Astro-MoE: Mixture of Experts for Multiband Astronomical Time Series

Martina Cádiz-Leyton [1] **Guillermo Cabrera-Vives** [2 3 4 5] **Pavlos Protopapas** [6] **Daniel Moreno-Cartagena** [2] **Ignacio Becker** [6]

Abstract

Multiband astronomical time series exhibit heterogeneous variability patterns, sampling cadences, and signal characteristics across bands. Standard transformers apply shared parameters to all bands, potentially limiting their ability to model this rich structure. In this work, we introduce Astro-MoE, a foundational transformer architecture that enables dynamic processing via a Mixture of Experts module. We validate our model on both simulated (ELAsTiCC-1) and real-world datasets (Pan-STARRS1).

1. Introduction

After a decade of breakthroughs enabled by single-epoch surveys, astronomy is transitioning toward a new era defined by multi-epoch observations. This shift has fueled the rapid growth of time-domain astronomy, which focuses on studying celestial objects and phenomena whose properties evolve over time through wide-field surveys that repeatedly image large areas of the sky (Graham et al., 2012; Kasliwal et al., 2019). The upcoming Legacy Survey of Space and Time (LSST; Ivezić et al. 2019), beginning full operations by the end of 2025, will accelerate this trend by generating multiband time series data for approximately 20 billion sources with unprecedented depth, cadence, and volume.

The intrinsic complexity of time-domain data (characterized by heterogeneous sampling, irregular cadences, and inter-band dependencies) has motivated significant advances in representation learning for astronomical time series (e.g. Protopapas 2017; Charnock & Moss 2017; Naul et al. 2018; Park et al. 2021; Donoso-Oliva et al. 2023; Pan et al. 2024; Becker et al. 2025). Transformer architectures (Vaswani et al., 2017) have shown particular promise for modeling such irregular sequential data, achieving state-of-the-art results within astronomy across tasks such as denoising (Morvan et al., 2022), classification (Pimentel et al., 2022; Allam Jr & McEwen, 2024), regression (Zhang et al., 2024), and uncertainty estimation (Cádiz-Leyton et al., 2024), even in data-scarce scenarios (Moreno-Cartagena et al., 2023; Donoso-Oliva et al., 2025). A core challenge in this domain is learning unified embeddings that compactly encode both temporal evolution and spectral characteristics across photometric bands. Such representations not only enhance performance on fundamental astronomical tasks (e.g., variable star classification, redshift regression) but also enable effective integration into multimodal frameworks (Rizhko & Bloom; Lanusse et al., 2023; Parker et al., 2024). For instance, recent models such as ATAT (Cabrera-Vives et al., 2024) demonstrate how combining light curve embeddings with auxiliary metadata can accelerate scientific discovery.

However, conventional transformer-based architectures face fundamental limitations when processing multiband time series. Transformers apply homogeneous processing to all inputs (i.e., reuse the same parameters for all inputs), which is suboptimal for astronomical data where each photometric band probes distinct astrophysical processes, exhibits unique sampling characteristics, and manifests independent variability behaviors. This diversity suggests that uniform weight-sharing mechanisms may not adequately capture band-specific features and complex inter-band interactions. Mixture of Experts (MoE) architectures offer a compelling solution through dynamic, input-dependent routing to specialized subnetworks (Masoudnia & Ebrahimpour, 2014; Shazeer et al., 2017). MoE-based models select different parameters for each example, resulting in sparsely-activated models with more parameters but constant computational cost. This capability is ideally suited to multiband astronomical data, where experts can learn band-specific representations while maintaining shared temporal knowledge. While recent works like Time-MoE (Shi et al., 2024) and Moirai-MoE (Liu et al., 2024) highlight the potential of ex-

[1]Edinburgh Futures Institute, University of Edinburgh, UK [2]Department of Computer Science, Universidad de Concepción, Edmundo Larenas 219, Concepción, Chile [3]Center for Data and Artificial Intelligence, Universidad de Concepción, Edmundo Larenas 310, Concepción, Chile [4]Heidelberg Institute for Theoretical Studies, Heidelberg, Baden-Württemberg, Germany [5]Millennium Institute of Astrophysics (MAS), Nuncio Monseñor Sotero Sanz 100, Of. 104, Providencia, Santiago, Chile [6]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Boston, MA, USA. Correspondence to: Martina Cádiz-Leyton <m.a.cadiz@sms.ed.ac.uk>, Guillermo Cabrera-Vives <guillecabrera@inf.udec.cl>.

ML4Astro 2025, Vancouver, CA.

pert routing in temporal data, their adaptation to the specific challenges of irregular, multiband astronomical time series remains an area of ongoing interest, with opportunities for further methodological development.

In this work, we introduce Astro-MoE, a pretrained transformer that incorporates sparsely-gated MoE modules in the architecture. This design enables the processing of multiband time series, producing embeddings that are both robust and informative for downstream tasks. Our architecture addresses the challenges of astronomical time series by allowing different experts to specialize in different variability patterns while maintaining the ability to model complex inter-band correlations.

