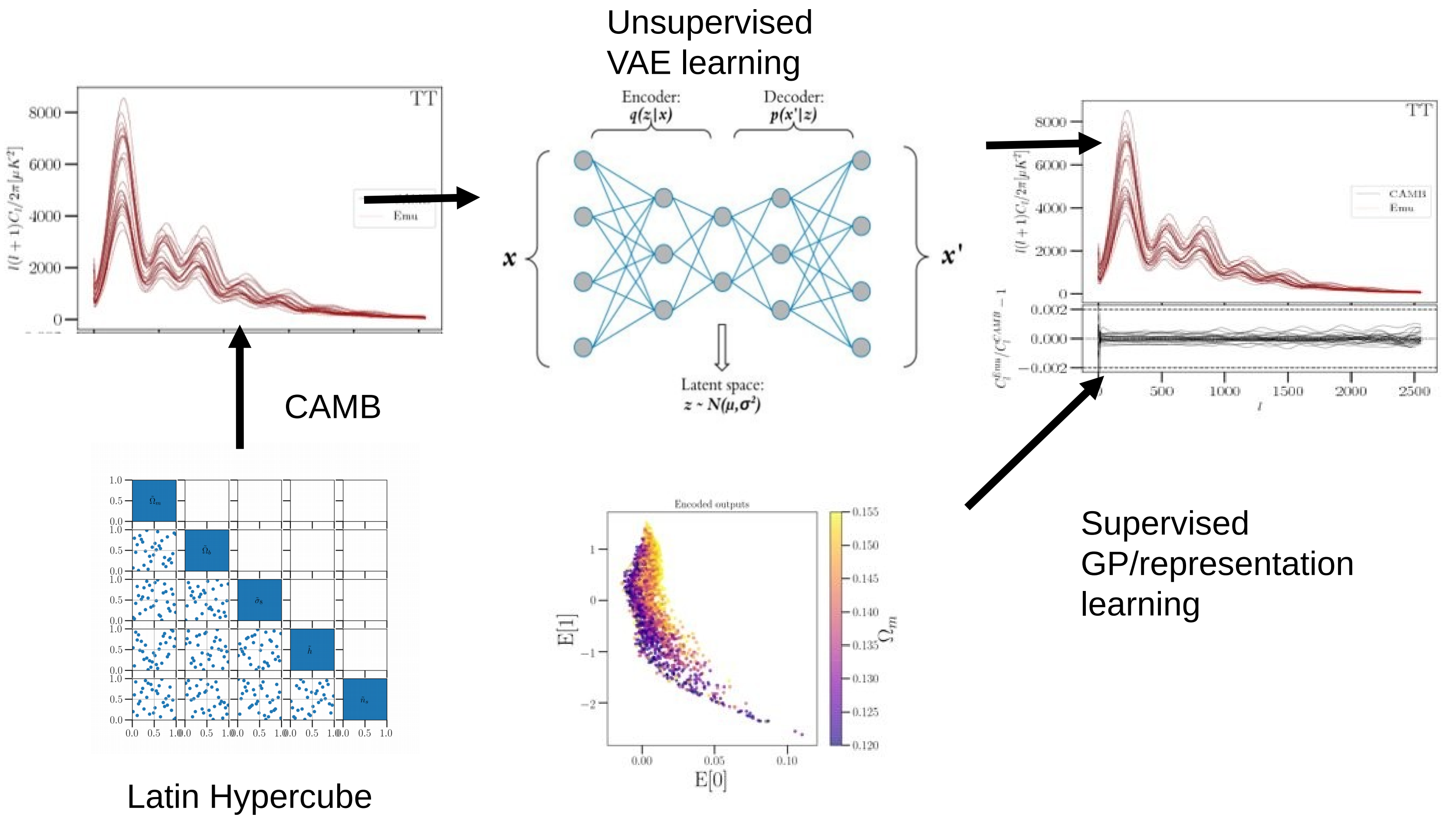
Accelerating HEP Science: Inference and Machine Learning at Extreme Scales

