AUGMENTING X-RAY ASTRONOMICAL REPRESENTATIONS WITH SCIENTIFIC KNOWLEDGE THROUGH CONTRASTIVE LEARNING

Rafael Martínez-Galarza[1,2,*] **Nicolò Oreste Pinciroli Vago**[3] **Shivam Raval**[4]
Carolina Cuesta-Lázaro[5] **Melanie Weber**[6] **David Alvarez-Melis**[6]
Alberto Accomazzi[2] **Cecilia Garraffo**[1,2] **Joshua Knutson**[7] **Ryan Thill**[7]
Christopher B. Green[7] **Imantha Ahangama**[7]
[1]AstroAI, Cambridge, Massachusetts, USA
[2]Center for Astrophysics | Harvard & Smithsonian, Cambridge, Massachusetts, USA
[3]Politecnico di Milano, Milan, Italy
[4]Department of Physics, Harvard University, Cambridge, Massachusetts, USA
[5]NSF Institute for Artificial Intelligence and Fundamental Interactions, USA
[6]Harvard SEAS, Cambridge, Massachusetts, USA
[7]Astromind, Austin, Texas, USA
*jmartine@cfa.harvard.edu

ABSTRACT

Astronomers have produced large multimodal datasets that include images, spectra, and time series, and that encode physical information about the observed objects. In addition, a large amount of physics-specific knowledge about these objects has been accumulated in the astronomical literature. We introduce a physics-informed representation alignment framework that matches X-ray observations of astrophysical objects and text summaries describing the physical properties of those sources. We perform contrastive learning between data representations learned using a Poisson process autodecoder and text summary representations generated with a Large Language Model. We demonstrate the generalization capabilities of the system and evaluate the performance of the post-alignment shared representations for regression tasks. We present a use case for anomaly detection.

1 INTRODUCTION

In preparation for the era of petabyte-scale astronomical datasets (Greenstreet et al., 2024), enabled by the next generation of multimodal (images, spectra, astronomical time-series) surveys, astronomers have recently embraced the concept of *foundation models* in astronomy (Parker et al., 2024; Leung & Bovy, 2024). This refers to deep neural networks trained using self-supervised learning on certain conventional tasks (e.g., light curve reconstruction) and operating on very large (terabyte to petabyte) repositories of astronomical data (Angeloudi et al., 2024). The learned representations can readily be used to perform a number of different downstream tasks, such as regression on relevant astrophysical parameters, classification of the sources according to their underlying astrophysical class, and inference of physical parameters through comparison with simulations.

Astronomical datasets are particularly well suited for representation learning approaches for at least two reasons: first, the data volume is already large, and is expected to increase dramatically in the near future with survey facilities such as the Vera Rubin Observatory and the Roman Space Telescope (Hernandez et al., 2024; Gezari et al., 2022) coming online in the next 5 years; second, astronomical datasets are, by design, multimodal: that is, for a single astronomical object, several modalities of data exist (e.g. images, spectra and light curves) that share physical information about the object carried by the recorded photons, even if obtained with different instruments. Representations learned using pre-trained networks operating on the different modalities can be aligned using contrastive learning approaches.

Multimodal approaches to representation learning in astronomy have so far been limited to bi-modal experiments (e.g., galaxy image/spectra, text/image retrieval) (Parker et al., 2024; Mishra-Sharma et al., 2024). They are primarily based on the assumption that the recorded photons encode similar information across modalities. Apart from CLIP-like approaches to associate astronomical images with text summaries of the observational proposals, no experiments have been carried out to attempt an alignment between physically informative text summaries and representations learned from numerical data objects other than images in raster formats. It is reasonable to expect that there is shared information between the natural language descriptions of astrophysical properties from sources recorded in the astronomical literature and the data structures that encode the physical information carried by the photons.

In this paper, we demonstrate representation alignment between a set of latent embeddings learned from X-ray band data of individual astrophysical sources and text summaries that are descriptive of astrophysically relevant information about those sources. We generate and embed text summaries using respectively OpenAI's gpt-4o-mini and ada-002 models. We also generate astronomical data embeddings using an auto-decoding neural network (Song et al., 2024) that operates on lists of X-ray photon recordings and predicts the time-dependent Poisson rate of photon arrivals for the associated X-ray source. We then perform contrastive learning between the two representations using an InfoNCE loss. We investigate the following: 1) Can we design Large Language Model (LLM) prompting strategies that result in physics-encoding text summaries? 2) Does cross-modal representation alignment preserve the prediction power of the embeddings for regression? 3) Can we connect physically meaningful natural language to data structures that contain information about specific astrophysical environments? The code is publicly available at `https://anonymous.4open.science/r/contrastiveregression-CFEE/`.

