Prediction and Causality of functional MRI and synthetic signal using a Zero-Shot Time-Series Foundation Model

Alessandro Crimi[1][0000−0001−5397−6363] and
Andrea Brovelli[2][0000−0002−5342−1330]

[1] AGH University of Krakow, Poland
alecrimi@agh.edu.pl
[2] Institut de Neurosciences de la Timone UMR 7289, Aix Marseille Université, CNRS, 13005, Marseille, France
andrea.brovelli@univ-amu.fr

Abstract. Time-series forecasting and causal discovery are key tools for studying brain dynamics and disease. With the rise of foundation models, it remains unclear how they compare to classical methods and whether they generalize in zero-shot settings.
Here, we compared a foundation model to classical methods for inferring directional interactions from human data. More specifically, we evaluated forecasting in zero-shot and fine-tuned settings, and compared its Granger-like estimates with standard Granger causality. We validated results about forecast on real time series from stroke patients from functional magnetic resonance imaging, and causality on synthetic time series from ground-truth models (logistic map coupling and Ornstein–Uhlenbeck processes). The foundation model achieved competitive zero-shot forecasting (mean absolute percentage error of 0.55 in controls and 0.27 in patients) and competitive result on the effective connectivity experiments. Overall, these findings suggest that foundation models offer versatility, strong zero-shot performance, and potential utility for forecasting and causal discovery in time-series data.

Keywords: ARIMA · Granger causality · fMRI · Foundation models

1 Introduction

Time-series analysis in neuroscience is of considerable importance, as it enables the characterization of dynamic brain processes and the inference of underlying mechanisms; however, it remains challenging due to the high dimensionality, noise, and intrinsic complexity of neural signals [2]. Time series are also the basis for network-level analyses of brain activity and inference of effective and functional connectivity [11], quantifying relationships that can ultimately be exploited as biomarkers. Recent advances in foundation models for time series, such as TimesFM [8], Time-MOE [16], and Llama-lag [14], promise zero-shot forecasting capabilities that could benefit brain science. Analogous to large language

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_47

models, these models encode time series into latent embeddings and predict future trajectories without task-specific training. Domain-specific transformers trained on EEG [5,4], fMRI [17], or combined EEG–fMRI data [1] face recurring challenges from data heterogeneity and inter-subject variability, suggesting that general-purpose foundation models may provide a more viable solution. Among traditional approaches, ARIMA [3] remains widely used and has demonstrated robust performance in neuroimaging, often outperforming neural network–based methods [12]; we therefore use it as a primary baseline. To date, most studies have focused exclusively on forecasting, while extensions to causality and effective connectivity remain unexplored. Here we investigate whether a foundation model can be adapted for causal inference in analogy to Granger causality, the most widely used method for estimating directional interactions from neural time series. Since ground-truth causality is absent in real data, we validate our approach on two synthetic systems with known causal structure: coupled logistic maps and multivariate Ornstein–Uhlenbeck processes, which additionally allow testing of excitatory versus inhibitory interactions. We selected TimesFM for this purpose as it natively supports time-varying covariates, a requirement not fulfilled by other foundation models.

2 Methods

2.1 Dataset and Pre-processing

We used three datasets in this study. The first two dataset are synthetic datasets designed to test causal discovery: one with causal relationships defined by coupled logistic maps, and the other based on multivariate Ornstein–Uhlenbeck

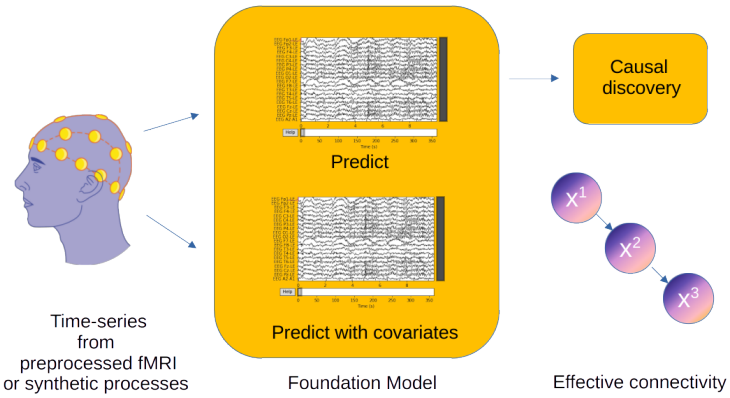


Fig. 1: Overview of the experiments: first we investigate the predictive power of the TimeSeries foundation model with brain signals, and then we evaluate if the time series predicted with it, can also be used for causal discovery.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_47

processes. The third dataset is a real-world dataset comprising both healthy and patient participants, used to evaluate differences in the prediction of healthy versus pathological fMRI time series.

