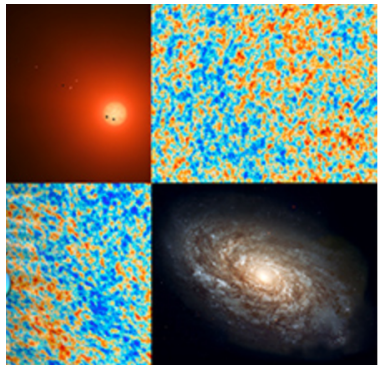
Machine Learning with the THREE HUNDRED* simulations.

Daniel de Andrés

Universidad Autónoma de Madrid

* https://the300-project.org



ML-IAP 2021

19 Oct, 2021

daniel.deandres@uam.es

Collaborators: Machine Learning group in the THREE HUNDRED Project.

- Gustavo Yepes, Marco De Petris, Weiguang Cui, Florian Ruppin, Federico De Luca, Giulia Gianfagna, Jesús Vega Ferrero (+EURANOVA* people: Mahmoud Jarraya, Gianmarco Aversano, Ichraf Lahouli, Romain Dupuis)

*https://euranova.eu/

Galaxy Clusters: The crossroads between Cosmology and Astrophysics.

Cosmology:

- Biggest virialized objects in the universe.
- Study of abundance and mass to test cosmological models
- Powerful tool to estimate cosmological parameters (Ω_m , $\boldsymbol{\sigma}_8$)

Astrophysics:

- Isolated system: giant astrophysical laboratories
- Many physical processes involving the baryons of the ICM: cooling, star formation, SN feedback, AGN etc

- Zoomed regions of 15/h Mpc radius around the **324 most massive clusters** of the 1Gpc **Multidark-Planck** simulation formed at z=0 (Mass: 3.2×10^{15} – 8×10^{14} Msun/h).
- **DATA SAMPLE:** 3 different versions of the same objects (324 regions)+ 4 void regions
  - ***GADGET-MUSIC*** (standard SPH, SN Feedback, Stellar winds)
  - ***GADGET-X*** (modern SPH, AGN feedback, Trieste Model)
  - ***GIZMO-SIMBA*** (modern SPH + AGN feedback Dave's Model)
- **128 snapshots** equally spaced in redshift stored for each simulated region, merger trees for all of them based on AHF halo finding.
- **Mock observations provided**: *X-ray (XMM , Athena) , tSZ(y-maps) , CCD (SDSS bands), lensing maps (thanks to Max and Carlo*)
- Participate in CHECK-MATE and NIKA2 LPSZ observational collaborations as simulation provider.

Ongoing ML projects in the300 collaboration

The 300th data base of simulated clusters is an excellent tool for the training of other machine learning (ML) algorithms.

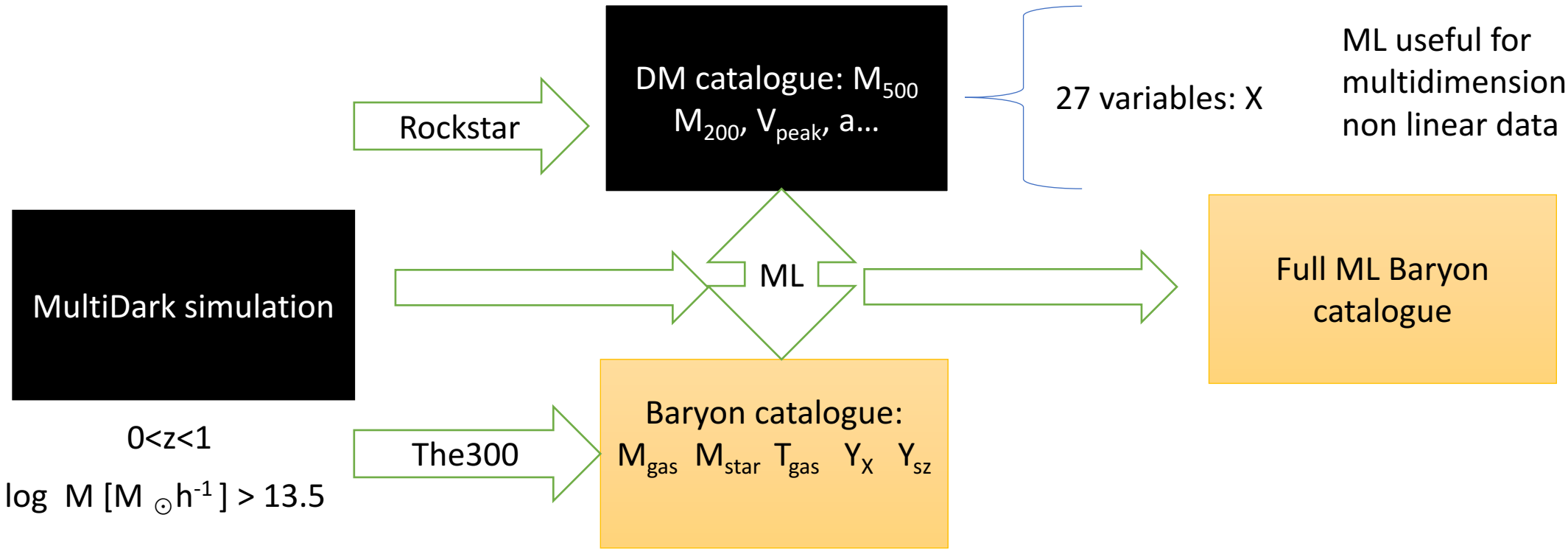
In this talk:

