The Astronomical Journal, 170:28 (13pp), 2025 July https://doi.org/10.3847/1538-3881/adcbad

AstroM^3: A Self-supervised Multimodal Model for Astronomy

M. Rizhko[1] and J. S. Bloom[1,2]

[1] University of California, Berkeley, Department of Astronomy, Berkeley, CA, USA
[2] Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Received 2024 November 19; revised 2025 March 27; accepted 2025 April 9; published 2025 June 10

Abstract

While machine-learned models are now routinely employed to facilitate astronomical inquiry, model inputs tend to be limited to a primary data source (namely images or time series) and, in the more advanced approaches, some metadata. Yet with the growing use of wide-field, multiplexed observational resources, individual sources of interest often have a broad range of observational modes available. Here we construct an astronomical multimodal dataset and propose AstroM^3, a self-supervised pretraining approach that enables a model to learn from multiple modalities simultaneously. We extend the Contrastive Language-Image Pretraining (CLIP) model to a trimodal setting, allowing the integration of time-series photometry data, spectra, and astrophysical metadata. In a fine-tuning supervised setting, CLIP pretraining improves classification accuracy, particularly when labeled data is limited, with increases of up to 14.29% in spectra classification, 2.27% in metadata, and 10.20% in photometry. Furthermore, we show that combining photometry, spectra, and metadata improves classification accuracy over single-modality models. In addition to fine-tuned classification, we can use the trained model in other downstream tasks that are not explicitly contemplated during the construction of the self-supervised model. In particular we show the efficacy of using the learned embeddings to identify misclassifications, for similarity search, and for anomaly detection. One surprising highlight is the “rediscovery” of Mira subtypes and two rotational variable subclasses using manifold learning and dimensionality reduction algorithms. To our knowledge this is the first construction of an $n > 2$ mode model in astronomy. Extensions to $n > 3$ modes are naturally anticipated with this approach.

Unified Astronomy Thesaurus concepts: Astrostatistics techniques (1886); Variable stars (1761)

1. Introduction

Despite the vast volumes of publicly available raw astronomical data, with a few notable subfield exceptions, the application of machine learning to discovery and inference has yet to broadly permeate the field. One impediment stems from the challenge of fusing data across heterogeneous modes of collection. Off-the-shelf architectures do not easily accommodate a mixture of irregularly sampled multispectral multiscale heteroskedastic time-series data, images, spectra, and metadata. Another issue, arising in the classification context, is that very few ground-truth labels exist. This “small label” problem arose, for example, in J. W. Richards et al. (2012), who sought to probabilistically classify 50,124 variable stars using only 810 labels over 28 classes. Last, models learned on a dataset from one survey do not easily transfer to other data collected on the same objects from different surveys (e.g., J. P. Long et al. 2012; D.-W. Kim et al. 2021). Our self-supervised multimodal architecture addresses the first two challenges, establishing methods and milestones for a more generalized foundation model applicable to inference tasks on unseen survey data.

Our work builds upon the Contrastive Language-Image Pretraining (CLIP) framework, originally introduced by A. Radford et al. (2021): CLIP demonstrated the power of contrastive learning on large-scale image and text datasets to learn joint representations. Since its introduction, CLIP has been extensively researched and improved in various ways. For example, Y. Li et al. (2021) enhanced data efficiency through supervision, while L. Yao et al. (2021) focused on improving semantic alignment. M. Cherti et al. (2023) introduced scaling laws, and Q. Sun et al. (2023) optimized the model for faster training. Additionally, CLIP has been combined with other pretraining objectives: N. Mu et al. (2022) incorporated image self-supervision, and A. Singh et al. (2022) along with J. Li et al. (2022) added masked multimodal, image, and language modeling. Furthermore, CLIP has been extended to other modalities: audio-text (Y. Wu et al. 2023), video-text (H. Luo et al. 2021; H. Xu et al. 2021; Y. Ma et al. 2022), and point cloud-text (R. Zhang et al. 2022). In the astronomical context, L. Parker et al. (2024) used dual-mode CLIP on static-sky galaxy images and spectra. Closest to the approach of our work outside of astronomy, A. Guzhov et al. (2022) adapted CLIP for use with three modalities: audio, image, and text. Given the proven versatility and success of CLIP in different domains, we build upon it herein. We extend CLIP to work on three modalities: time-series photometry, spectra, and metadata (see Figure 1). Our work, and a recent preprint from G. Zhang et al. (2024), are the first efforts to incorporate time-series data with CLIP, and our three-mode model represents a critical step toward the development of a foundational multimodal model for time-domain astronomy.

In this work, we introduce AstroM^3 as both a methodological framework and a practical tool for the broader astronomy community. Our approach demonstrates how self-supervised multimodal learning can integrate diverse astronomical data types, providing a foundation for models that generalize across different datasets and classification tasks. To ensure accessibility and usability, we release the full implementation of AstroM^3 along with pretrained and fine-tuned models, enabling researchers to leverage our framework for their specific applications. The source code is available at GitHub.[3] We provide two versions of the

[3] https://github.com/MeriDK/AstroM3/

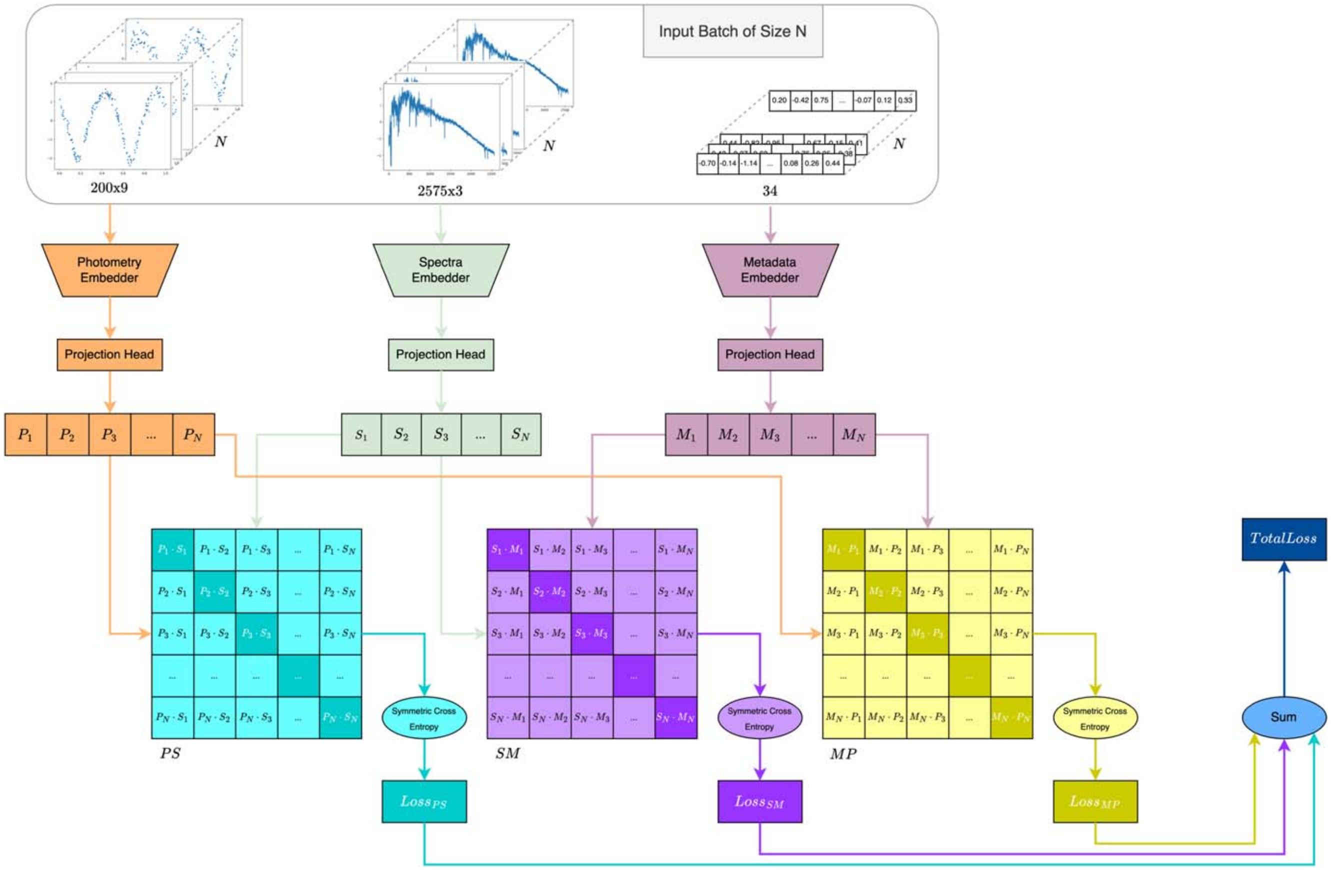

Input Batch of Size N
N
200x9
2575x3
34
0.20 -0.42 0.75 ... -0.07 0.12 0.33
-0.70 -0.14 -1.14 ... 0.08 0.26 0.44
Photometry Embedder
Spectra Embedder
Metadata Embedder
Projection Head
P_1 P_2 P_3 ... P_N
S_1 S_2 S_3 ... S_N
M_1 M_2 M_3 ... M_N
$P_1 \cdot S_1$ $P_1 \cdot S_2$ $P_1 \cdot S_3$ ... $P_1 \cdot S_N$
$P_N \cdot S_1$ $P_N \cdot S_2$ $P_N \cdot S_3$ ... $P_N \cdot S_N$
$S_1 \cdot M_1$ $S_1 \cdot M_2$ $S_1 \cdot M_3$ ... $S_1 \cdot M_N$
$S_N \cdot M_1$ $S_N \cdot M_2$ $S_N \cdot M_3$ ... $S_N \cdot M_N$
$M_1 \cdot P_1$ $M_1 \cdot P_2$ $M_1 \cdot P_3$ ... $M_1 \cdot P_N$
$M_N \cdot P_1$ $M_N \cdot P_2$ $M_N \cdot P_3$ ... $M_N \cdot P_N$
PS
SM
MP
Symmetric Cross Entropy
$Loss_{PS}$
$Loss_{SM}$
$Loss_{MP}$
Sum
TotalLoss