2. Methods

The Astro-MoE model extends the Astromer framework (Donoso-Oliva et al., 2023), a self-supervised light curve transformer originally designed for single-band data, to a multiband setting. Each astronomical object is represented as a sequence of observations across multiple photometric bands. For each band b and time step j, the input includes a flux measurement $\mu_{j,b}$ and its associated uncertainty $\sigma_{j,b}$. These values are combined into an input vector $x_{j,b} = (\mu_{j,b}, \sigma_{j,b})$, which represents the brightness and its error at time j in band b. To construct the model input, each band is encoded as a fixed-length sequence of $x_{j,b}$ vectors, with zero-padding applied as needed. The resulting sequences are then concatenated across bands to form a unified representation x, which preserves temporal ordering and captures band-specific information.

2.1. Mixture of Experts

As illustrated in Figure 1, our model adopts an encoder-only transformer architecture enhanced with sparse Mixture-of-Experts (MoE) layers (Shazeer et al., 2017; Mu & Lin, 2025). The MoE module is integrated into two key components: (1) the input embedding stage, and (2) the attention blocks, where it replaces the standard feedforward network (FFN) sublayer. Each MoE layer contains N_{experts} parallel experts, with sparse routing controlled by a learnable gating function. Here, N_{experts} denotes the number of experts in the layer, which can differ between components.

Given an input vector $x \in \mathbb{R}^{d_{\text{in}}}$, where d_{in} is the dimensionality of the input (e.g., 2 for brightness and uncertainty pairs), the gating network computes a score for each expert:

$$g(x) = W_g x + b_g, \quad g : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{N_{\text{experts}}}, \tag{1}$$

where W_g and b_g are the learnable weights and bias of the gating network. To induce sparsity, only the top-k scoring experts are selected using a TopK operator, and a softmax

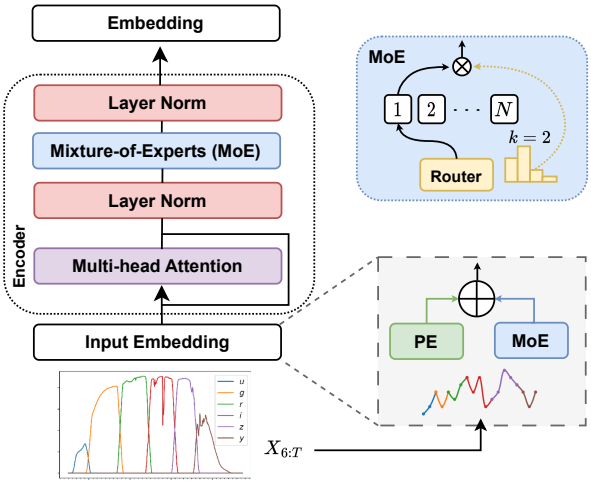

Figure 1. Astro-MoE architecture diagram with the traditional positional encoding.

is applied to compute normalized selection weights:

$$G(x) = \text{softmax}(\text{TopK}(g(x), k)), \tag{2}$$

where k is the number of experts selected per input. The operator $\text{TopK}(v, k)_i$ retains the top-k values in v and masks the rest with $-\infty$:

$$\text{TopK}(v, k)_i = \begin{cases} v_i, & \text{if } v_i \text{ is among the top-}k \text{ elements,} \\ -\infty, & \text{otherwise.} \end{cases}$$

This yields a sparse selection vector $G(x) \in \mathbb{R}^{N_{\text{experts}}}$ with nonzero weights only for the selected experts (i.e., $G(x)_e \neq 0$ if and only if expert e is selected). The final MoE output is then computed as:

$$\text{MoE}(x) = \sum_{e \in \text{Top-k}} G(x)_e \cdot E^{(e)}(x), \tag{3}$$

where $E^{(e)}(x)$ is the output of expert e.

In the input embedding stage, each expert $E^{(e)}$ is implemented as a linear transformation:

$$E^{(e)}(x) = W^{(e)} x, \quad W^{(e)} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{in}}}. \tag{4}$$

where d_{model} is the internal model dimension used by the transformer. Unlike prior approaches that apply a uniform transformation across all bands (Donoso-Oliva et al., 2023; Cabrera-Vives et al., 2024), we project the brightness-uncertainty pairs $(x_{j,b}, \sigma_{j,b}) \in \mathbb{R}^2$ into d_{model} using a MoE module. We set $N_{\text{experts}} = 6$ (one per photometric band) and $k = 2$, matching the six-band pretraining setup. This configuration promotes expert specialization across different observational regimes.

Within the attention blocks, each expert is implemented as a two-layer FFN, replacing the standard FFN sublayer. This

choice is inspired by prior large-scale MoE transformer models (Lepikhin et al., 2020; Fedus et al., 2022), which demonstrate that FFN layers often exhibit sparse and task-specific activation patterns, making them well-suited for expert-based specialization. In this setting, we use $N_{\text{experts}} = 8$ and select the top-$k = 2$ experts per token.

To encourage balanced expert utilization and avoid expert collapse, we incorporate a simplified version of the load balancing loss from Shazeer et al. (2017):

$$\mathcal{L}_{\text{aux}} = N_{\text{experts}} \cdot \sum_{e=1}^{N_{\text{experts}}} \bar{p}_e \cdot \bar{f}_e, \tag{5}$$

where $\bar{p}_e$ is the mean routing probability assigned to expert e, and $\bar{f}_e$ is the empirical fraction of tokens routed to that expert. This auxiliary loss is computed independently for each MoE layer and summed across all such layers. The total auxiliary loss is then scaled by a fixed coefficient $\lambda = 0.01$ and added to the main task loss (e.g., classification or regression). The scaling factor ensures that the auxiliary term promotes expert diversity without overpowering the primary learning objective.

