Probabilistic inference of galaxy properties using multi-modal variational autoencoders

Kirill Grishin

Eric Aubourg, Cécile Roucelle, Cyrille Rosset

Bayesian Deep Learning Workshop ed. 3, 2025

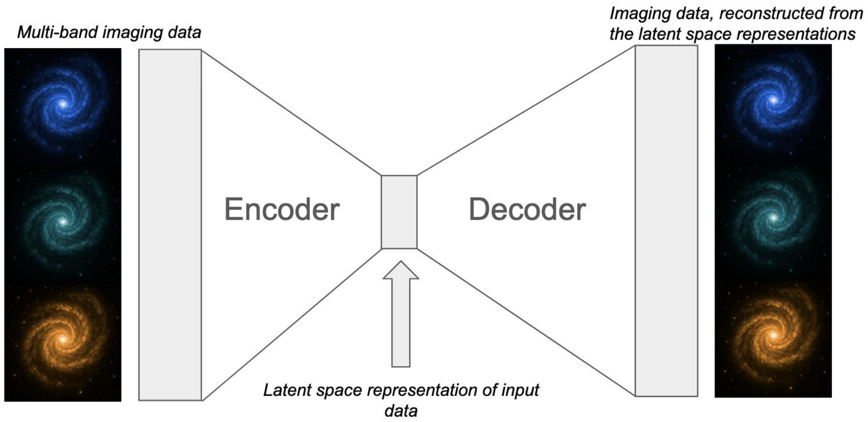
Motivation

The goal is to obtain probabilistic estimate for the properties of galaxies, including structural ones, such as size, ellipticity, position angle and photometric ones, including possible redshift, spectral shape, basic information about stellar population properties.

This approach allows probabilistic analysis of the large samples of objects identified in the deep astronomical images.

Architecture of the network

The structural and photometric parameter distributions can be derived from the latent space representations encoded from galaxy imaging data. This approach is part of a broader probabilistic analysis pipeline that already includes object deblending (Biswas+24). The resulting parameter distributions can be directly used in the probabilistic weak lensing studies.



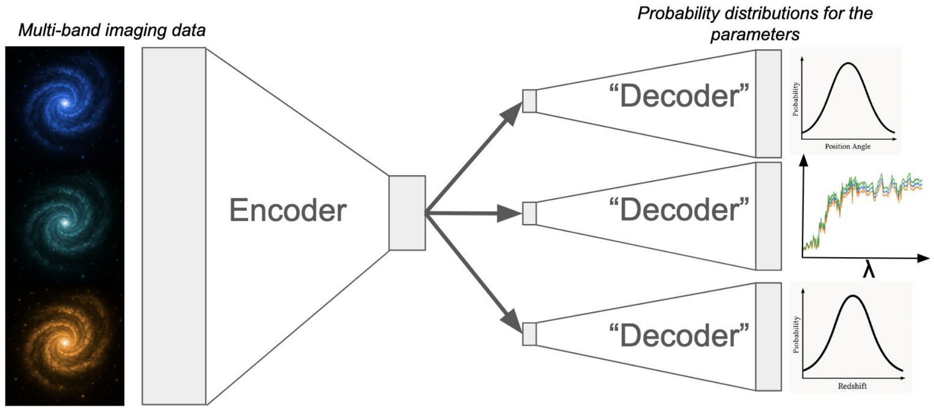
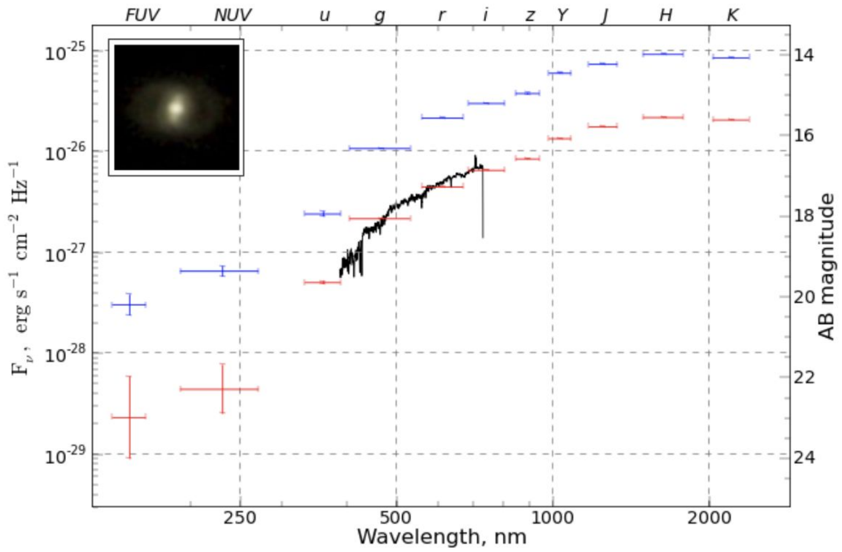