Synthetic data We generate synthetic data sets with known ground-truth causality: **Logistic Map Coupling.** Three unidirectionally coupled discrete-time processes $\{X_t^{(1)}\}$, $\{X_t^{(2)}\}$, and $\{X_t^{(3)}\}$ were generated for $t = 1, \ldots, n$ with $n = 100$ time points. The initial conditions were defined as $X_0^{(j)} = c_j + \epsilon_j, \quad j \in \{1, 2, 3\}$, where $c_1 = 0.1$, $c_2 = 0.2$, and $c_3 = 0.3$, and $\epsilon_j \sim \mathcal{U}(-0.01, 0.01)$ represents small random perturbations. In this way the dynamics evolve according to the following coupled logistic map recurrences:

$$
\begin{aligned}
X_t^{(1)} &= rX_{t-1}^{(1)}(1 - X_{t-1}^{(1)}), \\
X_t^{(2)} &= rX_{t-1}^{(2)}(1 - X_{t-1}^{(2)}) + \alpha X_{t-1}^{(1)}, \\
X_t^{(3)} &= rX_{t-1}^{(3)}(1 - X_{t-1}^{(3)}) + \alpha X_{t-1}^{(2)},
\end{aligned}
$$

where r and α are coupling coefficients. We generated 10 simulations with α ranging from 0.1 to 0.9, and kept r=3.9 from previous literature. **Multivariate Ornstein–Uhlenbeck (MOU):** We simulated $N = 10$-node MOU processes $d\mathbf{X}_t = C\mathbf{X}_t\, dt + \Sigma^{1/2}\, d\mathbf{W}_t$, where $\mathbf{W}_t$ is an N-dimensional Wiener process with $d\mathbf{W}_t \sim \mathcal{N}(0, I_N\, dt)$. The connectivity matrix $C \in R^{N\times N}$ was random with density $d \in (0, 1)$ and nonzero entries drawn uniformly from $\left[-\frac{1}{Nd}, \frac{1}{Nd}\right]$ and rescaled for stability. Simulation used the Euler–Maruyama scheme. . Noise was implemented as $\sigma\sqrt{dt}\,\boldsymbol{\epsilon}_t$, with $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, I_N)$ and $\sigma = 0.01$, yielding $\Sigma = \sigma^2 I_N = 10^{-4} I_N$. For each $d \in \{0.1, \ldots, 0.9\}$, we generated 10 independent realizations with different random seeds and varying time-step dt between 0.1 and 0.2.

Human fMRI data The neuroimaging data were previously acquired by the School of Medicine at Washington University in St. Louis, with full acquisition and clinical procedures described in [6]. Briefly, the dataset includes 26 healthy control participants and 104 stroke patients who underwent fMRI scanning in the acute post-stroke phase. For the present study, we selected 26 control subjects and randomly sampled 26 stroke patients to obtain a balanced cohort. Preprocessing of the fMRI data was performed using fMRIPrep 23.1.3 [9]. The pipeline included skull stripping, spatial normalization to a standard brain template, and nuisance regression with 36 confounding parameters. The voxel-wise 4D signal was then parcellated into 117 regions of interest (ROIs) using the Schaefer atlas [15], yielding 117 regional time series per subject. Series were also MinMax scaled [0,1] prior to analysis. For each subject, 600 time points (20 minutes) were extracted and split into training (first 540 time points) and testing (remaining 60 time points) sets, corresponding to a 90%–10% split.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_47

2.2 Forecasting Models

We compared TimesFM—a 200M-parameter pre-trained model, evaluated with default hyperparameters (batch size=32) against several baselines:

- naive forecasters (mean strategy $\hat{y}_{t+1} = \frac{1}{T}\sum i = 1^T y_i$ and last-value strategy $\hat{y}_{t+1} = y_t$);
- linear regression (LR) (window length=60);
- ARIMA(p,d,q=5; no seasonality);
- Error, Trend, and Seasonality (ETS) with automated trend and damping selection.

All models were evaluated using the mean absolute percentage error (MAPE), defined as $= \frac{100\%}{n}\sum_{i=1}^{n} |(y_i - \hat{y}_i)/y_i|$.