2.2. Positional Embedding

Encoding temporal information is a crucial component in modeling light curves, as observations are irregularly spaced in time. Recent work has shown that the choice of positional encoding (PE) can significantly affect the performance of transformer-based models for time series (Moreno-Cartagena et al., 2023). Therefore, we explore two temporal encoding strategies, both integrated at the input embedding stage, before the self-attention layers.

The first approach follows the original transformer formulation (Vaswani et al., 2017), in which observation times are encoded using fixed sine and cosine functions at varying frequencies:

$$\text{PE}_i(t_{j,b}) = \begin{cases} \sin(t_{j,b} \cdot \omega_i), & i \text{ even}, \\ \cos(t_{j,b} \cdot \omega_i), & i \text{ odd}, \end{cases} \quad \omega_i = \frac{1}{1000^{2i/d_{\text{pe}}}}, \tag{6}$$

where i is the dimension index, and d_{pe} is the total number of positional encoding dimensions.

As an alternative, we adopt the learnable Time Modulation (TM) approach proposed by Cabrera-Vives et al. (2024), which incorporates temporal information directly into the input features. Instead of using a fixed linear projection, we pass each input vector $x_{j,b}$ through a MoE layer to obtain an adaptive representation. The resulting vector is then modulated using band-specific, time-dependent Fourier functions:

The resulting vector is then modulated using band-specific, time-dependent Fourier functions:

$$\text{TM}(x_{j,b}, t_{j,b}) = \text{MoE}(x_{j,b}) \odot \gamma_b^{(1)}(t_{j,b}) + \gamma_b^{(2)}(t_{j,b}), \tag{7}$$

where $\gamma_b^{(1)}$ and $\gamma_b^{(2)}$ are learnable band-specific Fourier series. The operator $\odot$ denotes element-wise (Hadamard) multiplication.

3. Experiments

3.1. Data description

For pretraining and classification, we use data from the first round of the Extended LSST Astronomical Time-series Classification Challenge (ELAsTiCC-1), a large-scale simulation designed to emulate the observational characteristics of the Vera C. Rubin Observatory's LSST. The dataset contains 1,845,146 multiband light curves spanning 32 astrophysical classes, including both periodic variables and transient events. Each light curve is observed in six optical bands (*ugrizy*) with realistic cadences, noise levels, and detection limits. Following Cabrera-Vives et al. (2024), we regroup the classes into 20 categories, discard poor-quality measurements using `PHOTFLAG`, and extract forced photometry ranging from 30 days before the first alert to the final detection. Additionally, each object is associated with 64 metadata columns describing contextual and observational properties. These include redshift estimates, sky coordinates, host galaxy characteristics, and summary statistics of the light curves. As in previous work, these metadata are incorporated as complementary inputs for classification. The test set contains 1,000 objects per class, ensuring balance across categories, while the remaining data are split into five class-stratified folds with an 80/20 training-validation ratio.

To evaluate our approach on an alternative classification task, we use photometric light curves from the second data release of Pan-STARRS1 (PS1), which offers observations in five optical bands: g_{P1}, r_{P1}, i_{P1}, z_{P1}, and y_{P1}. Following the methodology of Becker et al. (2025), we retrieve the PS1 photometry from the Detections table via MAST CasJobs, convert fluxes to AB magnitudes using the standard zero-point of 3631 Jy, and apply quality filters (e.g., `psfQfPerfect`, `infoFlags`, `infoFlags2`, `infoFlags3`) to ensure clean photometry. We require a minimum of four observations per band. As in Becker et al. (2025), we apply class balancing by limiting the maximum number of objects per class to 10,000, mitigating overfitting due to the strong class imbalance, particularly for RR Lyrae stars. Our final dataset includes six variable star classes. The data are split into seven folds, stratified by class, using 70% for training, 10% for validation, and 20% for testing, with the test set kept fixed across all folds. Appendix A shows the classes and number of objects in ELAsTiCC-1 and PS1, grouped as transients, stochastic variables, and periodic variables.

Table 1. Pretraining performance on the ELAsTiCC-1 test set.

MODEL	TE	R^2	RMSE
MULTIBAND-ASTROMER	PE	0.349	2.511
MOE-ASTRO	PE	0.403	2.398
MOE-ASTRO	TM	**0.438**	**2.327**

3.2. Training details

All model variants are based on an encoder-only transformer architecture with three self-attention blocks. Pretraining is performed using a masked reconstruction objective, following the strategy introduced by Donoso-Oliva et al. (2023). In each training step, 90% of the light curves are randomly selected for training. Within each selected light curve, 30% of the input tokens are masked, 30% are replaced with random values, and the remaining 40% are left unchanged. A dropout rate of 0.1 is applied throughout the network. For ELAsTiCC-1 classification, we concatenate the light curve embeddings with tabular embeddings extracted from a tabular transformer, following the ATAT architecture proposed by Cabrera-Vives et al. (2024). For the PS1 classification task, we apply a linear classifier directly on the light curve embeddings, without incorporating metadata. In both tasks, models are trained using a cross-entropy loss. Training is performed with the Adam optimizer, using a batch size of 256 and a learning rate of 1×10^{-4}.