- **Regressions of baryon properties from dark matter halo catalogs. Application to large volume dark matter only simulations (to construct fast all sky cluster number counts in X-ray and SZ). Paper In prep.**

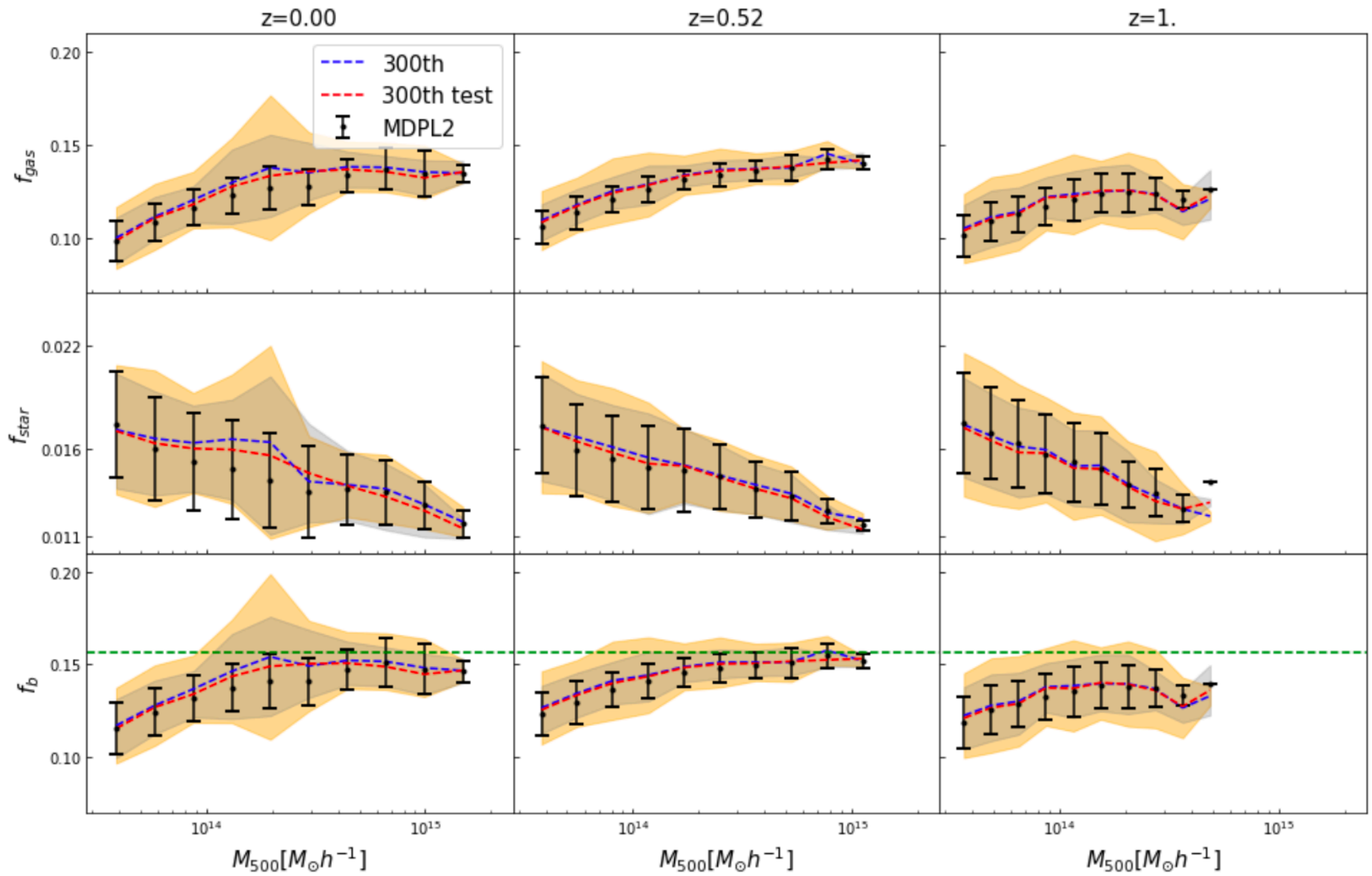
- **Deep Learning to Infer galaxy cluster masses in Planck Compton parameter maps. Paper in prep. -> IMAGING, for this session.**

