Don't Pay Attention to the Noise: Learning Self-supervised Representations of Light Curves with a Denoising Time Series Transformer

Mario Morvan [1] **Nikolaos Nikolaou** [1] **Kai Hou Yip** [1] **Ingo P. Waldmann** [1]

Abstract

Astrophysical light curves are particularly challenging data objects due to the intensity and variety of noise contaminating them. Yet, despite the astronomical volumes of light curves available, the majority of algorithms used to process them are still operating on a per-sample basis. To remedy this, we propose a simple Transformer model –called Denoising Time Series Transformer (DTST)– and show that it excels at removing the noise and outliers in datasets of time series when trained with a masked objective, even when no clean targets are available. Moreover, the use of self-attention enables rich and illustrative queries into the learned representations. We present experiments on real stellar light curves from the Transiting Exoplanet Space Satellite (TESS), showing advantages of our approach compared to traditional denoising techniques[1].

1. Introduction

Time series of observed flux –so called 'light curves'– are one of the most common data products of space observation. Their analysis enables the precise study of distant objects and phenomena within and beyond the solar system and the Milky Way including stars (e.g. Christensen-Dalsgaard et al., 2007), planets (e.g. Charbonneau et al., 2000; Di Stefano et al., 2021) asteroids (Warner et al., 2009) or black holes (e.g. Beskin & Tuntsov, 2002). However, light curves are often affected by instrumental, photon and background noise. In addition, the target itself often shows an undesirable variability of similar frequencies to the underlying scientific signal, making an optimal noise filter difficult to achieve. All these factors render the analysis of light curves challenging, often requiring technical expertise to build specialised pre-processing pipelines before physical modelling and interpretation.

Although the use of deep learning has started to emerge to successfully address some problems related to light curves (e.g. Sarro et al., 2006; Wang et al., 2016; Hložek et al., 2020; Shallue & Vanderburg, 2018; Pearson et al., 2018; Morvan et al., 2020; Nikolaou et al., 2020), these often address only the later stages of data analysis and are limited to building supervised learning models. These models are indeed generally trained on scarcely labelled or simulated data and thus suffer from biases or small training sizes when applied to new or full datasets. On the other hand, there already exist large datasets consisting of thousands to billions of light curves (e.g. Bakos et al., 2004; Pollacco et al., 2006; Auvergne et al., 2009; Butters et al., 2010; Borucki et al., 2010) with many more being generated by existing and future space telescopes. We believe that tailored deep learning models will be able to leverage these large datasets to improve the efficacy and efficiency of light curve processing in a self-supervised, semi-supervised or unsupervised manner.

The self-attention mechanism (Parikh et al., 2016) and the Transformer architecture (Vaswani et al., 2017) have initiated a revolution in the field of natural language processing (e.g. Devlin et al., 2019; Brown et al., 2020) and later computer vision (Khan et al., 2021). Transformers exhibit good generalisation, and offer easier training and better scalability compared to Long Short-Term Memory Networks (Hochreiter & Schmidhuber, 1997). Work is under way to adapt the Transformer architecture for time series tasks such as forecasting (e.g. Li et al., 2020; Zhou et al., 2021; Woo et al., 2022). In their study Zerveas et al. (2021) successfully pre-trained a Time Series Transformer via a masked objective before fine-tuning it for classification and regression. Even though the use of masked objectives is common in the aforementioned works, here our main objective is to denoise the time series. The masked objective allows us to solve the problem by means of a proxy imputation task without requiring any fine-tuning.

Our main contributions consist in: (i) introducing a simple self-supervised framework to perform time series denoising

[1]Department of Physics and Astronomy, University College London, Gower Street, London, WC1E 6BT, UK. Correspondence to: Mario Morvan <mario.morvan.18@ucl.ac.uk>.

Machine Learning for Astrophysics Workshop, 39^{th} International Conference on Machine Learning (ICML), Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1]Our code is publicly available on GitHub.

without access to clean targets; (ii) demonstrating how a Transformer encoder with minimal modification can perform light curve denoising effectively, leveraging the number and diversity of available inputs, (iii) producing flexible[2] and interpretable predictions by visualising attention scores associated with imputation and denoising of sequences.

We present experiments on real light curves from the Transiting Exoplanet Survey Satellite (TESS, Ricker et al., 2015). This is the first time a deep learning model is proposed to try to address both imputation and denoising on a dataset of light curves.

2. Methodology

2.1. Problem Formulation

Given a univariate time series $x = \{x_1, .., x_t, .., x_T\} \in \mathbb{R}^T$ we seek to predict its *trend*[3] $y \in \mathbb{R}^T$ which has been corrupted by a noise process ϵ such as $x_t = y_t + \epsilon_t$ for each time step t. No assumption is made about the corruption process except its independence from the trend. In particular, ϵ can be heteroscedastic and non-Gaussian.