4. Results

Table 5 reports the mean and standard deviation of pretraining scores for our model configurations on the ELAsTiCC-1 test set. As a baseline, we consider a multiband extension of Astromer, in which brightness vectors are ordered by observation time and projected via a shared linear layer. Comparing this baseline with our proposed Astro-MoE architecture, we observe improvements when incorporating MoE modules for both the brightness encoder and the FFN within the attention blocks. Specifically, this configuration achieves an R^2 of 0.403 and an RMSE of 2.398. When replacing the PE with the TM encoder, performance further improves by approximately 3%, reaching an R^2 of 0.438 and reducing RMSE to 2.327. Overall, these results indicate that the combination of sparse MoE-based representations and temporally adaptive encoding may enhance the expressiveness and accuracy of light curve models. It is also important to note that ELAsTiCC-1 is a complex dataset, featuring a maximum sequence length of 65 per band and comprising a wide diversity of astronomical object classes.

Table 2. Classification performance on the ELAsTiCC-1 test set using both light curve and metadata information.

MODEL	PRETRAINED	F1-SCORE
MULTIBAND-ASTROMER	YES	0.727 ± 0.034
	NO	0.786 ± 0.013
ATAT	NO	0.826 ± 0.005
MOE-ASTRO (PE)	YES	0.822 ± 0.008
MOE-ASTRO (TM)	YES	0.832 ± 0.015
MOE-ASTRO (TM)	NO	$\mathbf{0.860 \pm 0.003}$

Table 3. Classification performance on the Pan-STARRS1 test set.

MODEL	PRETRAINED	F1-SCORE
MOE-ASTRO (TM)	NO	0.373 ± 0.007
MOE-ASTRO (TM)	YES	$\mathbf{0.542 \pm 0.023}$

After pretraining, we evaluate the models in a classification setting. Table 2 reports the mean and standard deviation of macro F_1-scores on the ELAsTiCC-1 dataset using a multimodal approach that integrates both light curve and metadata features. The pretrained Multiband-Astromer baseline underperforms compared to its non-pretrained version ($F_1 = 0.727 \pm 0.034$ vs. $F_1 = 0.786 \pm 0.013$), suggesting that pretraining on the same dataset may result in early convergence and reduced adaptability during downstream fine-tuning. We also include ATAT (Cabrera-Vives et al., 2024), a non-pretrained transformer tailored for multimodal inputs, which achieves a competitive score of ($F_1 = 0.826 \pm 0.005$) under identical evaluation conditions. Our MoE-Astro model, which incorporates sparse expert routing in place of standard transformer components, yields consistent performance improvements. Specifically, the pretrained MoE-Astro with positional encoding (PE) reaches ($F_1 = 0.822 \pm 0.008$), while its variant with time modulation improves to ($F_1 = 0.832 \pm 0.015$). Notably, when trained from scratch, MoE-Astro (TM) achieves the highest performance with ($F_1 = 0.860 \pm 0.003$). These results suggest that MoE-based architectures can provide enhanced capacity allocation and generalization (see Appendix B). While the benefits of pretraining may depend on task similarity and dataset diversity, our approach currently achieves state-of-the-art performance on ELAsTiCC-1.

To assess the transferability of the best pretrained model (Astro-MoE with TM), we evaluate it on the challenging and imbalanced PS1 dataset. As shown in Table 3, pretraining improves performance from an F_1 score of 0.373 to 0.542, suggesting that Astro-MoE is capable of generalizing to new domains. This preliminary result highlights its potential for cross-survey applications and motivates further exploration across diverse real-world datasets.

5. Conclusion

We have empirically found evidence that sparse MoE models offer clear benefits for multiband astronomical time se-

ries analysis, with advantages observable even at initial scales. Their ability to allocate capacity dynamically, combined with efficient computation, makes them well-suited for large-scale pretraining on the heterogeneous and growing datasets of time-domain astronomy. Beyond improving performance, these architectures also offer practical benefits: by activating only a subset of experts per input, they reduce the computational cost during inference. This property makes them particularly attractive for real-time applications, such as transient classification in astronomical alert streams. Overall, Astro-MoE represents a promising direction for developing scalable and adaptive models to support the next generation of astronomical discovery.

Acknowledgments

The authors acknowledge support from the National Agency for Research and Development (ANID) grants: FONDECYT regular 1231877 (GCV, DMC); Millennium Science Initiative Program ICN12_009 (GCV). The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

References

Allam Jr, T. and McEwen, J. D. Paying attention to astronomical transients: introducing the time-series transformer for photometric classification. *RAS Techniques and Instruments*, 3(1):209–223, 2024.

Becker, I., Protopapas, P., Catelan, M., and Pichara, K. Multiband embeddings of light curves. *Astronomy & Astrophysics*, 694:A183, 2025.

Cabrera-Vives, G., Moreno-Cartagena, D., Astorga, N., Reyes-Jainaga, I., Förster, F., Huijse, P., Arredondo, J., Arancibia, A. M., Bayo, A., Catelan, M., et al. Atat: Astronomical transformer for time series and tabular data. *Astronomy & Astrophysics*, 689:A289, 2024.

