NEURAL INFORMATION PROCESSING SYSTEMS

Paper Code Contact

Synthetic Series–Symbol Data Generation for Time Series Foundation Models

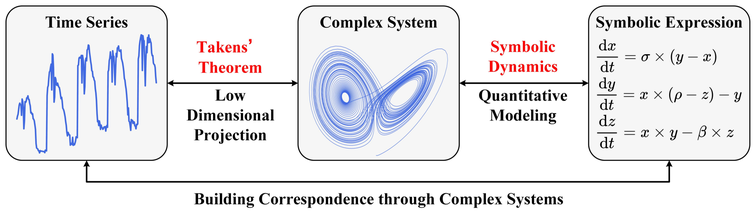
Wenxuan Wang, Kai Wu*, Yujian Betterest li, Dan Wang*, Xiaoyu Zhang

whenxuanwang@stu.xidian.edu.cn, {kwu, danwang}@xidian.edu.cn

Inspired by the complex dynamical systems, we introduce a dual-modal mechanism for unrestricted generation of high-quality time series and symbolic expressions, to mitigate the data scarcity and distribution imbalance. Pre-trained on this synthetic dataset, our model SymTime is competitive with real-data pre-trained models across various downstream tasks.

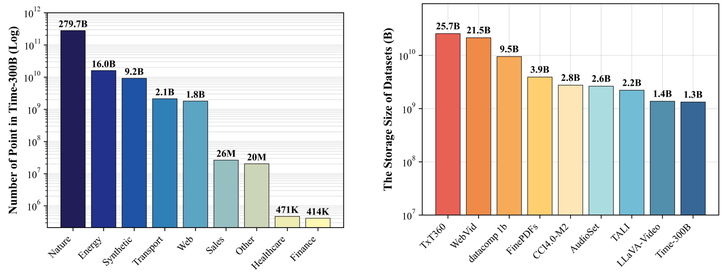
Introduction

Based on Takens' theorem and symbolic dynamics, we argue that a time series represents the external output of a complex dynamical system, which can be precisely described through mathematical symbols.



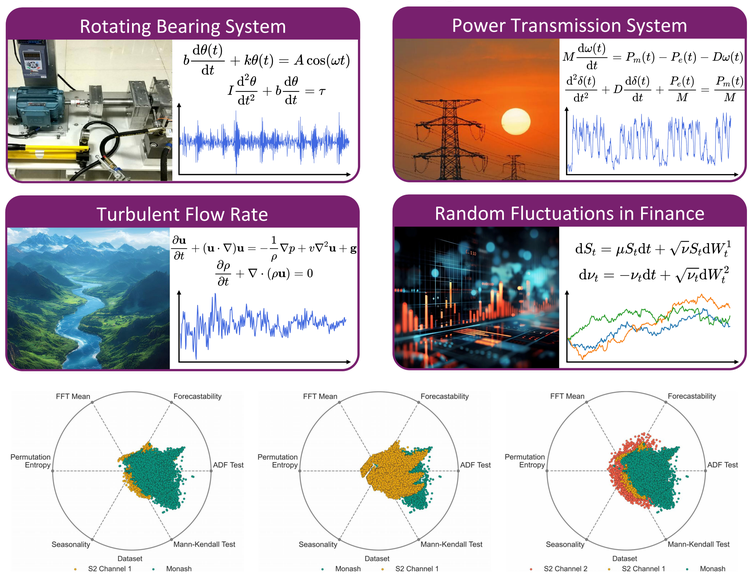
Motivations & Challenges

Time series foundaton models face serious problems of data scarcity and distribution imbalance compared with models in CV and NLP.

