Galaxy clustering and deep learning for photometric redshifts with Euclid data

Euclid-France Clustering 2022

Vincent Duret
Supervisors : Stéphanie Escoffier and William Gillard

Motivations :

Goals of the thesis : obtain better photometric redshifts and use their large number to make a photometric galaxy clustering analysis to derive cosmological constraints.

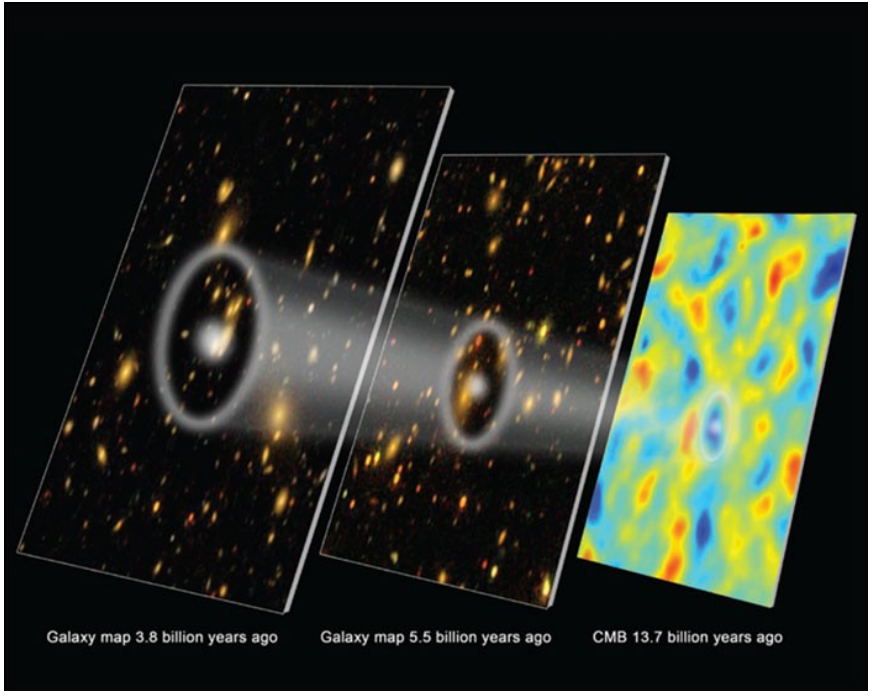
A photometric galaxy clustering approach : why ?

One of the main probe of Euclid will be galaxy clustering with 50 million spectroscopic redshifts with $\sigma = 0.001(1+z)$. Cosmological parameters constraints will be improved thanks to this probe.

However, Euclid will also provide photometry for 2 billion galaxies. These can also be used for galaxy clustering despite the lack of spectroscopy. This larger number of galaxies will help improve constraints from spectroscopic galaxy clustering.

Motivations :

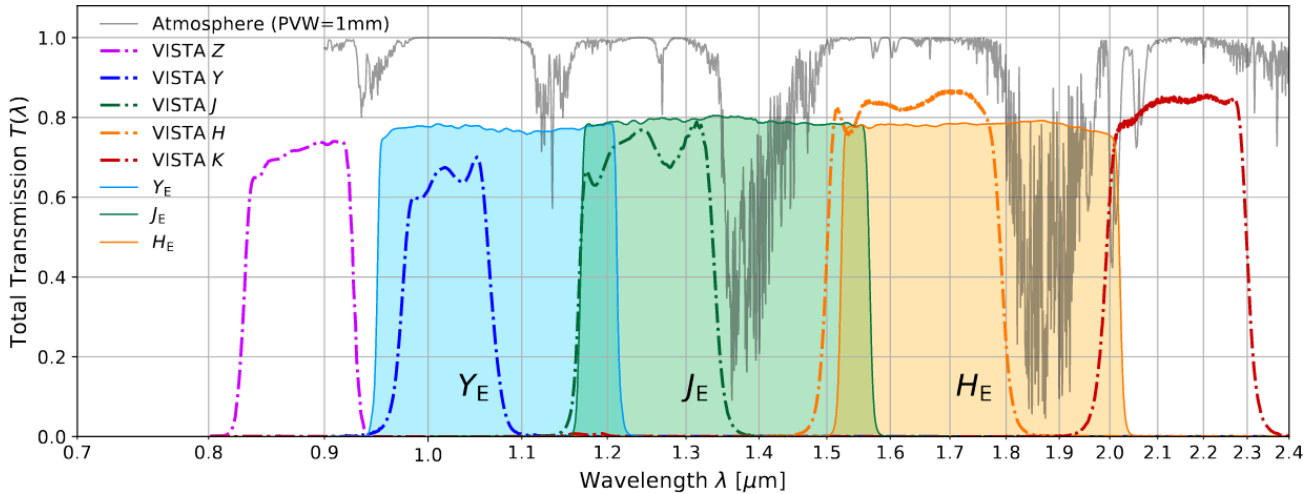
In my work, I extract the BAO signal from the 2-point correlation function.
The BAO scale studied at different times (redshifts) gives information on the expansion of the Universe, hence the choice of a tomographic analysis



https://astro.ucla.edu/~wright/BAO-cosmology.html

Motivations :

Photometric redshifts (photo-zs) are computed from the magnitude of the different bands. This information is much poorer than spectra with emission lines which explains the lower precision of these redshifts $\sigma = 0.05(1+z)$.



Euclid bands : Y, J, H + VIS (550-900 nm)
(Euclid preparation. XVIII. The NISP photometric system)

Motivations :

Existing methods to obtain photo-zs :
1) Template-fitting codes : BCNz, BPZ, CPz, EAzY, HyperZ, LePHARE, Phosphoros, Photo-z-SQL, ZEBRA, ZPEG

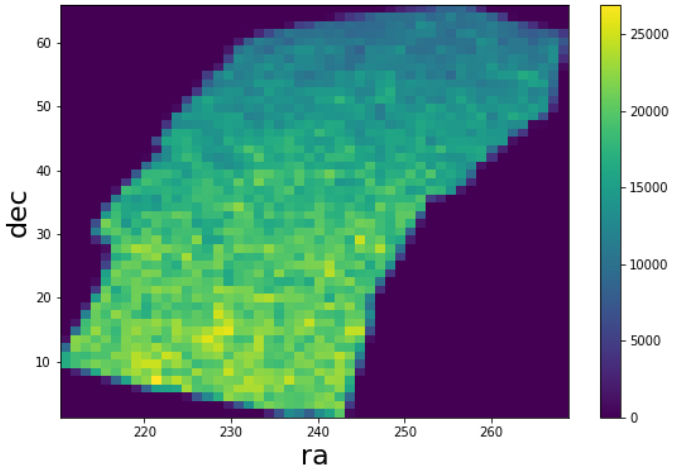
2) Machine learning : decision trees, random forests, kNN,...
Adaboost, ANNz2, BDT, CosmoPhotoz, CPz , DNF, DT, frankenz, GBRT, GPz, kNN, LR, METAPHOR, NNPZ, RF, SOMz, SPIDERz, TPZ

