Causal Foundation Models: Disentangling Physics from Instrument Properties

Jeroen Audenaert [*1] **Daniel Muthukrishna** [*1] **Paul F. Gregory** [1] **David W. Hogg** [2 3 4] **V. Ashley Villar** [5 6]

Abstract

Foundation models for structured time series data must contend with a fundamental challenge: observations often conflate the true underlying physical phenomena with systematic distortions introduced by measurement instruments. This entanglement limits model generalization, especially in heterogeneous or multi-instrument settings. We present a causally-motivated foundation model that explicitly disentangles physical and instrumental factors using a dual-encoder architecture trained with structured contrastive learning. Leveraging naturally occurring observational triplets (i.e., where the same target is measured under varying conditions, and distinct targets are measured under shared conditions) our model learns separate latent representations for the underlying physical signal and instrument effects. Evaluated on simulated astronomical time series designed to resemble the complexity of variable stars observed by missions like NASA's Transiting Exoplanet Survey Satellite (TESS), our method significantly outperforms traditional single-latent space foundation models on downstream prediction tasks, particularly in low-data regimes. These results demonstrate that our model supports key capabilities of foundation models, including few-shot generalization and efficient adaptation, and highlight the importance of encoding causal structure into representation learning for structured data.

[*]Equal contribution [1]Massachusetts Institute of Technology, Camridge, MA, USA [2]New York University, New York, NY, USA [3]Flatiron Institute, New York, NY, USA [4]Max-Planck-Institut für Astronomie, Heidelberg, Germany [5]Harvard University, Cambridge, MA, USA [6]The NSF AI Institute for Artificial Intelligence and Fundamental Interactions. Correspondence to: Jeroen Audenaert <jeroena@mit.edu>, Daniel Muthukrishna <danmuth@mit.edu>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025.

1. Introduction

Observational datasets across scientific and industrial domains often conflate two distinct sources of variation: (i) the true underlying signal of interest, and (ii) distortions introduced by measurement tools, such as sensor drift, calibration offsets and environmental or observing conditions. This entanglement poses a fundamental challenge to building foundation models that can generalize across instruments, domains, or modalities.

In astronomy, the availability of petabyte-scale, open-access data (e.g., The Multimodal Universe Collaboration et al., 2024) has recently spurred a rapid development of foundation models (e.g., Parker et al., 2024; Rizhko & Bloom, 2024; Zhang et al., 2024; Euclid Collaboration et al., 2025). Time series foundation models are also emerging across broader domains (e.g., Goswami et al., 2024; Das et al., 2024). However, astrophysical signals are often deeply entangled with systematic instrumental effects. This entanglement limits the interpretability and generalization of learned representations. For example, in Euclid Collaboration et al. (2025) the instrumental properties were found to be separated in the latent space. Fortunately, many astronomical surveys provide natural experimental structure: the same star is frequently observed under different instrument configurations, and the same configuration observes many stars. This recurring observational pattern offers a unique opportunity to disentangle underlying physical dynamics from instrumental signatures.

In this work, we leverage these structural properties to develop a foundation model that explicitly separates physical and instrumental factors. Our method is inspired by causal representation learning (Schölkopf et al., 2016; 2021) and contrastive learning (Chen et al., 2020). We validate our approach on a simulated dataset designed to resemble the complexity of variable star light curves (brightness over time) observed by the Transiting Exoplanet Survey Satellite (TESS; Ricker et al., 2015). By learning general-purpose, causally disentangled representations that support few-shot learning and transfer across conditions, our model exhibits key capabilities of foundation models for structured time series. We can train downstream tasks from either of our physics or instrument latent representations. The remainder

of this paper is organized as follows: Section 2 introduces the simulation framework. Section 3 describes the model architecture and contrastive training strategy. Section 4 evaluates the learned latent spaces and their utility in downstream prediction tasks.

2. Data

We construct a simulated dataset of time series observations designed to emulate the causal structure of real-world astronomical surveys, where each observed light curve reflects both intrinsic stellar variability and systematic effects introduced by measurement instruments. This controlled setting enables direct evaluation of a model's ability to disentangle physical and instrumental factors in the latent space.