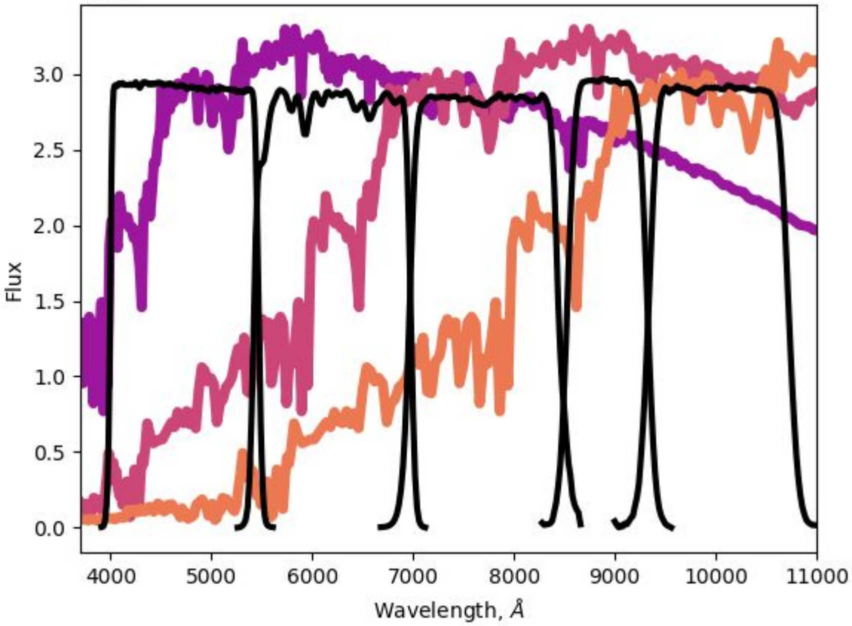
Photo-z part

In this work, I focus on the component responsible for inferring redshift and stellar population properties. As an initial step in validating the training strategy, we train the network on integrated photometric measurements. Transferring this approach to image cutouts will follow in a next phase.



Training dataset

For training the photometric component, we use a mock dataset of broadband photometric measurements. This dataset is generated using PEGASE.2 stellar population models, assuming the simplest possible star formation history: a single starburst with fixed age and metallicity. To produced magnitudes (and fluxes) we add Gaussian noise to the simulated data.



Training procedure

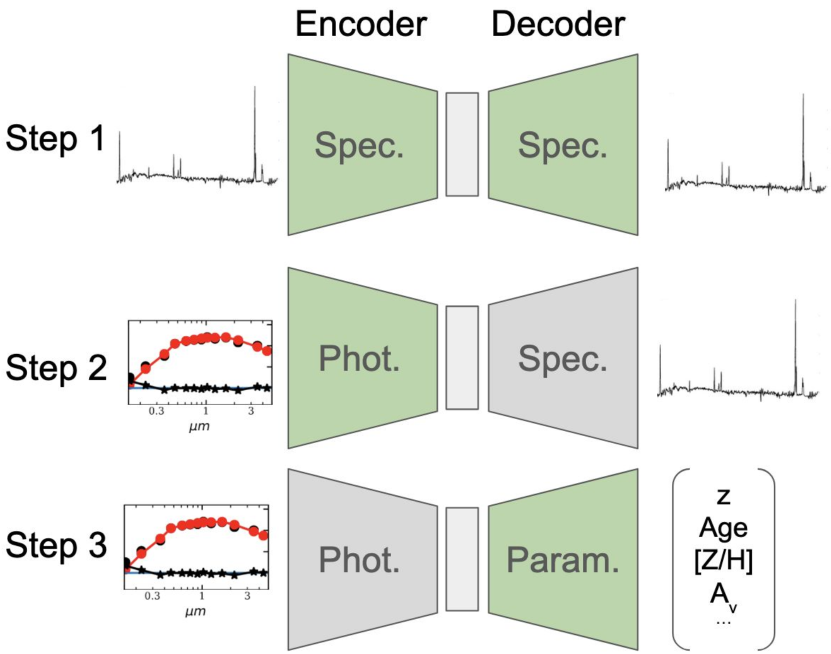
The training was carried out in the following stages:

1. Classical training of the spectral VAE.

2. Training only the encoder of the photometry VAE, while keeping the spectral decoder fixed.

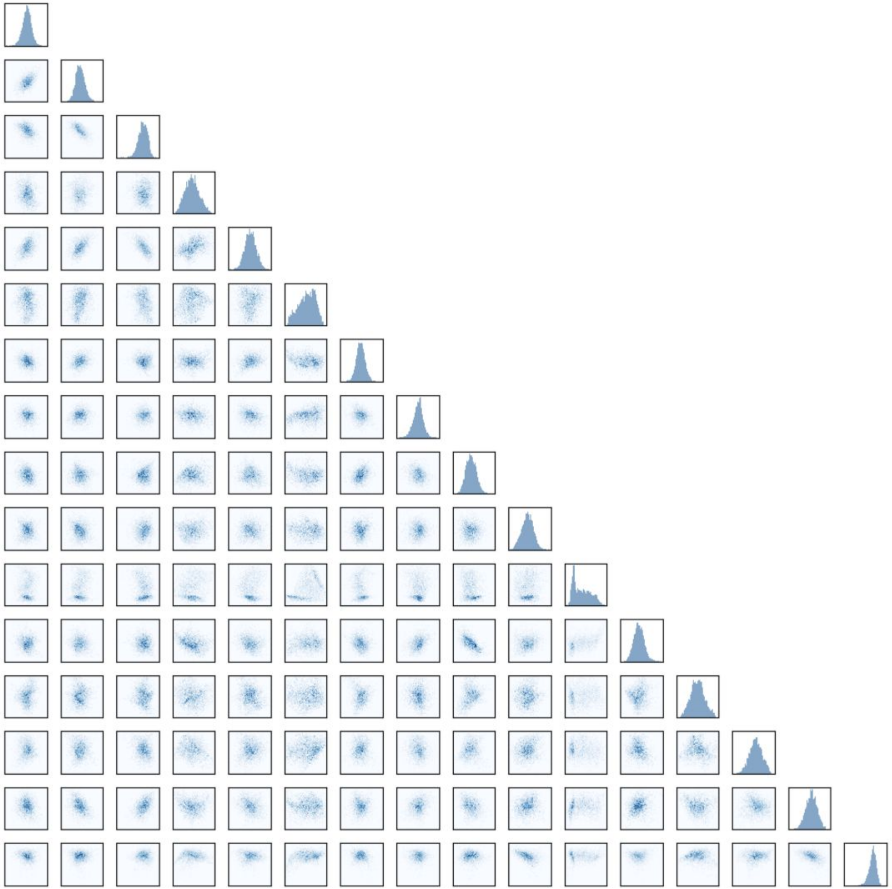
3. Training the parameter decoder, with the photometry VAE encoder kept fixed.

This stepwise procedure helps to mitigate degeneracies in the latent parameter space and results in more realistic posterior distributions.