3) Deep Learning : ANNz2 , CuBANz, DCMDN, Pasquet et al. 2018, Henghes et al. 2021B
→ most recent approach and the least explored

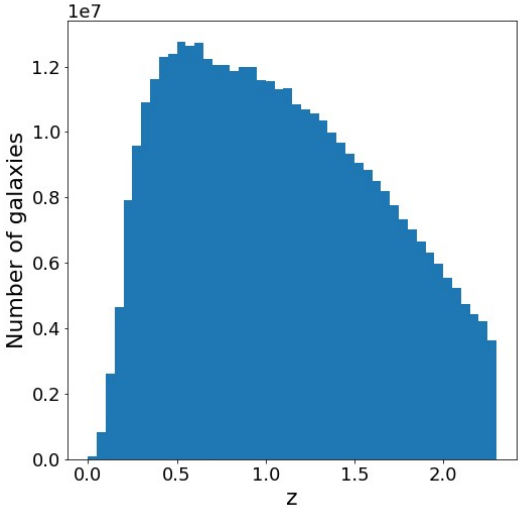
→Proposition : use galaxy images to derive photo-zs instead of the magnitudes alone.

Flagship :

Flagship 1.10.26 : full octant of 1684 square degrees, 405 912 848 galaxies with photo-z.



Footprint of Flagship



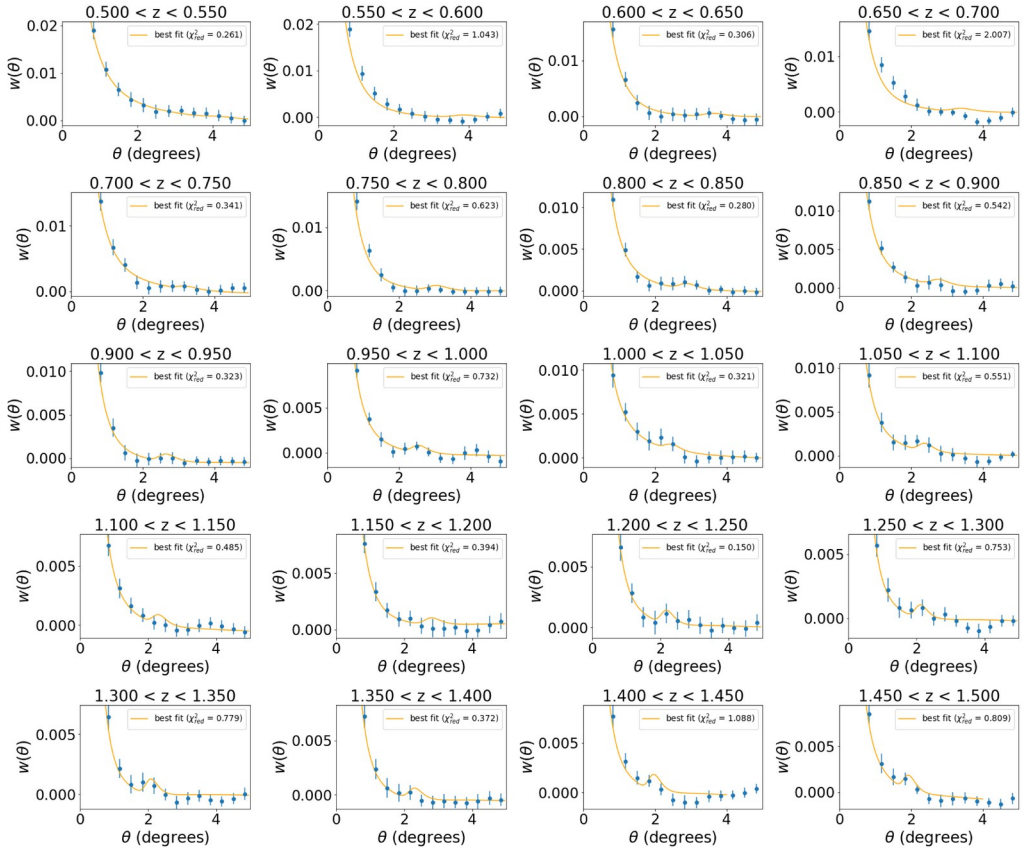
Redshift distribution

→ Angular 2-point correlation function measured with the Landy-Szalay estimator :

$$w(\theta) = \frac{DD - 2\,DR + RR}{RR}$$

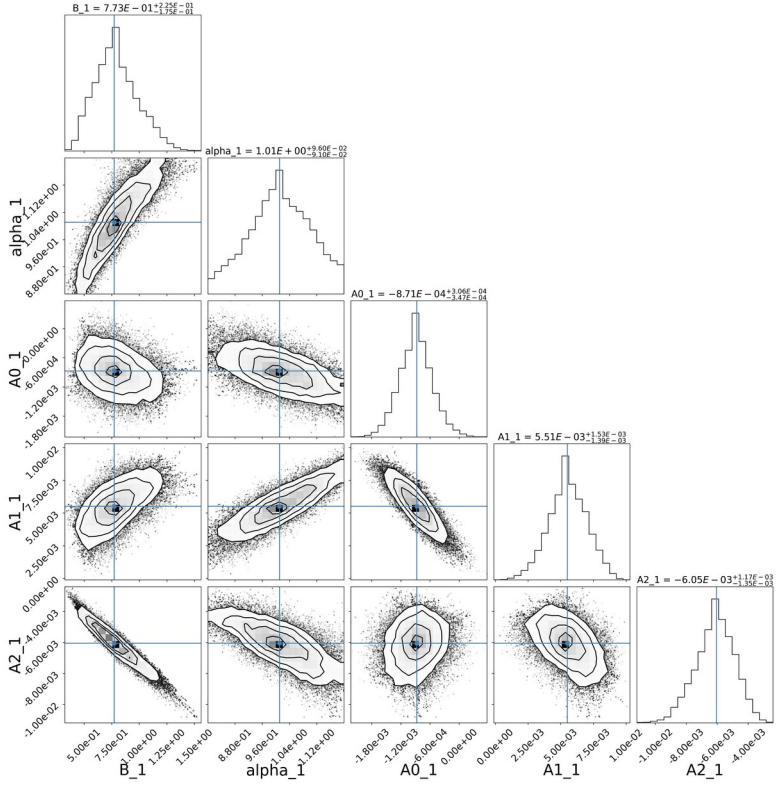
Galaxy clustering :