2.3 Causality analysis

Traditional approaches rely on the Wiener–Granger causality principle, which is based on predictability: if past values of one time series improve the prediction of another (beyond the latter's past values alone), the first is said to Granger-cause the second [7]. Granger causality can be computed by comparing a restricted autoregressive (AR) model of Y against a full model that also includes lagged values of X. The significance of the improvement is tested using an F-test on the residual variances. To expand this reasoning to time series modeled with a foundation model, we consider additional time series as covariates to build a full model and inspect the residuals. Here, the foundation model also generates predictions $\hat{Y}_t$ based on historical data $\hat{Y}_t = \text{TimesFM}(Y_{t-w:t-1})$, where w is the window size (context length) and $Y_{t-w:t-1} = \{Y_{t-w}, Y_{t-w+1}, \ldots, Y_{t-1}\}$. The residuals between the observed and predicted data of the foundation model are: $r_t = Y_t - \hat{Y}_t$. To computed Granger causality from the foundation model, we tested whether lagged covariates explain the residuals that the foundation model cannot capture. In the reported synthetic experiments, we fixed the total length of the MOU time series to 100 time points. Thus, the context window for TimesFM was set to $w = 30$, which represents one third of the total length of the series, constituting a sufficiently long interval. We fit a linear regression $r_{t+\ell} = \delta + \theta_\ell X_t + \eta_t$ and computed both the Pearson correlation ρ_ℓ and the coefficient of determination R^2. The best lag $\ell \in [1, 5]$ was selected after Benjamini–Hochberg false discovery rate correction. In summary, for the **classical Granger test**, X is said to Granger-cause Y if the F-statistic is significant, indicating that including lagged values of X significantly improves the prediction of Y. In contrast, under the **TimesFM residual method**, X is considered to have a causal influence on Y if lagged values of X are significantly correlated with, or explain a significant portion of, the residuals r_t—i.e., the component of Y not captured by the foundation model. This indicates that X contains predictive information about Y beyond what TimesFM could account for. In practice, directionality is deemed significant at a threshold of $\alpha = 0.05$, correcting for multiple testing if we choose among many lags. The intuition behind the residual-based approach

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_47

is as follows. TimesFM models the autoregressive component of a signal using a large pretrained context encoder; its residuals therefore represent variance not explained by the signal's own history. If a second time series X systematically predicts those residuals at some lag, this is analogous to the Granger criterion — X contains information about Y beyond Y's past alone. The key difference from classical Granger causality is that the autoregressive component is captured by a nonlinear, attention-based model rather than a linear autoregressor, potentially allowing the method to detect interactions that linear models miss. The residual regression then serves as a lightweight linear probe on top of this nonlinear baseline. The causal discovery was evaluated by computing the mismatch in directionality, whether a true causality was detected or not, and whether it was for example $X^{(1)} \to X^{(2)}$ or vice-versa. We tested mostly the synthetic data, as even if reported the average total causality for the fMRI data, we cannot evaluate it against a ground-truth. For the logistic coupling accuracy, precision and recall are sufficient. For the MOU, we need to take into account the sign of causality (excitatory or inhibitory). The foundation model used was TimesFM 1.2.0 accessed using Python 3.11 and PyMOU to generate the MOU processes [13]. The code is accessible at URL `https://github.com/alecrimi/timesFM_stroke`.

3 Results

Table 1 summarizes the precision of the forecast between the methods and the subject groups. TimesFM produced lower MAPE. However, Treating measurements across brain regions as unpaired observations, we found non-significant the results for the control subjects, and only significant ($pval < 0.05$) the results for the patient dataset. Fine-tuning the TimesFM model led to an improvement 8% for control subjects and 14% for stroke patients, quantified as 0.50 ± 0.17 for control and 0.23 ± 0.01 for stroke patients. Causality detection performance for the 3-node synthetic data is reported in Table 2. Qualitatively, it was observed that the mismatch using Granger causality was very often caused by not detecting the causality, while for the foundation model approach, the mismatch was given by introducing a spurious causality between $X_t^{(1)}$ and $X_t^{(3)}$. For MOU-based networks, the results are shown in Figure 2, where this behavior was even more pronounced, with qualitatively the mismatch given for Granger due to miss causality, and by TimesFM-based causality having more false positive. The metrics are calculated by varying the density of causality as described in Section 2.1.1. While TimesFM does not require training time (zero-shot), its inference time is nearly 10× higher than ARIMA. Memory usage follows a similar pattern, with TimesFM requiring 2.1GB versus ARIMA's 350MB. On the human fMRI data, TimesFM detected a mean of 4,650 causal relationships compared to 3,390 with classical Granger causality. Without ground truth, it is not possible to determine whether the additional connections reflect genuine causal influences or false positives. This is consistent with the precision–recall trade-off observed in the synthetic experiments, where TimesFM showed higher recall but also more false positives than classical Granger causality.