Prasanna Balaprakash, Jonas Chaves-Montero, Arindam Fadikar, Robert Gramacy, Salman Habib, Katrin Heitmann, David Higdon, Earl Lawrence, Yuewei Lin, Zarija Lukic, Sandeep Madireddy, Dimitriy Morozov, Nesar Ramachandra, Anze Slosar, Stefan Wild, Shinjae Yoo

- **Emulation of the Cosmic Microwave Background angular power spectrum using Probabilistic Generative Models**

The Cosmic Microwave Background (CMB) is a remnant radiation from the early Universe. It can be accurately characterized by its angular power spectrum, which is a high-dimensional function that depends on all the parameters of our current cosmological model. A dimensionality reduction is thus necessary for making Gaussian process-based interpolation of the angular power spectrum numerically tractable. Traditionally, this has been done using Principal Component Analysis. We developed a Variational Autoencoder-based dimension reduction for angular power spectrum and demonstrated that emulators build using this approach are very precise, in particular our emulator presents 1% errors over the dynamic range of state-of-the-art surveys CMB surveys like Planck. In addition, our emulator results in a ~2000 speed-up compared to traditional methods for generating the CMB angular power spectrum. As a consequence, Markov Chain Monte Carlo cosmological analyses will be speed-up several orders of magnitude, allowing very efficient parameter estimation.

In addition to this, we proposed an end-to-end emulating approach using variational information bottleneck, which is capable of customizing the dimension reduction to achieve better predictive models with quantified uncertainty in a single step.


