spriteworld, we do not apply filtering. We reapply 3% noise, which gives the files similar characteristics to time series from other domains. For video we do not apply either filtering or noise.

OpenWebText

EAST ST. LOUIS, Ill., Nov. 1 (UPI) — Two women and a man have been charged with abusing two young girls during a sex party at an Indiana hotel.

Federal prosecutors in southern Illinois say Tabitha Robinson of Midlothian, Ill., brought a 3-year-old in her care to the party, the Chicago Tribune reported. Louise Helen Masulla of New Athens, Ill., allegedly brought an 11-year-old.

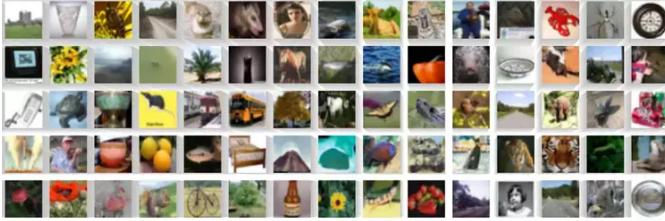
The two women and William Milligan of Bloomington Ind., were charged with conspiracy to produce child pornography and conspiracy to transport minors to engage in criminal sexual activity.

Investigators say the three met online and then at a Best Western in Terre Haute. They allegedly found pictures on the two women's computers that showed the children engaged in sexual activity.

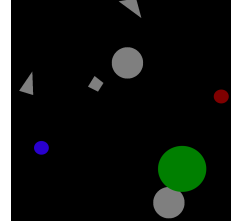
Shakespeare famously wrote that people are the stuff dreams are built on. For Sacramento, its dreams of Major League Soccer will be built on 244 acres of undeveloped land in the city's downtown RailYards.

The historic site, a pocket of low-lying, swampy land located where the American and Sacramento Rivers meet, was once the western terminus of the Transcontinental Railroad in the 1860s, but has lain dormant since Union Pacific Railroad workers ceased operations in 1995. A proposed effort to redevelop the area in 2007 was foiled by the economic recession that struck the country at the time, but true to the city's indomitable spirit, a new group of investors has resumed the task of bringing life to the area.

CIFAR-100



Spriteworld



KITTI



MuJoCo

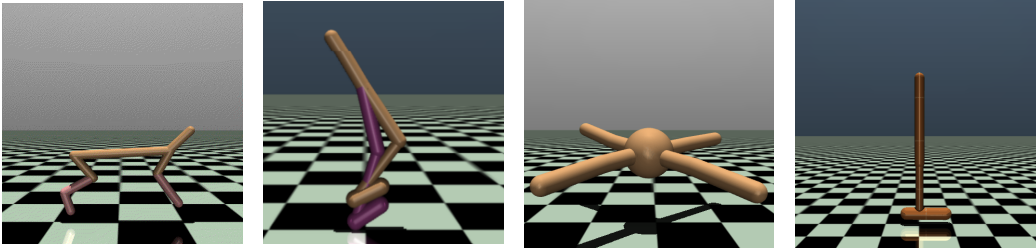


Figure 13: Examples on selected datasets, including OpenWebText [48], CIFAR-100 [65], Spriteworld [26], KITTI [45], and MuJoCo [40].

Altogether, this set of three steps can be applied on a wide range of data. In fact, the datasets we selected were smaller than the general corpus of image, video, language or simulator data, and future work can explore applying this process to generate more multivariate time series data.

E STRUCTURAL CAUSAL MODELS FOR SYNTHETIC DATA

The structural causal model strategy for representation learning involves constructing a graph, where each node is a variate and the value at a node is a function of its parents in the graph. Then the root nodes take on values according to some sequentially correlated system (ex. ARIMA model, random fourier features, etc.). At each timestamp the values are then propagated through the graph. A consistent subset of the nodes are used as metadata, and one node is selected as the target variable. The metadata nodes are then lagged in time by a random amount to ensure that the target variable is a function of the history of the metadata.