Fit of the 2pcf with a template : $B \times w_{camb}(\alpha\theta) + A_0 + A_1\,\theta^{-1} + A_2\,\theta^{-2}$



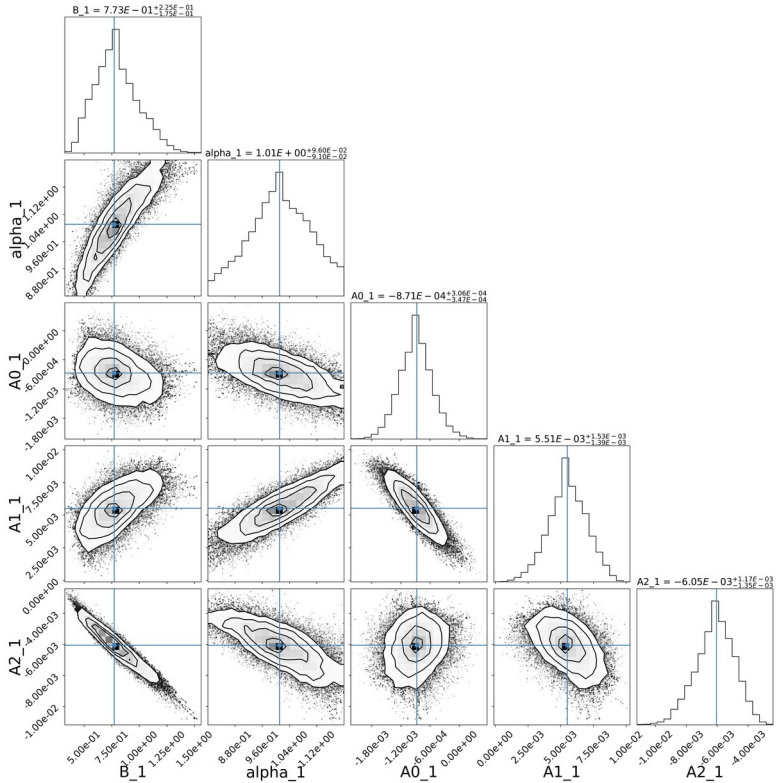
Galaxy clustering :

MCMC with this template :

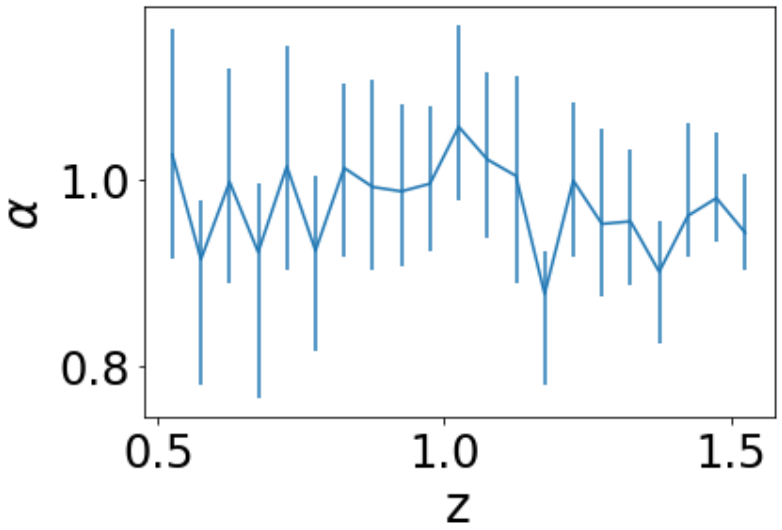


Galaxy clustering :

MCMC with this template :



Repeated in every bin to get α(z) with its error :



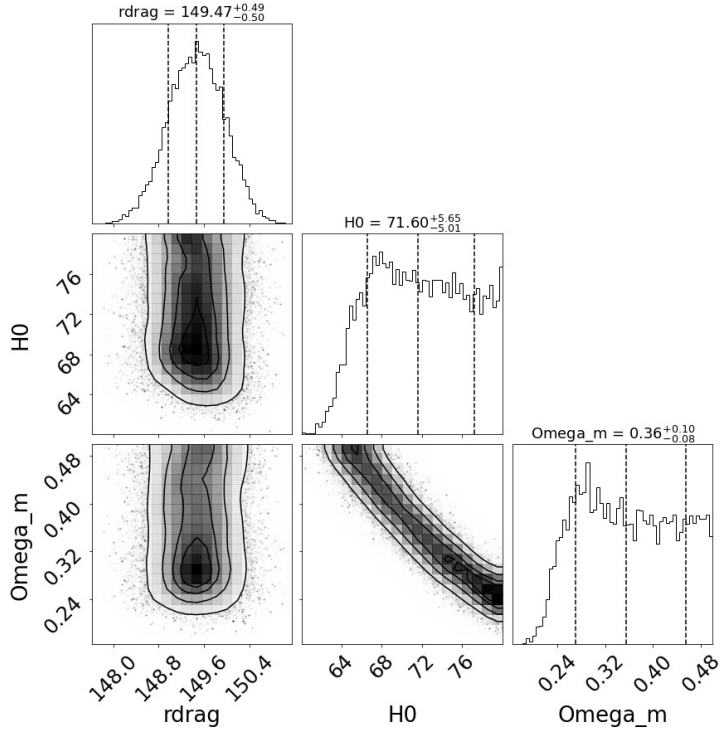
Galaxy clustering :

$$\alpha(z)=\frac{D_V(z)/r_{drag}}{D_V^{fid}(z)/r_{drag}^{fid}}$$ + prior on r_{drag} = 149.47 ± 0.48 Mpc

MCMC varying H_0 and Ω_m :