Unsupervised VAE learning

CAMB

Supervised GP/representation learning

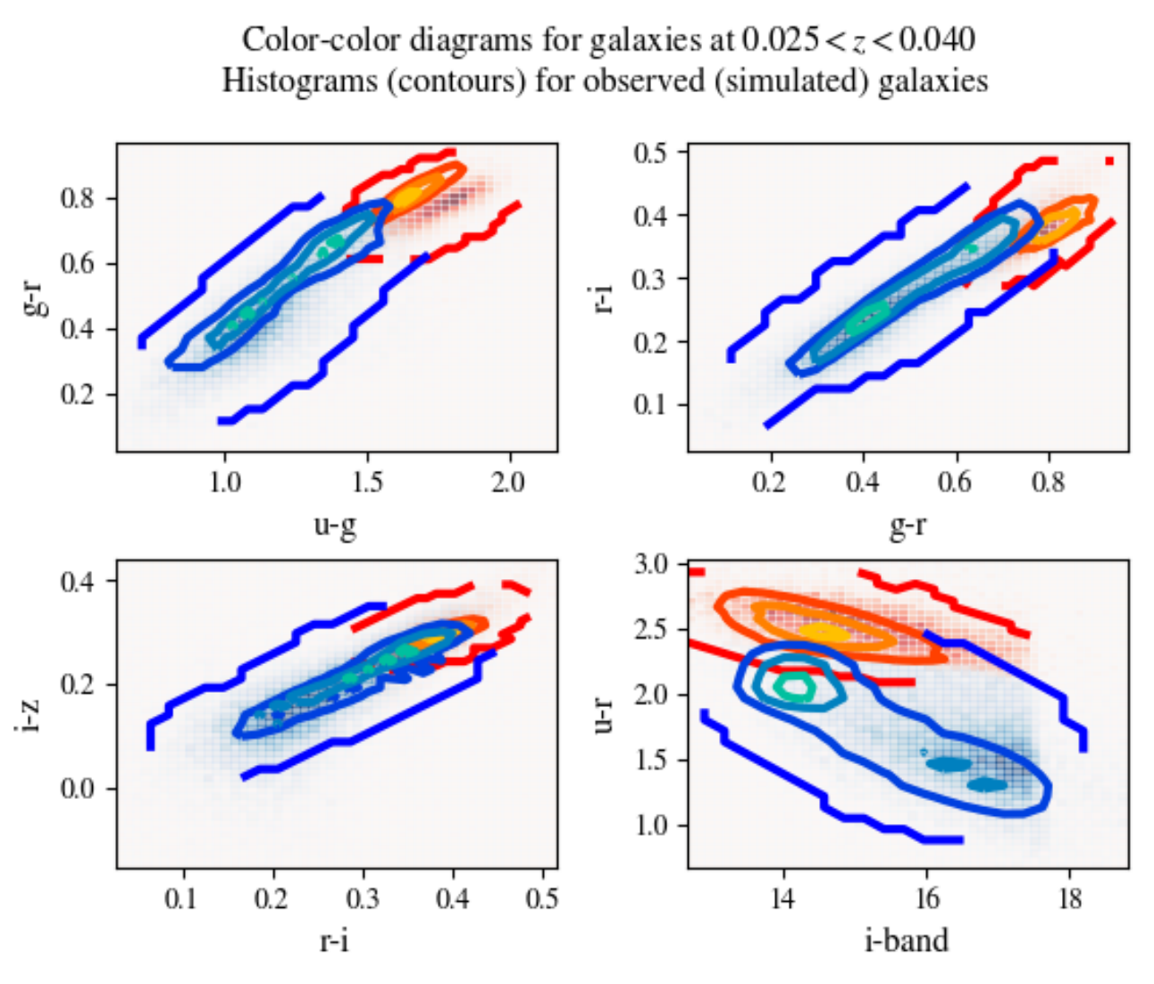