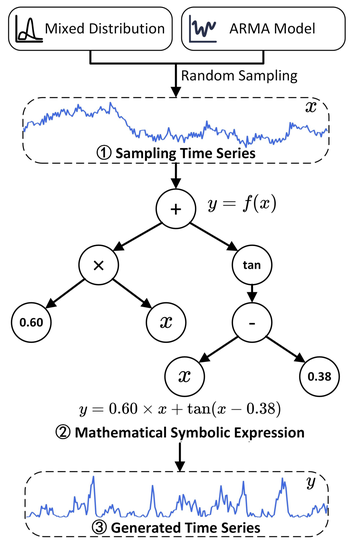


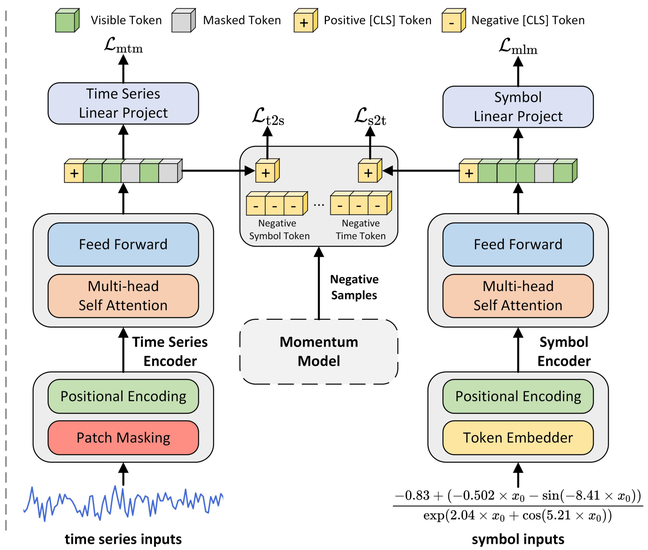
RQ 1: The Statistical Representation Coverage

We synthesise time series and symbolic expressions that reflect real physical processes, which are similar to large real time series datasets.



Series-Symbol Generation and SymTime Architecture





Pre-training Objectives

① Masked Time Series and Language Modeling: $\mathcal{L}_{\text{mtm}} = \frac{1}{N}\sum_{j\in\mathbf{M}_\mathbf{T}} \|p_j - \hat{p}_j\|^2$, $\mathcal{L}_{\text{mlm}} = \frac{1}{N}\sum_{j\in\mathbf{M}_\mathbf{S}} \mathbf{H}\left(y_j, p_j^{mask}(\hat{s})\right)$,

② Contrastive Learning between Series and Symbol: $\mathcal{L}_{\text{tsc}} = \frac{1}{2}\mathbb{E}\left[\mathbf{H}\left(y^{t2s}(t), p^{t2s}(t)\right) + \mathbf{H}\left(y^{s2t}(s), p^{s2t}(s)\right)\right]$

③ Cross-modal Momentum Distillation: $\mathcal{L}_{\text{tsc}}^{\text{mod}} = \frac{1}{2}\mathbb{E}\left[\mathbf{KL}\left(q^{t2s}(t)\|p^{t2s}(t)\right) + \mathbf{KL}\left(q^{s2t}(s)\|p^{s2t}(s)\right)\right]$

Stimulus Time Series X ⇨ Sampling of stimulus time series via mixed distributions and ARMA models

$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$

Symbolic Expression $f(\cdot)$ ⇨ Constructing symbolic expressions using binary trees as a medium where binary/unary operators are nodes with two/one child.

Response time series Y ⇨ $Y = f(X)$ Treat symbolic expressions as complex systems and input the stimulus time series into them to obtain the dynamic response of systems.

Research Questions in This Paper

RQ1: Can the unrestrictedly generated dataset **cover** diverse **representation** types of time series data?

RQ2: Can **SymTime** pre-trained on the **series-symbol dataset** achieve competitive results across five major time series analysis (TSA) tasks (forecasting, classification, imputation and anomaly detection)?

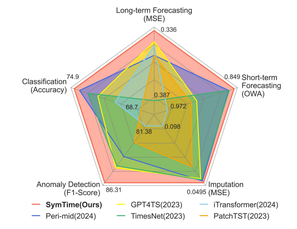
RQ3: Can **SymTime** learn **fundamental representations** of time series data on the **synthetic series-symbol dataset** to alleviate the data scarcity in time series analysis?