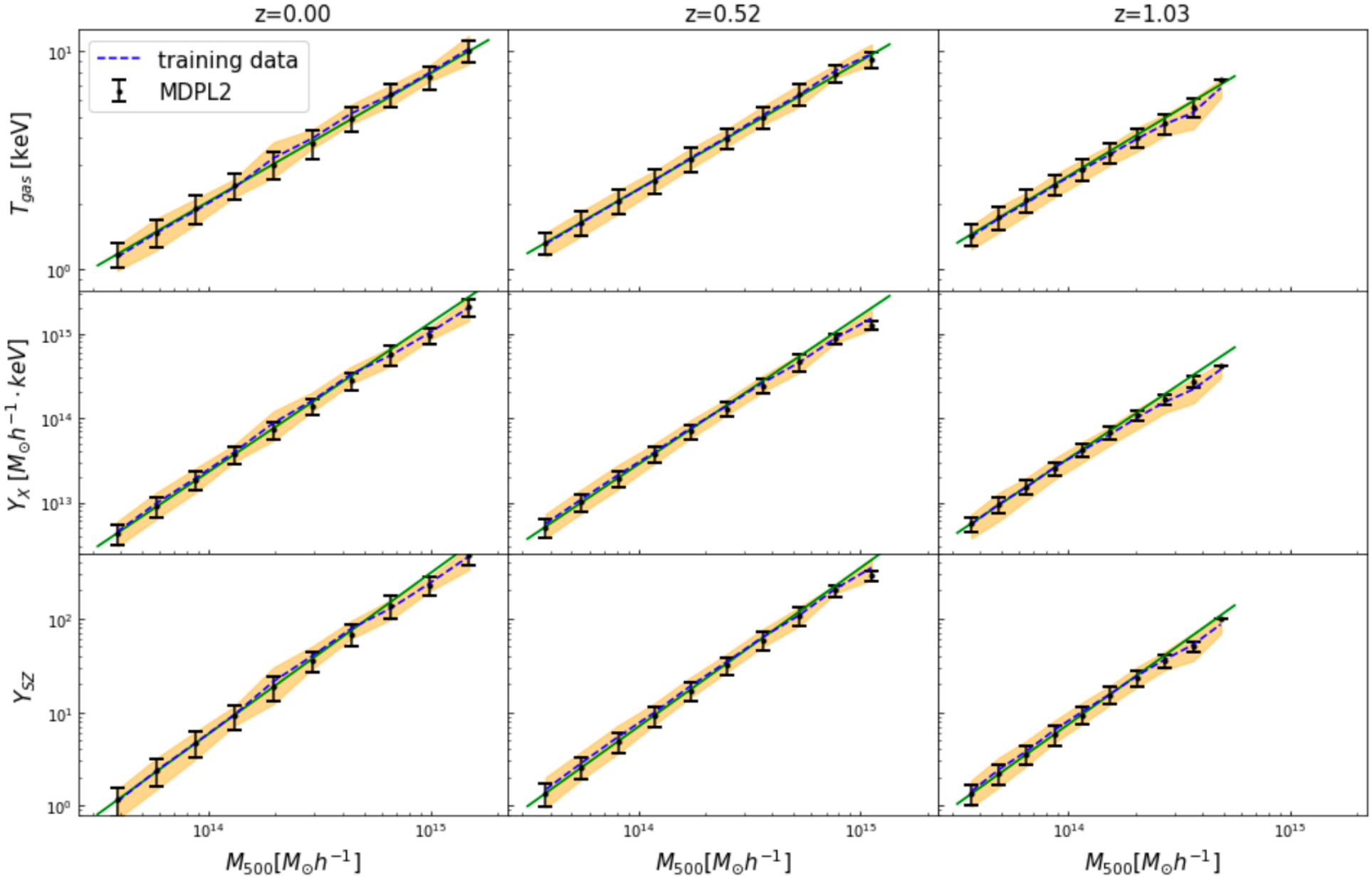
Regressions of baryon properties from dark matter halo catalogs.

Regressions of baryon properties from dark matter halo catalogs.



ML algorithms : Random Forests (RF), **Extreme gradient boosting (XGB),** feed forward Neural Networks (multilayer perceptron, MLP).