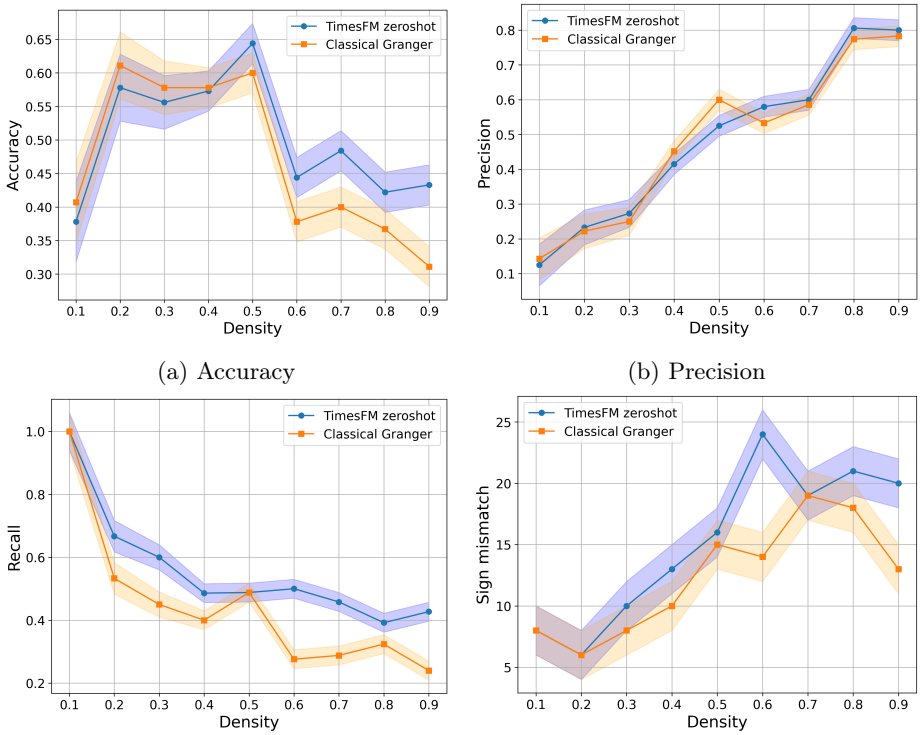
ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_47

Table 1: Forecasting performance (MAPE ± variance).

	TimesFM(zero-shot)	**LR**	**Mean**	**Last**	**ARIMA**	**ETS**
Ctrl	0.55 ± 0.42	0.59 ± 0.51	0.57 ± 0.46	0.59 ± 0.37	0.61 ± 0.64	0.63 ± 0.51
Pat	0.27 ± 0.01	0.39 ± 0.01	0.32 ± 0.01	0.32 ± 0.02	0.35 ± 0.04	0.32 ± 0.02

Table 2: Performance comparison for 3-node networks. Mean ± Variance

Method	**Accuracy**	**Precision**	**Recall**
TimesFM (zero-shot)	0.875 ± 0.0016	1.000 ± 0.0000	0.750 ± 0.0064
Granger	0.875 ± 0.0016	0.8033 ± 0.0026	1.000 ± 0.0000



(a) Accuracy (b) Precision

(c) Recall (d) Sign mismatch

Fig. 2: Accuracy, precision, recall, and causality sign mismatch for both methods, varying causal density in 10-node MOU networks.

4 Discussion

TimesFM in a zero-shot setting consistently achieved the lowest reconstruction error across both groups, suggesting that large foundation models trained on diverse time-series corpora can generalize to domains not explicitly represented

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_47

during pretraining. The non-significant difference in controls likely reflects the relatively stationary, near-linear characteristics of healthy fMRI signals, which limit the discriminative power of any single method. Conversely, patient time series exhibit stronger nonlinear patterns—irregular fluctuations and abrupt shifts driven by pathological mechanisms [10]—that violate linear model assumptions, potentially explaining the observed degradation in LR performance in this group. Fine-tuning TimesFM further improved performance, particularly in patients, although our primary focus here is on zero-shot capability. For causal discovery, no meaningful differences between methods were observed in the simple 3-node case. As shown in the results, in the logistic coupling experiments the TimesFM-based Granger-like approach exhibited high precision and low recall, whereas classical Granger causality showed higher recall and lower precision, leading to comparable overall accuracy. Advantages of the foundation model became more apparent in MOU experiments at higher causal densities. Classical Granger causality, constrained by its reliance on linear vector autoregressions, tends to miss true causal relationships (false negatives). In contrast, TimesFM, by leveraging attention and sequence modeling, is able to capture nonlinear and long-range dependencies, yielding higher recall at the cost of increased false positives and occasional sign mismatches—a trade-off that becomes more pronounced as network density increases. This sensitivity–specificity trade-off has practical implications. Classical Granger causality may be preferable when specificity is paramount, such as in small-n clinical hypothesis testing. The TimesFM residual-based method may instead be better suited as a first-pass screening tool in exploratory analyses, particularly when nonlinear dynamics are suspected, with identified connections subsequently validated through complementary approaches. To improve specificity, future work could explore L_1-regularization during residual regression to enforce sparsity, as well as permutation-based null distributions to yield better-calibrated significance thresholds. Finally, increasing noise levels (higher σ) in the MOU setting were associated with an overall deterioration in performance.