Cádiz-Leyton, M., Cabrera-Vives, G., Protopapas, P., Moreno-Cartagena, D., Donoso-Oliva, C., and Becker, I. Uncertainty estimation for time series classification: Exploring predictive uncertainty in transformer-based models for variable stars. *arXiv preprint arXiv:2412.10528*, 2024.

Charnock, T. and Moss, A. Deep recurrent neural networks for supernovae classification. *The Astrophysical Journal Letters*, 837(2):L28, 2017.

Donoso-Oliva, C., Becker, I., Protopapas, P., Cabrera-Vives, G., Vishnu, M., and Vardhan, H. Astromer-a transformer-based embedding for the representation of light curves. *Astronomy & Astrophysics*, 670:A54, 2023.

Donoso-Oliva, C., Becker, I., Protopapas, P., Cabrera-Vives, G., Cádiz-Leyton, M., and Moreno-Cartagena, D. Astromer 2. *arXiv preprint arXiv:2502.02717*, 2025.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Graham, M. J., Djorgovski, S. G., Mahabal, A., Donalek, C., Drake, A., and Longo, G. Data challenges of time domain astronomy. *Distributed and Parallel Databases*, 30:371–384, 2012.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., Alonso, D., AlSayyad, Y., Anderson, S. F., Andrew, J., et al. Lsst: from science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, 2019.

Kasliwal, M., Cannella, C., Bagdasaryan, A., Hung, T., Feindt, U., Singer, L., Coughlin, M., Fremling, C., Walters, R., Duev, D., et al. The growth marshal: a dynamic science portal for time-domain astronomy. *Publications of the Astronomical Society of the Pacific*, 131 (997):038003, 2019.

Lanusse, F., Parker, L. H., Golkar, S., Bietti, A., Cranmer, M., Eickenberg, M., Krawezik, G., McCabe, M., Ohana, R., Pettee, M., et al. Astroclip: Cross-modal pre-training for astronomical foundation models. In *NeurIPS 2023 AI for Science Workshop*, 2023.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Liu, X., Liu, J., Woo, G., Aksu, T., Liu, C., Savarese, S., Xiong, C., and Sahoo, D. Mixture of experts for time series foundation models. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

Masoudnia, S. and Ebrahimpour, R. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42: 275–293, 2014.

Moreno-Cartagena, D., Cabrera-Vives, G., Protopapas, P., Donoso-Oliva, C., Pérez-Carrasco, M., and Cádiz-Leyton, M. Positional encodings for light curve transformers: Playing with positions and attention. In *Machine Learning for Astrophysics Workshop, 40th International Conference on Machine Learning (ICML), PMLR 202*, Honolulu, Hawaii, USA, 2023.

Morvan, M., Nikolaou, N., Yip, K., and Waldmann, I. Don't pay attention to the noise: Learning self-supervised representations of light curves with a denoising time series

transformer. *Machine Learning for Astrophysics*, pp. 11, 2022.

Mu, S. and Lin, S. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*, 2025.

Naul, B., Bloom, J. S., Pérez, F., and van der Walt, S. A recurrent neural network for classification of unevenly sampled variable stars. *Nature Astronomy*, 2(2):151–155, 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, 2024.

Park, J. W., Villar, A., Li, Y., Jiang, Y.-F., Ho, S., Lin, J. Y.-Y., Marshall, P. J., and Roodman, A. Inferring black hole properties from astronomical multivariate time series with bayesian attentive neural processes. In *Uncertainty and Robustness in Deep Learning Workshop, 38th International Conference on Machine Learning (ICML), PMLR 139*, 2021.

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 Collaboration, T. P. A. AstroCLIP: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011, 06 2024. ISSN 0035-8711. doi: 10.1093/mnras/stae1450. URL https://doi.org/10.1093/mnras/stae1450.

Pimentel, Ó., Estévez, P. A., and Förster, F. Deep attention-based supernovae classification of multiband light curves. *The Astronomical Journal*, 165(1):18, 2022.

Protopapas, P. Recurrent neural network applications for astronomical time series. In *American Astronomical Society Meeting Abstracts# 230*, volume 230, pp. 104–03, 2017.

Rizhko, M. and Bloom, J. S. Self-supervised multimodal model for astronomy. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Shi, X., Wang, S., Nie, Y., Li, D., Ye, Z., Wen, Q., and Jin, M. Scaling to billion parameters for time series foundation models with mixture of experts. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhang, M., Wu, F., Bu, Y., Li, S., Yi, Z., Liu, M., and Kong, X. Spt: Spectral transformer for age and mass estimations of red giant stars. *Astronomy & Astrophysics*, 683:A163, 2024.

A. Classification scheme

Table 4 provides an overview of the classification taxonomy used in our work, summarizing the number of objects per class across two datasets: ELAsTiCC-1 and Pan-STARRS1 (PS1). The ELAsTiCC-1 taxonomy is structured into three primary variability types (transient, stochastic, and periodic) encompassing a wide range of astrophysical phenomena including supernovae subtypes (e.g., Ia, II, Iax), cataclysmic variables (e.g., Dwarf Novae), and pulsating stars (e.g., Delta Scuti, RR Lyrae). The PS1 dataset, by contrast, focuses exclusively on periodic variables, reflecting its strengths in long-term monitoring of the sky. The pronounced class imbalance poses challenges for training robust machine learning classifiers and underscores the importance of methods capable of handling skewed data distributions.