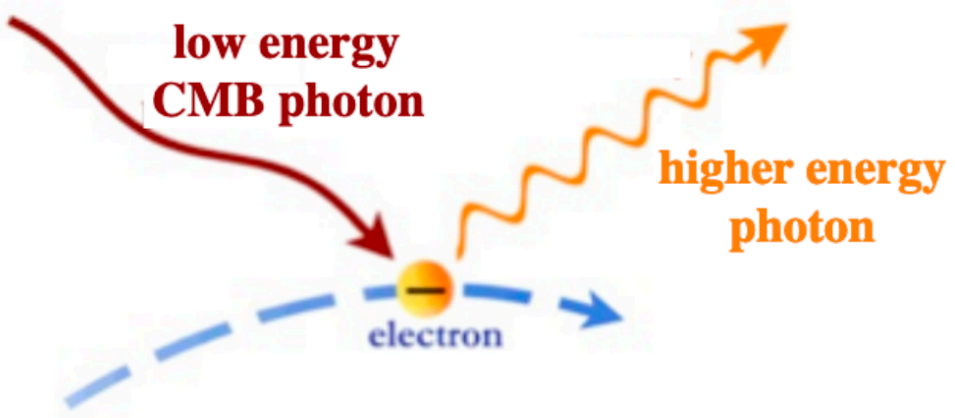

z=0.00
z=0.52
z=1.
300th
300th test
MDPL2
f_{gas}
f_{star}
f_b
0.20
0.15
0.10
0.022
0.016
0.011
10^{14}
10^{15}
$M_{500}[M_\odot h^{-1}]$

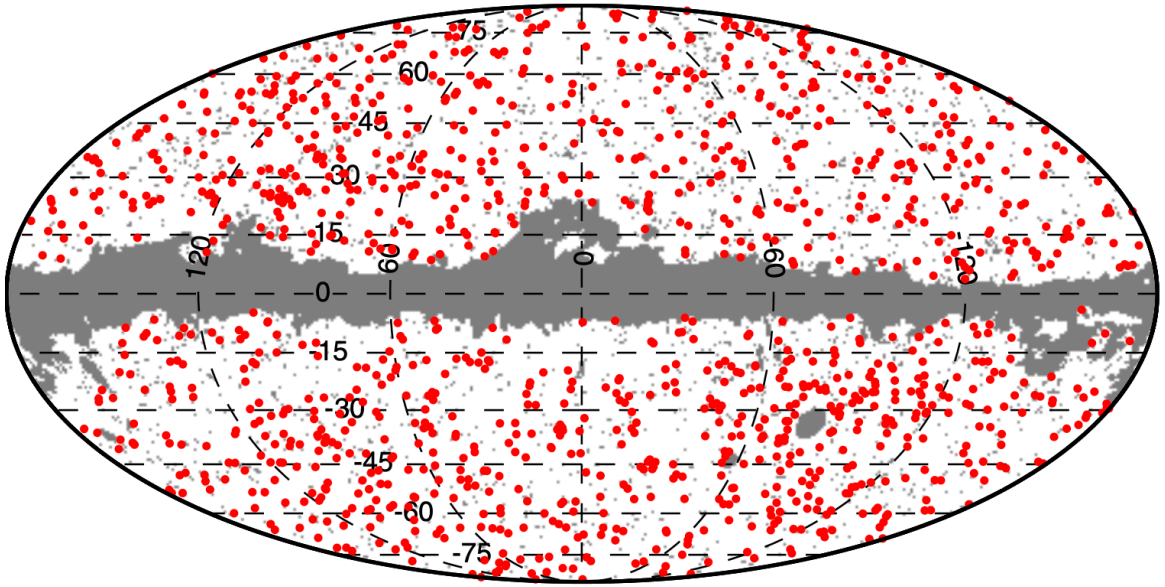

z=0.00
z=0.52
z=1.03
training data
MDPL2
T_{gas} [keV]
Y_X [$M_\odot h^{-1} \cdot keV$]
Y_{SZ}
$M_{500}[M_\odot h^{-1}]$
$M_{500}[M_\odot h^{-1}]$
$M_{500}[M_\odot h^{-1}]$

A Deep Learning Approach to Infer Galaxy Cluster Masses from Planck Compton parameter maps

- Galaxy clusters are the biggest gravitationally bound objects in the universe and they can be observed through the inverse-Compton scattering of the cosmic microwave background (CMB) photons by free electrons of the ICM, i.e. the Sunyaev-Zel'dovich (SZ) effect.