Figure 1. Overview of the multimodal CLIP framework adapted for astronomy, incorporating three data modalities: photometric time-series, spectra, and metadata. Each modality is processed by a dedicated encoder to create embeddings, which are then mapped into a shared embedding space through projection heads. Pairwise similarity matrices align the embeddings across modalities, and a symmetric cross-entropy loss, computed over these matrices, optimizes the model. The total loss, derived from all pairwise losses, guides the model's trimodal learning.

dataset: the raw multimodal dataset AstroM3Dataset[4] and a processed version optimized for training AstroM3Processed.[5] Additionally, we release the weights[6] of the pretrained AstroM^3 and fine-tuned versions specialized for different modalities: multimodal classification,[7] photometry classification,[8] spectra classification,[9] and metadata classification.[10]

2. Related Work

Early classification-focused research used hand-crafted features of time-series photometry and metadata with decision forests in a supervised context (J. Debosscher et al. 2007; P. Dubath et al. 2011; J. W. Richards et al. 2011; L. Palaversa et al. 2013). Neural network approaches to learn representations of time-series photometry (both in supervised and self-supervised contexts) then achieved state of the art, first with flavors of Recurrent Neural Networks (e.g., long short-term memory: B. Naul et al. 2018; Gated Recurrent Unit: D. Muthukrishna et al. 2019; I. Becker et al. 2020) and more recently with Convolutional Neural Networks (CNN; S. Jamal & J. S. Bloom 2020; K. Boone 2021) and Transformers (C. Donoso-Oliva et al. 2023; H. W. Leung & J. Bovy 2024). CNNs have been used to achieve state of the art classification on galaxy spectra (e.g., GalSpecNet: Y. Wu et al. 2024). M. A. Hayat et al. (2021) use CNN autoencoders with contrastive learning for self-supervised embedding of galaxy images.

AstroCLIP (L. Parker et al. 2024) fused pretrained embeddings of galaxy spectra and images with contrastive learning and showed the trained model to be competitive with purpose-built classification models. Our work differs from AstroCLIP in that (1) our primary objects are individual sources that vary in time (ie., not static like galaxies); (2) we explicitly build embeddings for three different modes of data; (3) our approach does not rely upon pretraining of embeddings for the different modes, but instead learns all embeddings simultaneously; and (4) we examine the efficacy of the model with missing modes at test time. Like with AstroCLIP, we find our model outperforms purpose-built supervised models for downstream tasks. To our knowledge, MAVEN (G. Zhang et al. 2024) is the only other CLIP-centric model applied in the astronomical time domain. It is a dual-mode model built for "one off" explosive supernovae events, whereas ours is focused on persistently variable sources. MAVEN first learns spectroscopic and photometric embeddings from synthetic data and then requires a fine-tuning step on real survey data. Our model is trained directly on real observational data.

[4] https://huggingface.co/datasets/AstroMLCore/AstroM3Dataset
[5] https://huggingface.co/datasets/AstroMLCore/AstroM3Processed
[6] https://huggingface.co/AstroMLCore/AstroM3-CLIP
[7] https://huggingface.co/AstroMLCore/AstroM3-CLIP-all
[8] https://huggingface.co/AstroMLCore/AstroM3-CLIP-photo
[9] https://huggingface.co/AstroMLCore/AstroM3-CLIP-spectra
[10] https://huggingface.co/AstroMLCore/AstroM3-CLIP-meta

3. Dataset Assembly

The basis of our observational dataset is the variable star catalog (T. Jayasinghe et al. 2019) observed and curated by the All-Sky Automated Survey for SuperNovae (ASAS-SN) project (B. J. Shappee et al. 2014). We downloaded the light curve data from the 2021 assembly of the 687,695 v-band variables and the 2022 assembly of the 378,861 g-band variables, along with the associated metadata catalogs. These catalogs contain crossmatched photometry information for each source from WISE (E. L. Wright et al. 2010), GALEX (P. Morrissey et al. 2007), 2MASS (M. F. Skrutskie et al. 2006), and Gaia EDR3 (Gaia Collaboration et al. 2021); variability statistics derived from the lightcurves in each bandpass (such as period and peak-to-peak amplitude); astrometric information from Gaia (such as parallax and proper motion); and a machine-learned classification from the ASAS-SN group (T. Jayasinghe et al. 2019). We deduplicated and merged these data using the crossmatched `source_id` from Gaia EDR3, with the merged catalog serving as the basis of the `metadata` mode.

To facilitate the use of positional information in the models, we transformed the galactic latitude $b \rightarrow \sin(b)$ and galactic longitude $l \rightarrow \cos(l)$. We also transformed all catalog apparent photometry m to absolute magnitude using the Gaia EDR3 parallax π (units of milliarcseconds) using $M = m + 5\log_{10}\pi - 10$. We did not deredderen any values. To cleanly delineate the `time-series` mode from the `metadata` mode, we removed features derived from photometric time-series data from the `metadata` catalog (and later used such features as auxiliary inputs in the `time-series` channel; see Section 4.1 below). We also removed any columns from the `metadata` catalog related to indices (such as source names). Last, we removed the assigned classification of each source (later used to test downstream tasks; see Section 5). The final metadata feature set consists of 34 parameters (see Table 6 for details).

To build the `spectral` mode, we crossmatched the sources with the v2.0 DR9 Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST; X.-Q. Cui et al. 2012) public catalog using the Gaia EDR3 ID. We downloaded the 41,204 1D spectra identified in the crossmatch and constructed a lookup table matching specific variable sources to LAMOST spectra. Most variable sources had zero associated spectra but a small subset had multiple spectra of the same source obtained over multiple epochs. For sources with multiple spectra, we randomly selected only one spectrum per source.

We filtered the dataset based on the following criteria: (1) each object must have data for all three modalities-time-series photometry, spectra, and metadata; (2) objects with any missing metadata were excluded completely; and (3) the object must belong to one of the 10 most common classes to ensure there are sufficient samples for effective CLIP training (H. Xu et al. 2023; I. Alabdulmohsin et al. 2024). The initial dataset contained 687,695 unique sources from the ASAS-SN catalog. After crossmatching and selecting sources with photometry, spectra, and complete metadata, the sample was reduced to 22,008 objects. Finally, filtering to the top 10 most common classes resulted in the final dataset of 21,440 objects. The selected classes and the corresponding number of objects are listed in Table 1.

Table 1
Summary of Variable Star Classes, Including Abbreviations, Descriptions, and Total Object Counts for Each Class Used in the Dataset

Class	Description	Total
EW	W Ursae Majoris type binaries	6168
SR	Semiregular variables	4590
EA	Detached Algol-type binaries	2916
RRAB	Fundamental Mode RR Lyrae variables	2351
EB	β Lyrae-type binaries	1976
ROT	Spotted Variables with rotation	1839
RRC	First Overtone RR Lyrae variables	796
HADS	High-amplitude δ Scuti type variables	281
M	Mira variables	268
DSCT	δ Scuti type variables	255
Total		21,440

To advance research in multimodal learning and modality alignment in astronomy, we release the AstroM^3 dataset in two formats: the original dataset with a custom loading script for exploratory data analysis, and a preprocessed Parquet version optimized for training. To ensure reproducibility, we provide predefined training, validation, and test splits, along with multiple subset configurations across different dataset sizes. The dataset is hosted on Hugging Face Datasets and can be easily accessed using the `datasets` library (see Appendix C).

4. Method

Our objective is to develop a self-supervised multimodal model that can learn from astronomical data across three distinct modalities: time-series photometry, spectra, and astrophysical metadata. Since each modality has unique data characteristics, different encoder architectures are required: sequence models for time-series photometry to capture temporal dependencies, convolutional networks for spectra to extract local features, and dense neural networks for structured metadata. To integrate these diverse data types, we extend the CLIP framework (A. Radford et al. 2021) to a trimodal setting, enabling simultaneous learning from multiple data types. This framework is inherently flexible and can be extended to additional modalities, such as imaging, by incorporating an image encoder into the contrastive learning setup. In this section, we describe the models used for each modality and how they are integrated into our multimodal CLIP framework.

4.1. Photometric Time-series Model

Photometric time-series data are flux measurements of astronomical objects over time. To effectively capture the temporal dependencies and handle sequences of varying lengths, we employ the Encoder component from the Informer model (H. Zhou et al. 2021).

Model Architecture. The photometric time-series encoder consists of:

1. Input embedding layer: projects the input features to a higher-dimensional space.
2. Informer encoder layers: eight encoder layers with a hidden dimension of 128, four attention heads, and a feedforward dimension of 512.
3. Output layer: produces a fixed-length embedding representing the input time-series data.