Latin Hypercube

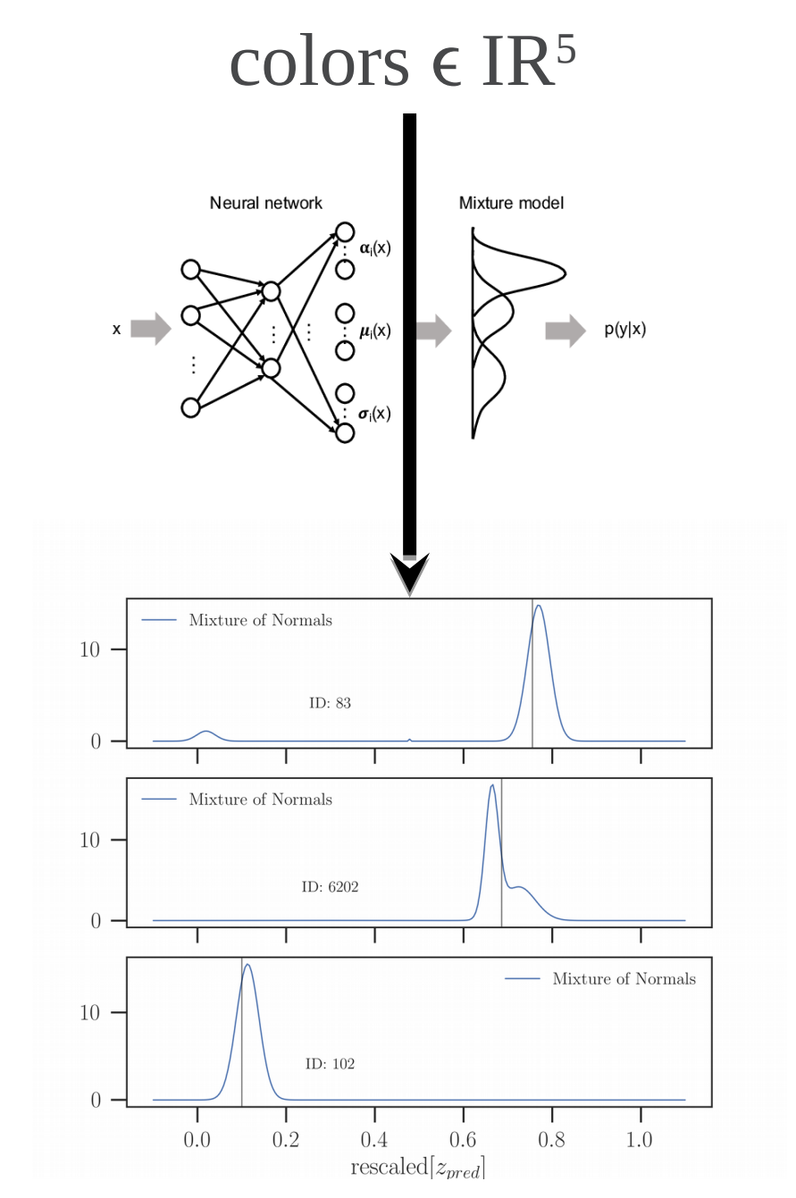
- **Photometric redshift estimation using Bayesian Neural Networks**