2 Related Work

Self-supervised approaches that learn representations from astronomical data have been used for similarity search in optical spectra (Stein et al., 2021), galaxy distance estimation (Hayat et al., 2021), and anomaly detection (Walmsley & Scaife, 2023). In the AstroCLIP project (Parker et al., 2024), the authors perform contrastive learning between learned representations of galaxy images and spectra, and perform accurate zero-shot prediction of the galaxy redshift from the image alone using cross-modal nearest neighbor searches in the shared representation space. In Mishra-Sharma et al. (2024), the authors associate astronomical images obtained from the Hubble Space Telescope with text in the corresponding observing proposal abstracts by fine-tuning a pre-trained CLIP model and achieving image retrieval using natural language. In contrast, the present paper aligns numerical structures representing Poisson-like photon recordings that codify spatial, spectral, and time variability properties of the sources, to summaries extracted from full papers describing the physical properties of the corresponding sources.

For a review on contrastive learning in astronomy, see Huertas-Company et al. (2023). Regarding LLMs applied to astrophysical literature, in Dung Nguyen et al. (2023) the authors fine-tune the LLAMA-2 model (Touvron et al., 2023) using astronomical paper abstracts and demonstrate significant domain adaptation. More recently, Iyer et al. (2024) present a LLM-enabled framework for literature review and knowledge discovery in astronomy, focusing on semantic searching.

3 Dataset Construction

We use a set of observations taken by the Chandra X-ray Observatory targeting at a broad range of X-ray emitting astrophysical sources[1], and a corpus of text data curated by NASA's Astrophysical Data Systems (ADS), and consisting of the titles, abstracts, and full bodies of academic papers written by experts on topics of high energy astrophysics, and that specifically refer to at least one of the Chandra observations. The Chandra Data Archive has created an association between each Chandra observation (identified by an ObsID) and all the papers that refer to it, allowing us to create positive pairs between individual observations of astrophysical sources and corresponding text descriptions.

[1] `https://cxc.harvard.edu/cda/`

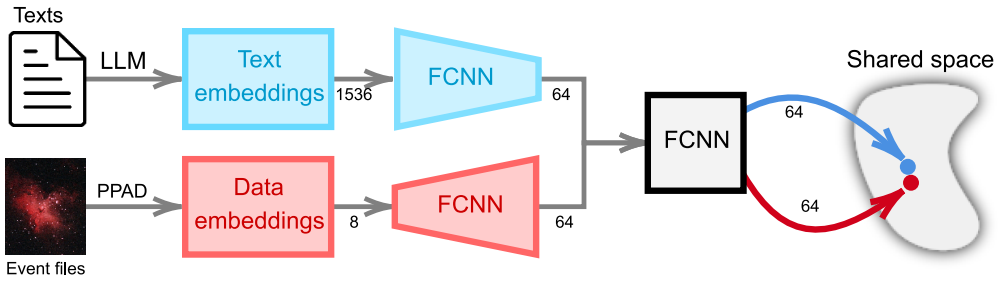

Figure 1: The contrastive learning network processes two types of embeddings: text embeddings of size 1536 and data embeddings of size 8. These embeddings are each transformed into a 64-dimensional space through a pair of fully connected neural networks, and then concatenated. A final fully connected network then integrates both modalities, learning a shared, aligned representation.

Each Chandra observation consists of a list of individual X-ray photon recordings (events) collected by the telescope detector over a small area of the sky. The field of the observation usually contains many detected astrophysical sources, and so for each X-ray source, we isolate its associated photon events. Each training example thus corresponds to event lists associated with an individual source, not the entire field, and can be understood as the outcome of a Poisson process of the detected photons from a single source, with time of arrival and photon energy recorded for each event.

For each source, we have created a data latent representation using the Poisson Process Auto Decoder (PPAD) presented in Song et al. (2024), a neural field decoder that maps fixed-length latent features to time-dependent, continuous Poisson rate functions for any range of photon energies. The PPAD is trained in a self-supervised fashion, starting from the photon event data, to predict the reconstructed Poisson rate function by minimizing a loss function with a continuous Poisson likelihood. Astronomers call this varying Poisson rate the light curve of the source, as it indicates the change of X-ray flux over time. The model also yields the latent embeddings for each source, which are optimized during training or inference. The resulting embeddings codify spectral and time-domain information and are useful for other downstream regression and classification tasks, as demonstrated in Song et al. (2025).

To create text summaries describing corresponding X-ray sources, we use the Chandra Source Catalog (Evans et al., 2024) to get the sky coordinates for all the X-ray sources in the field of each observation, above a certain signal to noise level. We then use the SIMBAD database[2] to find the list of all possible identifiers under which each source can appear in the literature. We use these identifiers to prompt an LLM (GPT 4o-mini) to search for information about the particular source in the associated papers using a physically motivated prompt. The text summaries were validated based on domain knowledge in X-ray astrophysics, by selecting a random set of 100 summaries and evaluating the accuracy of the descriptions. In subsection A.1, we show the prompt used. Finally, we embed the text summary using OpenAI's ada-002 embedding model. As a result of this process, for each astrophysical X-ray source, we have a text embedding of the text summary and data embeddings from the list of photon events.