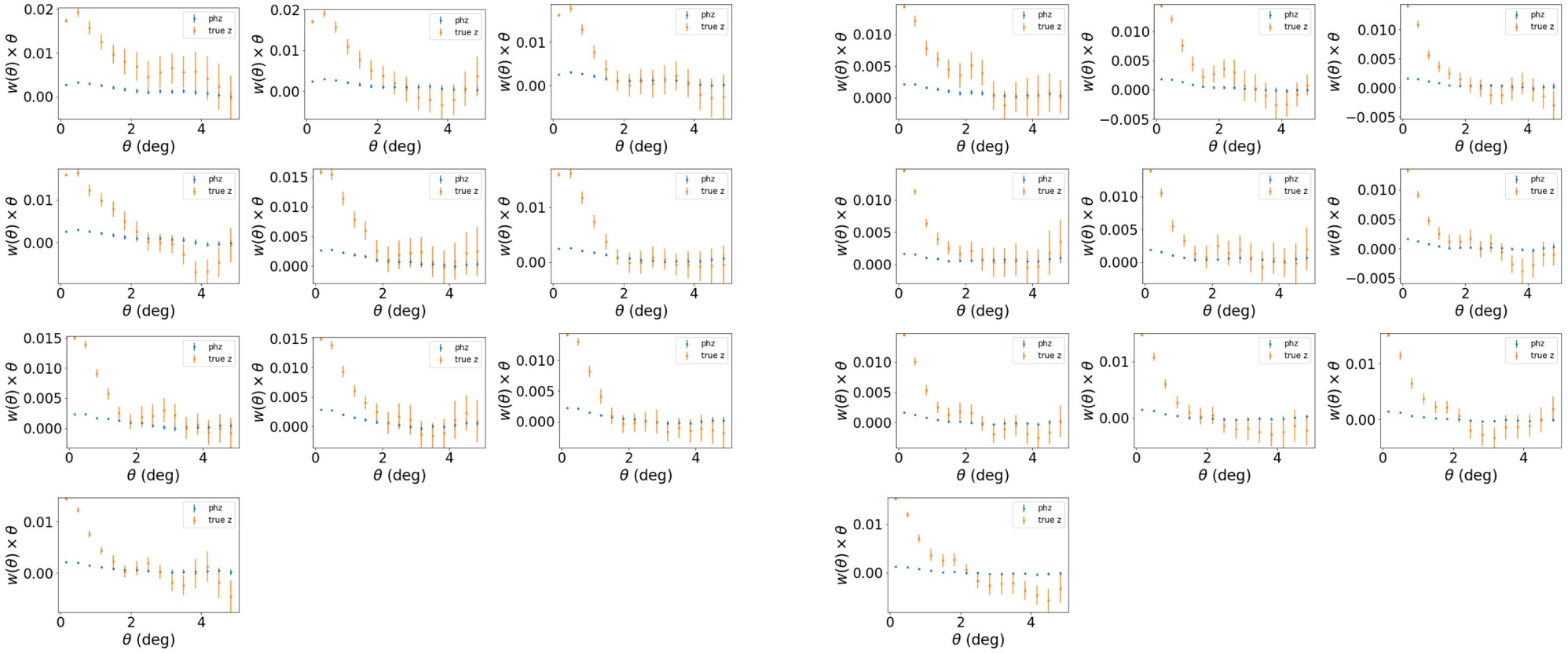
$$H_0=71.60^{+5.65}_{-5.01}$$
$$\Omega_m=0.36^{+0.10}_{-0.08}$$

→ preliminary constraints
The fitting of the 2pcf conditions α and will be improved.



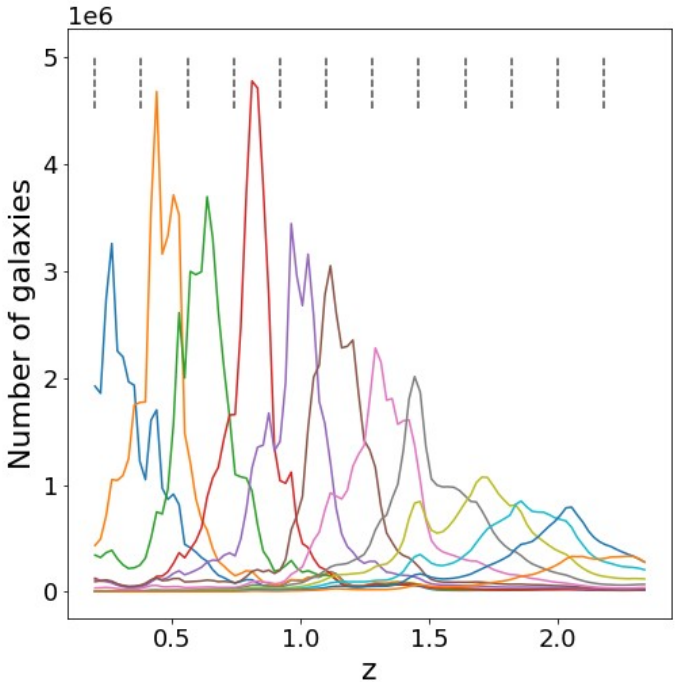
Galaxy clustering :

Up until now, results obtained when binning with true z. Comparison with a photo-z binning :



Galaxy clustering :

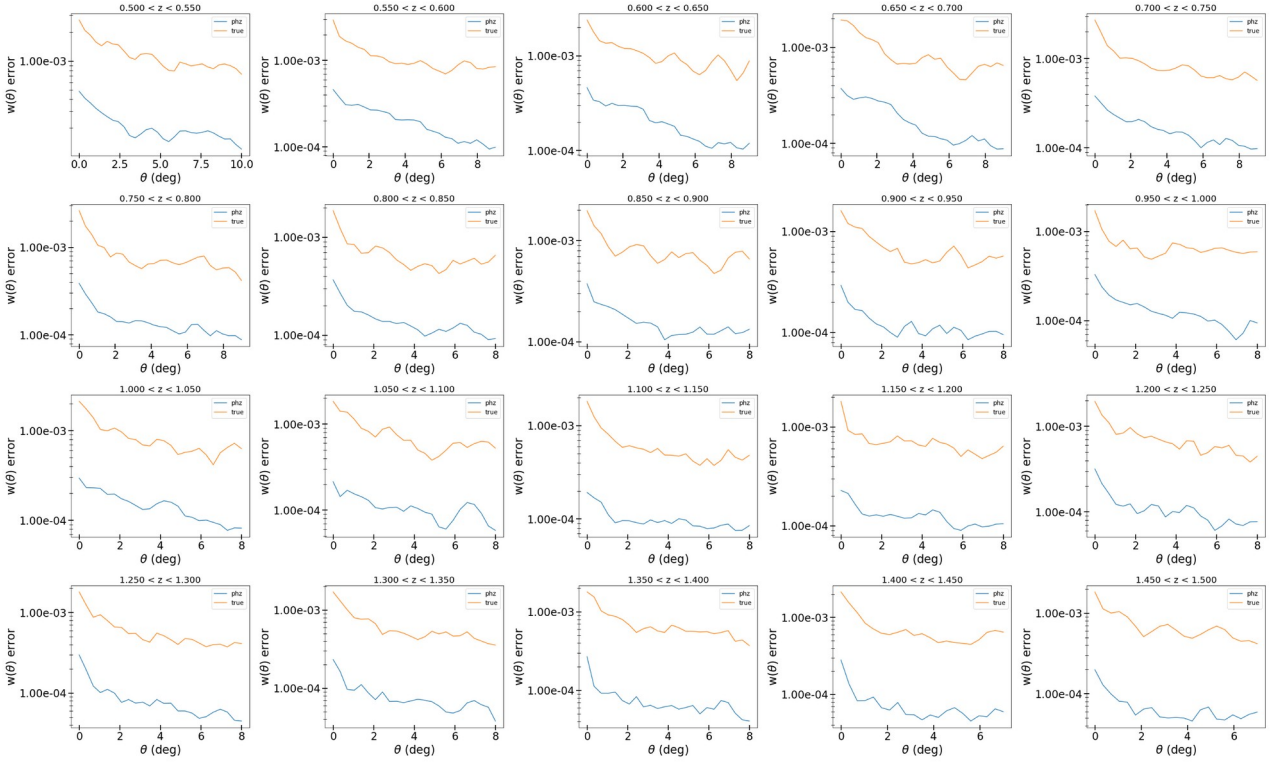
Spread of the photo-z distribution with respect to the true-z bins (dashed lines) :



→ explains the smoothing of the 2pcf variations.