Let us consider a generic model f solely fed with corrupted time series, i.e. trained without clean targets in a *Noise2Self* setting (Batson & Royer, 2019). After masking a fraction of each input x with randomly generated masks m, f produces predictions $\hat{y} = f(x)$ of the same length as the input but is trained with a regression loss computed solely on the masked values: $\mathcal{L}(f(x), x, m)$. This masked objective guarantees the independence of the predictions with respect to the local values and their associated noise. If missing values are present in the dataset, they are treated in the same way as randomly masked values, making the method robust to missing values. The only difference is that predictions for truly missing values are not included in the calculation of the training loss.

2.2. Denoising Time Series Transformer

An overview of the DTST is shown on Figure 1. For each input time series an input mask is generated combining missing values and artificially masked values (at training only). Masked and standardised inputs are linearly projected into input embeddings $z \in \mathbb{R}^{T \times D}$ of the model's dimension D. Input embeddings corresponding to masked positions are replaced by a learnable vector of dimension D, inspired by the mask token used in Devlin et al. (2019). This is a robust way of informing the model of the masked input positions. Additionally, we have found a learnable vector to perform better than replacing masked values by zero as is often done for time series imputation models (e.g. Cao et al., 2018; Zerveas et al., 2021; Yi et al., 2020). However using this scheme on its own affects the quality of the predictions outside the masks and for this reason we replace 10% of masked values in input with uniformly sampled values between -2 and 2 and do not replace their projected input embedding by the mask embedding. This setting was the most effective we have tried amongst several ones listed in Appendix B.

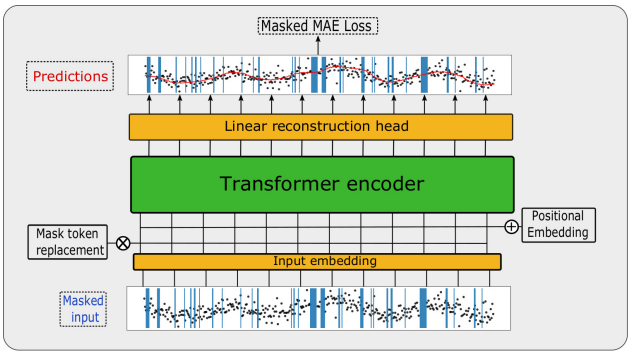


Figure 1. Schematic overview of the DTST model learning trend representations of inputs with a masked objective. Masks represented in shaded blue areas include both missing and randomly masked time steps during training. At test time, only truly missing values are masked. Yellow modules represent the time-distributed linear embedding and the prediction head respectively.

Positional embeddings are then added to the input embeddings to provide positional information. Since they produced better results than trainable positional embeddings in our experiments we used the same fixed positional embeddings as in Vaswani et al. (2017). We use a light version of the original Transformer encoder with the hyperparameters fixed to the values shown in Appendix C. Each encoder's output is finally projected back into the input dimension using a distributed $D \times 1$ linear layer.

3. Experiments

3.1. Dataset

We present experiments on a dataset of light curves from the TESS satellite, acquired during the first visit of its first sector in 2018. TESS light curves are challenging because of their length ($20,076$ time steps for short cadence data spread over 30 days), their noise level, residual instrumental systematics and missing blocks.

We select 2 minutes cadence light curves at the Presearch Data Conditioning stage, i.e. after removal of the main instrument systematics, cosmic rays and background noise with the standard TESS pipeline (Jenkins et al., 2016). After

[2]The flexibility of the model lies in its capability to handle inputs with missing values, variable sizes and generating processes characterised by different variances.

[3]The 'trend' here can contain low frequency variability, aperiodic or periodic patterns.

rejection of 50 light curves with negative flux, the dataset contains 15839 light curves. We selected 20% of all light curves uniformly at random for testing and the remaining 80% for training and validation.

3.2. Training and evaluation

Because of their length we randomly crop each light curve to select 400 consecutive time steps. A random mask is then generated before subtracting the mean and dividing by the standard deviation of the non-masked values for each input segment. This procedure can be seen as a data augmentation step, as the combination of cropping and masking operations will produce different inputs at each epoch.

For training the DTST we use the noise-scaled masked mean absolute error (NMMAE) defined as: $\text{NMMAE}(\hat{y}, x, m) = \frac{1}{M*n(x)} \sum_{t=1}^{T} m_t |x_t - \hat{y}_t|$, where m is a binary mask equal to 1 for masked time steps and 0 otherwise, $\hat{y}$ is the model's output prediction, $M = \sum_{t=1}^{T} m_t$ is the total number of masked steps in x and $n(x)$ is an estimate of the local noise by computing the average moving standard deviation with a window of size 10 and a step of 5. Compared to the mean-squared error, the mean absolute error (MAE) is more robust to outliers while rescaling using $n(x)$ helps to account for different variabilities in the training data. Predictions for the full light curves are then obtained by stitching together the predictions for segments of 400 time steps. In practice, evaluation segments are designed so as to allow overlaps of 50 steps and remove the outer 25 steps for each prediction.

As evaluation metrics we use the MAE and the inter-quartile range (IQR) of the detrended light curve as a measure of the residual noise, both expressed as percentages of the stellar flux. For both measures, lower values are desirable.