Latent space distributions

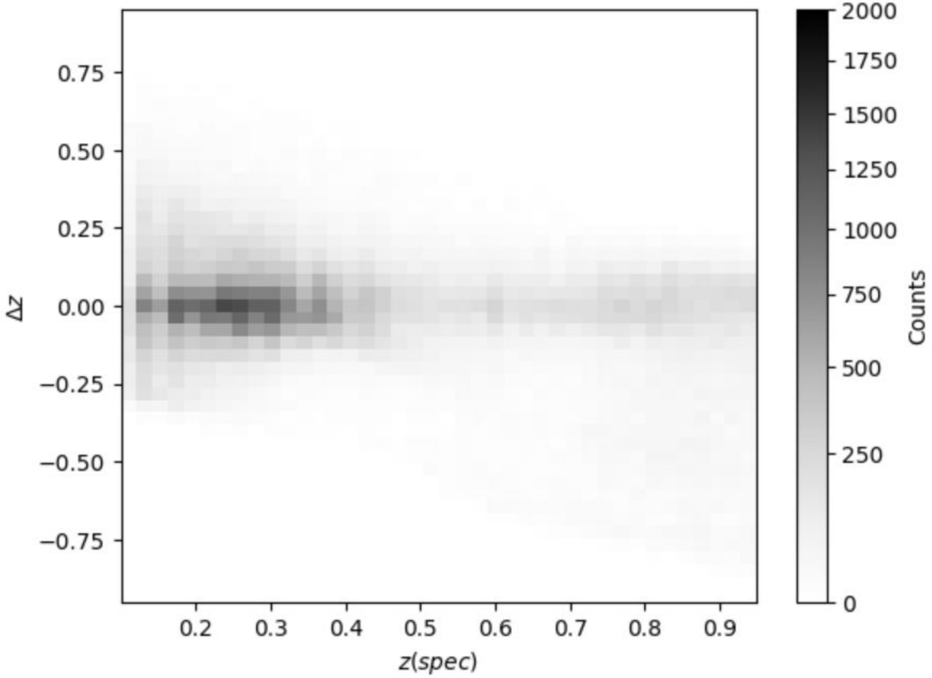
We achieved a relatively compact latent space, consisting of just 16 dimensions. This number closely aligns with the actual number of physical parameters required to describe galaxy SEDs, indicating that the model effectively captures the underlying structure of the data.



Validation of the results

We validated the trained network using the DESI DR1 spectroscopic sample for galaxies within the HSC SSP field in the redshift range 0 < z < 1.

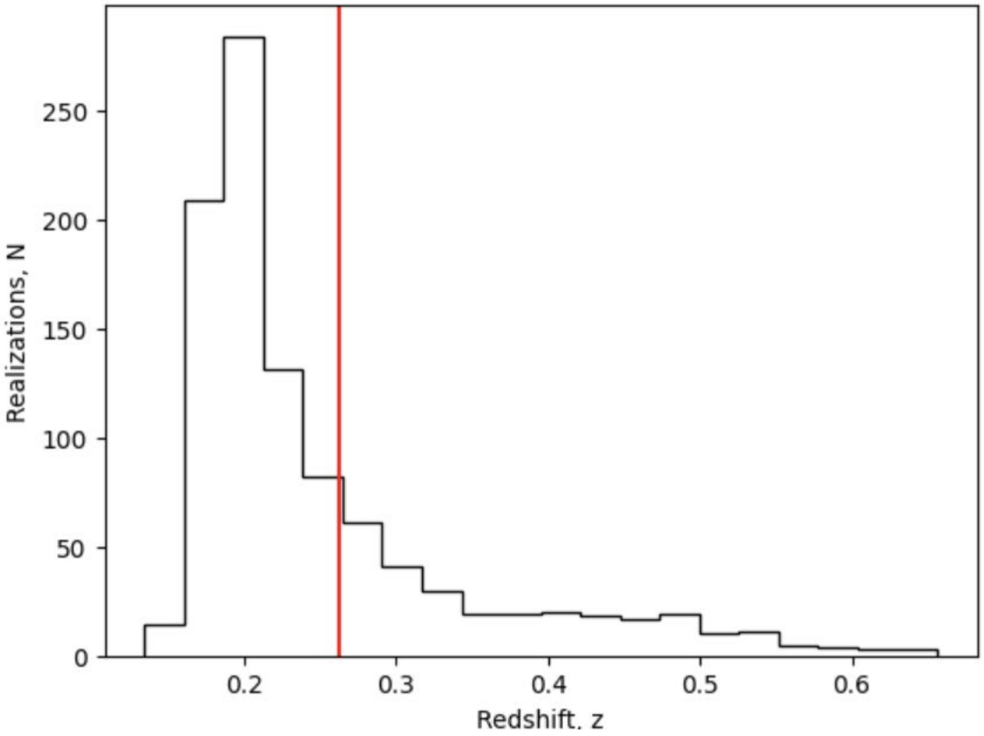
The achieved scatter in redshift estimation was Δz/(1+z)~0.07. The performance varied across two distinct subsamples: quiescent galaxies exhibited significantly lower scatter, while star-forming galaxies showed substantially higher redshift uncertainties.