4 Methodology

4.1 Contrastive Learning

Figure 1 presents the architecture of the network used for contrastive learning. The contrastive learning network processes two types of embeddings: text embeddings of size 1536 and data embeddings of size 8. We align the representations by first using a pair of fully connected neural networks to bring both the text embeddings and the PPAD embeddings to the same dimension size of 64. We then use a third fully connected network to generate a shared representation for both embeddings. The network is trained using InfoNCE loss to align the embedding pairs for each source. We regularize the loss to preserve the original distances in the pre-alignment embedding spaces (see Equation 1).

[2] https://simbad.u-strasbg.fr/simbad/

Table 1: Target variables for the regression tasks. *Variable* is the name used in this work and *Name* is the name used in the CSC. *Soft*, *medium* and *hard* refer to the average energy carried by the photons in the corresponding band.

Variable	Name	Description
Hard_{HS}	`hard_hs`	hard - soft energy band hardness ratio
Hard_{MS}	`hard_ms`	medium - soft energy band hardness ratio
Hard_{HM}	`hard_hm`	hard - medium energy band hardness ratio
p_{var}	`var_prob_b`	Gregory-Loredo variability probability
F_{sig}	`flux_significance_b`	flux significance

We train for 1000 epochs, using a batch size of 128, a temperature of 0.05, and an initial learning rate of 10^{-4}. After training, we evaluate alignment by performing regression on five physical variables, summarized in Table 1.

The choice of the embeddings sizes (8 for the PPAD embeddings and 64 for the fully connected networks) was based on our analysis of the natural trade-off between reconstruction quality and representation quality. For the PPAD, an embedding size of size 4 resulted in poor light curve reconstruction, whereas an embedding size of 16 resulted in a decrease in performance for downstream tasks. A grid search for the optimal size of the fully-connected neural network yielded 64 as the optimal value for the alignment task. We also note that the addition of the third concatenating fully connected neural network was necessary to achieve better generalization in the validation set. However, this limits our ability to perform cross-modal retrieval. We plan to release a version of the model that improves on this aspect.

4.2 Regression

We use linear regression, with the default `sklearn` hyperparameters for the task of predicting the summary statistics in Table 1. We explore three different training scenarios: 1) using the text embeddings only; 2) using the photon event latents only; and 3) using the aligned embeddings in the shared space. The regressor is trained separately for each scenario and for each target variable for 100 epochs and a learning rate of 10^{-3}.

4.3 Loss function

The loss function for alignment has two components: a contrastive loss and a regularizer. The contrastive loss is InfoNCE and is used to assess the alignment in the shared space. The regularizer is based on MSE. Overall, the loss is defined as $\mathcal{L} = \mathcal{L}_{InfoNCE} + \gamma\mathcal{L}_{reg}$, where $\mathcal{L}_{InfoNCE}$ is the InfoNCE loss, $\mathcal{L}_{reg}$ is a regularization loss and γ (here, 0.3, obtained applying a grid search in the hyperparameters space) is its weight. The regularization loss is defined as:

$$\mathcal{L}_{reg} = \sum_{s \in \{\text{text,data}\}} \text{MSE}(d_{s,latent}, d_{s,original}) \tag{1}$$

where $d_{s,original}$ is the distance between pairs of points in the original representation space for a modality s, $d_{s,latent}$ is the distance between the corresponding pairs in the final shared latent space and $\text{MSE}(x, y) = \mathbb{E}[(x-y)^2]$. The regularization term aims to preserve the initial pairwise distances between points in the shared latent space.

5 Results

5.1 Regression on astrophysical properties

Figure 2 presents the training (in blue) and validation (in orange) losses during contrastive training. Both decrease consistently over the epochs, indicating a successful optimization, and are nearly aligned, showing minimal overfitting and good generalization abilities. Convergence is observed

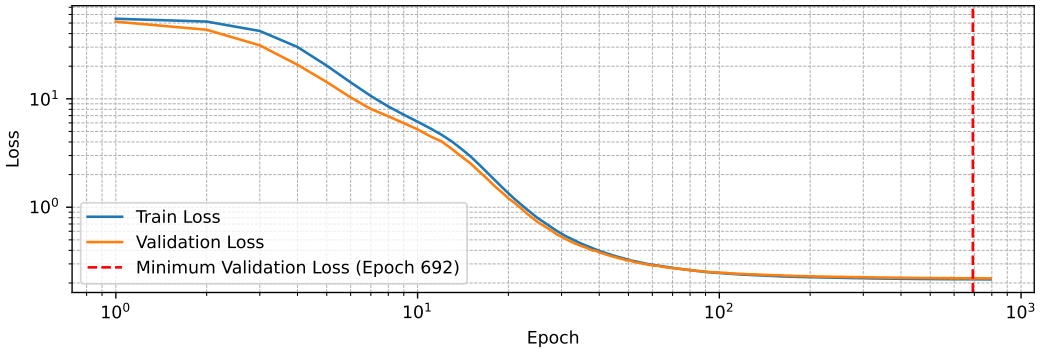

Figure 2: Train and validation losses as a function of epoch. The red dashed line indicates the epoch at which the validation loss is minimized.