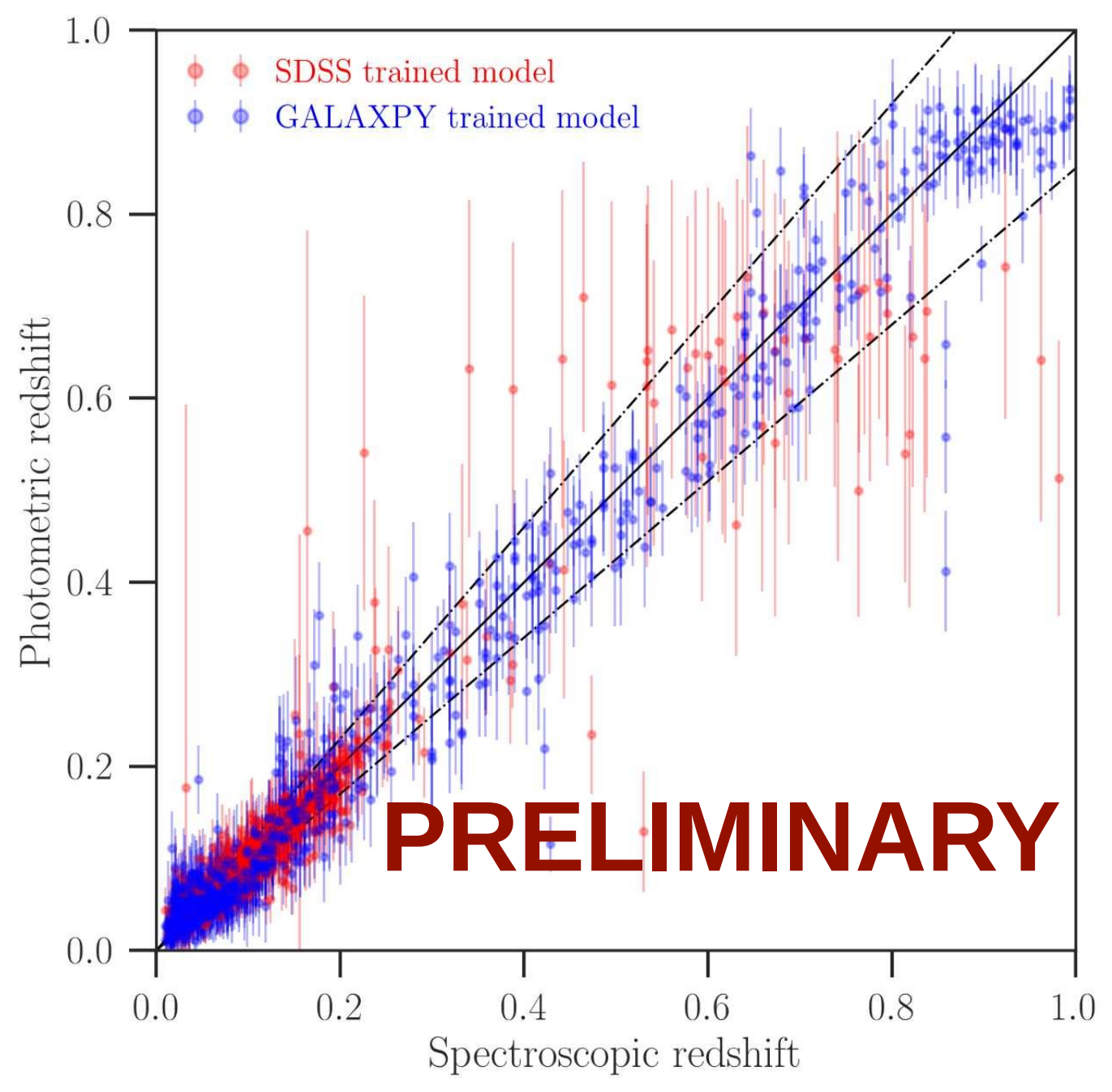
Future galaxy surveys such as LSST and SPHEREx will sample the spectral energy distribution of hundreds of millions of galaxies using a reduced number of elements. As information encoded in these elements ("colors") is very limited, it is very challenging to estimating the distance ("redshift") to these galaxies. This is necessary to go from two-dimensional images of the sky to a three-dimensional map of the Universe, required for cosmological analysis.