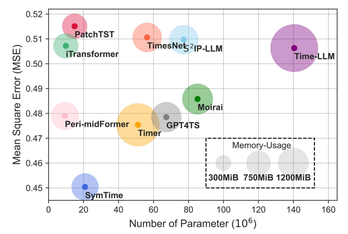
RQ4: Are the multiple pre-training objectives and **symbolic information** in **SymTime** effective?

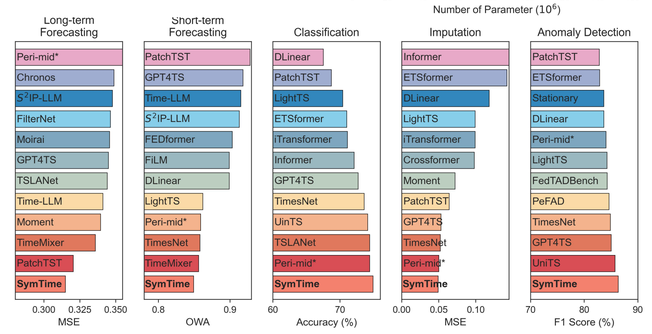
RQ5: How to demonstrate that **SymTime** learns **semantic information** of symbols?

RQ2: Main and Benchmark Results

The pre-trained SymTime achieve SOTA results in 5 general time series analysis tasks, surpassing previous models!

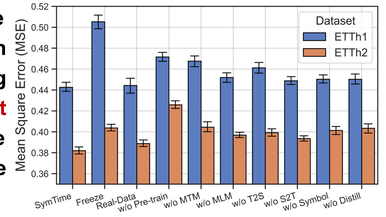






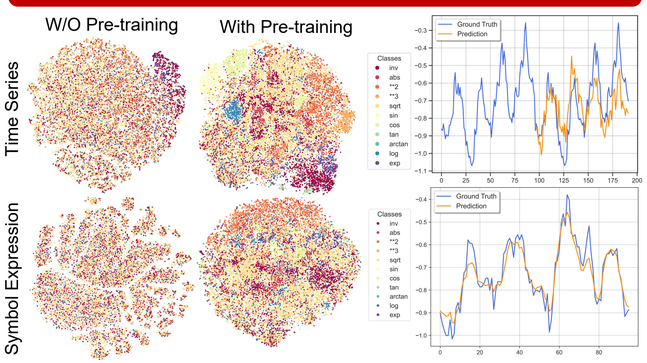
RQ 3 & 4: Ablation on Pre-training

We conducted extensive ablation experiments on various pre-training objectives and dataset sizes, demonstrating the effectiveness of the SymTime architectures and the pre-training paradigm on synthetic datasets.



Datasets	ETTm1		ETTm2		ETTh1		ETTh2		Weather		Electricity		Traffic		Exchange	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
0B	0.401	0.409	0.293	0.339	0.487	0.474	0.376	0.412	0.257	0.289	0.193	0.284	0.471	0.310	0.383	0.415
1B	0.376	0.398	0.292	0.331	0.461	0.459	0.403	0.419	0.257	0.282	0.199	0.285	0.473	0.303	0.370	0.410
10B	0.376	0.393	0.281	0.329	0.444	0.444	0.376	0.408	0.250	0.279	0.196	0.286	0.473	0.294	0.368	0.407
25B	0.378	0.393	0.278	0.325	0.434	0.438	0.371	0.405	0.253	0.282	0.195	0.288	0.467	0.299	**0.357**	**0.401**
50B	**0.371**	**0.390**	**0.274**	**0.321**	**0.430**	**0.436**	**0.365**	**0.402**	**0.247**	**0.276**	**0.187**	**0.276**	**0.457**	**0.291**	0.359	**0.401**

RQ 5: Representation Learning



It is found that after pre-training, SymTime can make data of the same type (positive samples of each other) as close as possible in the representation space, forming obvious clusters.