Each synthetic observation $F^{(n)}_{\text{observed}}(t)$ is generated by modulating a true stellar signal $F^{(s)}_{\text{true}}(t)$ with an instrument-specific scale $S_m(t)$ and offset $O_m(t)$, followed by the addition of Gaussian noise:

$$F^{(n)}_{\text{observed}}(t) = \text{clip}_{[-1,1]}\left(S_m(t) \cdot F^{(s)}_{\text{true}}(t) + O_m(t)\right) + \varepsilon, \tag{1}$$

where s and m index the star and instrument, respectively, $\varepsilon \sim \mathcal{N}(0, 0.03^2)$ adds observational noise, and $\text{clip}_{[-1,1]}$ bounds the signal amplitude to the range $[-1, 1]$.

The true stellar signal is modeled as a complex Fourier series:

$$F^{(s)}_{\text{true}}(t) = \text{Re}\left[\sum_{k=1}^{K-1} \frac{\theta_{s,k} \cdot e^{\text{i}\phi_{n,k-1}}}{k^{\alpha}} \cdot e^{\text{i}2\pi k t/L_s}\right], \tag{2}$$

where $\theta_{s,k}$ are stellar parameters, $\phi_{n,k}$ are observation-specific phases, α controls the power-law decay of frequency components, and $L_s = T \cdot e^{\theta_{s,0}} \cdot \lambda$ is the star's period modulated by a reduction factor λ.

Instrumental distortions are also constructed as Fourier series:

$$S_m(t) = 1 + 0.05 \cdot \text{Re}\left[\sum_{j=0}^{M-1} \beta_{m,j} \cdot e^{\text{i}\pi j t/T}\right], \tag{3}$$

$$O_m(t) = 0.05 \cdot \text{Re}\left[\sum_{j=0}^{M-1} \gamma_{m,j} \cdot e^{\text{i}\pi j t/T}\right], \tag{4}$$

with $\beta_{m,j}, \gamma_{m,j} \sim \mathcal{N}(0, 1)$ serving as instrument-specific parameters.

Each observation n is uniquely defined by a star–instrument pair (s, m). We generate large datasets with multiple repeated measurements of the same star under different instruments, and different stars observed with the same instrument. This structure mirrors the observing strategy of surveys like TESS, and is crucial to enabling the contrastive learning framework used in our method.

The resulting dataset contains 40,000 light curves, 2000 unique stars, 17 instrument configurations ($M = 17$), and 13 stellar parameters ($K = 13$). Each observation consists of $T = 100$ time steps. This scale provides sufficient statistical diversity while allowing for efficient training and evaluation. Example simulations are illustrated in Fig. 4.

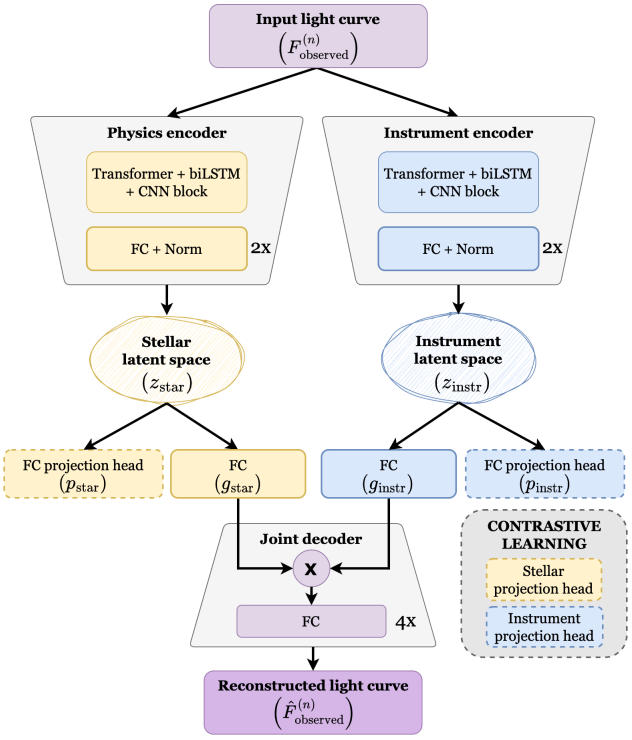

Figure 1. Causal autoencoder architecture for light curve disentanglement. The model employs dual encoders (stellar and instrumental) that process the same input independently to learn separate latent representations. The decoder combines these representations through an element-wise multiplicative interaction. Projection heads (dashed lines) enable contrastive learning during training but are not used for inference.