already after ≈ 100 epochs, with minor subsequent improvements. The red dashed line in Figure 2 indicates the best epoch based on the validation loss. Overall, the low loss (≈ 0.2) indicates a high level of alignment between the two modalities in the shared 64-dimensional space.

Table 2 presents the results for regression on 5 variables when considering only the text modality, only numerical data and the shared latent space considering both modalities. Combining both modalities yields better results compared to using a single modality. The variable showing the most substantial relative improvement is Hard_{HS} ($\approx 24\%$ MAE with respect to data), suggesting that the information from both modalities is complementary for predicting the spectral shape of the source.

In general, the most significant improvements are observed for hardness ratios comprising the hard component (on average, $\approx 23\%$ improved MAE with respect to data alone). This result suggests that contextual information about the source (typically included in text in the form of a description of a spectral model fitted to the source) enriches the purely numerical data from the photon events. In the case of p_{var}, instead, the information provided in the text and the data are similar, and the results do not improve significantly using both modalities. Moreover, most of the F_{sig} information is contained in the data, suggesting that text summaries do not contain relevant spectral information when it comes to the significance of the X-ray detection. Results also demonstrate that relevant information from both modalities can be captured effectively in a small latent space, with only $\approx 4\%$ of the dimensions of the initial latent spaces combined.

5.2 Nearest neighbor search and anomaly detection

As an evaluation metric for alignment, we look at the top-k retrieval accuracy in the validation set, defined here as the fraction of true associated pairs of either modality that fall within the nearest k neighbors in the shared embedding space, using a 2-norm distance metric. After training, this the top-k retrieval accuracy is 93.3% for $k = 1$, 97.5% for $k = 5$, and 98.1% for $k = 10$.

To evaluate the semantic meaning of the aligned representations in the validation set beyond the top-k retrieval metric, we look at the 10 nearest cross-modal (text) neighbors for the X-ray data embedding of a relatively rare type of object (X-ray binary 2CXO J100157.9+553945), and compare the associated aligned text descriptions to the text descriptions of the pre-alignment nearest data embeddings to the test source. We find that the aligned text descriptions associated with the text source comprise a narrower semantic domain compared to the pre-alignment associations, with descriptions of X-ray binaries being included in 6 out of the 10 nearest neighbors, compared to only two mentions of X-ray binaries in the pre-alignment neighbors.

To ensure that the model used more than just the text embedding information to align the representations, we also look at the nearest cross-modal (data) neighbors for the text embedding of the summary corresponding to the same source, and compare the mean and standard deviation of their associated hardness ratios Hard_{HS} with those of the pre-alignment text neighbors. We find that the aligned neighbors have a significantly smaller standard deviation for this spectral property compared to the pre-alignment case (see subsection A.2). In cases where the alignment is not perfect, nearest

Table 2: MAE comparison for different modalities and 5 variables (Hard_{HS}, Hard_{MS}, Hard_{HM}, p_{var} and F_{sig}). The best results are indicated in bold. Absolute and percentage improvements are shown relative to the best single-modality result for each variable.

Variable	**Modality**			**Improvement**	
	Text	**Data**	**Both**	**Absolute**	**%**
Hard_{HS}	0.40	0.29	**0.22**	0.07	24%
Hard_{MS}	0.32	0.27	**0.25**	0.02	7%
Hard_{HM}	0.27	0.22	**0.17**	0.05	23%
p_{var}	0.23	0.22	**0.21**	0.01	5%
F_{sig}	5.15	2.43	**2.27**	0.16	7%

neighbors to an astrophysical source of a given class may correspond to descriptions of objects of a different class, but with similar spectral properties. Such is the case, for example, between Active Galactic Nuclei and X-ray binaries, both of which are accreting compact objects only different in their mass scales.

We also run the aligned embeddings from the validation set through the Unsupervised Random Forest (URF) anomaly detector (Baron & Poznanski, 2017). We rank the sources according to their anomaly score, and find that the system is effective at isolating truly unique sources, such as a highly variable and spectrally hard Ultra-Luminous X-ray source (ULX), among other sources with extreme properties. These are relatively rare objects that may represent a transition between stellar-mass black holes and intermediate-mass black holes (Bachetti et al., 2014; Feng et al., 2010). This highlights the potential for discovery of our framework. In subsection A.3 we list some of the anomalies.