Galaxy clustering :

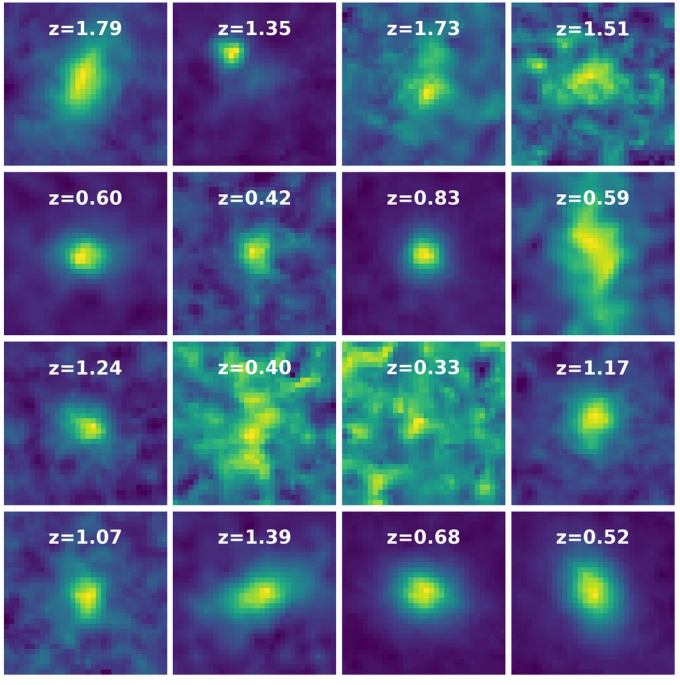
Comparison of the 2pcf errors with true-zs and photo-zs :



True z errors are 6.5 times larger (average on θ and z bins) with an increase at high z up to a factor 9-10 in the last three z bins. No explanation yet.

Photometric redshifts :

Data used : Euclid SC8
337420 galaxy images + fluxes in J,H,Y,VIS bands
Redshift range : 0 < z < 2.3

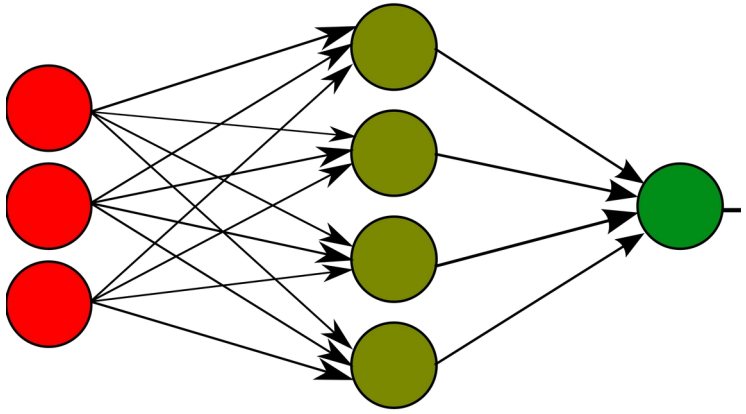


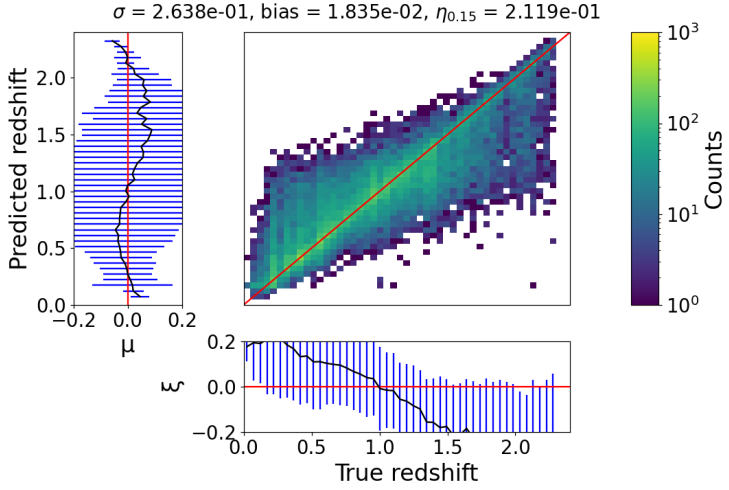
H band

Photometric redshifts :

Neural network used when using fluxes only : multi-layer perceptron (MLP)
Trained 200 epochs in ~ 2h20

Inputs :
flux in J,H,Y,VIS bands
Labels : z_{true}
Output : z_{phot}
Loss : MSE





Statistics :

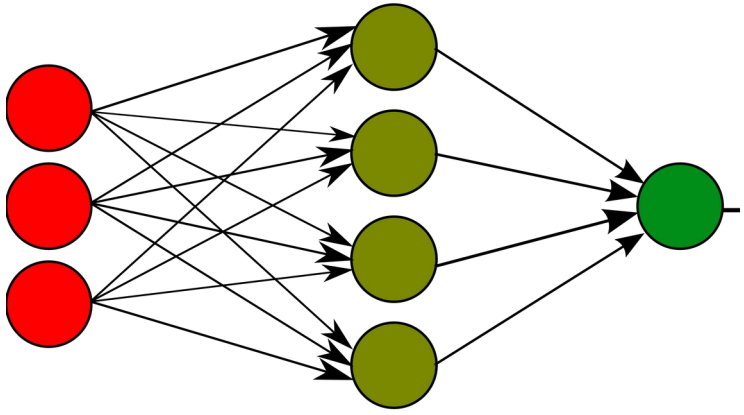
$$\text{bias} = \frac{z_{spec} - z_{phot}}{1 + z_{spec}} \qquad \text{outlier fraction } \eta_{0.15} = \frac{N(\text{bias} > 0.15)}{N_{total}}$$