3. Methods

We propose a causal foundation model for structured time series that disentangles physical properties of the observed system from systematic effects introduced by the measurement process. Our method builds on the assumption common in astronomical and other sensor-driven domains that observations are conditionally independent mixtures of underlying generative sources (e.g., stellar parameters) and instrument-specific transformations. Our method uses the common structure of many observations, in which the same

underlying physical object (e.g. a star) is observed repeatedly under different sensor configurations, and the same sensor observes multiple physical systems. This data structure enables a novel form of contrastive supervision.

Triplet-based contrastive learning. We leverage the causal structure of astronomical observations through triplet-based contrastive learning. Each training example consists of three observations: $(F^{(\text{anchor})}, F^{(\text{same\_star})}, F^{(\text{same\_inst})})$, where the anchor and same-star observations share stellar identity but differ in instrumentation, while the anchor and same-instrument observations share instrumentation but differ in stellar identity. This triplet construction directly encodes our modeling assumptions: stellar latent representations should be invariant across instrument changes, while instrumental representations should be invariant across stars.

Architecture. The architecture of our model is shown in Fig. 1. Given an observed time series $F^{(n)}_{\text{observed}}(t)$ generated from star s observed by instrument m, the model learns two latent representations: a stellar encoding $\boldsymbol{z}^{(n)}_{\text{star}}$ and an instrumental encoding $\boldsymbol{z}^{(n)}_{\text{instr}}$. Both are produced from the same input via two separate encoders with identical architecture but untied parameters:

$$\boldsymbol{z}^{(n)}_{\text{star}} = E_{\text{star}}(F^{(n)}, t), \quad \boldsymbol{z}^{(n)}_{\text{instr}} = E_{\text{instr}}(F^{(n)}, t). \tag{5}$$

Each encoder follows a Conformer-based architecture (Gulati et al., 2020), incorporating temporal self-attention, LSTM recurrence, and depthwise convolution to capture patterns across a range of timescales. The outputs are globally pooled and projected into a fixed-dimensional latent space (see Appendix B for details). To reconstruct the original input, the decoder first projects each latent representation into a shared hidden space via learned nonlinear transformations, and then combines them multiplicatively:

$$\hat{F}^{(n)}_{\text{observed}} = D\left(\boldsymbol{g}_{\text{star}} \odot \boldsymbol{g}_{\text{instr}}\right), \tag{6}$$

where $\boldsymbol{g}_{\text{star}}$ and $\boldsymbol{g}_{\text{instr}}$ are projections of the respective latent parameters $\boldsymbol{z}_{\text{star}}$ and $\boldsymbol{z}_{\text{instr}}$ mapped a small MLP projection head, and $\odot$ denotes element-wise multiplication (Hadamard product). The decoder then maps the fused embedding back to the time series domain.

Contrastive objectives. To encourage disentanglement, we apply contrastive losses in both latent spaces. For each anchor embedding $\mathbf{a}$, we define sets of positives $\mathcal{P}(a)$ and negatives $\mathcal{N}(a)$ based on metadata: for the stellar latent space, positives are light curves of the same star observed under different instruments; for the instrument latent space, positives are from the same instrument observing different stars.

We minimize a generalized InfoNCE loss $\mathcal{L}_{\text{InfoNCE}}(\mathbf{a})$:

$$-\log \frac{\sum_{\mathbf{p}\in\mathcal{P}(a)} \exp(\mathbf{a}^\top \mathbf{p}/\tau)}{\sum_{\mathbf{p}\in\mathcal{P}(a)} \exp(\mathbf{a}^\top \mathbf{p}/\tau) + \sum_{\mathbf{n}\in\mathcal{N}(a)} \exp(\mathbf{a}^\top \mathbf{n}/\tau)}, \tag{7}$$

where τ is a temperature parameter and all vectors are ℓ_2 normalized. We compute this loss across all anchor samples in the batch for both latent spaces:

$$\mathcal{L}_{\text{star}} = \frac{1}{B}\sum_{n=1}^{B} \mathcal{L}_{\text{InfoNCE}}(\boldsymbol{p}^{(n)}_{\text{star}}), \tag{8}$$

$$\mathcal{L}_{\text{instr}} = \frac{1}{B}\sum_{n=1}^{B} \mathcal{L}_{\text{InfoNCE}}(\boldsymbol{p}^{(n)}_{\text{instr}}), \tag{9}$$

where B is the number of anchor samples in the training batch, and $\boldsymbol{p}_{\text{star}}$ and $\boldsymbol{p}_{\text{instr}}$ are projections of the respective latent parameters $\boldsymbol{z}_{\text{star}}$ and $\boldsymbol{z}_{\text{instr}}$ mapped by a small MLP projection head as described in Chen et al. (2020).