Table 4. Number of objects per class in ELAsTiCC-1 and PS1, grouped by variability type. Each group lists the included classes and the number of objects in parentheses.

Group	**Classes (number of objects)**
ELAsTiCC-1 Transient (12)	CART (15,719), Iax (53,727) 91bg (53,414), Ia (211,892) Ib/c (310,328), II (445,419) SN-like/Other (103,683), SLSN (105,238) PISN (105,446), TDE (103,067) ILOT (14,253), KN (8,122)
ELAsTiCC-1 Stochastic (4)	M-dwarf Flare (2,640), uLens (27,263) Dwarf Novae (12,385), AGN (99,461)
ELAsTiCC-1 Periodic (4)	Delta Scuti (29,840), RR Lyrae (21,100) Cepheid (25,371), EB (96,778)
Pan-STARRS1 Periodic (6)	RRab (10,000), RRc (10,000) RRd (266), MIRA_SR (3,937) DSCT_SXPHE (1,906), T2CEP (189)

B. Astro-MoE confusion matrices

Figures 2, 3, and 4 present detailed visualizations of our proposed sparse MoE model's classification performance across various astronomical object types. To ensure robust and interpretable insights, we analyze confusion matrices aggregated over multiple evaluation runs. Each matrix is row-normalized so that each row sums to 100%, reflecting the distribution of predicted classes for each true class label. We report the median values across runs to provide a stable central estimate, which is less affected by outliers compared to the mean. To capture variability, we also compute the 25th and 95th percentiles, offering a clear view of the range in classification performance.

Each cell in the visualization conveys three metrics: the median prediction percentage (center), the 95th percentile (top right), and the 25th percentile (bottom right). Diagonal cells represent correct classifications, while off-diagonal cells indicate misclassifications. The color intensity encodes the prediction percentage, with darker green denoting higher accuracy. For visual clarity, cells with values below 0.05% are left blank.

C. Parameter count comparison

We introduce a Mixture-of-Experts variant of our model to enable dynamic routing through specialized feedforward layers. This architecture substantially increases model capacity, with approximately 3× more total parameters (4.9M vs. 1.5M). However, thanks to its sparse activation mechanism, the GPU inference time per batch increases only modestly, from 5.3ms to 7.1ms.

Critically, the MoE model activates only a subset of experts per forward pass, specifically, 2 out of 8 experts in the feedforward network and 6 in the positional embedding layer. This sparsity allows the model to retain high expressiveness without incurring the full computational cost of using all parameters. Moreover, this design opens up avenues for further