Data Preprocessing. Each light curve is a sequence of flux measurements $f = \{f_1, f_2, ..., f_T\}$ and flux errors $\sigma_f = \{\sigma_{f_1}, \sigma_{f_2}, \ldots, \sigma_{f_T}\}$ at corresponding times $t = \{t_1, t_2, ..., t_T\}$. We normalize the flux by subtracting the mean μ_f and dividing by the median absolute deviation MAD_f: $\tilde{f}_i = \frac{f_i - \mu_f}{\mathrm{MAD}_f}$. Flux errors are normalized by the flux median absolute deviation division: $\tilde{\sigma}_{f_i} = \frac{\sigma_{f_i}}{\mathrm{MAD}_f}$. Time is scaled between 0 and 1 for each light curve: $\delta_t = t_{\max} - t_{\min}$; $\tilde{t}_i = \frac{t_i - t_{\min}}{\delta_t}$. Auxiliary features such as amplitude, period, Lafler–Kinman string length statistic (J. Lafler & T. D. Kinman 1965), peak-to-peak variability, delta time $\frac{\delta_t}{365}$, and logarithm of median absolute deviation $\log \mathrm{MAD}_f$ are included as additional inputs.

Handling Variable Sequence Lengths. We set a maximum sequence length of L = 200. Sequences longer than this are randomly cropped during training and center cropped during validation and testing. Shorter sequences are padded with zeros, and an attention mask is used to differentiate between valid data and padding. This allows for a balance of computational efficiency and the ability to capture temporal trends.

4.2. Spectra Model

Spectral data provides detailed information about the composition and physical properties of astronomical objects. We adapt the GalSpecNet architecture (Y. Wu et al. 2024), which is specifically designed for processing one-dimensional astronomical spectra.

Model Architecture. The spectra encoder consists of:

1. Convolutional layers: four layers (64, 64, 32, 32 channels) followed by ReLU activations.
2. Pooling layers: max-pooling layers after each convolutional layer except for the last one.
3. Dropout layer: applied after the last convolutional layer for regularization.
4. Output layer: generates a fixed-length embedding of the spectral data.

Modifications. We reduce the last three fully connected layers to a single one for classification or omit it entirely when using the model as a feature extractor. We also add additional input channels for spectra errors and auxiliary data.

Data Preprocessing. Spectra are limited to the wavelength range of 3850–9000 Å and resampled at regular intervals of 2 Å using linear interpolation. Each spectrum $s = \{s_1, s_2, ..., s_W\}$ and its uncertainties $\sigma_s = \{\sigma_{s_1}, \sigma_{s_2}, \ldots, \sigma_{s_W}\}$ at corresponding wavelengths $w = \{w_1, w_2, ..., w_W\}$ are normalized in a similar way as photometry data; values are normalized by subtracting the mean μ and dividing by the median absolute deviation MAD_s: $\tilde{s}_i = \frac{s_i - \mu_s}{\mathrm{MAD}_s}$; while uncertainties are divided by MAD_s: $\tilde{\sigma}_{s_i} = \frac{\sigma_{s_i}}{\mathrm{MAD}_s}$. The logarithm of the median absolute deviation $\log \mathrm{MAD}_s$ is included as an auxiliary feature.

4.3. Metadata Model

The metadata modality consists of astrophysical parameters and observational data not included in the other two modalities. This includes features like absolute magnitudes in various bands, astrometric information, and other crossmatched catalog data. A full list of features and their descriptions is provided in Table 6.

*Model Architecture***.** The metadata encoder is a Multilayer Perceptron consisting of:

1. Input layer: accepts the 34 preprocessed features.
2. Hidden layers: two hidden layers with 512 units each followed by ReLU activations.
3. Dropout layers: applied after hidden layers for regularization.
4. Output layer: provides a fixed-length metadata embedding.

Data Preprocessing. Except for the steps already mentioned during the dataset assembly (see Section 3), we apply logarithm to period and then standardize each feature to have zero mean and unit variance.

4.4. AstroM^3: Multimodal CLIP Model

To integrate the three modalities we extend the CLIP model to a trimodal setting and name the entire architectural approach as AstroM^3. The goal is to learn a shared embedding space where representations from different modalities corresponding to the same astronomical object are close together (see Figure 1).

Projection Heads. Each modality has its own architecture, producing embeddings of different sizes. To bring these embeddings into a shared space, we apply a projection head to each modality. The projection head is a fully connected layer that maps the embeddings to a fixed size of 512. Let the original embeddings of photometry, spectra, and metadata be denoted as $\tilde{P}_i$, $\tilde{S}_i$, and $\tilde{M}_i$, where i denotes the i-th sample in a batch of size N. The projection heads transform these original embeddings as follows:

$$P_i = W_P \tilde{P}_i + b_P \tag{1}$$

$$S_i = W_S \tilde{S}_i + b_S \tag{2}$$

$$M_i = W_M \tilde{M}_i + b_M, \tag{3}$$

where W_P, W_S, and W_M are the weight matrices, and b_P, b_S, and b_M are the bias terms for the projection head of each modality. After applying these transformations, the projected embeddings P_i, S_i, and M_i all have a fixed size of 512, making them suitable for comparison in the shared embedding space.

Pairwise Similarity Matrices. For each pair of modalities (photometry-spectra, spectra-metadata, metadata-photometry) we compute similarity matrices using cosine similarity:

$$PS_{ij} = \frac{P_i \cdot S_j}{\|P_i\| \|S_j\|} \tag{4}$$

$$SM_{ij} = \frac{S_i \cdot M_j}{\|S_i\| \|M_j\|} \tag{5}$$

$$MP_{ij} = \frac{M_i \cdot P_j}{\|M_i\| \|P_j\|}. \tag{6}$$

Contrastive Loss. We use a symmetric cross-entropy loss to align the embeddings:

$$\mathcal{L}^{PS} = \mathcal{L}_{\mathrm{CE}}(PS, Y) + \mathcal{L}_{\mathrm{CE}}(PS^{\top}, Y) \tag{7}$$

$$\mathcal{L}^{SM} = \mathcal{L}_{\mathrm{CE}}(SM, Y) + \mathcal{L}_{\mathrm{CE}}(SM^{\top}, Y) \tag{8}$$

Table 2
Classification Accuracy Comparison Across Different Modalities, with and Without CLIP Pretraining

Data Type	No CLIP	CLIP
Spectra	76.278 ± 0.931	77.056 ± 1.033
Metadata	85.667 ± 0.858	85.864 ± 0.739
Photometry	90.724 ± 1.464	91.101 ± 0.589
All	94.127 ± 0.240	93.501 ± 0.737

Note. CLIP models were pretrained on the full dataset in a self-supervised manner before fine-tuning for classification, while no-CLIP models were trained directly on the classification task using the same dataset.

$$\mathcal{L}^{MP} = \mathcal{L}_{\text{CE}}(MP, Y) + \mathcal{L}_{\text{CE}}(MP^{\top}, Y) \tag{9}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss and Y is the label matrix defined as:

$$Y_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Total Loss. The overall loss is the sum of the individual pairwise losses:

$$\mathcal{L} = \mathcal{L}^{PS} + \mathcal{L}^{SM} + \mathcal{L}^{MP}. \tag{11}$$

By minimizing this loss, the model learns to align the embeddings across all three modalities, bringing representations of the same object closer together in the embedding space while pushing apart those of different objects.

5. Results

We evaluated the models on downstream classification across four modes: photometry only, spectra only, metadata only, and all modalities combined. For single modalities, we added a fully connected layer on top of the respective encoders for classification. In the multimodal setting, we averaged the embeddings from all three modalities and then applied a fully connected layer for classification. Each model was trained in two ways: with CLIP pretraining, where the model was initially trained using the CLIP framework and then fine-tuned for the downstream task, and without CLIP pretraining, where models were trained directly on the task with randomly initialized weights. Importantly, model architecture and setup were identical across all conditions, differing only in the initialization of weights. The training setup and hyperparameter search process are detailed in Appendix B. All models were cross validated using five random seeds and data splits for robust evaluation.

5.1. Fully Labeled Data

We compare classification accuracy across different modalities between models trained with and without CLIP pretraining, using the full dataset (21,440 objects, 10 classes from Table 1) for both pretraining and supervised classification. The results in Table 2 indicate that CLIP pretraining does not show statistically significant improvements across any of the modalities. Since CLIP is designed to leverage large, diverse datasets, its benefit may be limited when pretraining is performed solely on this dataset without external data. In this case, the representations learned during self-supervised pretraining may not provide additional useful information beyond what the supervised fine-tuning phase can capture, leading to similar performance between the models. With or without CLIP, we also show that by using all three modalities at the same time, we achieve better accuracy than by using any single modality alone.

5.2. Limited Labeled Data

To evaluate the effectiveness of CLIP pretraining when the availability of labeled data is limited, we conducted experiments on smaller subsets of the original dataset. Specifically, we created subsets containing 10%, 25%, and 50% of the data by downsampling the most common classes, ensuring a balanced class distribution. Table 3 provides details on the class distribution across these subsets. Note that we choose to downsample the overrepresented sources at random. An interesting alternative to this, to approximate the ways in which brighter sources preferentially are easier to label on new survey data, would be to select only the brightest (or highest signal-to-noise) sources to include in the training data.

Models. For each subset, we retrained all models, with and without CLIP pretraining, using the same optimization settings and hyperparameter search as previously applied. It is important to note that the CLIP model used for these experiments was the same as before: pretrained on the full dataset without using any labels. This setup is designed (for future applications) to leverage large amounts of unlabeled data for pretraining and then fine-tuning the model on smaller labeled datasets.