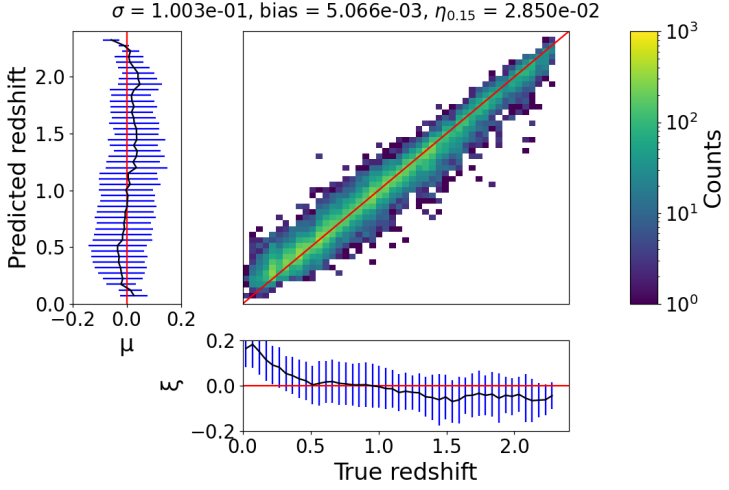
$$\sigma = \text{std}\left(z_{spec} - z_{phot}\right)$$

Photometric redshifts :

Neural network used when using images only : convolutional neural network (CNN)

Inputs :
images in J,H,Y,VIS
Labels : z_{true}
Output : z_{phot}
Loss : MSE



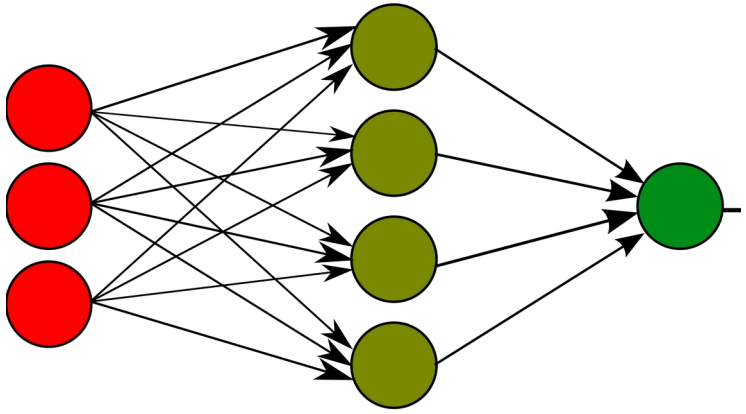


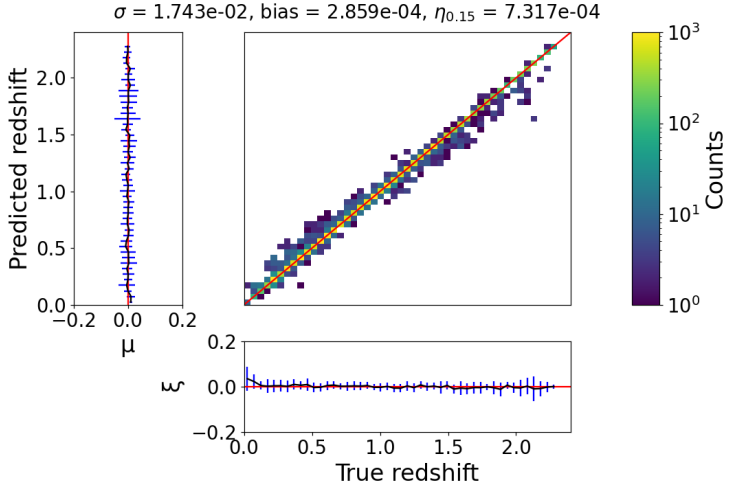
Improvement with respect to flux alone :
σ divided by 2.6 ; bias divided by 3.6 ; $\eta_{0.15}$ divided by 7

Photometric redshifts :

Neural network used when both images and fluxes : CNN + MLP

Inputs :
Images and fluxes
in J,H,Y,VIS
Labels : z_{true}
Output : z_{phot}
Loss : MSE

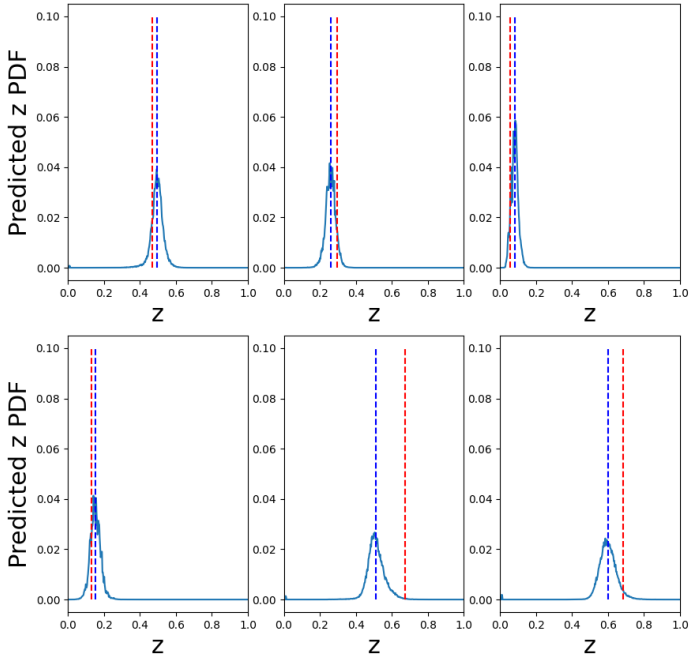




Improvement with respect to images alone :
σ divided by 6 ; bias divided by 18 ; $\eta_{0.15}$ divided by 39

Photometric redshifts :

Example of other results obtained with a different dataset : SDSS DR12
Redshift range : 0 < z < 1
Classifier network to get photo-z PDFs :



→ to be applied on Euclid SC8 data

Conclusion :

- the 2pcf has been measured on Flagship but new measurements will be done to study the effect of a bias on photo-zs as part of the KP3 of the GCph WP.

- preliminary constraints on H_0 have been extracted but the fitting procedure should be improved. To do so, a different fitting formula is currently tested.