This formulation generalizes the InfoNCE loss used in SimCLR (Chen et al., 2020) to allow multiple positives per anchor selected via structured metadata.

Full objective. We train the model with a weighted sum of reconstruction and contrastive objectives:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{star}}\mathcal{L}_{\text{star}} + \lambda_{\text{instr}}\mathcal{L}_{\text{instr}}, \tag{10}$$

where the reconstruction loss is defined as a masked mean squared error:

$$\mathcal{L}_{\text{recon}} = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{F}^{(n)}_{\text{observed}}[t] - F^{(n)}_{\text{observed}}[t]\right)^2. \tag{11}$$

At inference time, the projection heads used for contrastive training are discarded. Downstream tasks operate on the disentangled latent representations, which we demonstrate are interpretable and predictive of physical parameters, particularly in low-data regimes.

4. Results & Discussion

We evaluate our causal foundation model by exploring the quality of its learned latent spaces and its performance on downstream predictive tasks. We compare against a foundation model baseline with identical architecture but a single shared latent space, trained with contrastive learning over same-star pairs only (no instrumental disentanglement).

4.1. Quality of the embedding

To assess whether the model successfully disentangles stellar and instrumental factors, we analyze the learned latent spaces $\boldsymbol{z}_{\text{star}}$ and $\boldsymbol{z}_{\text{instr}}$ using UMAP (McInnes et al., 2018)

and their correlation with ground-truth parameters. Each embedding is projected to 2D, and color-coded by either instrument ID (m) or stellar properties ($\theta_{s,0}$).

As illustrated in Fig. 2, the stellar latent space exhibits strong alignment with intrinsic physical property $\theta_{s,0}$ and shows minimal clustering by instrument. In contrast, the instrument latent space is well-structured by instrument configuration m, as expected. However, we also observe a moderate correlation between $\theta_{s,0}$ and the instrument latent space, indicating some leakage of physical information into $\boldsymbol{z}_{\text{instr}}$.

It may be that the recorded brightness can exhibit dependencies that couple stellar and instrumental properties (e.g., brighter stars may be more affected by certain instrumental distortions). In future work, we will explore minimizing the mutual information between the latent spaces to improve the seperations.

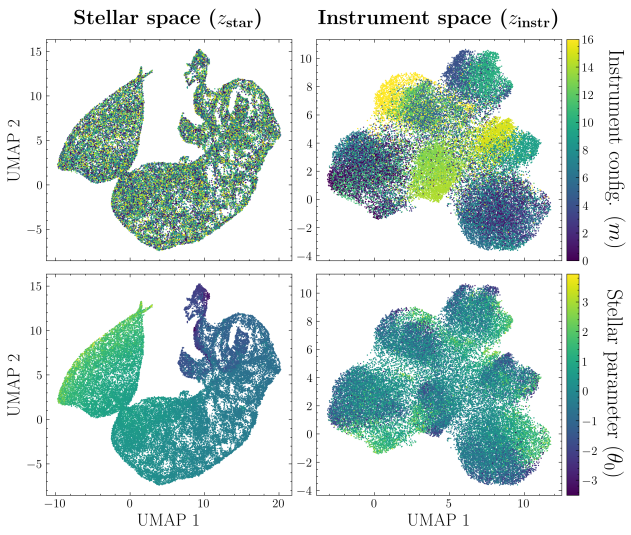

Figure 2. UMAP projections of stellar ($\boldsymbol{z}_{\text{star}}$, left) and instrumental ($\boldsymbol{z}_{\text{instr}}$, right) latent spaces. Top: colored by instrument configuration. Bottom: colored by primary stellar parameter $\theta_{s,0}$. The stellar space captures physical properties while minimizing instrument clustering. The instrument space reveals strong sector structure but shows partial leakage of stellar information.