We compare the DTST to the median filter and Tukey's biweight algorithms with implementations from Hippke et al. (2019) as baselines. These have shown optimal or near-optimal performance in removing the noise prior to detecting exoplanets in Kepler and TESS data. Both methods require to set the window length –in time units for Tukey's algorithm and in number of cadences for the median filter. For comparison we select two window lengths: a long window of 6 hours ($\sim$ 300 time steps) which is adapted for exoplanet transit detection and a short window of $\sim$ 2 hours which provides comparable denoising scores to the DTST but overfits some of the high frequency variability.

3.3. Results

After experimenting with various architectures and masking scenarios (see appendix B) on the training set, we evaluated the DTST and the baselines on the test set. Results are presented in Table 1. On average, the DTST provides the smallest residual noise and auto-correlation out of the several baselines evaluated here. The difficulty for traditional techniques here lies in reconciling the diversity of the stellar processes composing the dataset, and it is therefore understandable that a single cadence-based or window-based filter with a fixed window size will either fail to denoise targets with high variability or overfit the noise on those with low variability.

Table 1. Denoising performance on 3168 test light curves from Sector 1. Averaged errors are given in percentage of the stellar flux. Window sizes considered by the three algorithms to make predictions are shown on the second line.

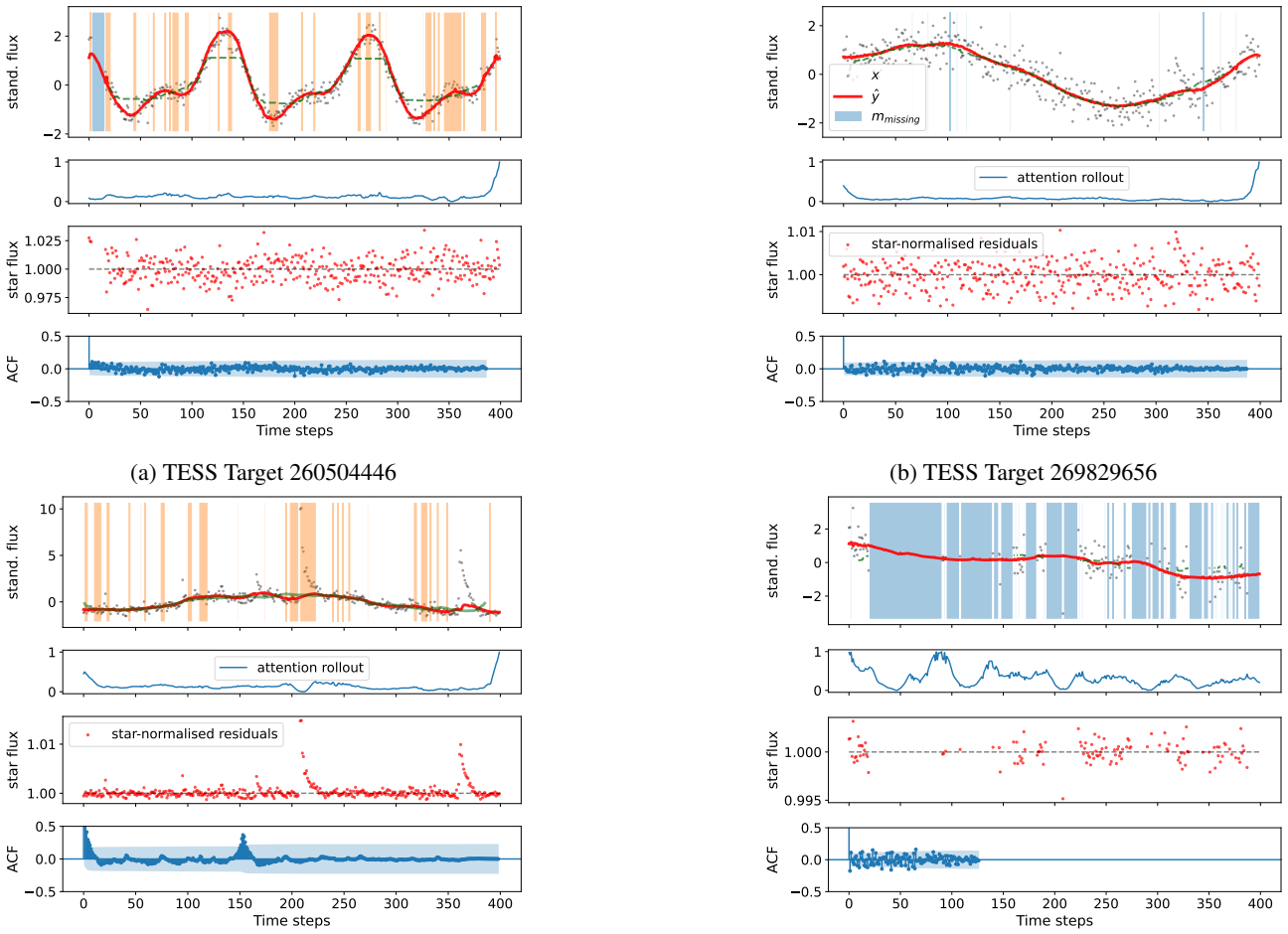
	MEDIAN FILTER		**BIWEIGHT**		**DTST**
	65 steps	181 steps	2 h	6 h	400 steps
IQR	0.393%	0.465%	0.398%	0.469%	$\mathbf{0.385\%}$
MAE	0.244%	0.286%	0.245%	0.286%	$\mathbf{0.235\%}$