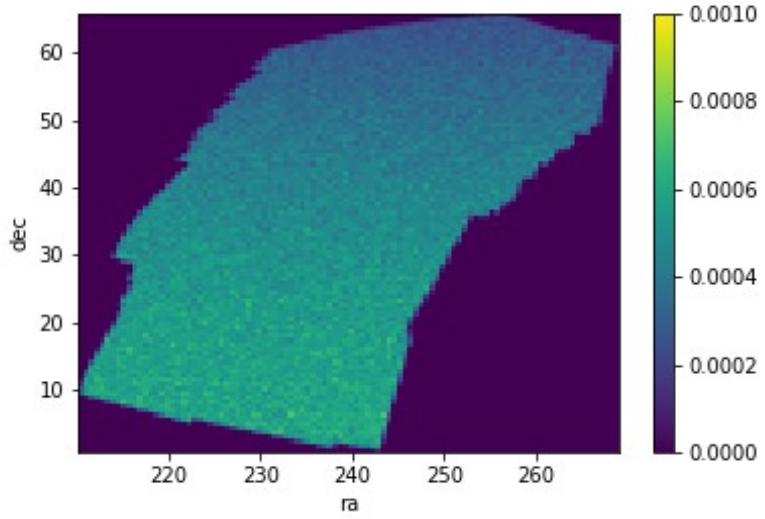
- the prediction of photo-zs with deep learning has been tested on SDSS galaxy images with good performances. Preliminary photo-zs obtained using true redshifts as labels show very interesting performances evolution when using images instead of fluxes.

- acquiring more data from SC8 will be necessary to determine an efficient training sample selection while keeping the number of galaxies large enough.

Thank you for your attention !
Questions ?

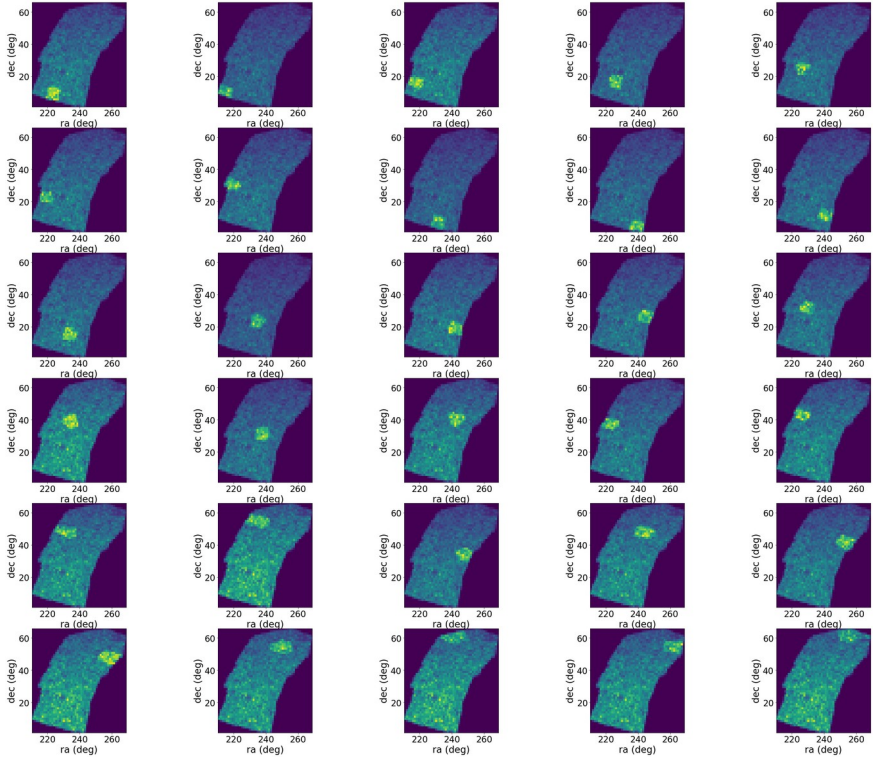
Galaxy clustering :

Randoms :
- $N_{randoms}$ / $N_{galaxies}$ = 50
- générés à l'aide d'un mask angulaire Healpix avec nside = 4096



Galaxy clustering :

Patches for jackknife samples :
- Generated by TreeCorr with a k-means clustering algorithm to have patches of similar area.



Patches superimposed to the entire area

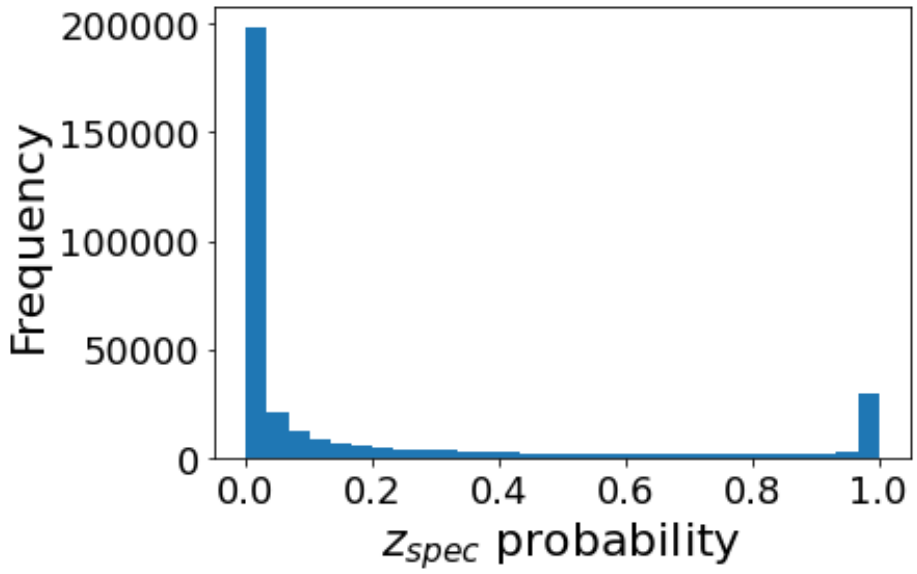
Galaxy clustering :

Empirical fitting formula :

$$A + B\theta^{\gamma} + Ce^{-(\theta-\theta_{fit})^2/(2\sigma^2_{fit})}$$

BAO peak modelized by a gaussian of mean θ_{fit} and width σ_{fit} + a power law.

SC8 spectroscopic redshifts :



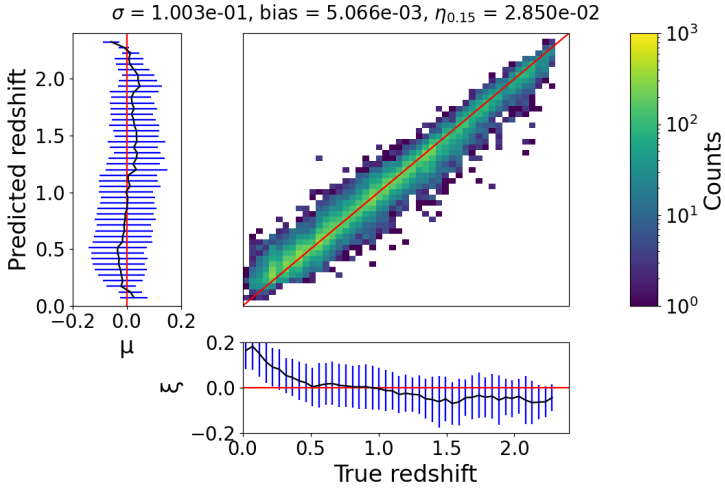
Distribution of z_{spec} probability

→ large peak close to 0.
Selecting only galaxies with z_{spec} probability > 0.9 leaves only 10 % of the data and 16 % with a threshold at 0.5.