4.2. Downstream tasks

To assess the utility of the learned representations, we evaluate performance on a supervised downstream task: predicting the primary stellar parameter $\theta_{s,0}$ from limited labeled data. We train a lightweight MLP regressor using four different input representations: (i) raw light curves (baseline), (ii) latent embeddings from a baseline foundation model with a single shared latent space trained using a contrastive loss on same-star pairs, and (iii) our proposed disentangled representations from the stellar latent space ($\boldsymbol{z}_{\text{star}}$) and (iv) the instrument latent space ($\boldsymbol{z}_{\text{instr}}$).

Fig. 3 shows the R^2 scores between the predicted and ground-truth stellar parameters for each input across different training set sizes. Models trained on $\boldsymbol{z}_{\text{star}}$ consistently outperform those using raw data or baseline latent spaces, especially in the few-shot regime. The model trained on our stellar or physics latent space performs as well or better than the normal foundation model with ten times less training data. These results demonstrate that our model captures meaningful stellar properties in its latent space, enabling effective few-shot learning and strong generalization from unlabeled, heterogeneous observational data to predictive downstream tasks.

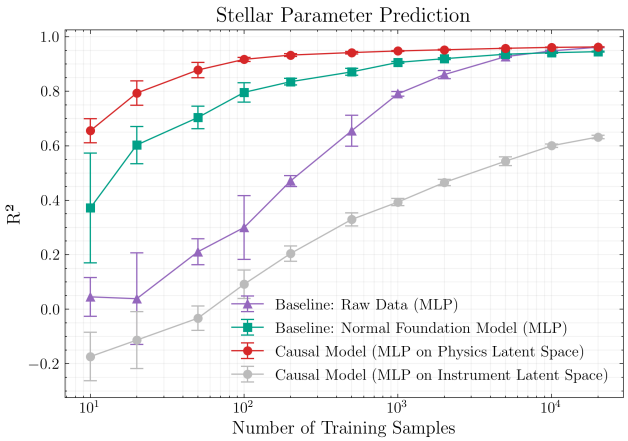

Figure 3. Comparison of prediction performance across limited training sample sizes. The plots show the R^2 for predicting the stellar parameter using four approaches: MLP on raw light curves (baseline), MLP on a baseline foundation model (single latent space with contrastive loss on same-star pairs), MLP on the stellar latent space, and MLP on the instrument latent space. Data points indicate mean and standard deviation across five evaluation runs. The latent space-based methods demonstrate superior performance, especially with limited training data.

5. Conclusions

Our results suggest that incorporating causal structure into foundation models–specifically through a dual-latent design that separately encodes physical and instrumental components–can significantly outperform conventional foundation models with a single latent space, particularly in limited-data regimes. The model's use of structured triplet-based contrastive learning enables effective disentanglement of generative factors, leveraging observational metadata rather than requiring explicit supervision. Although our results are based on a conformer encoder, we observed similar results with simpler MLP-based architectures, suggesting that the disentanglement is due to the contrastive learning

objective rather than architectural complexity.

Beyond astronomy, this methodology is broadly applicable to domains where observational confounding is a concern and where structured triplet relationships can be inferred or constructed, such as biomedicine, remote sensing, or climate forecasting. Our preliminary experiments on NASA TESS light curves demonstrate the method's potential for real-world application. Future work will explore improved disentanglement strategies, deployment on large mission-scale datasets, and extensions to multimodal and cross-domain foundation models.

Software and Data

We used Pytorch and JAX to develop the model and simulations and will make the full code publicly available after acceptance.

Acknowledgements

Funding for the TESS and Kepler missions is provided by NASA's Science Mission Directorate. The Villar Astro Time Lab acknowledges support through the David and Lucile Packard Foundation, National Science Foundation under AST-2433718, AST-2407922 and AST-2406110, as well as an Aramont Fellowship for Emerging Science Research. This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/).

Impact Statement

This paper presents work whose goal is to advance the field of machine learning, signal processing and astrophysics. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020. doi: 10.48550/arXiv.2002.05709.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2310.10688.