We show examples of predictions in Figure 2 for different test samples showing a range of variability patterns. We corrupted the two inputs on the left (Figures 2a and 2b) with random masks similar to those used during training. The predicted time series shown in red on each upper sub-plot plot shows very good agreement with the expected trend for both masked and unmasked input time steps. In dashed green line is shown a the result of a median filter on each light curve with a window of 65 cadences (equivalent to 2 hours). While it provides good results for slowly varying stellar processes (Figure 2b), this setting fails to account for faster processes (Figures 2a and 2c) or inputs with many missing values (Figure 2d).

On each third sub-plot we show the residual light curve in units of stellar flux. The associated autocorrelation function (ACF) of the model-fit residual is plotted on the last subplot of each figure.The ACF is a useful tool to analyse the significance of residual time correlations as in Figure 2c. Target 140045538 indeed shows bursts of flaring activity which are not predicted by the DTST and therefore leave a significant signature in the ACF. As short transients or planetary transits may lie at the border between noise, outliers and signals, further fine-tuning of the model may be needed for either predicting or ignoring them consistently.

3.4. 1D Attention Maps

We use Rolling Attention (Abnar & Zuidema, 2020) to combine the attention scores of all layers and heads. We direct the reader to Appendix A for more details and examples. This enables us to visualise which parts of the inputs received more attention for producing the outputs, both during training and validation. Thus we are using the generated attention maps both for orienting the model's development and interpreting its predictions.



(a) TESS Target 260504446

(b) TESS Target 269829656

(c) TESS Target 140045538

(d) TESS Target 92257800

Figure 2. Diagnosis of predictions for four different test stars. On the left are two examples with random artificial masks (in orange) mimicking the training process. On the right are two uncorrupted inputs, where the blue shaded masks indicate truly missing data in input. Each sub-figure contains from top to bottom: (i) inputs as black dots, the DTST's predictions as red line, and median filter with window of 65 cadences as green dashed line, (ii) rolling attention time series scaled between 0 and 1, (iii) the star-normalised residual errors and (iv) the auto-correlation function with missing data ignored.

Our first observation is that both input tails often receive high rolling attention scores. This is understandable as these lack context on either their left or right and therefore prove more challenging to predict. We also observed that large masked regions receive generally less attention than non masked regions. This is in fact a useful check during the model's development to verify if the model manages to distinguish between the mask representation and the real values. Furthermore, values surrounding the identified gaps often show greater attention than average, probably as they are particularly relevant for the prediction of masked values. Finally it is often interesting to look at the attention patterns for time steps corresponding to unexpected flux values. Those are sometimes ignored such as the rightmost flaring event on Figure 2c or conversely receive more attention than average when they can inform predictions.

4. Conclusion

In this work we presented a conceptually simple framework to denoise time series via a proxy imputation task. We performed experiments and showed how such an approach based on a Transformer encoder architecture is effective at removing the noise in light curves from the TESS satellite. Compared to traditional techniques, this model can offer flexibility and increased performance when pre-processing large datasets of light curves. Further works will extend these experiments to other real and simulated datasets while assessing the generalisation power and possible gain from using a pre-trained model. Finally we would like to use this approach as a basis for downstream tasks such as event detection, imputation and upsampling.

References

Abnar, S. and Zuidema, W. Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://www.aclweb.org/anthology/2020.acl-main.385.