Formally, the data distribution is dictated by g_{SCM} , constructed in practice using MLPs and loosely derived from [56; 102]. Each dataset is described by a graph, where each node in the graph is a variate. The set of variates that are not caused by other variates are the inputs s_{in} , a set of intermediate nodes, some of which are hidden s_{hidden} , a set of observed nodes x and target nodes y , and a set of noise nodes u_{hidden} . An SCM is defined by $g_{SCM}(\mathcal{S}, \mathcal{U}, \mathcal{E})$, where $\mathcal{E}$ are the set of directed edges arranged in a directed acyclic graph. The input nodes are independently sampled according to some noise distribution S_{in} , the noise nodes are sampled according to U_{hidden} , and the hidden nodes are computed as a function of the parents in the graph, $s_i := g_i(\text{PA}_i(\mathbf{s}), u_i)$. In practice, g_i

is represented with nonlinearity activation function $a : \mathbb{R} \rightarrow \mathbb{R}$, a weight vector $w^{|\text{PA}|}$ according to $s_i = a(w^\top \text{PA}_i(\mathbf{s})) + u_i$. For this work, we sample the noise correlated in time, where we use the following set of functions:

1. ARMA functions with random AR and MA components
2. Random piecewise linear functions with between 8-20 pieces.
3. Impulse functions where the impulses are sampled at random times, between 8-20 impulses
4. Random fourier features, with between 8-20 frequencies sampled
5. Weiner Distribution with trend component
6. Random piecewise splines, with between 8-20 pieces.

For non-input nodes, the noise parameter is 0.01% Gaussian noise.

The observed nodes and target nodes are selected randomly from the graph, and a lag operator is applied to the observations of the observed node, so that without loss of generality if metadata node x_i is the only parent of y , then $y = f(x_i^t)$, thus ensuring a lagged relationship between the target and the metadata node.

F DYNAMICAL SYSTEM SYNTHETIC DATA (SKEW-PRODUCT GENERATION)

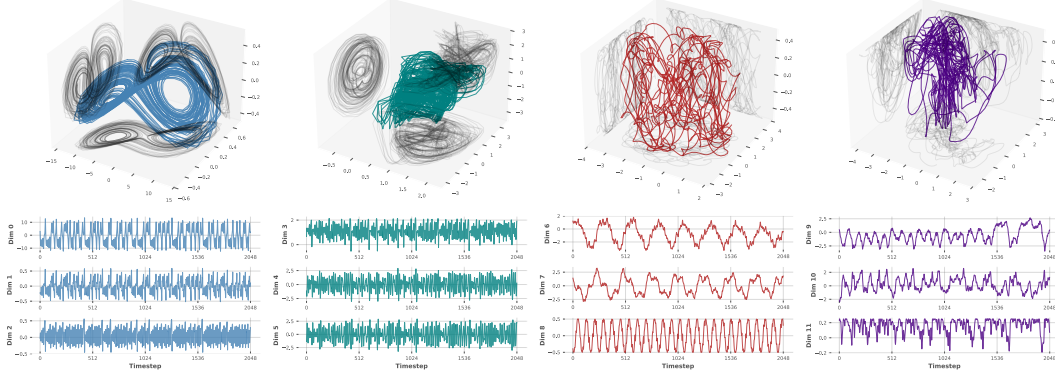


Figure 14: Example skew-product dynamical system after two generations of recombination.

Starting with a founder pool of 129 chaotic dynamical systems from the `dysts` dataset [46], we randomly perturb the parameters of each in a Gaussian ball around the default value. In Generation 1, we sample pairs of perturbed systems (which we denote the driver and the response system) and couple their flows together in a skew-product form. In particular, the coupling is unidirectional and defines the influence of the driver system on the response system. For each candidate perturbed skew system that successfully integrates, we apply a suite of attractor tests to eliminate trajectories that are converging to a fixed point, diverging to infinity, limit cycles, or simple straight lines. We then apply the 0-1 test [37] to eliminate periodic and quasiperiodic systems. We check for a continuous broadband power spectrum, as well as positive leading Lyapunov exponent via the data-driven Rosenstein estimator [107]. Lastly, we check for stationarity via a combination of the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) [66] and Augmented Dickey-Fuller (ADF) [34] tests. In Generation 2, we sample pairs of successful perturbed skew systems and couple them together in the same fashion as before. By construction, we have a combinatorial growth in the number of possible distinct pairings of systems to couple together. This extends the framework of [68] and allows us to create higher-dimensional systems with arbitrary causal relationships. Additional dynamical diversity can be introduced via the choice of coupling maps.