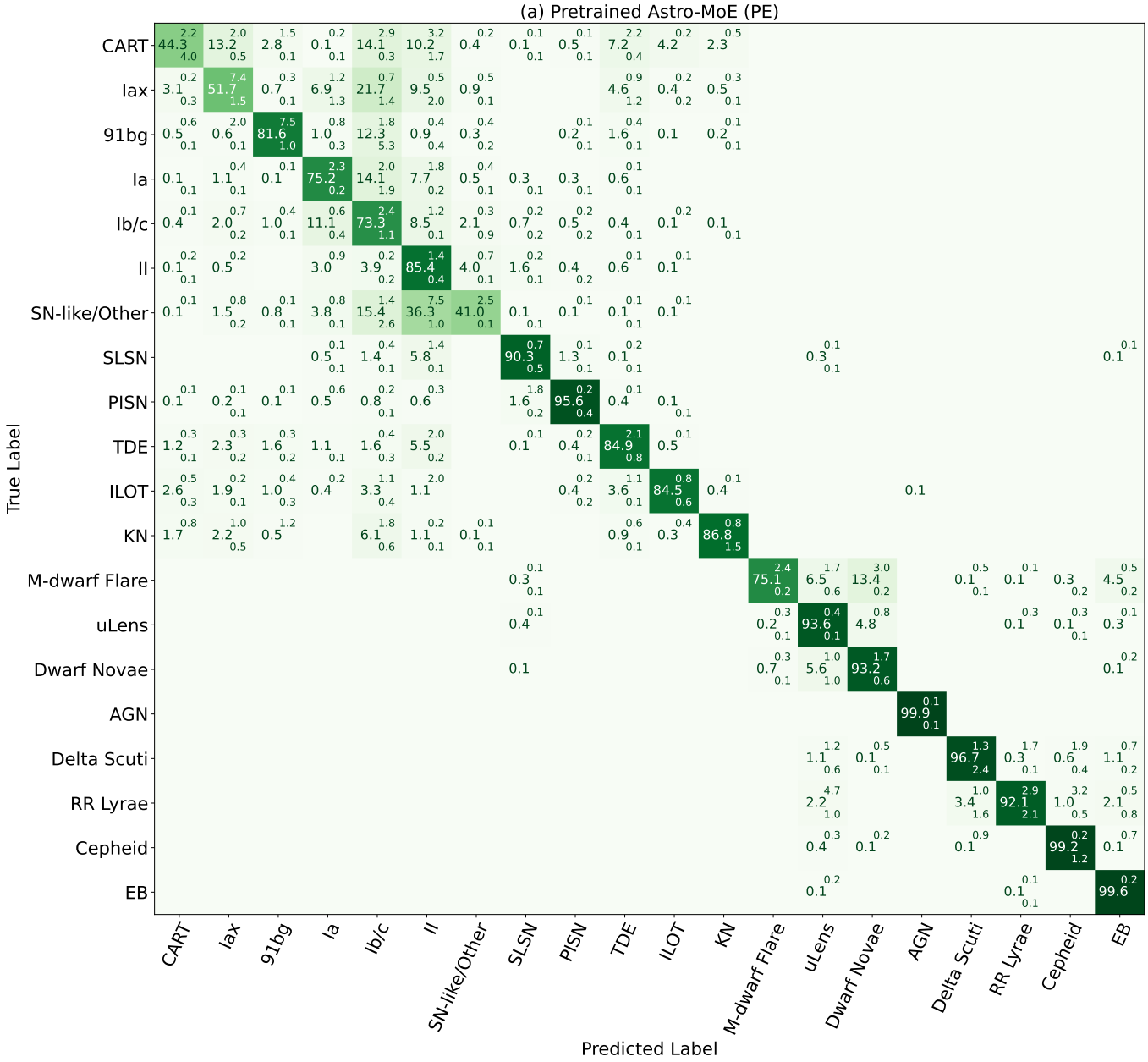

Figure 2. Pretrained Astro-MoE (PE) confusion matrix.

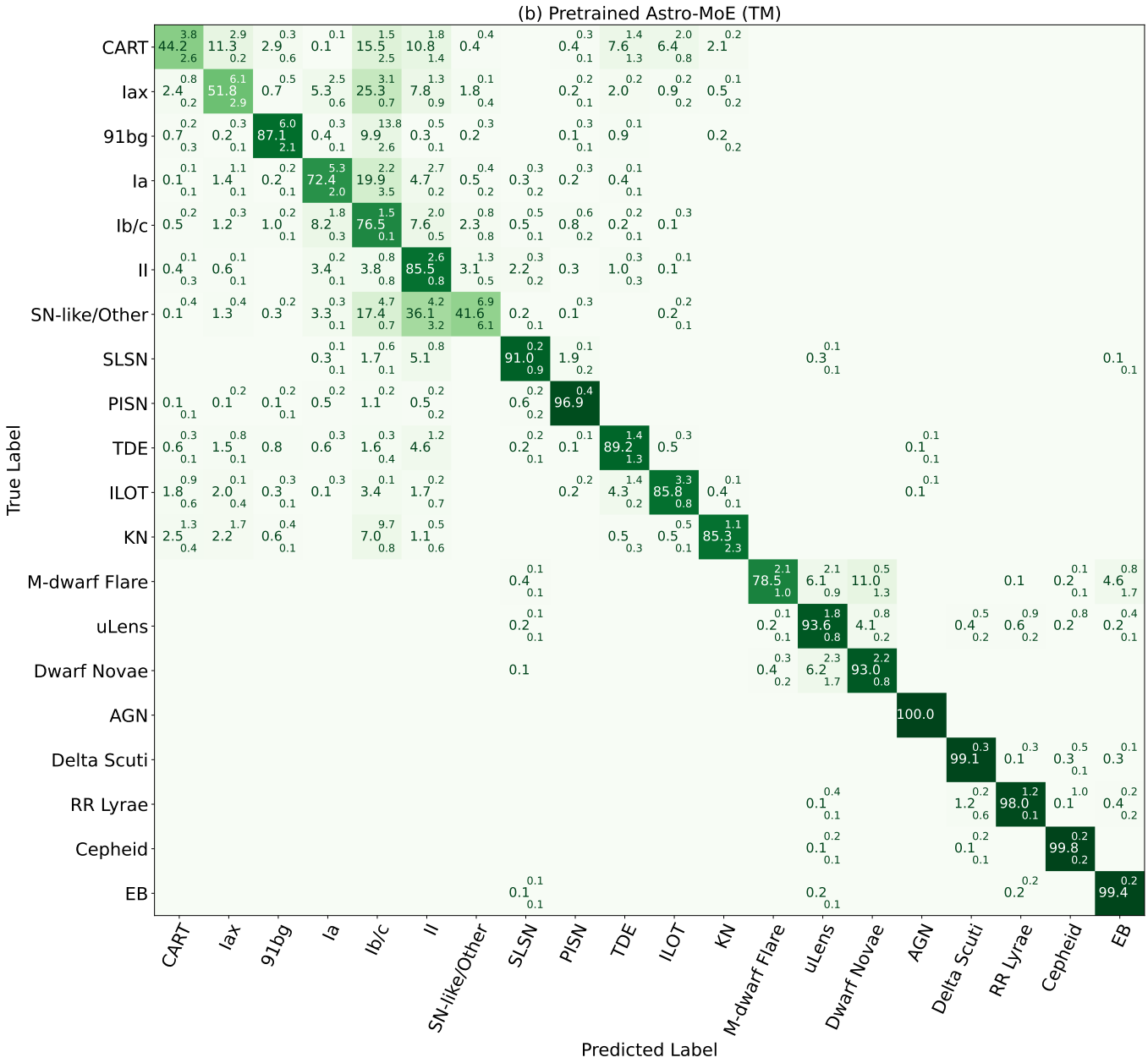

Figure 3. Pretrained Astro-MoE (TM) confusion matrix.