Results. The results in Table 4 demonstrate that CLIP pretraining improves model performance when labeled data is limited. For example, at the 50% data split, CLIP increased the average accuracy of the spectra model by 3.30% (from 68.07% to 71.38%), and by 14.29% at the 10% data split (from 46.68% to 60.96%). Metadata and photometry show a similar trend, with statistically significant accuracy gains of 2.73% and 10.20%, respectively, at the 25% data split. At the 10% split, CLIP improved metadata accuracy by 2.27% and photometry accuracy by 6.74%. For all modalities combined, although the difference in accuracy between models with and without CLIP pretraining was not statistically significant, CLIP models generally performed better. These findings suggest that CLIP is beneficial, especially when labeled training data is limited, making it an effective approach for leveraging large unlabeled datasets in future work.

5.3. UMAP Analysis

We use Uniform Manifold Approximation and Projection (UMAP) method (L. McInnes et al. 2018) to visualize how well our model distinguishes among classes in the embedding space. UMAP is fit on the averaged embeddings across all modalities from the training set, and projections are generated for both the training (Figure 2(a)) and the test (Figure 2(b)) sets. The results show that:

1. Most classes are well separated, though Detached Algol-type binaries (EA), β Lyrae-type binaries (EB), and W Ursae Majoris type binaries (EW) partially overlap. This is expected on a physical basis, as these are all types of binary stars and share similar characteristics.[11]

[11] Our photometry embedding works in time-flux space; future extensions to phase-flux space may help disambiguate the eclipsing binary classes.

Table 3
Class Distribution Across Training, Validation, and Test Sets for Different Dataset Splits (Full, 50%, 25%, 10%), Created by Downsampling the Most Common Classes to Balance Subsets

Class	Train				Val				Test			
	Full	50%	25%	10%	Full	50%	25%	10%	Full	50%	25%	10%
EW	4890	1209	516	166	597	149	64	21	681	160	69	22
SR	3647	1209	516	166	479	149	64	21	464	160	69	22
EA	2343	1209	516	166	272	149	64	21	301	160	69	22
RRAB	1886	1209	516	166	231	149	64	21	234	160	69	22
EB	1571	1209	516	166	207	149	64	21	198	160	69	22
ROT	1454	1209	516	166	189	149	64	21	196	160	69	22
RRC	624	624	516	166	93	93	64	21	79	79	69	22
HADS	226	226	226	166	29	29	29	21	26	26	26	22
M	216	216	216	166	30	30	30	21	22	22	22	22
DSCT	206	206	206	166	25	25	25	21	24	24	24	22

Table 4
Accuracy Comparison Across Data Splits (50%, 25%, 10%) with and Without CLIP Pretraining for Different Data Types (Spectra, Photometry, Metadata, All)

Data Type	Pretrain	50%	25%	10%
Spectra	No CLIP	68.072 ± 1.759	63.729 ± 1.637	46.677 ± 3.486
	CLIP	**71.376 ± 1.465**	66.676 ± 2.549	**60.963 ± 2.584**
Metadata	No CLIP	82.628 ± 1.048	78.933 ± 2.419	74.888 ± 1.856
	CLIP	83.076 ± 1.367	**81.667 ± 1.332**	**77.161 ± 1.476**
Photometry	No CLIP	85.197 ± 6.209	78.382 ± 2.359	84.527 ± 2.655
	CLIP	89.572 ± 0.652	**88.580 ± 0.948**	**91.264 ± 0.974**
All	No CLIP	91.977 ± 0.272	90.811 ± 1.030	88.261 ± 1.248
	CLIP	92.174 ± 0.719	91.243 ± 1.254	89.900 ± 2.000

Note. Statistically significant improvements in bold.

2. As expected, the test set follows the same UMAP projection structure as the training set. For instance, spotted variables with rotational modulation (ROT) from the test set align with their counterparts in the training set.

Outliers. Based on the UMAP projections, we observed that some objects were located outside their expected clusters. To investigate further, we trained a DBSCAN model (M. Ester et al. 1996) on each class, configuring it to identify a single major cluster per class, with all objects outside of that cluster marked as outliers. We manually reviewed the objects flagged as outliers and found that most objects are falling into two categories: (1) objects with incorrectly assigned classifications from the catalog, and (2) objects with the correct labels that are in-class outliers because of their unique features.

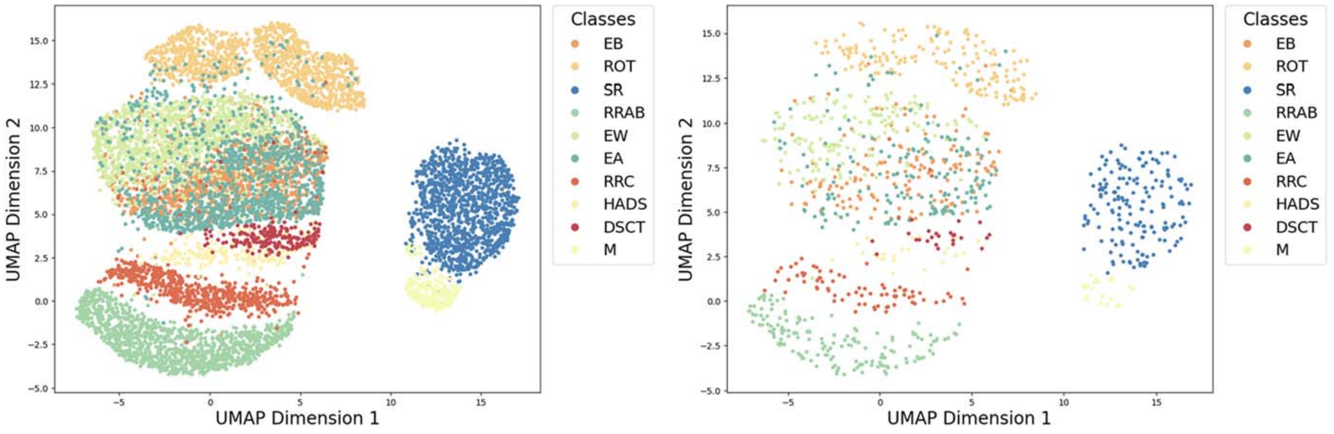
Misclassifications. Figure 3 highlights misclassification candidates, showing both the photometry and spectrum for representative examples summarized below:

1. EDR3 854619503161255424, likely EW binary: the reported Gaia period is twice that of the catalog (0.2780335 day), suggesting this source is likely an EW binary. The lack of the asymmetric shape typical of a High Amplitude Delta Scuti (HADS) star supports this reclassification.
2. EDR3 3161660872675410560, EB: this source, V* AC CMi, is a known EB, suggesting that the RR Lyrae classification is incorrect.
3. EDR3 270394132885715456, possible SR or Mira variable: Gaia lists half the period (102 day) compared to the catalog, but the catalog period appears correct. An SR or Mira classification is likely more appropriate.
4. EDR3 1993171318512584832, known Mira variable: this source, V0439 Cas, is a known Mira variable, indicating that its current SR classification is inaccurate.
5. EDR3 3411304197488061056, likely EW binary with incorrect catalog period: Gaia classifies this object as an eclipsing binary, which aligns better with an EW (W UMa-type contact binary) classification. The catalog period differs from that in Gaia (0.415448 day), likely contributing to the misclassification as an RRC.

In-class Outliers. Figure 4 displays objects that were flagged as outliers despite having correct labels. These stars were marked as outliers due to distinctive features:

1. EDR3 3017256242460492800, An EA-type star (Figure 4(a)): identified as V1174 Ori, is a special X-ray bright pre-main-sequence system in the Orion star-forming cluster (K. G. Stassun et al. 2022).
2. EDR3 3406832075676358912, correctly classified as EB (Figure 4(b)): shows unusual out-of-eclipse modulations, possibly from rotation.
3. EDR3 3372147259826502272 (V* DU Gem), a semi-detached binary with emission lines (Figure 4(c)).
4. EDR3 45787237593398144, both a misclassification and in-class outlier (Figure 4(d)): likely an EB rather than EA, with a light curve suggesting rotation or pulsation effects.

Two ROT Clusters. Interestingly, the spotted variables with rotational modulation (ROT) class appears to be divided into



(a) UMAP - Train (b) UMAP - Test

Figure 2. UMAP visualizations of multimodal embeddings: (a) training set and (b) test set, showing class separability and alignment between sets. Each source in the training and test set are colored by the class determined in (T. Jayasinghe et al. 2019) but these class labels are not used in the construction of the embeddings.