Traditionally, machine learning methods have been used to estimate photometric redshifts. However, they rely on training sets that are incomplete for faint sources, which introduces biases in the results. We avoid this problem by augmenting our training set using GALAXPY, a robust generative model for emulating galaxy colors using Gaussian Processes. GALAXPY captures effects of star formation histories, metallicities, initial mass functions, dust attenuations, and emission line ratios on colors and allow us to fill the regions of the parameter space not covered by data from observations.

We use Mixed Density Networks for mapping galaxy colors to redshifts, producing a probability redshift distribution for each source, and we handle degeneracies that may arise in the mapping using Gaussian Mixture models. Using SDSS galaxies, our preliminary results show that the number of outliers is reduced by using synthetic data produced by GALAXPY as training set.


Color-color diagrams for galaxies at $0.025 < z < 0.040$
Histograms (contours) for observed (simulated) galaxies



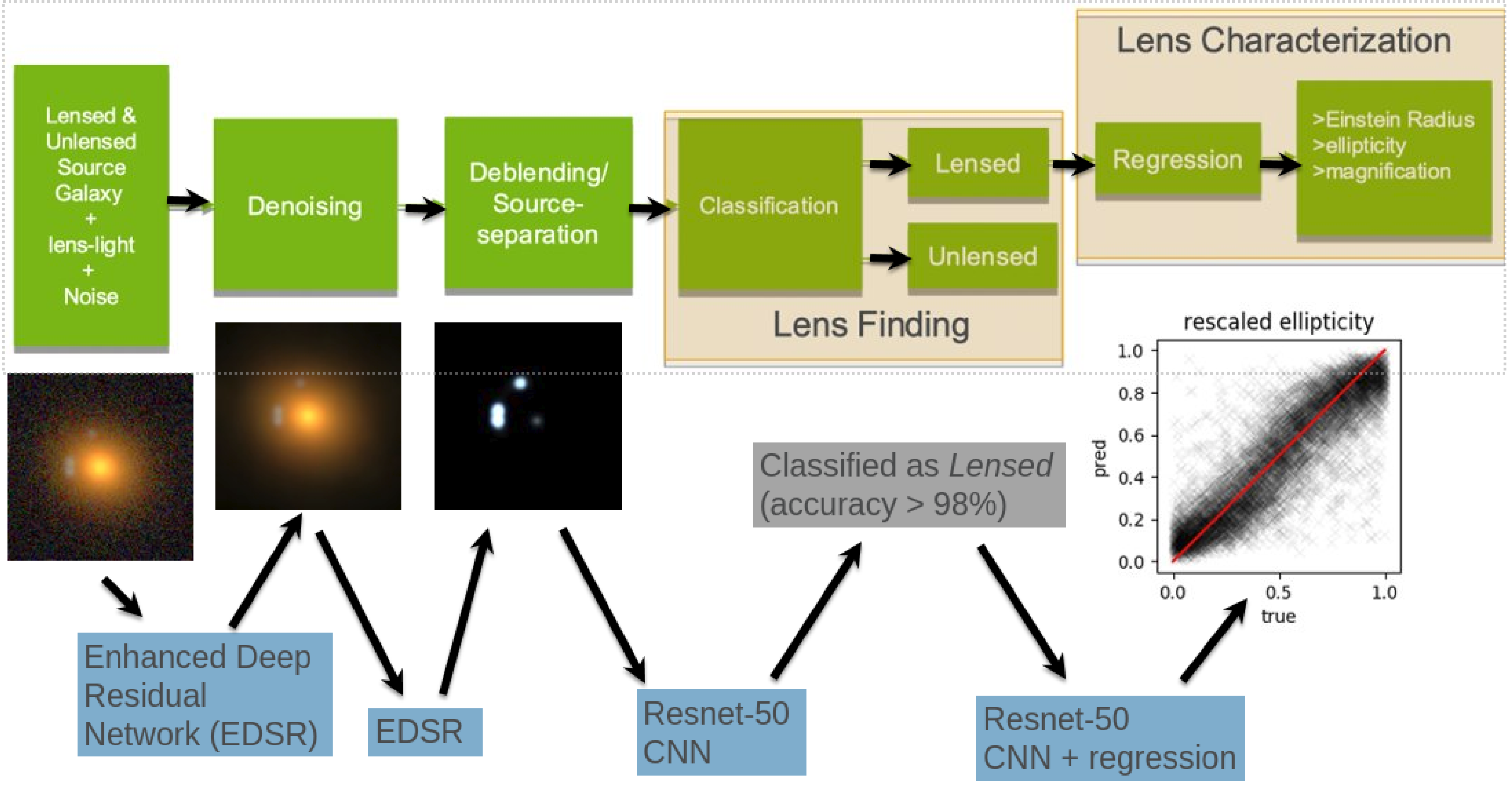


- **Classification and Characterization of strongly-lensed galaxies**

Gravitational lensing consists on the deflection of light when it passes near to a gravitational mass. When the strength of the deflection is so severe that the image of the background source is very distorted, it is classified as strong. Strong galaxy lensing can be used to accurately determine the distribution of matter on small scales, shedding light on the nature of dark matter. In addition, strong lensed variable sources allow to precisely measure the expansion history of the Universe.

Traditionally, strong lensed systems have been identified by eye. Nonetheless, the next generation of galaxy surveys such as Euclid and LSST will produce an enormous amount of data that is beyond human capability. Machine learning pipelines for the automatic identification and characterization of strong lenses are thus mandatory. We develop a deep learning based pipeline that identify strong lenses, remove possible contaminants, reduce the noise in the image, separate the light from the background and foreground sources, and characterize the properties of the system.

We start by generating one million simulated images for training using data from the CosmoDC2 project, one of the most complex synthetic catalogs ever constructed. It was designed for the second LSST DESC Data Challenge and led by the Argonne Cosmology Group. Using these images, we train an end-to-end, fully Convolutional Neural Network (CNN) based architecture - Enhanced Deep Super Resolution network - to identify, denoise, and deblend them. The classification is then performed using a deep residual CNN (Restnet-50), which achieves an accuracy of 98.3% on testing data. A similar model is used to estimate the properties of strong lensed systems, which propagates errors on the input data onto the inferred parameters.