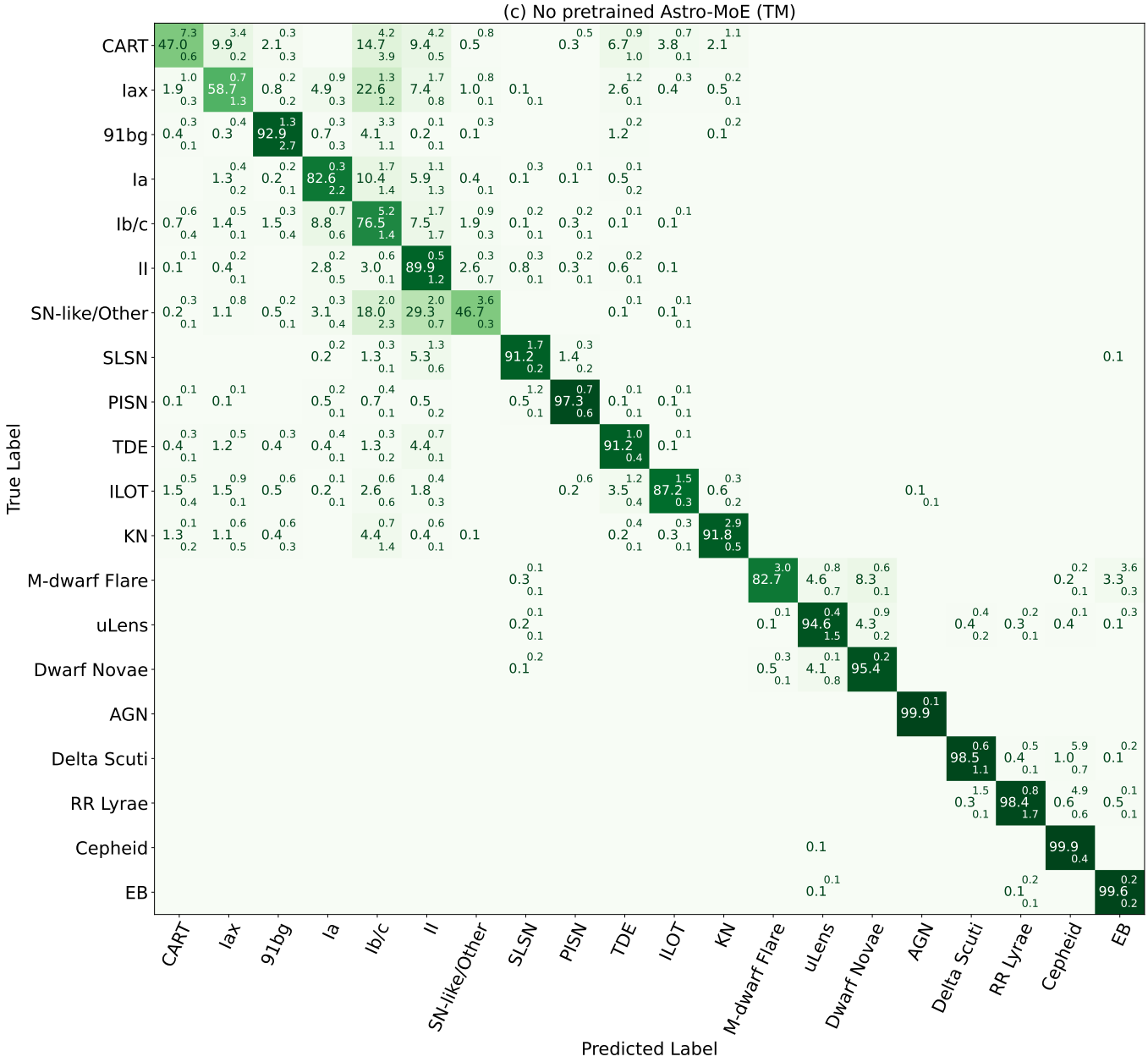

Figure 4. Non-pretrained Astro-MoE (TM) confusion matrix.

Table 5. Comparison of model capacity and inference efficiency across variants for the ELAsTiCC classification task.

MODEL	TE	PARAMS	INFERENCE TIME (MS)
MULTIBAND-ASTROMER	PE	1.5M	5.3 MS
MOE-ASTRO	PE	4.6M	6.8 MS
MOE-ASTRO	TM	4.9M	7.1 MS

optimization: reducing the number of active experts, for instance, could further lower inference time without necessarily compromising performance.

This tradeoff highlights a core advantage of MoE architectures; their ability to scale capacity without a linear increase in computation or latency. In the context of ELAsTiCC, where classification involves a wide range of astrophysical phenomena and complex temporal dynamics, such additional representational power is beneficial. Overall, our results suggest that the increased capacity of the MoE model justifies the minor overhead, providing a scalable and efficient solution for modeling heterogeneous multiband time series data.