Auvergne, M., Bodin, P., Boisnard, L., Buey, J.-T., Chaintreuil, S., Epstein, G., Jouret, M., Lam-Trong, T., Levacher, P., Magnan, A., Perez, R., Plasson, P., Plesseria, J., Peter, G., Steller, M., Tiphène, D., Baglin, A., Agogué, P., Appourchaux, T., Barbet, D., Beaufort, T., Bellenger, R., Berlin, R., Bernardi, P., Blouin, D., Boumier, P., Bonneau, F., Briet, R., Butler, B., Cautain, R., Chiavassa, F., Costes, V., Cuvilho, J., Cunha-Parro, V., De Oliveira Fialho, F., Decaudin, M., Defise, J.-M., Djalal, S., Docclo, A., Drummond, R., Dupuis, O., Exil, G., Fauré, C., Gaboriaud, A., Gamet, P., Gavalda, P., Grolleau, E., Gueguen, L., Guivarc'h, V., Guterman, P., Hasiba, J., Huntzinger, G., Hustaix, H., Imbert, C., Jeanville, G., Johlander, B., Jorda, L., Journoud, P., Karioty, F., Kerjean, L., Lafond, L., Lapeyrere, V., Landiech, P., Larqué, T., Laudet, P., Le Merrer, J., Leporati, L., Leruyet, B., Levieuge, B., Llebaria, A., Martin, L., Mazy, E., Mesnager, J.-M., Michel, J.-P., Moalic, J.-P., Monjoin, W., Naudet, D., Neukirchner, S., Nguyen-Kim, K., Ollivier, M., Orcesi, J.-L., Ottacher, H., Oulali, A., Parisot, J., Perruchot, S., Piacentino, A., Pinheiro da Silva, L., Platzer, J., Pontet, B., Pradines, A., Quentin, C., Rohbeck, U., Rolland, G., Rollenhagen, F., Romagnan, R., Russ, N., Samadi, R., Schmidt, R., Schwartz, N., Sebbag, I., Smit, H., Sunter, W., Tello, M., Toulouse, P., Ulmer, B., Vandermarcq, O., Vergnault, E., Wallner, R., Waultier, G., and Zanatta, P. The CoRoT satellite in flight: description and performance. *Astronomy & Astrophysics*, 506(1):411–424, October 2009. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/200810860. URL http://www.aanda.org/10.1051/0004-6361/200810860.

Bakos, G., Noyes, R. W., Kovács, G., Stanek, K. Z., Sasselov, D. D., and Domsa, I. Wide-Field Millimagnitude Photometry with the HAT: A Tool for Extrasolar Planet Detection. *Publications of the Astronomical Society of the Pacific*, 116:266–277, March 2004. ISSN 0004-6280. doi: 10.1086/382735. URL http://adsabs.harvard.edu/abs/2004PASP..116..266B.

Batson, J. and Royer, L. Noise2Self: Blind Denoising by Self-Supervision. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 524–533. PMLR, May 2019. URL https://proceedings.mlr.press/v97/batson19a.html. ISSN: 2640-3498.

Beskin, G. M. and Tuntsov, A. V. Detection of compact objects by means of gravitational lensing in binary systems. *Astronomy and Astrophysics, v.394, p.489-503 (2002)*, 394:489, November 2002. ISSN 0004-6361. doi: 10.1051/0004-6361:20021150. URL https://ui.adsabs.harvard.edu/abs/2002A%26A...394..489B/abstract.

Borucki, W. J., Koch, D., Basri, G., Batalha, N., Brown, T., Caldwell, D., Caldwell, J., Christensen-Dalsgaard, J., Cochran, W. D., DeVore, E., Dunham, E. W., Dupree, A. K., Gautier, T. N., Geary, J. C., Gilliland, R., Gould, A., Howell, S. B., Jenkins, J. M., Kondo, Y., Latham, D. W., Marcy, G. W., Meibom, S., Kjeldsen, H., Lissauer, J. J., Monet, D. G., Morrison, D., Sasselov, D., Tarter, J., Boss, A., Brownlee, D., Owen, T., Buzasi, D., Charbonneau, D., Doyle, L., Fortney, J., Ford, E. B., Holman, M. J., Seager, S., Steffen, J. H., Welsh, W. F., Rowe, J., Anderson, H., Buchhave, L., Ciardi, D., Walkowicz, L., Sherry, W., Horch, E., Isaacson, H., Everett, M. E., Fischer, D., Torres, G., Johnson, J. A., Endl, M., MacQueen, P., Bryson, S. T., Dotson, J., Haas, M., Kolodziejczak, J., Van Cleve, J., Chandrasekaran, H., Twicken, J. D., Quintana, E. V., Clarke, B. D., Allen, C., Li, J., Wu, H., Tenenbaum, P., Verner, E., Bruhweiler, F., Barnes, J., and Prsa, A. Kepler Planet-Detection Mission: Introduction and First Results. *Science*, 327:977, February 2010. doi: 10.1126/science.1185402. URL http://adsabs.harvard.edu/abs/2010Sci...327..977B.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in neural information processing systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Butters, O. W., West, R. G., Anderson, D. R., Cameron, A. C., Clarkson, W. I., Enoch, B., Haswell, C. A., Hellier, C., Horne, K., Joshi, Y., Kane, S. R., Lister, T. A., Maxted, P. F. L., Parley, N., Pollacco, D., Smalley, B., Street, R. A., Todd, I., Wheatley, P. J., and Wilson, D. M. The first WASP public data release. *Astronomy & Astrophysics*, 520:L10, September 2010. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201015655. URL https:

//www.aanda.org/articles/aa/abs/2010/12/aa15655-10/aa15655-10.html.

Cao, W., Wang, D., Li, J., Zhou, H., Li, L., and Li, Y. BRITS: Bidirectional Recurrent Imputation for Time Series. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 6775–6785. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7911-brits-bidirectional-recurrent-imputation-for-time-series.pdf.