Euclid Collaboration, Siudek, M., Huertas-Company, M., Smith, M., Martinez-Solaeche, G., Lanusse, F., Ho, S., Angeloudi, E., Cunha, P. A. C., Domínguez Sánchez, H., Dunn, M., Fu, Y., Iglesias-Navarro, P., Junais, J., Knapen, J. H., Laloux, B., Mezcua, M., Roster, W., Stevens, G., Vega-Ferrero, J., Aghanim, N., Altieri, B., Amara, A., Andreon, S., Auricchio, N., Aussel, H., Baccigalupi, C., Baldi, M., Bardelli, S., Battaglia, P., Biviano, A., Bonchi, A., Branchini, E., Brescia, M., Brinchmann, J., Camera, S., Cañas-Herrera, G., Capobianco, V., Carbone, C., Carretero, J., Casas, S., Castander, F. J., Castellano, M., Castignani, G., Cavuoti, S., Chambers, K. C., Cimatti, A., Colodro-Conde, C., Congedo, G., Conselice, C. J., Conversi, L., Copin, Y., Courbin, F., Courtois, H. M., Cropper, M., Da Silva, A., Degaudenzi, H., De Lucia, G., Di Giorgio, A. M., Dinis, J., Dolding, C., Dole, H., Dubath, F., Duncan, C. A. J., Dupac, X., Dusini, S., Escoffier, S., Farina, M., Farinelli, R., Faustini, F., Ferriol, S., Finelli, F., Fotopoulou, S., Frailis, M., Franceschi, E., Galeotta, S., George, K., Gillis, B., Giocoli, C., Gracia-Carpio, J., Granett, B. R., Grazian, A., Grupp, F., Gwyn, S., Haugan, S. V. H., Holmes, W., Hook, I. M., Hormuth, F., Hornstrup, A., Jahnke, K., Jhabvala, M., Keihänen, E., Kermiche, S., Kiessling, A., Kubik, B., Kümmel, M., Kunz, M., Kurki-Suonio, H., Le Boulc'h, Q., Le Brun, A. M. C., Le Mignant, D., Ligori, S., Lilje, P. B., Lindholm, V., Lloro, I., Mainetti, G., Maino, D., Maiorano, E., Mansutti, O., Marcin, S., Marggraf, O., Martinelli, M., Martinet, N., Marulli, F., Massey, R., Maurogordato, S., McCracken, H. J., Medinaceli, E., Mei, S., Melchior, M., Mellier, Y., Meneghetti, M., Merlin, E., Meylan, G., Mora, A., Moresco, M., Moscardini, L., Nakajima, R., Neissner, C., Niemi, S. M., Nightingale, J. W., Padilla, C., Paltani, S., Pasian, F., Pedersen, K., Percival, W. J., Pettorino, V., Pires, S., Polenta, G., Poncet, M., Popa, L. A., Pozzetti, L., Raison, F., Renzi, A., Rhodes, J., Riccio, G., Romelli, E., Roncarelli, M., Saglia, R., Sakr, Z., Sánchez, A. G., Sapone, D., Sartoris, B., Schewtschenko, J. A., Schneider, P., Schrabback, T., Scodeggio, M., Secroun, A., Seidel, G., Seiffert, M., Serrano, S., Simon, P., Sirignano, C., Sirri, G., Stanco, L., Steinwagner, J., Tallada-Crespí, P., Taylor, A. N., Tereno, I., Toft, S., Toledo-Moreo, R., Torradeflot, F., Tutusaus, I., Valenziano, L., Valiviita, J., Vassallo, T., Verdoes Kleijn, G., Veropalumbo, A., Wang, Y., Weller, J., Zacchei, A., Zamorani, G., Zerbi, F. M., Zinchenko, I. A., Zucca, E., Allevato, V., Ballardini, M., Bolzonella, M., Bozzo, E., Burigana, C., Cabanac, R., Cappi, A., Di Ferdinando, D., Escartin Vigo, J. A., Gabarra, L., Martín-Fleitas, J., Matthew, S., Mauri, N., Metcalf, R. B., and Pezzotta, A. Euclid Quick Data Release (Q1) Exploring galaxy properties with a multi-modal foundation model. *arXiv e-prints*, art. arXiv:2503.15312, March 2025. doi: 10.48550/arXiv.2503.15312.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S.,

and Dubrawski, A. Moment: A family of open time-series foundation models. In *Forty-first International Conference on Machine Learning International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.03885.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. Conformer: Convolution-augmented transformer for speech recognition, 2020. URL https://arxiv.org/abs/2005.08100.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

Pan, J.-S., Ting, Y.-S., and Yu, J. Astroconformer: The prospects of analysing stellar light curves with transformer-based deep learning models. *Monthly Notices of the Royal Astronomical Society*, 528(4):5890–5903, March 2024. doi: 10.1093/mnras/stae068.