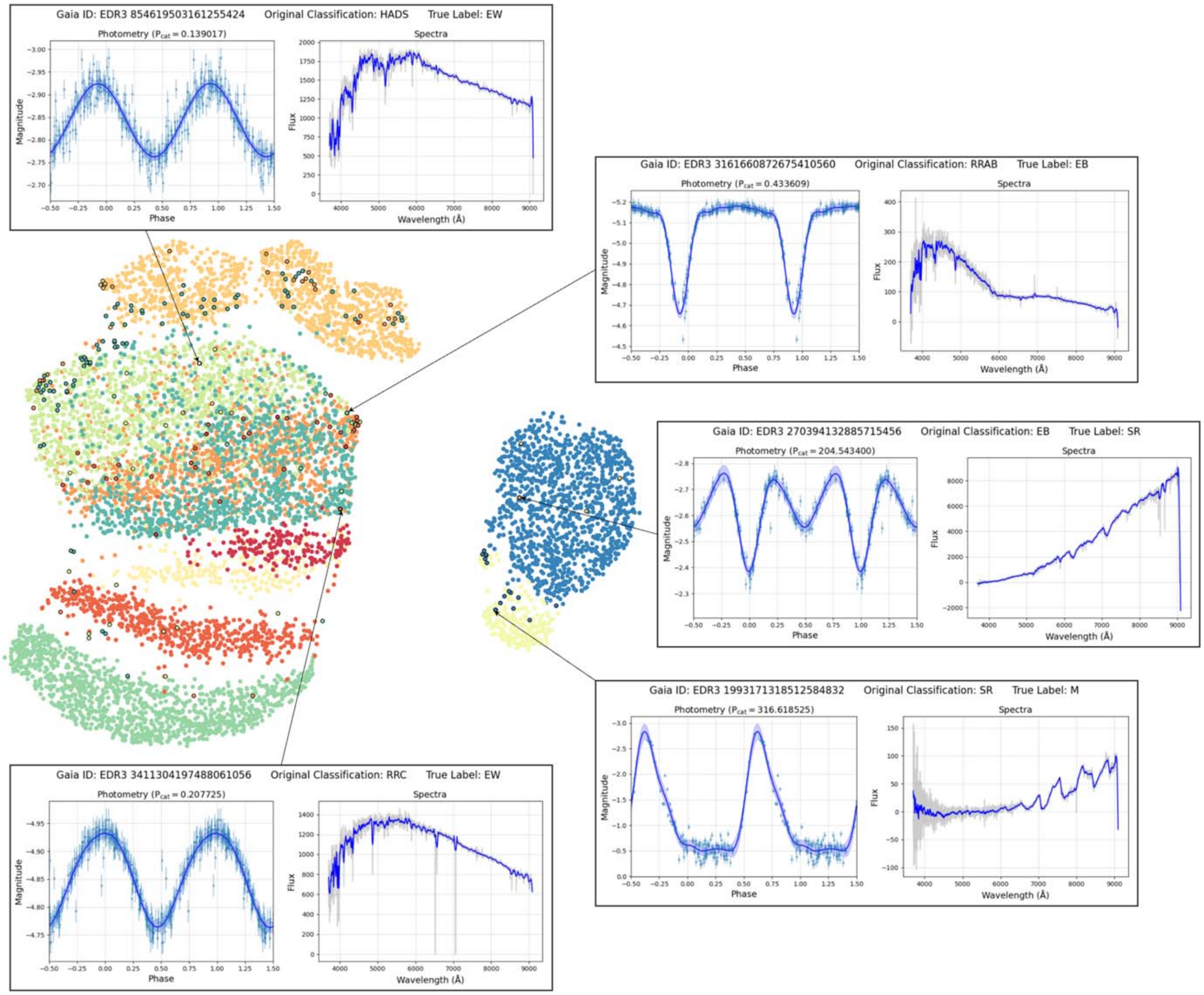


Figure 3. Examples of catalog misclassifications with photometry and spectrum for each object. Top to bottom: (1) likely EW misclassified as HADS; (2) V* AC CMi, a known semidetached binary misclassified as RR Lyrae; (3) possible SR or Mira variable with period alignment issues; (4) known Mira variable (V0439 Cas) misclassified as SR; (5) likely EW binary (N. Mowlavi et al. 2023) misclassified as RRC.

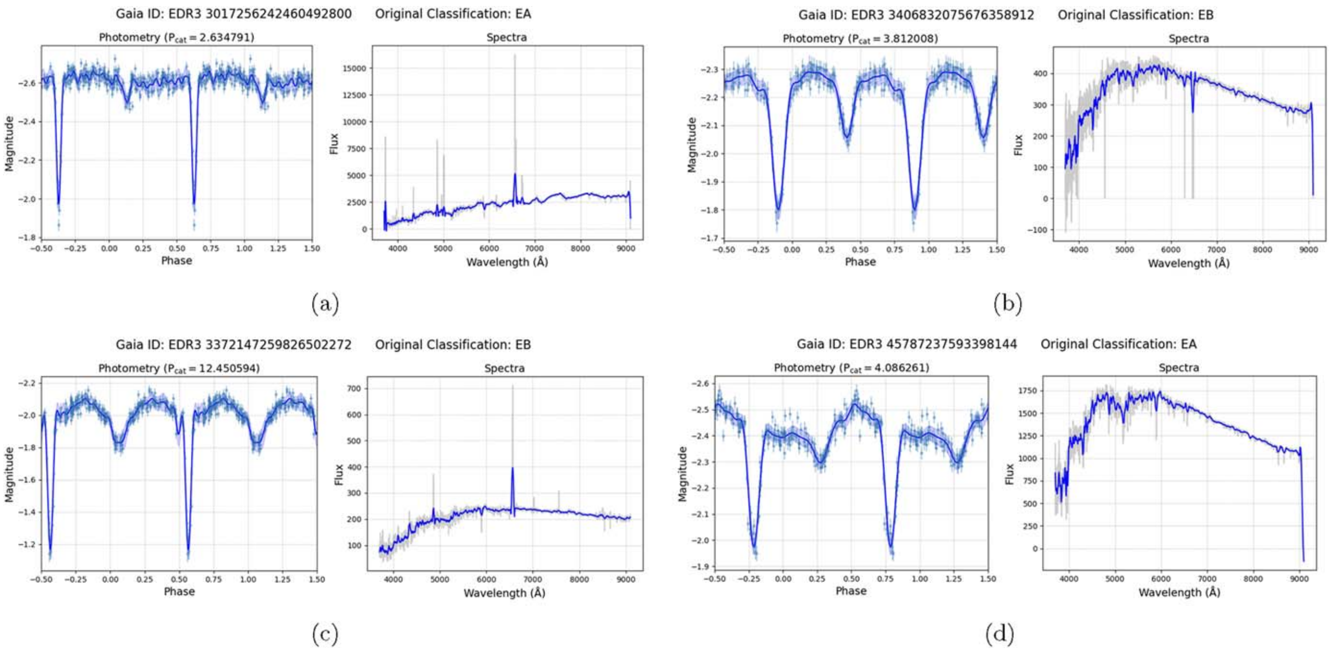


Figure 4. Examples of in-class outliers flagged by the model due to distinctive features, despite correct labels. (a) EA-type star, V1174 Ori, an X-ray bright pre-main-sequence system (K. G. Stassun et al. 2022). (b) EB-type star with unusual out-of-eclipse modulations, possibly due to rotation. (c) Semidetached binary with emission lines. (d) Likely an EB misclassified as EA, with light curve patterns indicating rotation or pulsation.

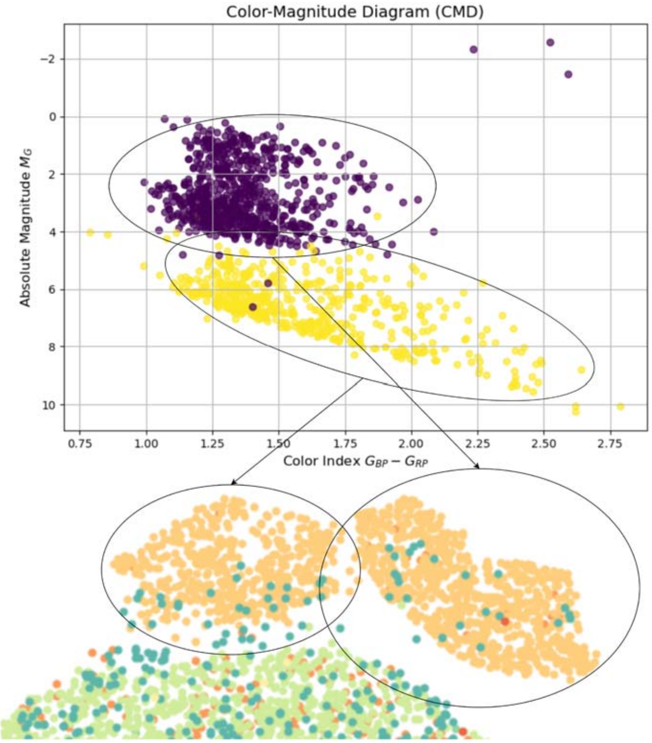


Figure 5. Color–magnitude diagram for ROT variables, with two clusters identified through unsupervised learning as giants and dwarfs.

two adjacent clusters, suggesting two physically distinct subtypes. To investigate further, we plotted these objects on a color–magnitude diagram (Figure 5). The plot revealed that the model had distinguished two subtypes within the ROT class: giants and dwarfs. Notably, the model discovered this distinction in an unsupervised learning process, without explicit labels for these subtypes.

Two Mira Clusters. As in Figure 6, the Miras were also split into two clusters—one larger and one significantly smaller. Upon closer inspection, we find that these clusters correspond to two distinct subtypes of Miras: M-type and C-type. This distinction was not explicitly available beforehand, as our dataset only included the general "Mira" label. This demonstrates the ability of the approach taken herein to uncover hidden patterns in astronomical data and its potential for enabling new scientific discoveries.

New Classes. During dataset creation, we filtered out classes with insufficient sample sizes. Now, with the learned embedding, we use these objects to test the ability of the model to project unseen classes. Figure 7 shows they are located as expected: (a) double mode RR Lyrae variables are located inside the cluster of RR Lyrae variables Type ab (RRAB); (b) uncertain ROTs within the certain ROT cluster; (c) yellow semiregular variables and long secondary period in the SR cluster; (d) first overtone cepheids and some fundamental-mode classical cepheids near δ Scuti variables (DSCT). Interestingly, most uncertain classifications fall within the Mira cluster.