Side plots :

ξ :
In each bin of the histogram, I compute the mean and standard deviation of the $z_{predicted,i}$ - z_{bin} for all $z_{spec,i}$ falling into that bin

μ :
In each bin of the histogram, I compute the mean and standard deviation of the $z_{spec,i}$ - z_{bin} for all $z_{predicted,i}$ falling into that bin



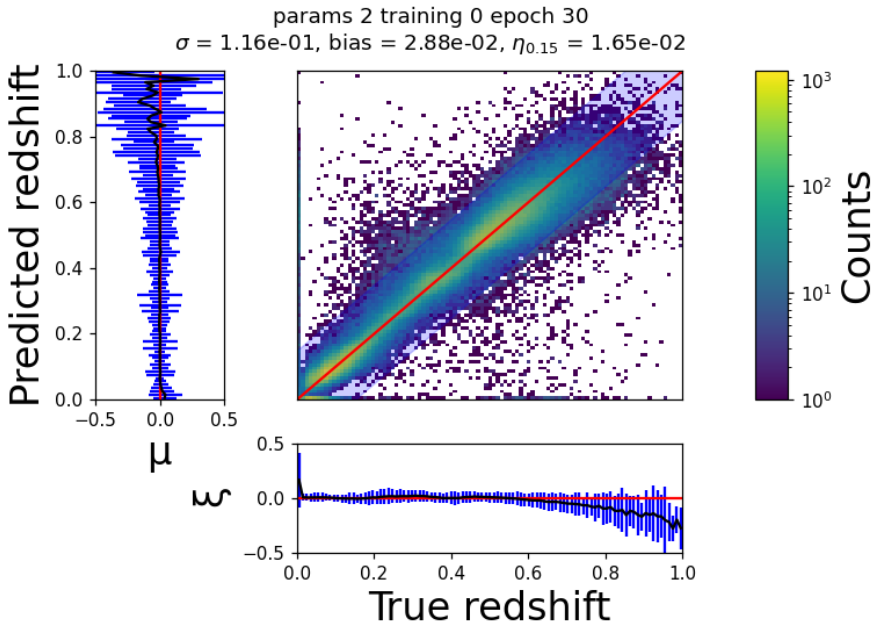
Then the mean is displayed as the black line and the standard deviation in blue.
Statistics are computed on these plots too and can be used to further compare performances :

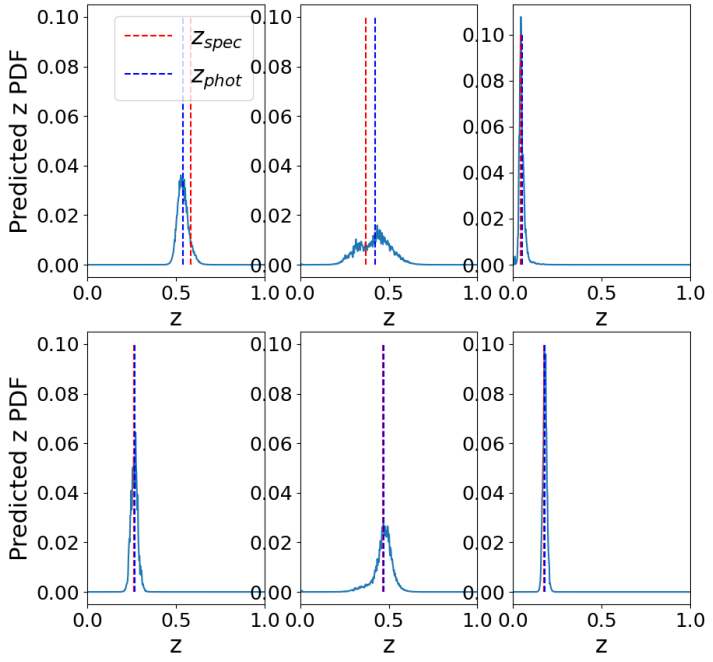
Avg % error	Min % error	Max % error	Avg % error without high z bins	Min % error without	Max % error without
12.33	0.71	69.59	24.56	6.90	69.59

Photo-zs on SDSS data :

10^6 galaxies, 0 < z < 1, u,g,r,i,z bands

Result with a ResNet34 in classification mode :

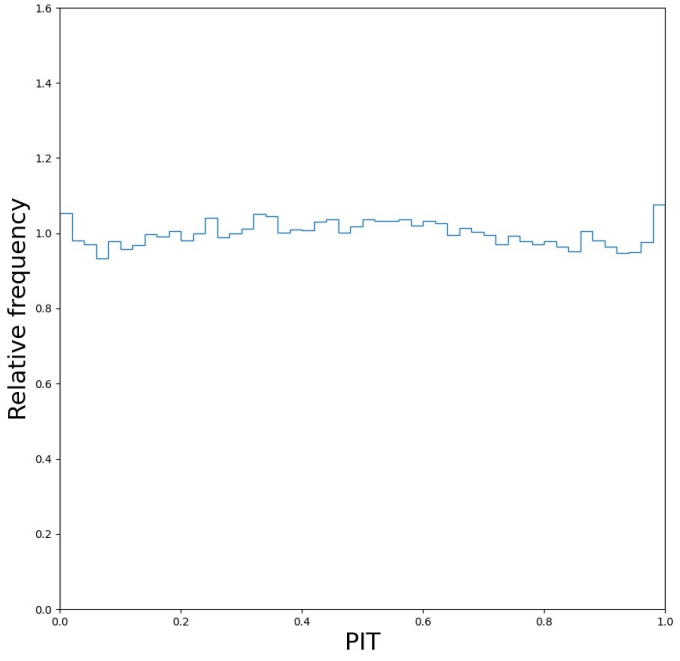




PIT distribution :

10^6 galaxies, 0 < z < 1, u,g,r,i,z bands

Result with a ResNet34 in classification mode :



Probability Integral Transform (PIT), for a galaxy i :

$$CDF_i(z_i) = \int_0^{z_i} PDF_i(z)\,dz$$

If the PDFs are too narrow then z_{true} is often under/overestimated and the PIT will be closer to 1.
If they are too wide then z_{true} will often be in the PDF, leading to intermediate PIT values.
If there is a bias then it creates a slope.
→ an ideal PIT distribution is flat.

PIT distribution :

Example of a bad PIT distribution :

Many PDFs miss z_{spec} → The PIT distribution is convex