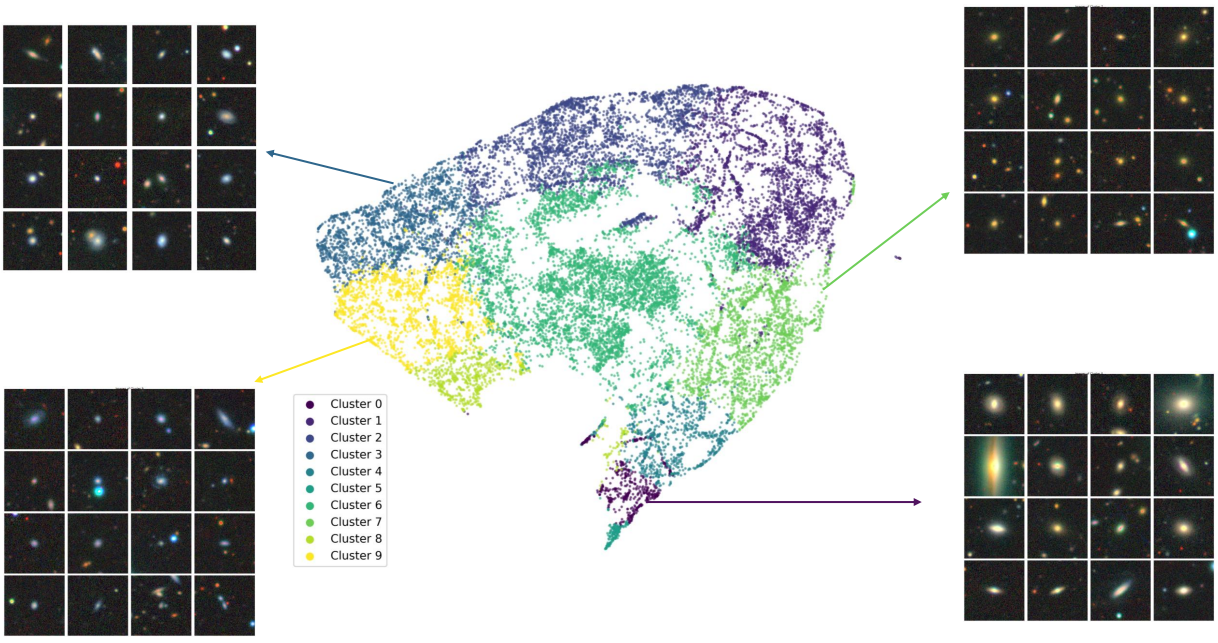
–Executive Summary–

AstroCLIP: A Cross-Modal Foundation Model for Galaxies

Reproduction and Extension

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 on 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 and spectra. Moreover, given that different modalities represent distinct perspectives of the same underlying physical entity, the embeddings of these modalities in a shared latent space should be able to be structured around shared semantics.

AstroCLIP is a pioneering model that can embed both galaxy images and spectra into a single, information-rich 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 structurally meaningful map of the data, in which locations in the embedding space correspond to physical properties of the galaxies, while distances reflect the similarity of the galaxies in the original data space. This enables a broad spectrum of applications, including precise in-modality and cross-modality semantic similarity search, physical property estimation, anomaly detection, and more, without requiring additional training or tuning. Hence, the embedding space can serve as a robust foundation for downstream tasks; these models are often referred to as *foundation models*. AstroCLIP represents one of the first such foundation models in astronomy, contributing to the growing field of foundational models for science.



2-Dimensional visualisation of AstroCLIP's embedding space images from different clusters.

Objectives and Methodology

The primary objectives of this project are:

- **Reproduce Original Results:** Follow the key principles of the original AstroCLIP paper (version 1) and obtain similar performance on the same dataset.
- **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.

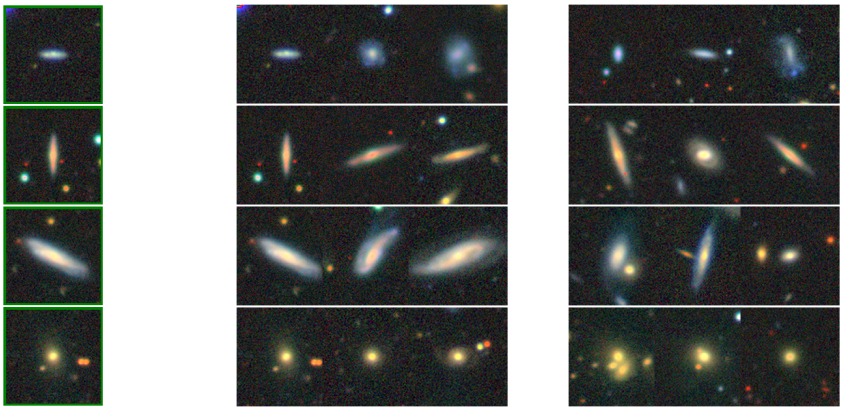
Raw galaxy images and spectra are high-dimensional and complex. To extract meaningful information from these data, we must compress them into a lower-dimensional embedding that captures the essential features and provides a shared space for both modalities. To achieve this, we leverage the concept of *transfer learning* by utilising two pre-trained models: one trained on galaxy images and another on galaxy spectra. These models take the raw data as input and were trained on secondary tasks that are related to our objective of embedding construction. The key principle behind this approach is that the pre-trained models have developed robust internal representations of the data, enabling them to construct richer embeddings. We then keep the weights of the pre-trained models fixed and append them with small, trainable fully connected neural networks that map the high-dimensional data to the shared low-dimensional embedding space.

Next, a new training phase is initiated, where the two encoders are jointly trained to align the embeddings of the two modalities. This is achieved using a contrastive loss function that maximises the similarity between cross-modal embeddings corresponding to the same galaxy while simultaneously minimising the similarity between cross-modal embeddings corresponding to different galaxies.

Key Findings for Downsteam Tasks & Extension

After training AstroCLIP, we embed the DESI Legacy Survey dataset into our shared latent 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: A random galaxy is assigned as the query, and the most similar galaxies are retrieved based purely on their embeddings. AstroCLIP demonstrates robust query retrieval capabilities, identifying visually and spectrally similar galaxies for both in-modality search (e.g., image query to image retrieval) and cross-modality search (e.g., image query to spectrum retrieval). This capability can be instrumental in 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 averaging the embeddings of the nearest neighbours of a galaxy in the embedding space, AstroCLIP accurately predicts its redshift and stellar mass. This holds true even when considering only the representations of one modality in isolation. Such results suggest that these physical properties naturally emerge as structuring principles in our model, despite not being explicitly trained for them. Consequently, this strongly indicates 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. This projection revealed distinct patterns in structuring based on redshift and stellar mass. Furthermore, distinct 'islands' were identified in the projection, including one composed of images that contain artificial artefacts, 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 their morphology, despite not being explicitly trained on these features.