Charbonneau, D., Brown, T. M., Latham, D. W., and Mayor, M. Detection of Planetary Transits Across a Sun-like Star. *The Astrophysical Journal*, 529(1): L45–L48, January 2000. ISSN 0004637X. doi: 10.1086/312457. URL http://arxiv.org/abs/astro-ph/9911436. arXiv: astro-ph/9911436.

Christensen-Dalsgaard, J., Arentoft, T., Brown, T. M., Gilliland, R. L., Kjeldsen, H., Borucki, W. J., and Koch, D. Asteroseismology with the Kepler mission. *Communications in Asteroseismology*, 150: 350, June 2007. ISSN 1021-2043. doi: 10.1553/cia150s350. URL https://ui.adsabs.harvard.edu/abs/2007CoAst.150..350C. ADS Bibcode: 2007CoAst.150..350C.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Di Stefano, R., Berndtsson, J., Urquhart, R., Soria, R., Kashyap, V. L., Carmichael, T. W., and Imara, N. A possible planet candidate in an external galaxy detected through X-ray transit. *Nature Astronomy*, 5(12):1297–1307, December 2021. ISSN 2397-3366. doi: 10.1038/s41550-021-01495-w. URL https://www.nature.com/articles/s41550-021-01495-w. Number: 12 Publisher: Nature Publishing Group.

Hippke, M., David, T. J., Mulders, G. D., and Heller, R. Wotan: Comprehensive time-series de-trending in Python. *The Astronomical Journal*, 158(4):143, September 2019. ISSN 1538-3881. doi: 10.3847/1538-3881/ab3984. URL http://arxiv.org/abs/1906.00966. arXiv: 1906.00966.

Hložek, R., Ponder, K. A., Malz, A. I., Dai, M., Narayan, G., Ishida, E. E. O., Allam Jr, T., Bahmanyar, A., Biswas, R., Galbany, L., Jha, S. W., Jones, D. O., Kessler, R., Lochner, M., Mahabal, A. A., Mandel, K. S., Martínez-Galarza, J. R., McEwen, J. D., Muthukrishna, D., Peiris, H. V., Peters, C. M., and Setzer, C. N. Results of the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC). *arXiv:2012.12392 [astro-ph]*, December 2020. URL http://arxiv.org/abs/2012.12392. arXiv: 2012.12392.

Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Jenkins, J. M., Twicken, J. D., McCauliff, S., Campbell, J., Sanderfer, D., Lung, D., Mansouri-Samani, M., Girouard, F., Tenenbaum, P., Klaus, T., Smith, J. C., Caldwell, D. A., Chacon, A. D., Henze, C., Heiges, C., Latham, D. W., Morgan, E., Swade, D., Rinehart, S., and Vanderspek, R. The TESS science processing operations center. In *Software and Cyberinfrastructure for Astronomy IV*, volume 9913, pp. 99133E. International Society for Optics and Photonics, August 2016. doi: 10.1117/12.2233418. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9913/99133E/The-TESS-science-processing-operations-cen 10.1117/12.2233418.short.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. Transformers in Vision: A Survey. *ACM Computing Surveys*, December 2021. ISSN 0360-0300. doi: 10.1145/3505244. URL https://doi.org/10.1145/3505244. Just Accepted.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *ICLR*, 2015.

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., and Yan, X. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. *arXiv:1907.00235 [cs, stat]*, January 2020. URL http://arxiv.org/abs/1907.00235. arXiv: 1907.00235.

Morvan, M., Nikolaou, N., Tsiaras, A., and Waldmann, I. P. Detrending Exoplanetary Transit Light Curves with Long Short-term Memory Networks. *The Astronomical Journal*, 159(3):109, February 2020. ISSN 1538-3881. doi: 10.3847/1538-3881/ab6aa7. URL https://doi.org/10.3847%2F1538-3881%2Fab6aa7. Publisher: American Astronomical Society.

Nikolaou, N., Waldmann, I. P., Tsiaras, A., Morvan, M., Edwards, B., Hou Yip, K., Tinetti, G., Sarkar, S., Dawson, J. M., Borisov, V., Kasneci, G., Petkovic, M., Stepisnik, T., Al-Ubaidi, T., Bailey, R. L., Granitzer, M., Julka, S., Kern, R., Ofner, P., Wagner, S.,

Heppe, L., Bunse, M., and Morik, K. Lessons Learned from the 1st ARIEL Machine Learning Challenge: Correcting Transiting Exoplanet Light Curves for Stellar Spots. *arXiv e-prints*, 2010:arXiv:2010.15996, October 2020. URL http://adsabs.harvard.edu/abs/2020arXiv201015996N.

Parikh, A., Täckström, O., Das, D., and Uszkoreit, J. A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1244. URL https://aclanthology.org/D16-1244.

Pearson, K. A., Palafox, L., and Griffith, C. A. Searching for exoplanets using artificial intelligence. *Monthly Notices of the Royal Astronomical Society*, 474(1):478–491, February 2018. ISSN 0035-8711. doi: 10.1093/mnras/stx2761. URL https://doi.org/10.1093/mnras/stx2761.