5.4. Modalities Importance

We evaluate the AstroM^3 multimodal classification model to assess the contribution of each modality by testing classification accuracy under different input conditions: when using only one modality, when one modality is missing, and when all three modalities are available. This flexibility is achieved by averaging embeddings before the fully connected layer, rather than concatenating them, and leveraging the shared embedding space.

Table 5 highlights two key takeaways:

1. The model is flexible and can operate with different amounts of input data, handling cases where one or more modalities are missing.

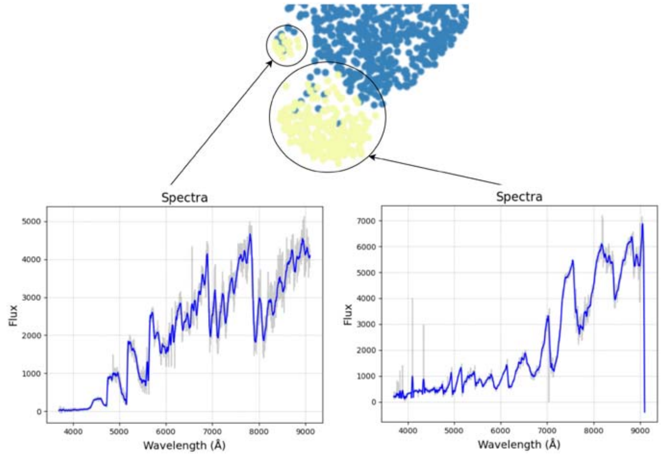


Figure 6. Spectral examples of Mira variables, showing two distinct clusters corresponding to M-type and C-type Miras, "rediscovered" through unsupervised learning.

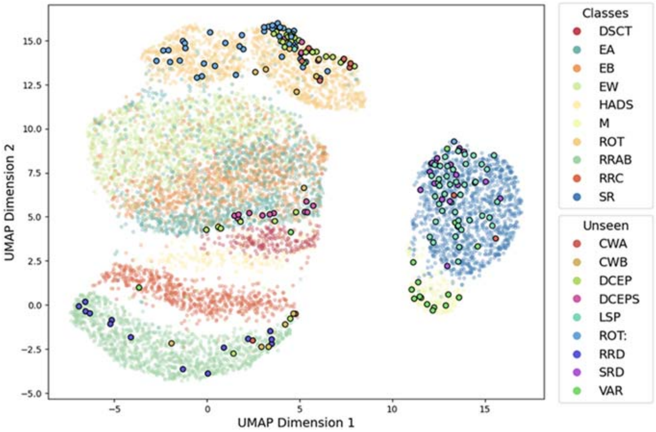


Figure 7. Projections of new, previously unused classes in the embedding space, aligning with related clusters and demonstrating the model's ability to position unseen classes accurately.

2. More modalities lead to better accuracy, with the best performance achieved when all three modalities are used together.

The results further show that different modalities contribute uniquely to classification performance. For instance, photometry plays a more important role for DSCT, EA, and EB, whereas metadata is particularly informative for EW. Across all classes, integrating two modalities improves performance over a single modality, and using all three yields the highest overall accuracy.

5.5. Similarity Search

An additional strength of our approach is the ability to perform similarity or dissimilarity searches within the embedding space. Users can leverage the pretrained AstroM^3 model to extract embeddings from photometry, spectra, and metadata without requiring additional fine-tuning. These embeddings can then be used for various tasks, including modality-specific similarity search, cross-modality retrieval, and outliers detection. This expands the utility of the CLIP-based model beyond classification to serve as a versatile tool for exploratory data analysis, anomaly detection, and multimodal inference. This capability holds promise for aiding the discovery of rare or unexpected phenomena in astronomical data.

Modality-Specific Similarity Search. Our model allows us to find similar objects based on a chosen modality. For example, if we want to find objects with spectral features similar to those in Figure 4(a), we can embed the spectrum of that object and compute the cosine similarity with other objects in the dataset (where a cosine similarity of 1 indicates maximum similarity). Figure 8 shows the two most similar objects based solely on spectral similarity, with cosine similarities of 0.8784 and 0.8451, respectively. As shown, they share clear visual similarities.

Cross-Modality Contrast Search. Our approach also allows for searches to find objects that are similar in one modality but differ in another. For instance, we can first identify the 10 objects most similar to Figure 4(d) based on photometric cosine similarity. Among these, we then select the object with the greatest spectral difference. This process results in the object shown in Figure 9, which has a photometric cosine similarity of 0.7749 but a much lower spectral similarity of 0.1430. Notably, this object is also a misclassification with an incorrect period; the correct classification should be an RS Canum Venaticorum variable, with the actual period being half the reported value (16.3401046 days).

Cross-Modality Similarity Search. When only photometric data is available, we can identify the closest matching spectra by calculating the cosine similarity between the photometric embedding and all the spectra embeddings in the dataset. This approach is possible because the model is trained to align photometry, spectra, and metadata in the same shared embedding space. For instance, using the photometry of the object shown in Figure 4(d), we find that the closest spectra in the dataset, as shown in Figure 9, has a cosine similarity of 0.4872. Although there is no guarantee that the predicted spectra will perfectly match the actual spectra—especially given the relatively moderate cosine similarity—this method allows us to form hypotheses about an object's composition without requiring direct spectroscopic data.

Outlier Detection. Beyond UMAP-based analysis, we can identify outliers using all 512 features of the embedding space. This allows us to detect (1) misclassifications, (2) in-class outliers, and (3) complete outliers that do not belong to any known class. To identify (1) and (2), we can calculate class centroids by averaging all embeddings for each class. We then build a cosine distance distribution for each class and set a threshold, such as the 99th percentile. Any object with a cosine distance from its class centroid exceeding this threshold can be labeled as an outlier. This process can be performed separately for each modality, and the results can be further refined by marking only those objects that are identified as outliers in more than one modality. For (3), we can apply DBSCAN clustering on the entire set of embeddings without using explicit labels, marking any object that falls outside the main clusters as a complete outlier.

5.6. Comparison with Prior Work

MAVEN (G. Zhang et al. 2024) and AstroCLIP (L. Parker et al. 2024) apply contrastive learning but use only two modalities: MAVEN incorporates spectra and photometry, while AstroCLIP combines spectra with images. In contrast, AstroM^3 integrates three modalities: photometry, spectra, and

Table 5
Class-wise Classification Accuracy for Different Modality Combinations in the AstroM^3 Model

Class	P	S	M	P + S	S + M	M + P	P + S + M
DSCT	91.67	70.83	75.00	95.83 ↑	75.00	95.83 ↑	91.67
EA	78.75	35.62	64.38	80.00 ↑	65.62 ↑	86.25 ↑	85.62 ↑
EB	80.00	58.75	47.50	76.25 ↓	62.50 ↑	76.88 ↓	76.88 ↓
EW	70.00	66.88	78.75	88.75 ↑	76.88 ↓	88.75 ↑	90.00 ↑
HADS	100.00	23.08	76.92	100.00	76.92	92.31 ↓	96.15 ↓
M	95.45	54.55	90.91	95.45	90.91	95.45	95.45
ROT	91.25	84.38	92.50	97.50 ↑	96.25 ↑	99.38 ↑	100.00 ↑
RRAB	96.88	64.38	94.38	97.50 ↑	95.00 ↑	100.00 ↑	100.00 ↑
RRC	86.08	63.29	86.08	92.41 ↑	86.08	94.94 ↑	94.94 ↑
SR	98.12	94.38	96.88	98.75 ↑	98.12 ↑	100.00 ↑	100.00 ↑
Average	88.82	61.61	80.33	92.24 ↑	82.33 ↑	92.98 ↑	93.07 ↑

Note. P, S, and M refer to Photometry, Spectra, and Metadata, respectively. Up arrows (↑) indicate accuracy improvement compared to the best-performing single modality for that class, while down arrows (↓) indicate a decrease. Results highlight that different classes benefit from different modalities, and overall accuracy improves when more modalities are used, with the highest performance achieved when all three modalities are combined.

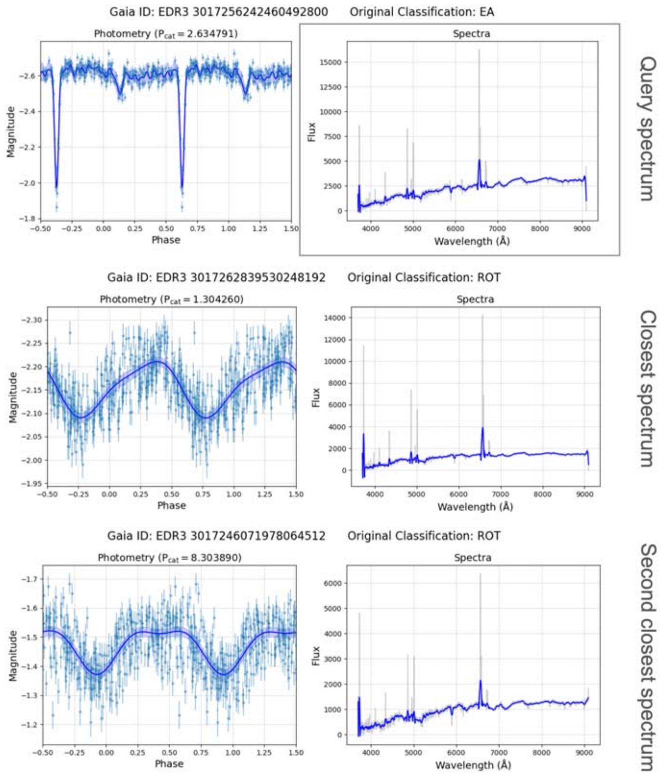


Figure 8. The query spectrum (top) and its two closest matches (middle and bottom) based on spectral cosine similarity.

metadata. Due to these differences, a direct comparison is not meaningful.

