Draft version December 6, 2024
Typeset using LaTeX default style in AASTeX631

The Multimodal Universe: 100 TB of Machine Learning Ready Astronomical Data

The Multimodal Universe Collaboration

Eirini Angeloudi,[1,2] Jeroen Audenaert,[3] Micah Bowles,[4,5] Benjamin M. Boyd,[6] David Chemaly,[6] Brian Cherinka,[7] Ioana Ciucă,[8,9,10] Miles Cranmer,[6,5] Aaron Do,[6] Matthew Grayling,[6] Erin E. Hayes,[6] Tom Hehir,[6,5] Shirley Ho,[11,12,13,5] Marc Huertas-Company,[1,2,9] Kartheik G. Iyer,[14,11,9] Maja Jablonska,[10,9] Francois Lanusse,[11,5,15] Henry W. Leung,[16] Kaisey Mandel,[6] Juan Rafael Martínez-Galarza,[17,18] Peter Melchior,[13] Lucas Meyer,[11,5] Liam H. Parker,[11,5,19] Helen Qu,[20] Jeff Shen,[13] Michael J. Smith,[21,9] Mike Walmsley,[16] John F. Wu,[7,22]

[1] *Instituto de Astrofisica de Canarias*
[2] *Universidad de La Laguna*
[3] *Massachusetts Institute of Technology*
[4] *University of Oxford*
[5] *Polymathic AI*
[6] *University of Cambridge*
[7] *Space Telescope Science Institute*
[8] *Stanford University*
[9] *UniverseTBD*
[10] *Australian National University*
[11] *Flatiron Institute*
[12] *New York University*
[13] *Princeton University*
[14] *Columbia University*
[15] *Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM*
[16] *University of Toronto*
[17] *Center for Astrophysics, Harvard & Smithsonian*
[18] *AstroAI*
[19] *University of California, Berkeley*
[20] *University of Pennsylvania*
[21] *Aspia Space*
[22] *Johns Hopkins University*

ABSTRACT

We present *the Multimodal Universe*, a new framework collating over 100 TB of multimodal astronomical data for its first release, spanning images, spectra, time series, tabular and hyper-spectral data. This unified collection enables a wide variety of machine learning applications and research across astronomical domains. The dataset brings together observations from multiple surveys, facilities, and wavelength regimes, providing standardized access to diverse data types. By providing uniform access to this diverse data, the Multimodal Universe aims to accelerate the development of machine learning methods for observational astronomy that can work across the large differences in astronomical datasets. The framework is actively supported and is designed to be extended whilst enforcing minimal self consistent conventions making contributing data as simple and practical as possible.

Keywords: Astronomical databases: miscellaneous — Methods: data analysis — Methods: statistical

The increasing volume and complexity of astronomical data has driven significant adoption of machine learning (ML) methods in astronomy. However, most applications are tailored to specific datasets, surveys, or instruments, requiring significant domain expertise and custom implementations. We introduce the Multimodal Universe, a curated collection of multimodal data designed to accelerate research in fields related to astronomy, astrophysics and machine learning. A full report on the dataset has been published by NeurIPS 2024 as The Multimodal Universe Collaboration et al. (2024). The most up to date version is documented at https://github.com/MultimodalUniverse/MultimodalUniverse/.

The Multimodal Universe combines data from major astronomical surveys, summarized by Table 1. The dataset is designed with several key principles, all of which are designed to make using the collated data in a machine learning context simpler than ever before.

- Multimodal alignment through careful cross-matching between surveys,
- Standardized data formats and access patterns,
- Comprehensive documentation of selection effects and biases,
- Public availability of all data download, collection and processing scripts.

Table 1. Adapted from Table 1 of The Multimodal Universe Collaboration et al. (2024), this is a summary of data included in the Multimodal Universe. Details of all samples, including full citations, are provided in the full paper. N_c indicates the number of channels of a given observation. Additionally, we provide a well documented python script on the landing page that enables the generation of BibTeX citations and acknowledgements for relevant datasets. [1] Indicates these are represented as 110 basis coefficients that can be resampled to an arbitrary wavelength grid. [2] Indicates simulated dataset.