Parker, L., Lanusse, F., Golkar, S., Sarra, L., Cranmer, M., Bietti, A., Eickenberg, M., Krawezik, G., McCabe, M., Morel, R., Ohana, R., Pettee, M., Régaldo-Saint Blancard, B., Cho, K., Ho, S., and Polymathic AI Collaboration. AstroCLIP: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531 (4):4990–5011, July 2024. doi: 10.1093/mnras/stae1450.

Ricker, G. R., Winn, J. N., Vanderspek, R., Latham, D. W., Bakos, G. Á., Bean, J. L., Berta-Thompson, Z. K., Brown, T. M., Buchhave, L., Butler, N. R., Butler, R. P., Chaplin, W. J., Charbonneau, D., Christensen-Dalsgaard, J., Clampin, M., Deming, D., Doty, J., De Lee, N., Dressing, C., Dunham, E. W., Endl, M., Fressin, F., Ge, J., Henning, T., Holman, M. J., Howard, A. W., Ida, S., Jenkins, J. M., Jernigan, G., Johnson, J. A., Kaltenegger, L., Kawai, N., Kjeldsen, H., Laughlin, G., Levine, A. M., Lin, D., Lissauer, J. J., MacQueen, P., Marcy, G., McCullough, P. R., Morton, T. D., Narita, N., Paegert, M., Palle, E., Pepe, F., Pepper, J., Quirrenbach, A., Rinehart, S. A., Sasselov, D., Sato, B., Seager, S., Sozzetti, A., Stassun, K. G., Sullivan, P., Szentgyorgyi, A., Torres, G., Udry, S., and Villasenor, J. Transiting Exoplanet Survey Satellite (TESS). *Journal of Astronomical Telescopes, Instruments, and Systems*, 1: 014003, January 2015. doi: 10.1117/1.JATIS.1.1.014003.

Rizhko, M. and Bloom, J. S. AstroM^3: A self-supervised multimodal model for astronomy. *arXiv e-prints*, art. arXiv:2411.08842, November 2024. doi: 10.48550/arXiv.2411.08842.

Schölkopf, B., Hogg, D. W., Wang, D., Foreman-Mackey, D., Janzing, D., Simon-Gabriel, C.-J., and Peters, J. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences*, 113(27):7391–7398, 2016. doi: 10.1073/pnas.1511656113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1511656113.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

The Multimodal Universe Collaboration, Angeloudi, E., Audenaert, J., Bowles, M., Boyd, B. M., Chemaly, D., Cherinka, B., Ciuca, I., Cranmer, M., Do, A., Grayling, M., et al. The multimodal universe: Enabling large-scale machine learning with 100tb of astronomical scientific data. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2412.02527.

Zhang, G., Helfer, T., Gagliano, A. T., Mishra-Sharma, S., and Ashley Villar, V. Maven: a multimodal foundation model for supernova science. *Machine Learning: Science and Technology*, 5(4):045069, December 2024. doi: 10.1088/2632-2153/ad990d.

A. Simulated data

We show an example triplet of our simulated time series in Fig. 4. The top row is the anchor, middle row the same star but observed by different instrument and the bottom row a random star observed by the same instrument as the anchor.