- The Planck Collaboration collected a all-sky survey of SZ galaxy cluster maps and estimated their masses through the Y_{500}-M_{500} scaling relation. However, these masses are expected to be bias low due to the fact that hydrostatic equilibrium hypothesis is assumed, reported a mean bias of **1-b=0.8**. The accurate determination of this bias is of paramount importance in cosmology.



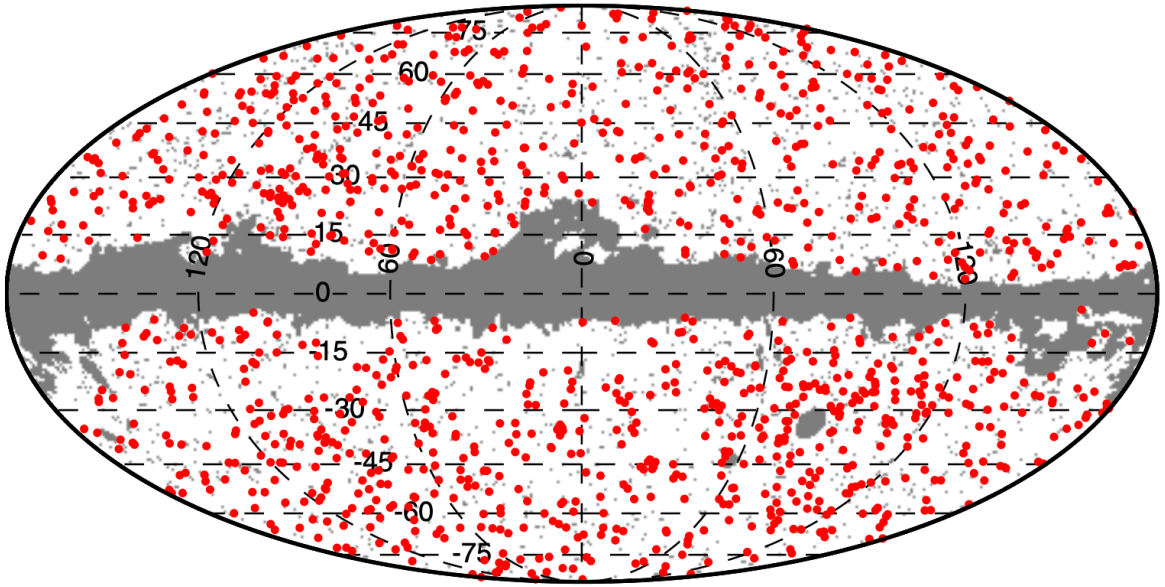


Planck collaboration

- Galaxy clusters are the biggest gravitationally bound objects in the universe and they can be observed through the inverse-Compton scattering of the cosmic microwave background (CMB) photons by free electrons of the ICM, i.e. the Sunyaev-Zel'dovich (SZ) effect.

- The Planck Collaboration collected a all-sky survey of SZ galaxy cluster maps and estimated their masses through the Y_{500}-M_{500} scaling relation. However, these masses are expected to be bias low due to the fact that hydrostatic equilibrium hypothesis is assumed, reported a mean bias of **1-b=0.8**. The accurate determination of this bias is of paramount importance in cosmology.

$$E(z)^{-2/3} d_A(z)^2 Y_{500} = B \left[\frac{M_{500}^{SZ}}{3 \times 10^{14} h_{70}^{-1} \, M_\odot} \right]^{\alpha}$$