Modality	Source Survey	N_c	Shape	Number of samples	Main science
Images	Legacy Surveys DR10	4	160×160	124M	Galaxies
	Legacy Surveys North	3	152×152	15M	Galaxies
	HSC	5	160×160	477K	Galaxies
	BTS	3	63×63	400K	Supernovae
	JWST	6-7	96×96	300K	Galaxies
Spectra	Gaia BP/RP	-	110[1]	220M	Stars
	SDSS-II	-	Variable	4M	Galaxies, Stars
	DESI	-	7081	1M	Galaxies
	APOGEE SDSS-III	-	7514	716k	Stars
	GALAH	-	Variable	325k	Stars
	Chandra	-	Variable	129K	Galaxies, Stars
	VIPERS	-	557	91K	Galaxies
Hyperspectral Image	MaNGA SDSS-IV	4563	96×96	12k	Galaxies
Time Series	PLAsTiCC[2]	6	Variable	3.5M	Time-varying objects
	TESS	1	Variable	1M	Exoplanets, Stars
	CfA Sample	5-11	Variable	1K	Supernovae
	YSE	6	Variable	2K	Supernovae
	PS1 SNe Ia	4	Variable	369	Supernovae
	DES Y3 SNe Ia	4	Variable	248	Supernovae
	SNLS	4	Variable	239	Supernovae
	Foundation	4	Variable	180	Supernovae
	CSP SNe Ia	9	Variable	134	Supernovae
	Swift SNe Ia	6	Variable	117	Supernovae
Tabular	Gaia	-	-	220M	Stars
	PROVABGS	-	-	221K	Galaxy
	Galaxy10 DECaLS	-	-	15K	Galaxy

We provide a number of benchmarks in The Multimodal Universe Collaboration et al. (2024) to highlight the utility and problems that these datasets can address. As a demonstration of the Multimodal Universe's utility, we reproduce AstroCLIP (Lanusse et al. 2023) using cross-matched Legacy Survey images and DESI spectra. Importantly the massive engineering effort needed for the original work is reduced to a few lines to select and cross match the appropriate data with the Multimodal Universe. This engineering cost is traditionally a large overhead when new machine learning projects are initiated. This release alleviates these costs, reduces errors of re-engineering the datasets, and unifies the underlying data framework enabling ML pipelines to be more easily transferred to data from other surveys. Additionally, this will enable all scales of ML models, including models trained on the full dataset across surveys, instruments, and domains.

The Multimodal Universe enables research into critical challenges in scientific ML, including distribution shifts, uncertainty quantification, and model calibration. The data is hosted in full by the flatiron institute[1]. The current release constitutes the data listed in Table 1. Extensions in the form of improved infrastructure and additional data are already being added to the Multimodal Universe, with incremental versioning expected over the coming years. Our collaboration looks forward to enabling access to data from observations, simulations and other sources in the near future.

Access The Multimodal Universe landing page to find details on the most recent version of the dataset, associated code, and contribution guide to add your data. We provide access to a simple script which can be run from the command line to retrieve the appropriate citations and acknowledgements for any and all of the available datasets.

Software: The Multimodal Universe (The Multimodal Universe Collaboration et al. 2024), huggingface datasets, Astropy (Astropy Collaboration et al. 2013, 2018, 2022), python, h5py, Globus (Foster & Madduri 2013)

REFERENCES

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33, doi: 10.1051/0004-6361/201322068

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123, doi: 10.3847/1538-3881/aabc4f

Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, ApJ, 935, 167, doi: 10.3847/1538-4357/ac7c74

Foster, I. T., & Madduri, R. K. 2013, in Proceedings of the 4th ACM Workshop on Scientific Cloud Computing, Science Cloud '13 (New York, NY, USA: Association for Computing Machinery), 1Ű2, doi: 10.1145/2465848.2480345

Lanusse, F., Parker, L., Golkar, S., et al. 2023, arXiv e-prints, arXiv:2310.03024, doi: 10.48550/arXiv.2310.03024

The Multimodal Universe Collaboration, Angeloudi, E., Audenaert, J., et al. 2024, in The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track. https://openreview.net/forum?id=EWm9zR5Qy1

[1] Available at: https://users.flatironinstitute.org/~polymathic/data/MultimodalUniverse/