# AstroCLIP: A Cross-Modal Foundation Model for Galaxies Reproduction and Extension

*–Executive Summary–*

In recent years, the field of astronomy has experienced a data flood through the advent of large-scale sky surveys like the Dark Energy Spectroscopic Instrument (DESI). These surveys generate vast, complex datasets that require sophisticated analysis techniques. Traditional machine learning models, which rely heavily ground truth labels, are often inadequate due to the scarcity of labeled astronomical data. This challenge has driven the development of self-supervised learning (SSL) methods, which can leverage the unlabelled data and generate rich, versatile representations of physical objects in the form of low-dimensional vectors called embeddings. While in this field, SSL so far has been primarily applied to single data types (modalities), observational astronomy is inherently multimodal, with astronomical objects being observed in various ways, such as images, spectra. Additionally, considering that different modalities are essentially different views of the same underlying physical object, it follows that the embeddings of these modalities in a shared latent space should be able to be structured in physically meaningful manner.

AstroCLIP is a pioneering model that can embed both galaxy images and spectra into a shared, stucturally meaningful latent space. The key principle behind this embedding space construction is *contrastive learning*: a simple framework in which embeddings of the same galaxy are pulled closer together and those of different galaxies are pushed further apart. This creates a semantically meaningful map of the data, in which locations in the embedding space correspond to physically meaningful properties of the galaxies, while distances reflect the similarity of the galaxies in the original data space. This allows for a wide range of applications, including accurate in-modality and cross-modality semantic similarity serac, physical property estimation, anomaly detection, and more, without any additional training or tuning. Hence, the embedding space can act as a powerful 'foundation' for downstream tasks; these types of models are often dubbed *foundation models*. AstroCLIP constitudes one of the first such foundation models in astronomy, in the growing field of research on foundational models in science.

## Objectives and Methodology

The primary objectives of this project are:

- **Reproduce Original Results:** Deploy pre-trained single-modal models for galaxy images and spectra from the DESI Legacy Survey and train a cross-modal model under a contrastive learning framework to align the embeddings of these modalities.
- **Extend Analysis:** Analyse the embedding space structure using clustering algorithms and dimensionality reduction techniques to demonstrate the model's capability to capture semantically meaningful information about galaxies across different modalities.

Raw galaxy images and spectra are high-dimensional and complex. To extract meaningful information from these data, we need to compress them into a lower-dimensional embedding that captures the essential features of the modality and is a shared space for both modalities. To that end, we leverage the concept of *transfer learning* by utilising two pre-trained models: one that was trained on galaxy images and another on galaxy spectra. These models take the raw data as input and were trained on a secondary task that is related to our task of embedding construction. The key principle behind this approach is that the pre-trained models have built a good internal representation of the data, and hence can construct richer embeddings. We

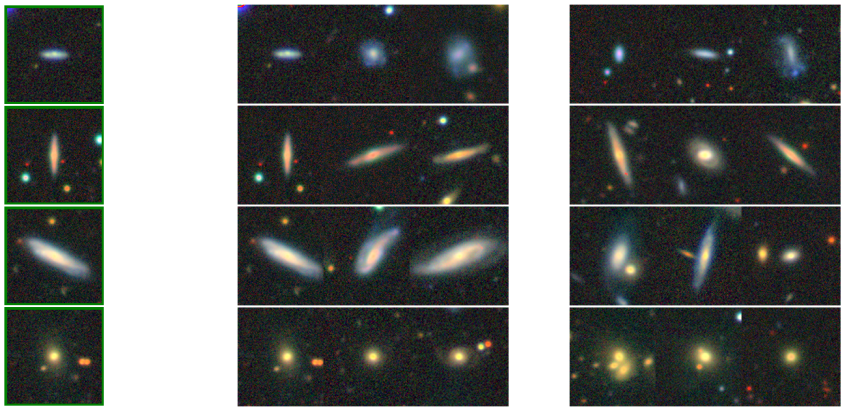
then keep the weights of the pre-trained models fixed and append both of them with small, trainable fully connected neural networks that map the high-dimensional data to the shared low-dimensional embedding space.

We then perform a new training phase, where the two encoders are jointly trained to align the embeddings of the two modalities. This is done by using a contrastive loss function that maximises the similarity between cross-modal embeddings that correspond to the same galaxy while simultaneously minimising the similarity between cross-modal embeddings that correspond to different galaxies.