Finally, we investigate which sources show the largest increase in anomaly score when the aligned multimodal representations are processed through the URF, as opposed to only the text embeddings or only the photon event embeddings. When ranked by their relative difference in anomaly score between multi-modal and text only, a very distinct type of object appears to be represented very often (three times more often than in the unimodal cases) at the top of the ranking: pulsars. Pulsars are rare astrophysical sources associated with highly magnetized neutron stars at high rotational speeds that often show a pattern of repeating X-ray flares.

6 CONCLUSIONS

We introduce a compact aligned representation, obtained through contrastive learning, to match learned embeddings of numerical data described by a Poisson process (X-ray photon detections from astrophysical sources), and text summaries describing the physical properties of those sources. We show that the data embeddings can be meaningfully enhanced with text embeddings from appropriately designed text summaries, resulting in an improved performance of regression downstream tasks. We also show that the learned representations can be generalized to previously unseen data, and that they are semantically meaningful in the sense that similarity in the aligned representation translates into similarity in both the physical properties of the sources as derived from the data, and the associated astrophysical concepts described in the text. These results are an encouraging first step in the design of a more general system that is able to perform cross-modal retrieval for Poisson-like datasets in other knowledge domains and allow the generation of data from text as well as the generation of physics-informed descriptions of unlabeled observations.

ACKNOWLEDGMENTS

The authors want to thank Ashley Villar, Yanke Song, Liam Parker, and Mike Smith for useful discussions and feedback. Funding for this research has been provided by Astromind.

REFERENCES

Eirini Angeloudi, Jeroen Audenaert, Micah Bowles, Benjamin M. Boyd, David Chemaly, Brian Cherinka, Ioana Ciucă, Miles Cranmer, Aaron Do, Matthew Grayling, Erin E. Hayes, Tom Hehir, Shirley Ho, Marc Huertas-Company, Kartheik G. Iyer, Maja Jablonska, Francois Lanusse, Henry W. Leung, Kaisey Mandel, Juan Rafael Martínez-Galarza, Peter Melchior, Lucas Meyer, Liam H. Parker, Helen Qu, Jeff Shen, Michael J. Smith, Mike Walmsley, John F. Wu, and The Multimodal Universe Collaboration. The Multimodal Universe: 100 TB of Machine Learning Ready Astronomical Data. *Research Notes of the AAS*, 8(12):301, December 2024. ISSN 2515-5172. doi: 10.3847/2515-5172/ad9a63. Publisher: The American Astronomical Society.

M. Bachetti, F. A. Harrison, D. J. Walton, B. W. Grefenstette, D. Chakrabarty, F. Fürst, D. Barret, A. Beloborodov, S. E. Boggs, F. E. Christensen, W. W. Craig, A. C. Fabian, C. J. Hailey, A. Hornschemeier, V. Kaspi, S. R. Kulkarni, T. Maccarone, J. M. Miller, V. Rana, D. Stern, S. P. Tendulkar, J. Tomsick, N. A. Webb, and W. W. Zhang. An ultraluminous X-ray source powered by an accreting neutron star. *Nature*, 514:202–204, October 2014. ISSN 0028-0836. doi: 10.1038/nature13791. URL https://ui.adsabs.harvard.edu/abs/2014Natur.514..202B. ADS Bibcode: 2014Natur.514..202B.

Dalya Baron and Dovi Poznanski. The weirdest SDSS galaxies: results from an outlier detection algorithm. *Monthly Notices of the Royal Astronomical Society*, 465:4530–4555, March 2017. ISSN 0035-8711. doi: 10.1093/mnras/stw3021. URL https://ui.adsabs.harvard.edu/abs/2017MNRAS.465.4530B. Publisher: OUP ADS Bibcode: 2017MNRAS.465.4530B.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, Josh Peek, Kartheik Iyer, Tomasz Różański, Pranav Khetarpal, Sharaf Zaman, David Brodrick, Sergio J. Rodríguez Méndez, Thang Bui, Alyssa Goodman, Alberto Accomazzi, Jill Naiman, Jesse Cranney, Kevin Schawinski, and UniverseTBD. AstroLLaMA: Towards specialized foundation models in astronomy. *arXiv e-prints*, pp. arXiv:2309.06126, September 2023. doi: 10.48550/arXiv.2309.06126. arXiv: 2309.06126 [astro-ph.IM] Number: arXiv:2309.06126 tex.adsnote: Provided by the SAO/NASA Astrophysics Data System.