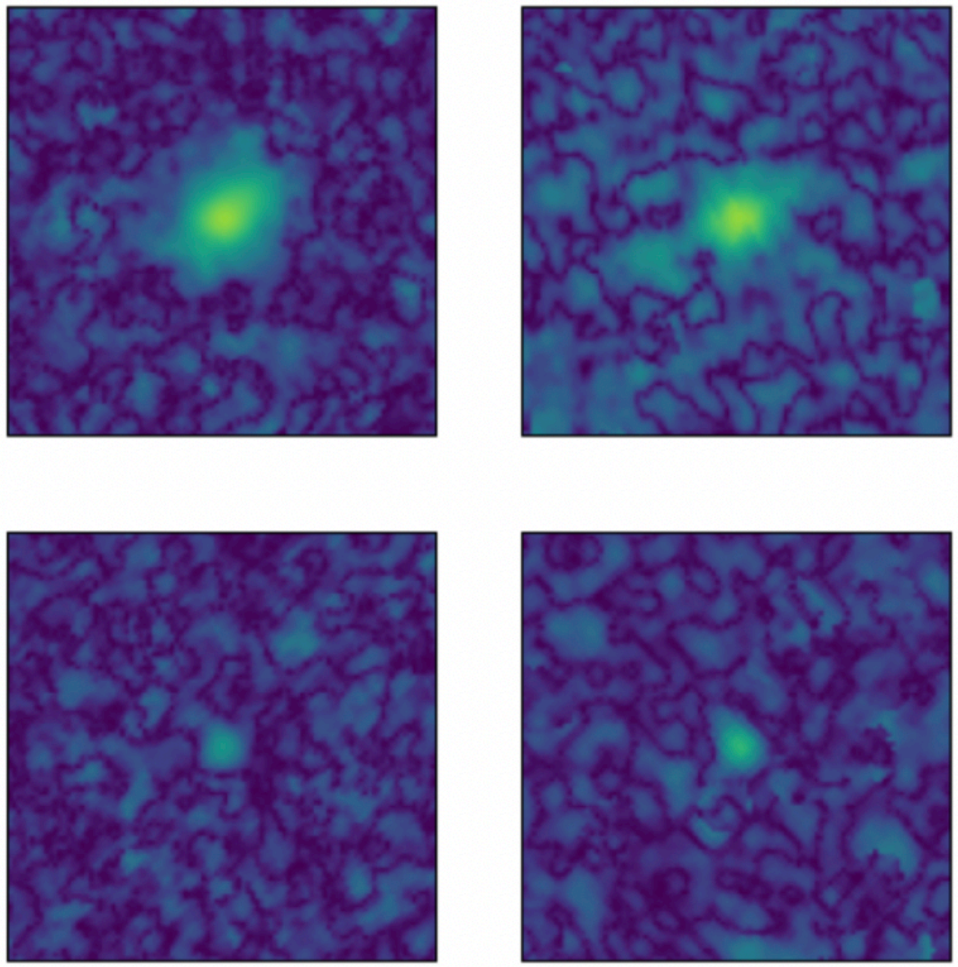
Planck collaboration

- We aim to address this issue by training a ***Convolutional Neural Network*** (CNN) on a large catalog of almost 200,000 simulated Planck-like SZ maps (with the same angular resolution and noise levels) from THE THREE HUNDRED simulations of galaxy clusters .

- This approach is based on finding a mapping between simulated SZ maps and the 3D dynamical mass M_{500} **without assuming any a priori symmetry or physical assumption (beside the simulation physical models)**.