AstroM^3 is a flexible framework rather than a fixed model, allowing researchers to swap its photometry, spectra, and metadata encoders with different architectures or even replace them with entirely new modalities such as images or text. Since its primary goal is to align multimodal embeddings rather than optimize a specific modality, comparing it to single-modality models is not meaningful. The strength of AstroM^3 lies in its ability to improve classification accuracy through modality alignment, regardless of the specific encoders or data types used.

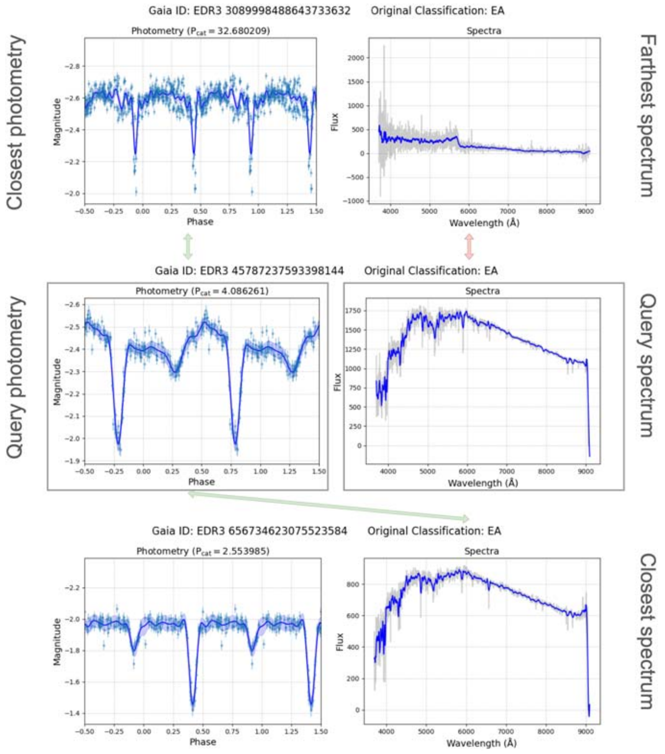


Figure 9. Examples of cross-modality contrast and similarity searches. The middle row represents the query object. We use its photometry and spectrum to find another object (top row) with the most similar photometry and the most dissimilar spectrum. For the bottom row, we use the query object's photometry to find the closest matching spectrum in the dataset, illustrating how a single modality can retrieve a similar object in a different modality. Green arrows indicate high similarity, while red arrows indicate low similarity.

AstroM^3 benefits most from pretraining on the full dataset and fine-tuning on smaller subsets, suggesting that other self-supervised techniques, such as masked modeling or restoration tasks, could also improve performance. Since no standard multimodal benchmarks exist in astronomy, we release a dataset and establish AstroM^3 as a baseline to support future comparisons and encourage further exploration of pretraining approaches.

6. Conclusion

We present the curation of a large labeled dataset suitable for building and testing next-generation multimodal self-supervised models. This includes 21,440 objects with time-series photometry, spectra, and metadata. We also introduce AstroM^3 self-supervised pretraining framework that leverages all three data modalities. By extending the Contrastive Language-Image Pretraining model to handle a trimodal setting, our approach effectively learns joint representations across diverse astronomical data types, enhances classification accuracy, and leverages unlabeled data to improve performance when labeled data is limited. Beyond classification, AstroM^3 demonstrates versatility in tasks such as misclassification detection and in-class outlier identification. Additionally, it shows promise for scientific discovery by "rediscovering" different Mira types and Rotational variables subclasses, and enables efficient searches by identifying similar objects, cross-modality contrasts, or cross-modality similarities-facilitating targeted exploration of specific sources. By releasing the dataset, code, and pretrained models, we provide a pathway for researchers to fine-tune AstroM^3 for different subfields and use cases. Whether used as a framework for developing new multimodal models or as a ready-to-use model for classification, similarity or outliers searches, AstroM^3 is designed to facilitate broader adoption of self-supervised learning in astronomy.

Model limitations and use cases: while AstroM^3 achieves strong performance on its training dataset, its generalization to other surveys and observational conditions remains untested. The model was trained on a curated dataset from one time-domain and one spectroscopic survey, and its applicability to additional classes and instruments requires further validation. Nevertheless, we release the pretrained model `AstroM3-CLIP`, which provides photometry, spectra, metadata, and multimodal embeddings. These embeddings can be used for fine-tuning on new surveys, particularly when labeled data is scarce, as well as for clustering, visualization, and exploratory data analysis using techniques like UMAP. The learned embedding space also enables similarity search, identifying astrophysical objects with shared characteristics even without labeled data, and anomaly detection, flagging outliers that may indicate misclassified sources or novel astrophysical phenomena. Fine-tuned versions of AstroM^3 support variable star classification across 10 classes using photometry `AstroM3-CLIP-photo`, spectra `AstroM3-CLIP-spectra`, metadata `AstroM3-CLIP-meta`, and multimodal input `AstroM3-CLIP-all`. These models are available on the Hugging Face Model Hub and are easy to use, with code examples provided for extracting embeddings and classification (see Appendix C).

Future work: to be clear, while our approach outperforms classification tasks on the dataset we have curated, we are not claiming that AstroM^3 has been shown to achieve state-of-the-art on classification of time-variable sources in general. The application of AstroM^3 to existing photometric benchmark datasets from other surveys is a clear next step. There are several other directions for extending our framework beyond AstroM^3. Given the abundance of photometry and metadata compared to spectra, one key area is to develop an algorithm capable of handling missing modalities *during training*, allowing us to leverage all available photometry and metadata. Additional directions include expanding the framework to integrate even more modalities, such as photometry from other bands and human comments on sources; learning to manage varying and missing metadata; and incorporating new classes, including nonperiodic ones. Building a larger, more diverse dataset and applying the models to tasks like prediction and anomaly detection are essential next steps toward creating a truly foundational multimodal model for astronomy.

Acknowledgments

M. R. and J. S. B. were partially supported by the Gordon and Betty Moore Foundation and an AAG grant (#2206744) from the National Science Foundation. J. S. B. thanks the Miller Institute for Basic Research in Science for supporting his research. The spectra used in this paper were from LAMOST (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope) on the Guoshoujing Telescope, a Major National Scientific Project built by the Chinese Academy of Sciences. Funding for the LAMOST project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences. Photometry used in this paper were from ASAS-SN, funded in part by the Gordon and Betty Moore Foundation and also funded in part by the Alfred P. Sloan Foundation grant. This work also made use of data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC; https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

Appendix A
Metadata Description

AstroM^3 metadata consists of 34 parameters extracted from Gaia, WISE, 2MASS, and ASAS-SN catalogs. These features include absolute magnitudes, proper motions, parallaxes, and multi-band photometry. A complete list of metadata features and descriptions is provided in Table 6.

Table 6
Descriptions of Metadata Features Used in the Dataset

Feature	Description
mean_vmag	Mean magnitude in the visible band
phot_g_mean_mag	Gaia G-band mean magnitude
e_phot_g_mean_mag	Uncertainty in Gaia G-band mean magnitude
phot_bp_mean_mag	Gaia BP band mean magnitude
e_phot_bp_mean_mag	Uncertainty in Gaia BP band mean magnitude
phot_rp_mean_mag	Gaia RP band mean magnitude
e_phot_rp_mean_mag	Uncertainty in Gaia RP band mean magnitude
bp_rp	BP mean magnitude minus RP mean magnitude
parallax	Gaia DR3 Parallax measurement
parallax_error	Uncertainty in parallax measurement
parallax_over_error	Signal-to-noise ratio for parallax measurement
pmra	Proper motion in the R.A. direction
pmra_error	Uncertainty in pmra
pmdec	Proper motion in the decl. direction
pmdec_error	Uncertainty in pmdec
j_mag	2MASS J-band magnitude
e_j_mag	Uncertainty in 2MASS J-band magnitude
h_mag	2MASS H-band magnitude
e_h_mag	Uncertainty in 2MASS H-band magnitude
k_mag	2MASS K-band magnitude
e_k_mag	Uncertainty in 2MASS K-band magnitude
w1_mag	WISE W1 band magnitude
e_w1_mag	Uncertainty in WISE W1 band magnitude
w2_mag	WISE W2 band magnitude
e_w2_mag	Uncertainty in WISE W2 band magnitude
w3_mag	WISE W3 band magnitude
w4_mag	WISE W4 band magnitude
j_k	J-band minus K-band magnitude
w1_w2	W1 band minus W2 band magnitude
w3_w4	W3 band minus W4 band magnitude
pm	Total proper motion
ruwe	Renormalized unit weight error
l	Galactic longitude
b	Galactic latitude

Appendix B
Training Setup and Hyperparameters

In this work, we used Optuna (T. Akiba et al. 2019) to perform hyperparameter optimization for our models. Our goal was to minimize the validation loss across multiple architectures and pretraining strategies. We tuned CLIP itself, as well as models for photometry, spectra, metadata, and multimodal data, with two initialization options: random initialization or pretrained CLIP weights. The search used Bayesian optimization, stopping when validation loss plateaued for six epochs, and the final parameters were selected based on the lowest validation loss across multiple trials. The listed parameter ranges were used for both random initialization and fine-tuning from pretrained CLIP weights, ensuring a consistent search space across all training configurations. The tuned hyperparameters for all models are stored as configuration files in the GitHub repository,[12] ensuring reproducibility of training runs.