5 Conclusion

Our study suggests that even without fine-tuning, foundation models applied to time series can achieve reasonable performance in early event prediction, even for clinically relevant cases. However, causal discovery remains challenging. Our experiments do not provide clear or consistent evidence of advantages for using the investigated foundation model over traditional Granger causality. Nevertheless, future work may explore incorporating sparsity-inducing approaches to mitigate false positives and improve the reliability of causal relationships.

Acknowledgments

The research presented in this article received partial funding from the Polish Ministry of Science and Higher Education assigned to the AGH University of

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_47

Science and Technology in Krakow. We thank the School of Medicine at Washington University in St. Louis for providing the data.

References

1. Bayazi, M.J.D., et al.: General-purpose brain foundation models for time-series neuroimaging data. In: NeurIPS Workshop on Time Series in the Age of Large Models (2024)
2. Biswal, B.B., Mennes, M., Zuo, X.N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., et al.: Toward discovery science of human brain function. Proceedings of the national academy of sciences **107**(10), 4734–4739 (2010)
3. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time series analysis: forecasting and control. John Wiley & Sons (2015)
4. Chen, C.S., Chen, Y.J., Tsai, A.H.W.: Large cognition model: Towards pretrained EEG foundation model. arXiv preprint arXiv:2502.17464 (2025)
5. Chen, Y., et al.: EEGformer: Towards transferable and interpretable large-scale EEG foundation model. arXiv preprint arXiv:2401.10278 (2024)
6. Corbetta, M., Ramsey, L., Callejas, A., Baldassarre, A., Hacker, C.D., Siegel, J.S., Astafiev, S.V., Rengachary, J., Zinn, K., Lang, C.E., et al.: Common behavioral clusters and subcortical anatomy in stroke. Neuron **85**(5), 927–941 (2015)
7. Crimi, A., Dodero, L., Sambataro, F., Murino, V., Sona, D.: Structurally constrained effective brain connectivity. NeuroImage **239**, 118288 (2021)
8. Das, A., Kong, W., Sen, R., Zhou, Y.: A decoder-only foundation model for time-series forecasting. In: Forty-first International Conference on Machine Learning (2024)
9. Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., et al.: fMRIPrep: a robust preprocessing pipeline for functional MRI. Nature methods **16**(1), 111–116 (2019)
10. Falcó-Roget, J., Cacciola, A., Sambataro, F., Crimi, A.: Functional and structural reorganization in brain tumors: a machine learning approach using desynchronized functional oscillations. Communications Biology **7**(1), 419 (2024)
11. Friston, K.J.: Functional and effective connectivity: a review. Brain connectivity **1**(1), 13–36 (2011)
12. Ganesan, A., Paul, A., Nagabushnam, G., Gul, M.: Human-in-the-loop predictive analytics using statistical learning. Journal of Healthcare Engineering (1) (2021)
13. Gilson, M., Moreno-Bote, R., Ponce-Alvarez, A., Ritter, P., Deco, G.: Estimation of directed effective connectivity from fMRI functional connectivity hints at asymmetries of cortical connectome. PLoS Comp Bio **12**(3), e1004762 (2016)
14. Rasul, K., Ashok, A., Williams, A.R., Khorasani, A., Adamopoulos, G., Bhagwatkar, R., Biloš, M., Ghonia, H., 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)
15. Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.N., Holmes, A.J., et al.: Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cerebral cortex **28**(9), 3095–3114 (2018)
16. Shi, X., Wang, S., Nie, Y., Li, D., Ye, Z., Wen, Q., Jin, M.: Time-MOE: Billion-scale time series foundation models with mixture of experts. arXiv:2409.16040 (2024)
17. Wang, C., Jiang, Y., Peng, Z., Li, C., Bang, C., Zhao, L., Lv, J., Sepulcre, J., Yang, C., He, L., et al.: Towards a general-purpose foundation model for fMRI analysis. arXiv preprint arXiv:2506.11167 (2025)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_47