Pollacco, D. L., Skillen, I., Collier Cameron, A., Christian, D. J., Hellier, C., Irwin, J., Lister, T. A., Street, R. A., West, R. G., Anderson, D. R., Clarkson, W. I., Deeg, H., Enoch, B., Evans, A., Fitzsimmons, A., Haswell, C. A., Hodgkin, S., Horne, K., Kane, S. R., Keenan, F. P., Maxted, P. F. L., Norton, A. J., Osborne, J., Parley, N. R., Ryans, R. S. I., Smalley, B., Wheatley, P. J., and Wilson, D. M. The WASP Project and the SuperWASP Cameras. *Publications of the Astronomical Society of the Pacific*, 118:1407–1418, October 2006. ISSN 0004-6280. doi: 10.1086/508556. URL http://adsabs.harvard.edu/abs/2006PASP..118.1407P.

Ricker, G. R., Winn, J. N., Vanderspek, R., Latham, D. W., Bakos, G. A., 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. URL http://adsabs.harvard.edu/abs/2015JATIS...1a4003R.

Sarro, L. M., Sánchez-Fernández, C., and Giménez, A. Automatic classification of eclipsing binaries light curves using neural networks. *Astronomy & Astrophysics*, 446 (1):395–402, January 2006. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361:20052830. URL http://arxiv.org/abs/astro-ph/0511346. arXiv: astro-ph/0511346.

Shallue, C. J. and Vanderburg, A. Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90. *The Astronomical Journal*, 155(2):94, January 2018. ISSN 1538-3881. doi: 10.3847/1538-3881/aa9e09. URL http://stacks.iop.org/1538-3881/155/i=2/a=94?key=crossref.373953d9fd4268d95bf7424bbc462372.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, D., Hogg, D. W., Foreman-Mackey, D., and Schölkopf, B. A Causal, Data-Driven Approach to Modeling the Kepler Data. *Publications of the Astronomical Society of the Pacific*, 128(967):094503, September 2016. ISSN 0004-6280, 1538-3873. doi: 10.1088/1538-3873/128/967/094503. URL http://arxiv.org/abs/1508.01853. arXiv: 1508.01853.

Wang, S., Li, B. Z., Khabsa, M., Fang, H., and Ma, H. Linformer: Self-Attention with Linear Complexity. *arXiv:2006.04768 [cs, stat]*, June 2020. URL http://arxiv.org/abs/2006.04768. arXiv: 2006.04768.

Warner, B. D., Harris, A. W., and Pravec, P. The asteroid lightcurve database. *Icarus*, 202:134–146, July 2009. ISSN 0019-1035. doi: 10.1016/j.icarus.2009.02.003. URL https://ui.adsabs.harvard.edu/abs/2009Icar..202..134W. ADS Bibcode: 2009Icar..202..134W.

Woo, G., Liu, C., Sahoo, D., Kumar, A., and Hoi, S. ETSformer: Exponential Smoothing Transformers for Time-series Forecasting. *ArXiv*, 2022.

Yi, J., Lee, J., Kim, K. J., Hwang, S. J., and Yang, E. Why Not to Use Zero Imputation? Correcting Sparsity Bias in Training Neural Networks. In *ICLR*, 2020.

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., and Eickhoff, C. A Transformer-based Framework for

Multivariate Time Series Representation Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 2114–2124, New York, NY, USA, August 2021. Association for Computing Machinery. ISBN 978-1-4503-8332-5. doi: 10.1145/3447548.3467401. URL https://doi.org/10.1145/3447548.3467401.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *The Thirty-Fifth {AAAI} Conference on Artificial Intelligence, {AAAI} 2021, Virtual Conference*, volume 35, pp. 11106–11115. {AAAI} Press, March 2021. URL http://arxiv.org/abs/2012.07436. arXiv: 2012.07436 version: 2.

A. 1D Attention Maps

Attention from the model's output to the input time series is computed using Attention Rollout (Abnar & Zuidema, 2020). This procedure consists in recursively multiplying matrices of attention weights through the transformer layers, thus accounting for mixing of attention in the network. Figure 3 shows more examples of predictions overlayed with their corresponding input with Rollout Attention used to highlight time steps with greater attention.