Let $m, n \in \mathbb{N}$ denote the driver and response system dimension, respectively. Suppose we have a driver system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ and a response system $\dot{\mathbf{y}} = \mathbf{g}(\mathbf{y})$. A simple coupling strategy is an *Additive Coupling Map* in which we randomize the dimension ordering of the driver system, normalize the flows by their RMS, and then simply add the driver flow to the response flow $\dot{\mathbf{y}} = \mathbf{g}(\mathbf{y}) + \dot{\mathbf{x}}$. Concatenating dimensions, the flow of the combined skew product system is simply $[\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{y})]$, modulo the normalization and randomization of dimension ordering.

Given a desired rank r , we provide the option to construct the skew coupling matrix C as a low-rank positive semidefinite perturbation. Randomizing the dimension ordering of the driver system,

$$C = \begin{bmatrix} D & I_n + VSV^\top \end{bmatrix}, \quad D = (I_{\max\{n, m\}})_{1:n, 1:(m)},$$

$$V \in \mathbb{R}^{n \times r} \text{ with } V_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \quad \|v_k\|_2 = 1 \text{ (column-normalized). And } S = \text{diag}(s_1, \dots, s_r), \quad s_k \sim \mathcal{U}[0, 1].$$

Our *Activated Coupling Map* first applies this coupling matrix to the driver and response flows. As before, we normalize each flow by its RMS, but before adding them together, we pass the transformed driver flow through a nonlinearity (we choose a hyperbolic tangent function).

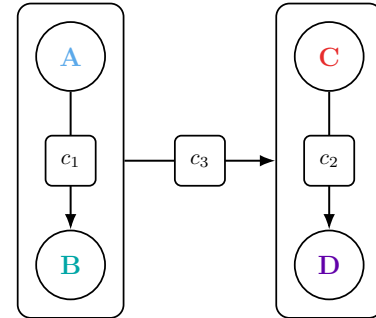
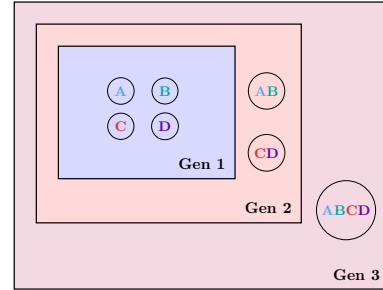


Figure 15: Skew-Product Generation

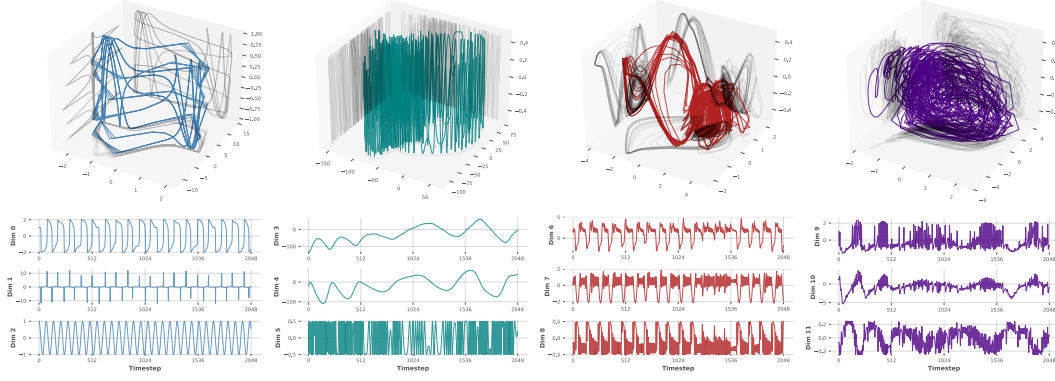


Figure 16: Example skew-product dynamical system after two generations of recombination.