For each model type, the hyperparameters we explored included:

1. Learning rate (`lr`): sampled from a logarithmic scale between 1×10^{-5} and 1×10^{-2}
2. Dropout rates for photometry (`p_dropout`), spectra (`s_dropout`), and metadata (`m_dropout`), all sampled from a uniform distribution between 0.0 and 0.4.
3. Adam optimizer parameters:
   - Beta1 (`beta1`): sampled from a uniform distribution between 0.7 and 0.99.
   - Weight decay (`weight_decay`): sampled from a logarithmic scale between 1×10^{-5} and 1×10^{-1}.
4. Learning rate scheduler factor (`factor`): sampled from a uniform distribution between 0.1 and 1.0 for the `ReduceLROnPlateau` scheduler.

Training Setup. For each trial, additional techniques were applied to ensure model stability and improved convergence.

1. Gradient clipping was applied to stabilize training. For CLIP, a clipping value of 45 was used, while for the photometry and spectra models, the clipping value was set to 5.
2. Training duration: The models were trained for a fixed number of epochs: 100 epochs for CLIP and 50 epoch for others
3. A warmup scheduler was employed to gradually increase the learning rate from a very low value to the target learning rate over the first 10 epochs.

Appendix C
Code Examples

C.1. Loading the Dataset

AstroM^3 is a time-series astronomy dataset containing photometry, spectra, and metadata for variable stars. AstroM^3 is available in two formats: `AstroMLCore/AstroM3Dataset`, which has the original files, and `AstroMLCore/AstroM3Processed`, a preprocessed version ready for training. Both versions allow loading subsets and different random seeds. Each dataset format provides predefined training, validation, and test splits across different subset sizes:

1. `full`: entire dataset.
2. `sub50`: 50% subset.
3. `sub25`: 25% subset.
4. `sub10`: 10% subset.

For reproducibility, subsets are available with different random seeds (`42`, `66`, `0`, `12`, `123`).

Loading examples: to load the dataset using the Hugging Face `datasets` library, specify the name in the format "{subset}_{seed}". For example:

```
from datasets import load_dataset
# Load the original dataset (default: full dataset, seed 42)
dataset = load_dataset("AstroMLCore/AstroM3Dataset", trust_remote_code=True)
# Load the original dataset, 25% subset with seed 123
dataset = load_data set("AstroMLCore/AstroM3Dataset", name="sub25_123", trust_remote_code=True)
# Load the preprocessed dataset, full set with seed 42
dataset = load_data set("AstroMLCore/AstroM3Processed", name="full_42")
```

[12] https://github.com/MeriDK/AstroM3/tree/main/configs

C.2. Extracting Embeddings

The pretrained model can generate embeddings from photometry, spectra, and metadata. If any modality is missing, the remaining embeddings can be averaged accordingly:

```
  from AstroM3.src.model import AstroM3
  # Create model and load the weights model =
AstroM3.from_pretrained("AstroMLCore/
AstroM3-CLIP")
  # Get embeddings
  p_emb, s_emb, m_emb = model.get_embeddings
(photometry, photometry_mask, spectra,
metadata)
  # If all three modalities are present
  emb = (p_emb + s_emb + m_emb) / 3
  # If spectra is missing
  emb = (p_emb + m_emb) / 2
```

C.3. Classification

Fine-tuned models can be used for classification across different modalities:

```
  from AstroM3.src.model import AstroM3,
Informer, GalSpecNet, MetaModel
  # Photometry classification
  photo_model = Informer.from_pretrained
("AstroMLCore/AstroM3-CLIP-photo")
  logits = photo_model(photometry, photometry_
mask)
  # Spectra classification
  spectra_model = GalSpecNet.from_pretrained
("AstroMLCore/AstroM3-CLIP-spectra")
  logits = spectra_model(spectra)
  # Metadata classification
  meta_model = MetaModel.from_pretrained
("AstroMLCore/AstroM3-CLIP-meta")
  logits = meta_model(metadata)
  # Multimodal classification
  all_model = AstroM3.from_pretrained("Astro-
MLCore/AstroM3-CLIP-all")
  logits = all_model(photometry, photome-
try_mask, spectra, metadata)
```

ORCID iDs

M. Rizhko https://orcid.org/0000-0003-3885-4661
J. S. Bloom https://orcid.org/0000-0002-7777-216X

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. 2019, in Proc. the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining (New York: ACM), 2623
Alabdulmohsin, I., Wang, X., Steiner, A., et al. 2024, arXiv:2403.04547
Becker, I., Pichara, K., Catelan, M., et al. 2020, MNRAS, 493, 2981
Boone, K. 2021, AJ, 162, 275
Cherti, M., Beaumont, R., Wightman, R., et al. 2023, in Proc. the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (New York: IEEE), 2818
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RAA, 12, 1197
Debosscher, J., Sarro, L. M., Aerts, C., et al. 2007, AAP, 475, 1159
Donoso-Oliva, C., Becker, I., Protopapas, P., et al. 2023, AAP, 670, A54
Dubath, P., Rimoldini, L., Süveges, M., et al. 2011, MNRAS, 414, 2602
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in Proc. of the Second Int. Conf. on Knowledge Discovery and Data Mining KDD (Washington, DC: AAAI Press), 226
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021, A&A, 649, A1
Guzhov, A., Raue, F., Hees, J., & Dengel, A. 2022, in ICASSP 2022-2022 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (New York: IEEE), 976
Hayat, M. A., Stein, G., Harrington, P., Lukić, Z., & Mustafa, M. 2021, ApJL, 911, L33
Jamal, S., & Bloom, J. S. 2020, ApJS, 250, 30
Jayasinghe, T., Stanek, K. Z., Kochanek, C. S., et al. 2019, MNRAS, 486, 1907
Kim, D.-W., Yeo, D., Bailer-Jones, C. A. L., & Lee, G. 2021, AAP, 653, A22
Lafler, J., & Kinman, T. D. 1965, ApJS, 11, 216
Leung, H. W., & Bovy, J. 2024, MNRAS, 527, 1494
Li, J., Li, D., Xiong, C., & Hoi, S. 2022, ICML, 1, 12888
Li, Y., Liang, F., Zhao, L., et al. 2021, arXiv:2110.05208
Long, J. P., El Karoui, N., Rice, J. A., Richards, J. W., & Bloom, J. S. 2012, PASP, 124, 280
Luo, H., Ji, L., Zhong, M., et al. 2021, arXiv:2104.08860
Ma, Y., Xu, G., Sun, X., et al. 2022, in Proc. the 30th ACM Int. Conf. on Multimedia (New York: ACM), 638
McInnes, L., Healy, J., & Melville, J. 2018, arXiv:1802.03426
Morrissey, P., Conrow, T., Barlow, T. A., et al. 2007, ApJS, 173, 682
Mowlavi, N., Holl, B., Lecoeur-Taïbi, I., et al. 2023, A&A, 674, A16
Mu, N., Kirillov, A., Wagner, D., & Xie, S. 2022, in European Conf. on Computer Vision (Berlin: Springer), 529
Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., & Hložek, R. 2019, PASP, 131, 118002
Naul, B., Bloom, J. S., Pérez, F., & van der Walt, S. 2018, NatAs, 2, 151
Palaversa, L., Ivezić, Ž., Eyer, L., et al. 2013, AJ, 146, 101
Parker, L., Lanusse, F., Golkar, S., et al. 2024, MNRAS, 531, 4990
Radford, A., Kim, J. W., Hallacy, C., et al. 2021, ICML, 1, 8748
Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, ApJ, 733, 10
Richards, J. W., Starr, D. L., Miller, A. A., et al. 2012, ApJS, 203, 32
Shappee, B. J., Prieto, J. L., Grupe, D., et al. 2014, ApJ, 788, 48
Singh, A., Hu, R., Goswami, V., et al. 2022, in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (New York: IEEE), 15617
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Stassun, K. G., Torres, G., Kounkel, M., et al. 2022, ApJ, 941, 125
Sun, Q., Fang, Y., Wu, L., Wang, X., & Cao, Y. 2023, arXiv:2303.15389
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Wu, Y., Chen, K., Zhang, T., et al. 2023, in ICASSP 2023-2023 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (New York: IEEE), 1
Wu, Y., Tao, Y., Fan, D., Cui, C., & Zhang, Y. 2024, MNRAS, 527, 1163
Xu, H., Ghosh, G., Huang, P.-Y., et al. 2021, arXiv:2109.14084
Xu, H., Xie, S., Tan, X. E., et al. 2023, arXiv:2309.16671
Yao, L., Huang, R., Hou, L., et al. 2021, arXiv:2111.07783
Zhang, G., Helfer, T., Gagliano, A. T., Mishra-Sharma, S., & Villar, V. A. 2024, MLS&T, 5, 045069
Zhang, R., Guo, Z., Zhang, W., et al. 2022, in Proc. the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (New York: IEEE), 8552
Zhou, H., Zhang, S., Peng, J., et al. 2021, in Proc. of the AAAI Conf. on Artificial Intelligence (Washington, DC: AAAI Press), 11106