## Downsteam Tasks, Extension & Key Findings

After training AstroCLIP, we embed the DESI Legacy Survey dataset into our shared embedding space and perform a series of downstream tasks to evaluate the model's performance. These tasks and their key findings are as follows:

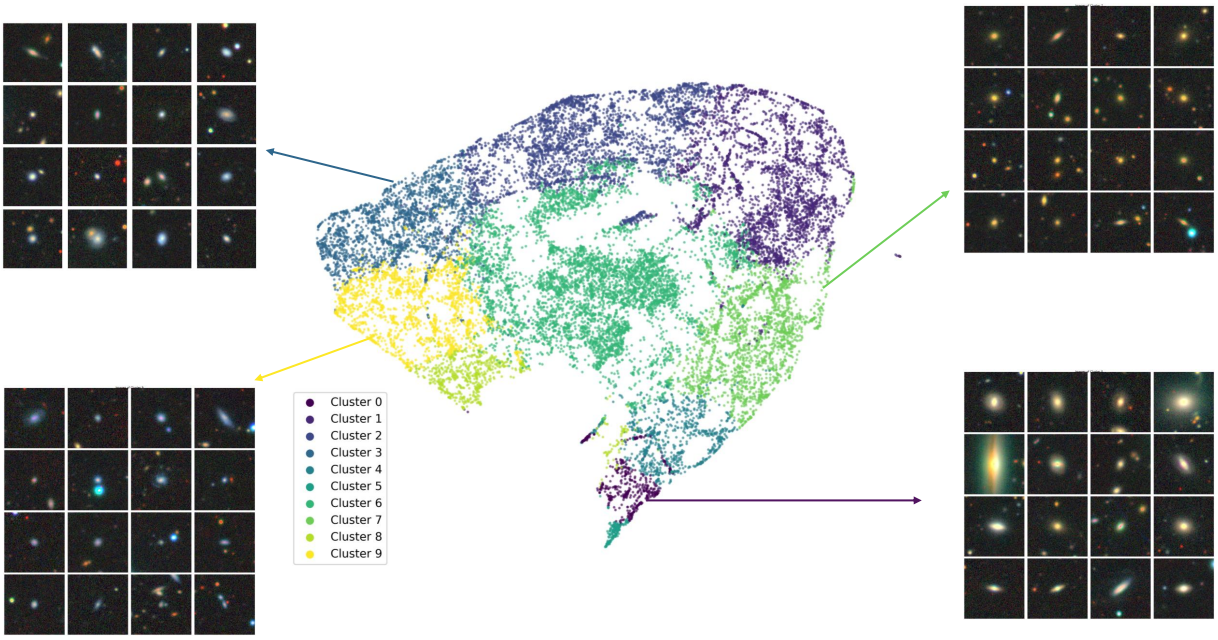
**Semantic Similarity Search:** We assign a random galaxy as our query and retrieve the most similar galaxies based purely on their embeddings. AstroCLIP demonstrates robust query retrieval capabilities, finding visually and spectrally similar galaxies both for in-modality search (for example image query to image retrieval) and cross-modality search (for example image query to spectrum retrieval). This capability can be instrumental for parsing large astronomical datasets and searching for rare or unusual astronomical objects.



(a) Query galaxies (b) **In-modality** retrieval (c) **Cross-modality** retrieval

**Physical Property Prediction:** By simply averaging the embeddings of the nearest neighbours of a galaxy in the embedding space, AstroCLIP predicts its redshift and stellar mass with great accuracy. This is also the case when we consider only the embeddings of one modality in isolation. This indicates that these physical properties emerge as natural structuring principles in our model, despite not being explicitly trained for them. This is a strong indication that the embedding space captures physically meaningful properties of galaxies.

**Extension to original work: Embedding Space Analysis:** We extend the original work by visualising the embedding space using a projection onto a 2D plane. Our projection revealed distinct patterns in structuring based on redshift and stellar mass. We also identified distinct 'islands' in the projection, including one composed of images with capture artifacts, and others containing mostly spiral galaxies with high brightness. These clusters were consistent across both image and spectrum embeddings, highlighting the model's ability to align and structure embeddings around physically meaningful properties like morphology, despite not being explicitly trained on these features.


Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9