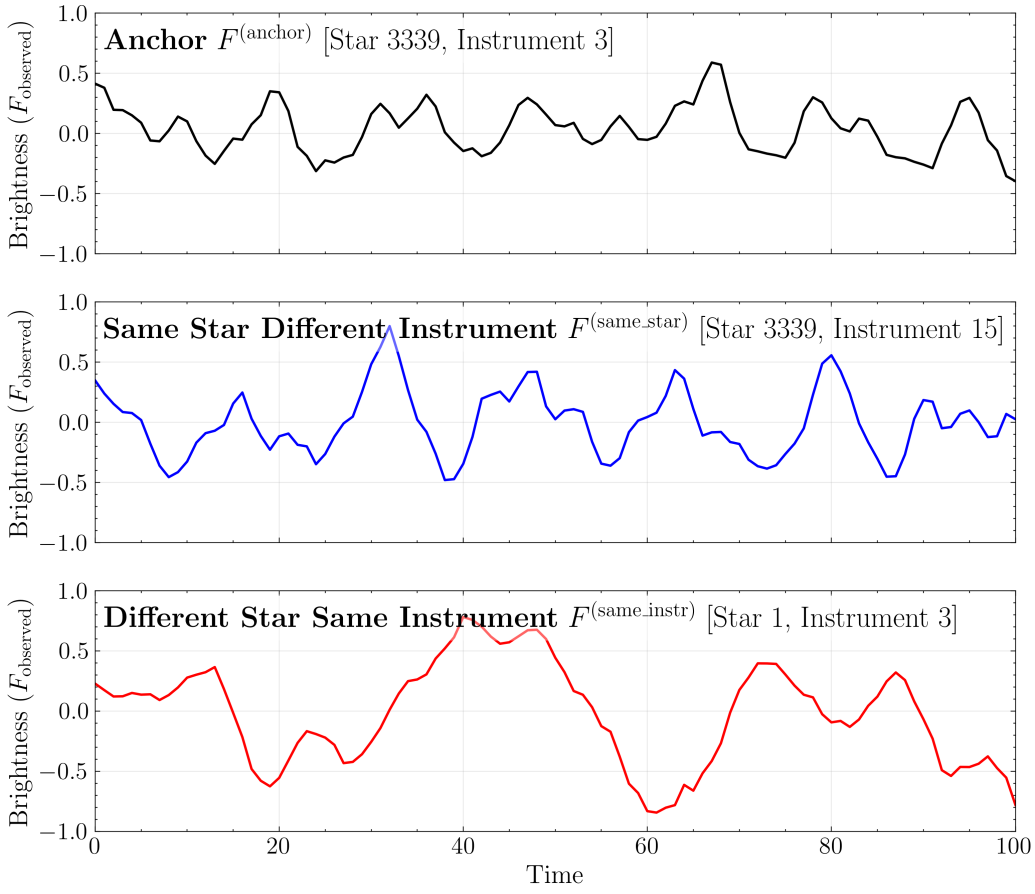

Figure 4. Example of simulated triplet data, as explained in Section. 2.

B. Encoder architecture

The detailed encoder architecture is shown in Fig. 5. The time series encoding architecture was originally developed as a time series classifier for the TESS mission (Gregory et al., in prep.), inspired by the conformer architecture (Gulati et al., 2020) for astronomical data from Pan et al. (2024).

Implementation Details: Each encoder (as illustrated in Fig. 5) consists of 4 conformer blocks with 64-dimensional embeddings, 4 attention heads, and 128-dimensional feed-forward networks. The stellar and instrumental latent spaces each have dimensionality 20. We train with Adam optimizer (learning rate=0.001), early-stopping based on the validation loss, and loss weights $\lambda_{\text{recon}} = 1.0$, $\lambda_{\text{star}} = 1.0$, $\lambda_{\text{instr}} = 1.0$.

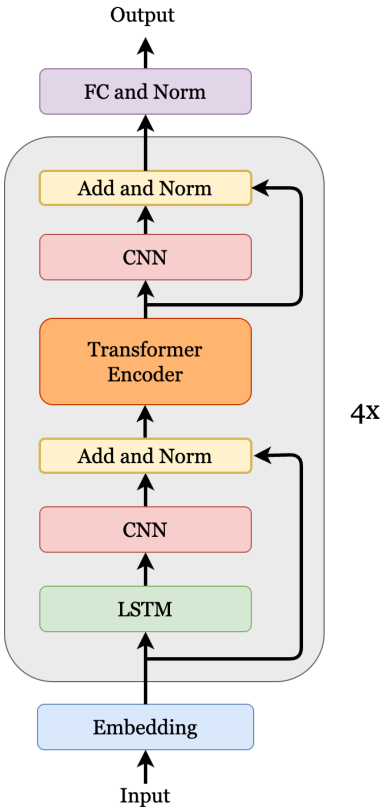

Figure 5. Encoder architecture.