Ian N. Evans, Janet D. Evans, J. Rafael Martínez-Galarza, Joseph B. Miller, Francis A. Primini, Mojegan Azadi, Douglas J. Burke, Francesca M. Civano, Raffaele D'Abrusco, Giuseppina Fabbiano, Dale E. Graessle, John D. Grier, John C. Houck, Jennifer Lauer, Michael L. McCollough, Michael A. Nowak, David A. Plummer, Arnold H. Rots, Aneta Siemiginowska, and Michael S. Tibbetts. The chandra source catalog release 2 series. 274(2):22, October 2024. doi: 10.3847/1538-4365/ad6319. arXiv: 2407.10799 [astro-ph.HE] Number: 22 tex.adsnote: Provided by the SAO/NASA Astrophysics Data System.

Hua Feng, Fengyun Rao, and Philip Kaaret. Discovery of Millihertz X-Ray Oscillations in a Transient Ultraluminous X-Ray Source in M82. *The Astrophysical Journal*, 710:L137–L141, February 2010. ISSN 0004-637X. doi: 10.1088/2041-8205/710/2/L137. URL https://ui.adsabs.harvard.edu/abs/2010ApJ...710L.137F. Publisher: IOP ADS Bibcode: 2010ApJ...710L.137F.

Suvi Gezari, Misty Bentz, Kishalay De, K. Decker French, Aaron Meisner, Michelle Ntampaka, Robert Jedicke, Ekta Patel, Daniel Perley, Robyn Sanderson, Christian Aganze, Igor Andreoni, Eric F. Bell, Edo Berger, Ian Dell'Antonio, Ryan Foley, Henry Hsieh, Mansi Kasliwal, Joel Kastner, Charles D. Kilpatrick, J. Davy Kirkpatrick, Casey Lam, Karen Meech, Dante Minniti, Ethan O. Nadler, Daisuke Nagai, Justin Pierel, Irene Shivaei, Rachel Street, Erik J. Tollerud, and Benjamin Williams. R2-D2: Roman and rubin – from data to discovery. *arXiv e-prints*, pp. arXiv:2202.12311, February 2022. doi: 10.48550/arXiv.2202.12311. arXiv: 2202.12311 [astro-ph.IM] Number: arXiv:2202.12311 tex.adsnote: Provided by the SAO/NASA Astrophysics Data System.

Sarah Greenstreet, Siegfried Eggl, Meredith Rawls, and Mario Juric. The Impact of Satellite Constellations on Rubin Observatory's Legacy Survey of Space and Time (LSST). *Bulletin of the AAS*, August 2024.

Abdul Hayat, Peter Harrington, George Stein, Zarija Lukić, and Mustafa Mustafa. Estimating galactic distances from images using self-supervised representation learning. *arXiv e-prints*, pp. arXiv:2101.04293, January 2021. doi: 10.48550/arXiv.2101.04293. arXiv: 2101.04293 [astro-ph.IM] Number: arXiv:2101.04293 tex.adsnote: Provided by the SAO/NASA Astrophysics Data System.

Fabio Hernandez, George Beckett, Peter Clark, Matt Doidge, Tim Jenness, Edward Karavakis, Quentin Le Boulc'h, Peter Love, Gabriele Mainetti, Timothy Noble, Brandon White, and Wei Yang. Overview of the distributed image processing infrastructure to produce the Legacy Survey of Space and Time. In *European physical journal web of conferences*, volume 295 of *European physical journal web of conferences*, pp. 01042. EDP, May 2024. doi: 10.1051/epjconf/202429501042. arXiv: 2311.13981 [astro-ph.IM] Number: 01042 tex.adsnote: Provided by the SAO/NASA Astrophysics Data System.

Marc Huertas-Company, Regina Sarmiento, and Johan H. Knapen. A brief review of contrastive learning applied to astrophysics. *RAS Techniques and Instruments*, 2(1):441–452, January 2023. doi: 10.1093/rasti/rzad028. arXiv: 2306.05528 [astro-ph.IM] tex.adsnote: Provided by the SAO/NASA Astrophysics Data System.

Kartheik G. Iyer, Mikaeel Yunus, Charles O'Neill, Christine Ye, Alina Hyk, Kiera McCormick, Ioana Ciucă, John F. Wu, Alberto Accomazzi, Simone Astarita, Rishabh Chakrabarty, Jesse Cranney, Anjalie Field, Tirthankar Ghosal, Michele Ginolfi, Marc Huertas-Company, Maja Jabłońska, Sandor Kruk, Huiling Liu, Gabriel Marchidan, Rohit Mistry, J. P. Naiman, J. E. G. Peek, Mugdha Polimera, Sergio J. Rodríguez Méndez, Kevin Schawinski, Sanjib Sharma, Michael J. Smith, Yuan-Sen Ting, and Mike Walmsley. pathfinder: a semantic framework for literature review and knowledge discovery in astronomy. 275(2):38, December 2024. doi: 10.3847/1538-4365/ad7c43. arXiv: 2408.01556 [astro-ph.IM] Number: 38 tex.adsnote: Provided by the SAO/NASA Astrophysics Data System.