Lensed & Unlensed Source Galaxy + lens-light + Noise

Denoising

Deblending/ Source-separation

Classification

Lensed

Unlensed

Lens Finding

Lens Characterization

Regression

>Einstein Radius
>ellipticity
>magnification

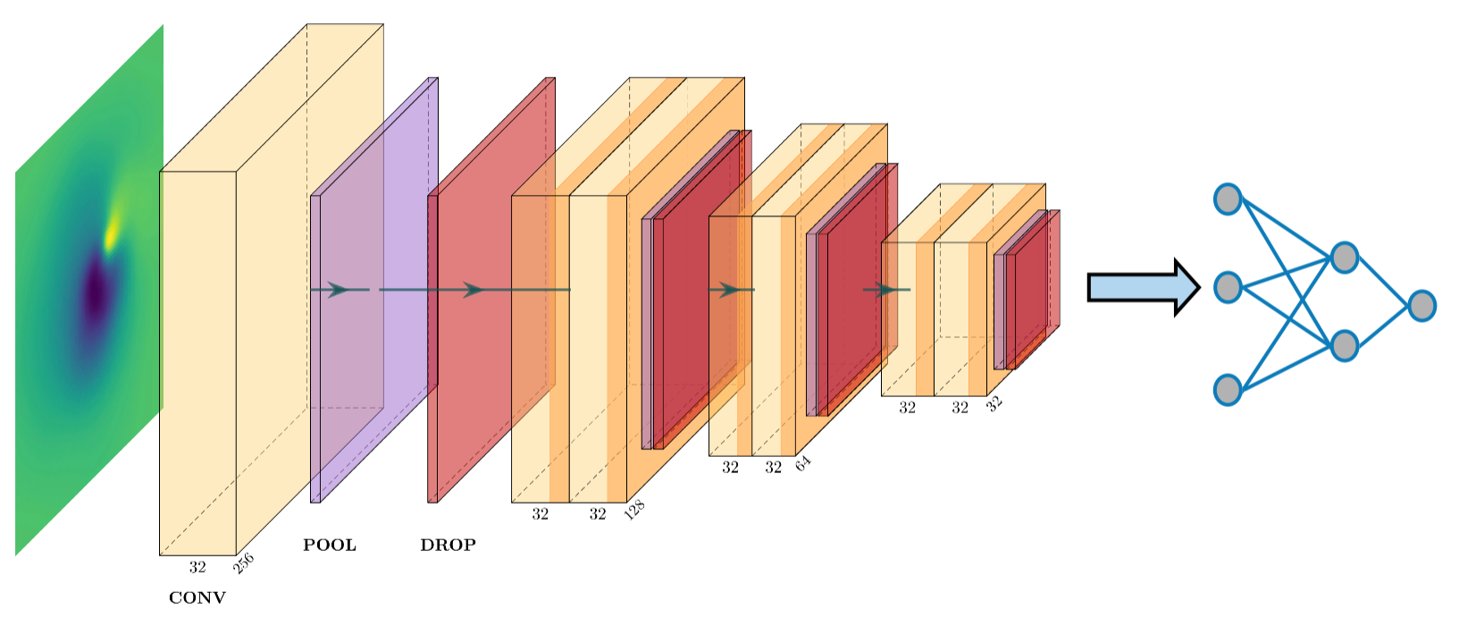
Classified as *Lensed* (accuracy > 98%)

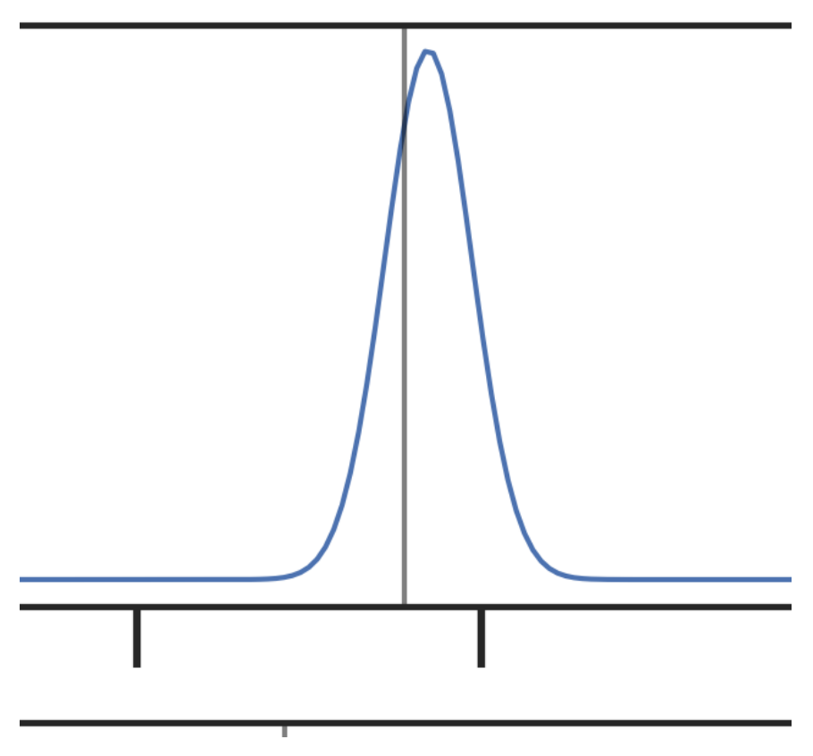
Enhanced Deep Residual Network (EDSR)

EDSR

Resnet-50 CNN

Resnet-50 CNN + regression





Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.