Simulation: training

Observation: prediction



DATA SET

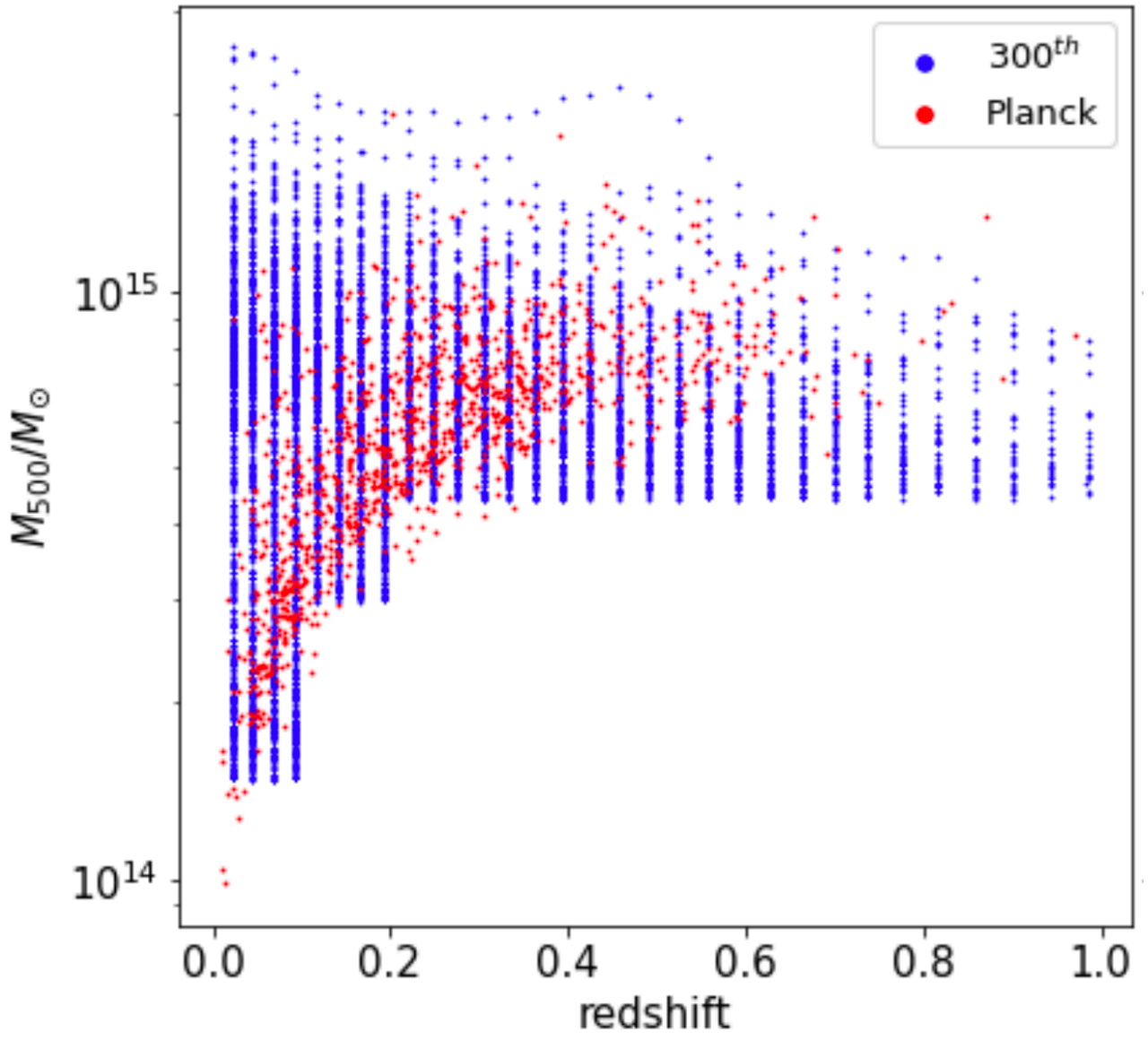
Training data set

We extract 7106 clusters from the 324 re-simulated regions. Furthermore, we take 27 different projections of every single cluster amounting to a total of 191,862 simulated maps.

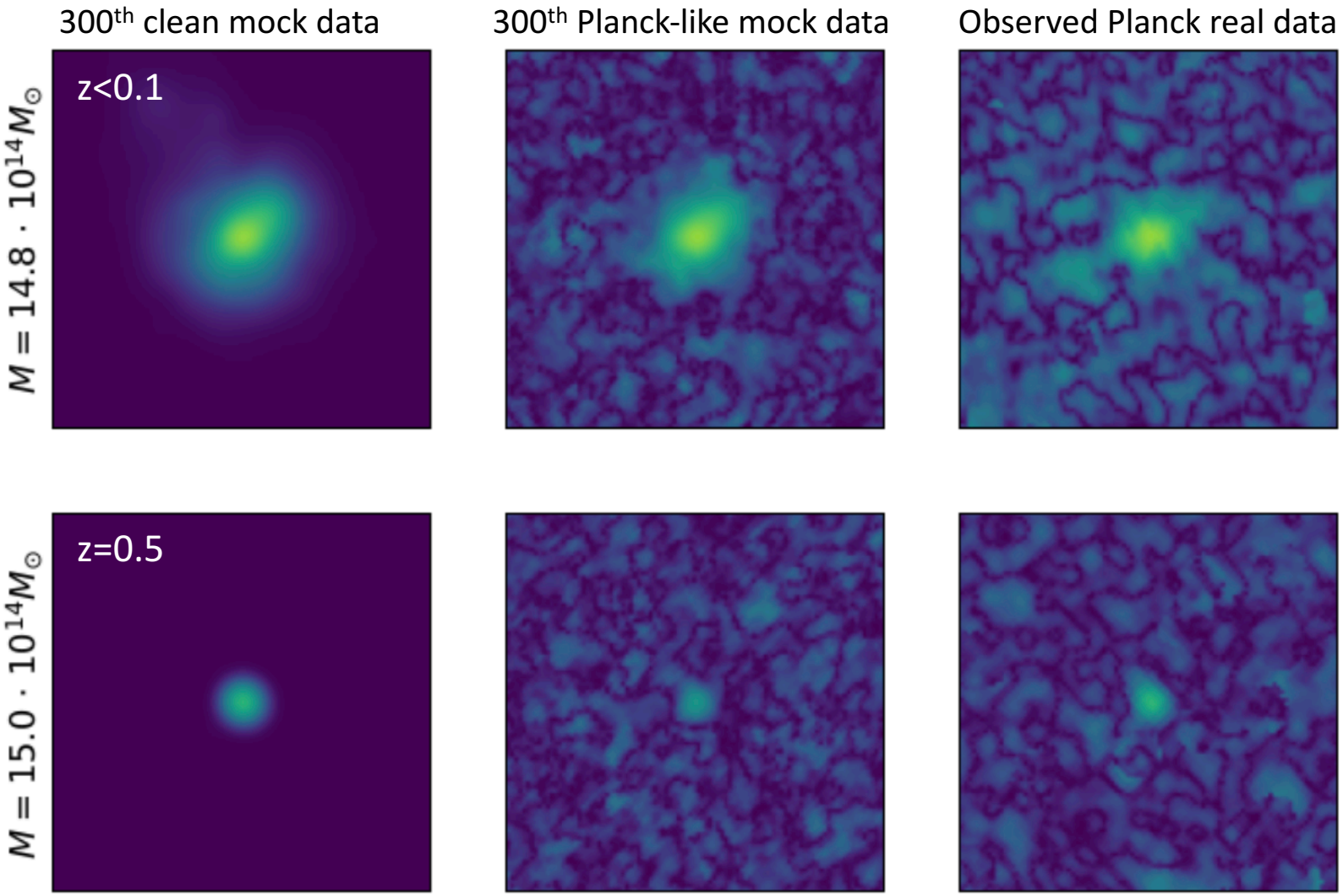
Particularly, we use this selection of clusters for covering the Planck PLSZ2 sample in redshift and mass.