Henry W Leung and Jo Bovy. Towards an astronomical foundation model for stars with a transformer-based model. *Monthly Notices of the Royal Astronomical Society*, 527(1):1494–1520, January 2024. ISSN 0035-8711. doi: 10.1093/mnras/stad3015.

Siddharth Mishra-Sharma, Yiding Song, and Jesse Thaler. PAPERCLIP: Associating astronomical observations and natural language with multi-modal models. *arXiv e-prints*, pp. arXiv:2403.08851, March 2024. doi: 10.48550/arXiv.2403.08851. arXiv: 2403.08851 [astro-ph.IM] Number: arXiv:2403.08851 tex.adsnote: Provided by the SAO/NASA Astrophysics Data System.

Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Rudy Morel, Ruben Ohana, Mariel Pettee, Bruno Régaldo-Saint Blancard, Kyunghyun Cho, Shirley Ho, and Polymathic AI Collaboration. AstroCLIP: a cross-modal foundation model for galaxies. 531(4):4990–5011, July 2024. doi: 10.1093/mnras/stae1450. arXiv: 2310.03024 [astro-ph.IM] tex.adsnote: Provided by the SAO/NASA Astrophysics Data System.

Yanke Song, Ashley Villar, and Juan Rafael Martínez-Galarza. A poisson-process AutoDecoder for astrophysical, time-variable, x-ray sources. In *Machine learning and the physical sciences workshop@ NeurIPS*, volume 2024, 2024.

Yanke Song, Victoria Ashley Villar, Juan Rafael Martínez-Galarza, and Steven Dillmann. A Poisson Process AutoDecoder for X-ray Sources, February 2025. URL `http://arxiv.org/abs/2502.01627`.

George Stein, Peter Harrington, Jacqueline Blaum, Tomislav Medan, and Zarija Lukic. Self-supervised similarity search for large scientific datasets. *arXiv e-prints*, pp. arXiv:2110.13151, October 2021. doi: 10.48550/arXiv.2110.13151. arXiv: 2110.13151 [astro-ph.IM] Number: arXiv:2110.13151 tex.adsnote: Provided by the SAO/NASA Astrophysics Data System.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation

language models. *ArXiv*, abs/2302.13971, 2023. URL https://api.semanticscholar.org/CorpusID:257219404.

Mike Walmsley and Anna M. M. Scaife. Rare galaxy classes identified in foundation model representations. *arXiv e-prints*, pp. arXiv:2312.02910, December 2023. doi: 10.48550/arXiv.2312.02910. arXiv: 2312.02910 [astro-ph.GA] Number: arXiv:2312.02910 tex.adsnote: Provided by the SAO/NASA Astrophysics Data System.

A Appendix

A.1 LLM prompting

The following is the prompt we use to provide the text summarizations:

```
prompt_question2 = f"""
Given the text provided, search for information about the source
identified with any of the following names:

                {', '.join(repr(item) for item in name_ids)}.

The source is a source of type {tipos[j]}.

Again based on the text provided, answer the following questions
regarding the source in question, without mentioning the name of
the source or the target:

Is the source specifically mentioned in the text, or is the source
the target of the observation? If the answer is 'yes' to any of
these questions, do the following. If not, say only "Not
discussed".

A) Summarize the X-ray properties of the source in question, as
inferred directly from the data. Focus on variability (transient
behavior, periodicity, etc.), and spectral features (models
fitted, hardness ratios, n_h, etc.), but provide values of any
relevant measured quantities if measured directly from the X-ray
data.
B) Describe how these properties or other X-ray data from the
source is used to test the scientific hypotheses being examined in
the text provided.
"""
```

A.2 Nearest neighbor retrieval table

Table 3 shows the results of our nearest neighbor search experiment for astrophysical source 2CXO J100157.9+553945, an X-ray binary. In the aligned representation, objects that are semantically related have higher similarity with respect to their pre-alignment data embeddings, indicating that additional context provided by the text aids in the association. On the other hand, text alone does not contain all the information about spectral properties, as indicated by the higher $\sigma_{\rm HR}$ of the 10 nearest neighbors in the pre-alignment text embeddings with respect to the aligned representation. Thus, both data and text embeddings are being used in creating physically meaningful associations of astrophysical sources.

Table 3: Nearest neighbor retrieval experiment for the aligned representation and the pre-alignment text and data representations. $n_{\rm context}$ is the number of neighbors within the 10 nearest that are consistent with the source type. $\mu_{\rm HR}$ and $\sigma_{\rm HR}$ are respectively the mean and standard deviation of the hardness ratio values among the 10 nearest neighbors.