Obtained distributions

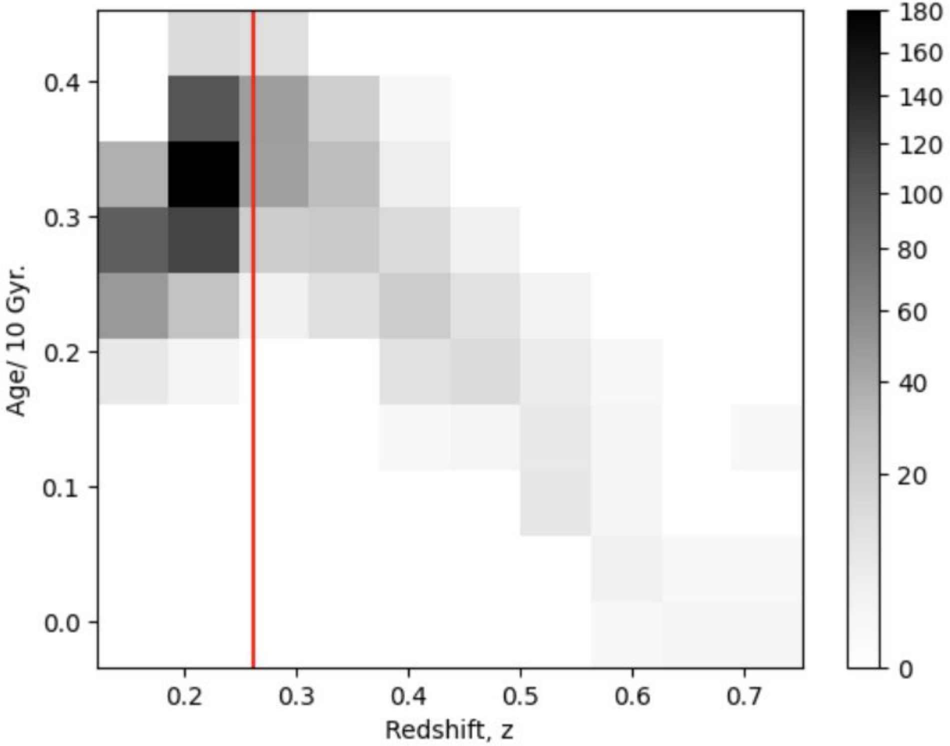
By sampling the latent space, the VAE enables the recovery of parameter distributions, even when they deviate from a Gaussian form.

These distributions can reveal the presence of multiple plausible solutions and indicate the potential range of alternative parameter values.



Obtained distributions

Photometric analysis using VAEs also enables the study of complex parameter distributions—for example, joint probability distributions as functions of redshift and stellar population age. These distributions for stellar population properties can further be incorporated into cosmological analyses, providing a more detailed understanding of galaxy evolution and large-scale structure.



Conclusions

The application of multi-modal VAEs for inferring photometric properties of galaxies demonstrates several key advantages:

- **Efficient training**: Multi-modal VAEs can be effectively trained using a staged approach, where individual components are trained separately.

- **Compact latent representation**: The dimensionality of the latent space can be significantly reduced, helping to avoid degeneracies while preserving meaningful physical information.

- **Training on mock data**: Training exclusively on mock datasets proves to be effective, offering greater flexibility and enabling straightforward transfer of network weights to real observations.