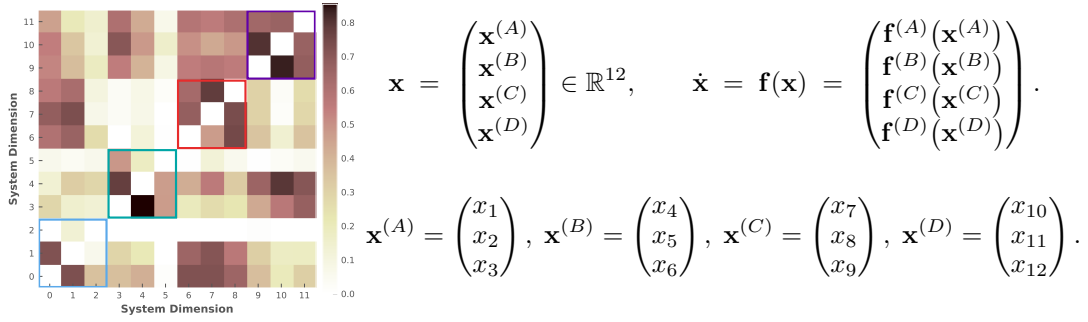


Figure 17: CCM Map

$$\phi(u) := \frac{1}{2}(|u+1| - |u-1|).$$

Flow RHS of Component Systems

$$\mathbf{f}^{(A)}(\mathbf{x}^{(A)}) = \begin{pmatrix} x_2 \\ \mu(1 - x_1^2)x_2 - x_1 + a \sin(x_3) \\ \omega \end{pmatrix} \quad (\text{ForcedVanDerPol}),$$

$$\mathbf{f}^{(B)}(\mathbf{x}^{(B)}) = \begin{pmatrix} a x_5 - a x_4 + d x_4 x_6 \\ k x_4 + f x_5 - x_4 x_6 \\ c x_6 + x_4 x_5 - \varepsilon x_4^2 \end{pmatrix} \quad (\text{Tsucs2}),$$

$$\mathbf{f}^{(C)}(\mathbf{x}^{(C)}) = \begin{pmatrix} -x_7 + d\phi(x_7) - b\phi(x_8) - b\phi(x_9) \\ -x_8 - b\phi(x_7) + c\phi(x_8) - a\phi(x_9) \\ -x_9 - b\phi(x_7) + a\phi(x_8) + \phi(x_9) \end{pmatrix} \quad (\text{CellularNeuralNetwork}),$$

$$\mathbf{f}^{(D)}(\mathbf{x}^{(D)}) = \begin{pmatrix} -x_{10} + t_x^{-1}x_{11} - \frac{a}{t_x}x_{10}^3 + \frac{b}{t_x}x_{10}^2 + t_x^{-1}x_{12} \\ -a x_{10}^3 - (d-b)x_{10}^2 + x_{12} \\ -\frac{s}{t_z}x_{10} - t_z^{-1}x_{12} + \frac{c}{t_z} \end{pmatrix} \quad (\text{HindmarshRose}).$$

System Parameters

$$\text{ForcedVanDerPol: } (a, \mu, \omega) = (1.2, 8.53, 0.63)$$

$$\text{Tsucs2: } (a, c, d, \varepsilon, f, k) = (40, 0.833, 0.5, 0.65, 20, 0)$$

$$\text{CellularNeuralNetwork: } (a, b, c, d) = (4.4, 3.21, 1.1, 1.24)$$

$$\text{HindmarshRose: } (a, b, c, d, s, t_x, t_z) = (0.49, 1.0, 0.0322, 1.0, 1.0, 0.03, 0.8)$$

G LLM USAGE STATEMENT

LLMs were not involved in the writing or ideation of this work. Cursor was used as a coding IDE, and LLM tab completions were used to implement various functions, but the data and concepts were generated without LLM usage (except for the openwebtext dataset, which used a language embedder as part of the system in Appendix D).