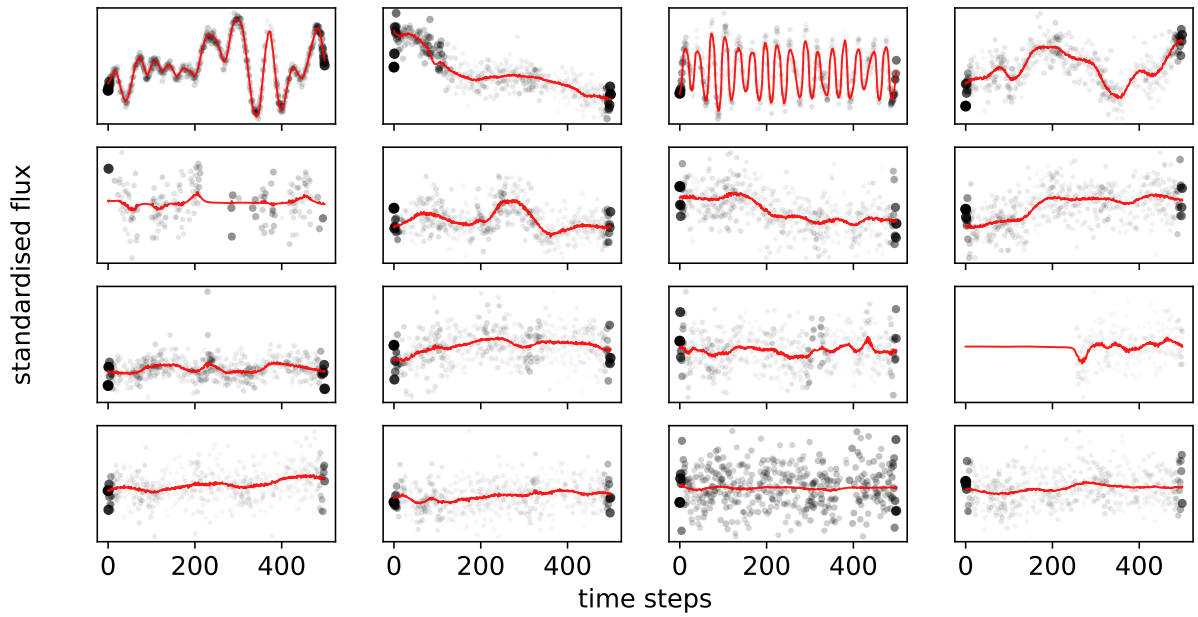


Figure 3. 1D attention maps computed with Attention Rollout and overlayed with predictions (red lines) for 16 random light curves from TESS dataset. The size and opacity of individual point inputs is directly proportional to the Attention Rollout values.

B. Masking Strategy

Masking Patterns We explored various generating mechanisms for the distributions of masked values in each input. Given a fixed ratio of values to mask, we tested using a Bernoulli distribution for time-independent masking and a geometric distribution over masked blocks lengths. The geometric distribution was used to impose longer masks and thus a more challenging imputation task to the model. Mean block lengths of 5, 10 and 20 were tested and final results were presented for a window of 10 as it offered the best compromise between the length of signals to impute and denoising performance.Intuitively, the length of masked blocks will control the degree of temporal locality of the noise processes to remove, and using wider masked regions will indeed force the model to make use of longer-term dependencies to make accurate predictions.

Masking Ratios We set the masking ratio to 30% after experimenting with 10%, 20%, 30%, 40% and 50%. Heuristically, increasing the masking ratio speeds up training but also affects performance as fewer inputs are available for prediction. When values are missing in the inputs, the masking ratio is considered with respect to the number of non-missing time steps. This maintains the ratio of data used for training constant while avoiding degenerate cases where the random mask would be empty or would cover the entirety of the non-missing input.

Replacement Strategy We considered various replacement strategies for masked input values: (i) replacing by zero, (ii) by a uniformly random value centred on zero, (iii) by a special learnable vector (inspired by Devlin et al., 2019) in the model's space, and (iv) keeping the original values. Case (iv) was quickly discarded as it led to overfitting the noise. While option (iii) offered the best imputation performance, we observed that it performed poorly on its own for denoising the full inputs, and that this issue was mitigated by using case (ii) for a random fraction of input time steps, even as small as 10%. This can be understood as an extra corruption operation on the input, thus forcing the model to provide coherent predictions

even outside the regions whose embeddings are more explicitly masked with a dedicated vector. Whilst we have compared these several cases, it would be interesting to investigate further the influence of the replacement strategy (e.g. different distributions) and ratios on the denoising performance.

C. Hyperparameters

Fixed hyperparameters for all presented experiments are shown in Table 2.

For training we used Adam (Kingma & Ba, 2015) optimiser with learning rate 0.001 and $\beta = (0.9, 0.999)$.

Table 2. Hyperparameters

PARAMETER	VALUE
Learning rate	0.001
Batch size	64
Dim. model	64
Dim. feedforward	128
Num layers	3
Num. heads	8
Masking ratio	30%
Average masking length	10

D. Computational Efficiency

Even though training the DTST on thousands of time series can take up to several hours on a single V100 GPU, its inference cost remains very low with around $10\ \mu s$ for a full TESS light curve unfolded in windows of length 400 passed as a batch of size ~ 60. This is to be compared with $\sim 50\ \mu s$ and $\sim 173\ \mu s$ per TESS light curve for the efficient Wotan implementations of biweight and median filter respectively.

The $O(T^2)$ complexity in space and time of vanilla attention could be mitigated by using sparse attention (see e.g. Zhou et al. (2021) in $O(T \log T)$, Wang et al. (2020) in $O(T)$). Additional experimental studies would need to be performed to evaluate their respective impact on performance and on the trade-off between long sequences and full attention.