Representation	$n_{\rm context}$	$\mu_{\rm HR}$	$\sigma_{\rm HR}$
Aligned	6	-0.025	0.330
Pre-alignment text	10	0.084	0.649
Pre-alignment data	2	-0.051	0.318

A.3 Anomaly score histogram and example

In Figure 3 we show the distribution of URF anomaly scores for the aligned embeddings of validation set sources, and in Table 4 we list some anomalies. Among the objects with the highest anomaly scores is 2CXO J095550.1+694046 (Obsid 10542), which is a fluctuating, transient ULX. These types of transient luminous objects are rather rare and indicate very unique physical conditions. In Figure 4 we show the LLM summary for this source.

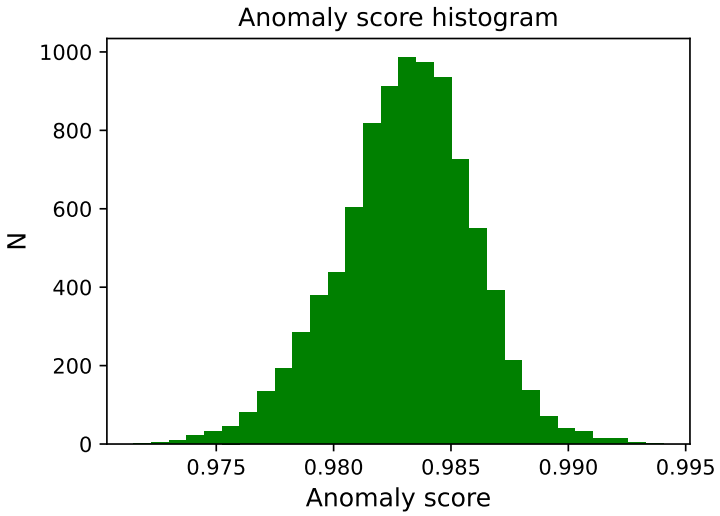

Figure 3: The distribution of anomaly scores obtained using the URF method applied to the aligned shared representations in the validation set.

Table 4: A list of highly variable anomalies resulting from applying the URF to the aligned embeddings. We list the identifiers, the hardness ratio, and the type.

Name	**ObsID**	**Hard**$_{\rm HS}$	**Type**
2CXO J095550.1+694046	10542	0.92	ULX
2CXO J223940.2+751321	8588	-0.13	YSO
2CXO J171632.6+430229	3778	-0.84	EmLine
2CXO J201536.9+371123	11092	0.64	Cataclysmic

'The source identified with "[WSG84] 2" is mentioned in the text provided. I summarize the X-ray properties and their implications: A) X-ray Properties\n\n1. **Variability**: The source has transient behavior with significant luminosity fluctuations. Specifically, periodicities have been highlighted. The periodic signals were found at approximately 55 d, although the stability of this modulation was questioned, implying that the source does not maintain consistent brightness levels over time.\n\n2. **Spectral Features**: - The source was fitted using models that include power-law and thermal components. The photon index for the power-law fit was constrained to be around Γ = 1.33 ± 0.15 in the energy range of 0.5–8 keV.\n - For the thermal model component, the temperatures were estimated at around 0.6 keV, suggesting an emission predominantly from hot plasma.\n - The absorption column density (N_H) was determined to be around 3 × 10^22 cm^−2, which suggests a significant amount of interstellar material obscuring the X-ray emissions from the source.\n\n3. **Luminosity**: The source\'s observed X-ray luminosity was reported to exceed 10^39 erg s^−1 during various observations, indicating that the source operates above the Eddington limit, which is characteristic of sources of type X.\n\n B) Scientific Hypotheses Testing\n\nThe properties of the source are critical in evaluating hypotheses regarding the nature of ultraluminous X-ray sources. The variability observed in the light curves across the different temporal analyses suggested that the source operates under the "propeller regime" of accretion. This regime indicates the presence of a strong magnetic field and can be key for understanding the behavior of neutron stars in binary systems.\n\nThe spectral features, including the determined N_H and photon index, help to classify the source's emission mechanism. They also provide insight into the environmental conditions surrounding the source, including the level of obscuration by interstellar matter. The correlation of the measured luminosity and the spectral characteristics with other known sources – particularly how they fit within the expected theoretical frameworks for ultraluminous X-ray sources – serves to validate or challenge prevailing models. The emission patterns can assist in distinguishing whether the source behaves more like an intermediate-mass black hole or if it is instead a neutron star system undergoing unusual accretion dynamics. In summary, the X-ray properties of the source serve to reinforce the arguments made regarding the diversity of ULXs and their potential to test existing astrophysical models related to black hole (or neutron star) formation and evolution through observations of their variability and spectral characteristics.'

Figure 4: The LLM-extracted summary for source 2CXO J095550.1+694046, an anomalous ULX.