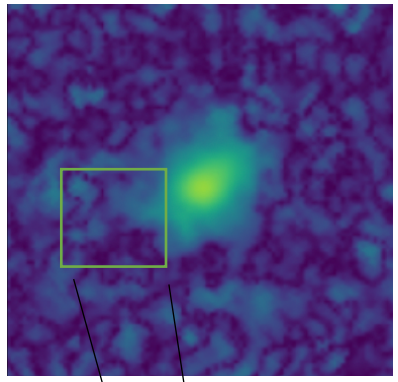
Furthermore , to improve the performance, we train four CNN models for different redshift ranges:

- 0<z<0.1
- 0.1<z<0.2
- 0.2<z<0.4
- 0.4<z<1

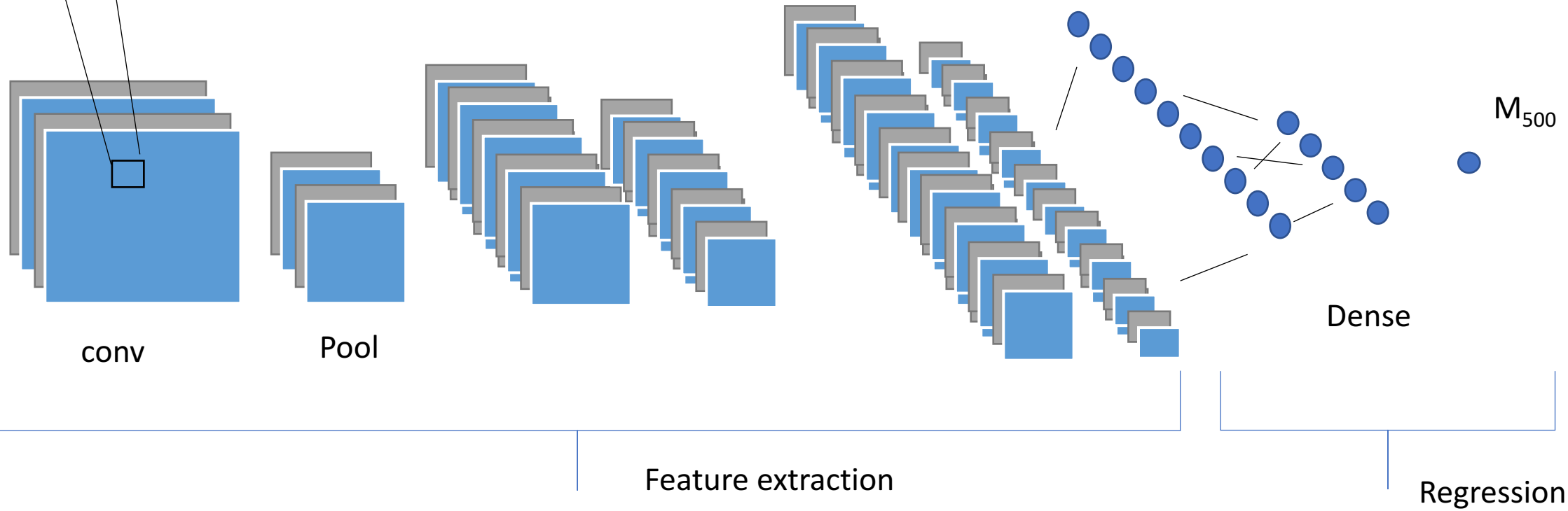


- **the300 clean mock data:** we simulated the SZ signal with the same angular resolution as Planck data 1.7 arcmin.
- **the300 Planck-like mock data:** We add instrumental noise with the same power spectrum as the Planck sky.
- **Observed Planck real data:** SZ clusters measured by Planck .

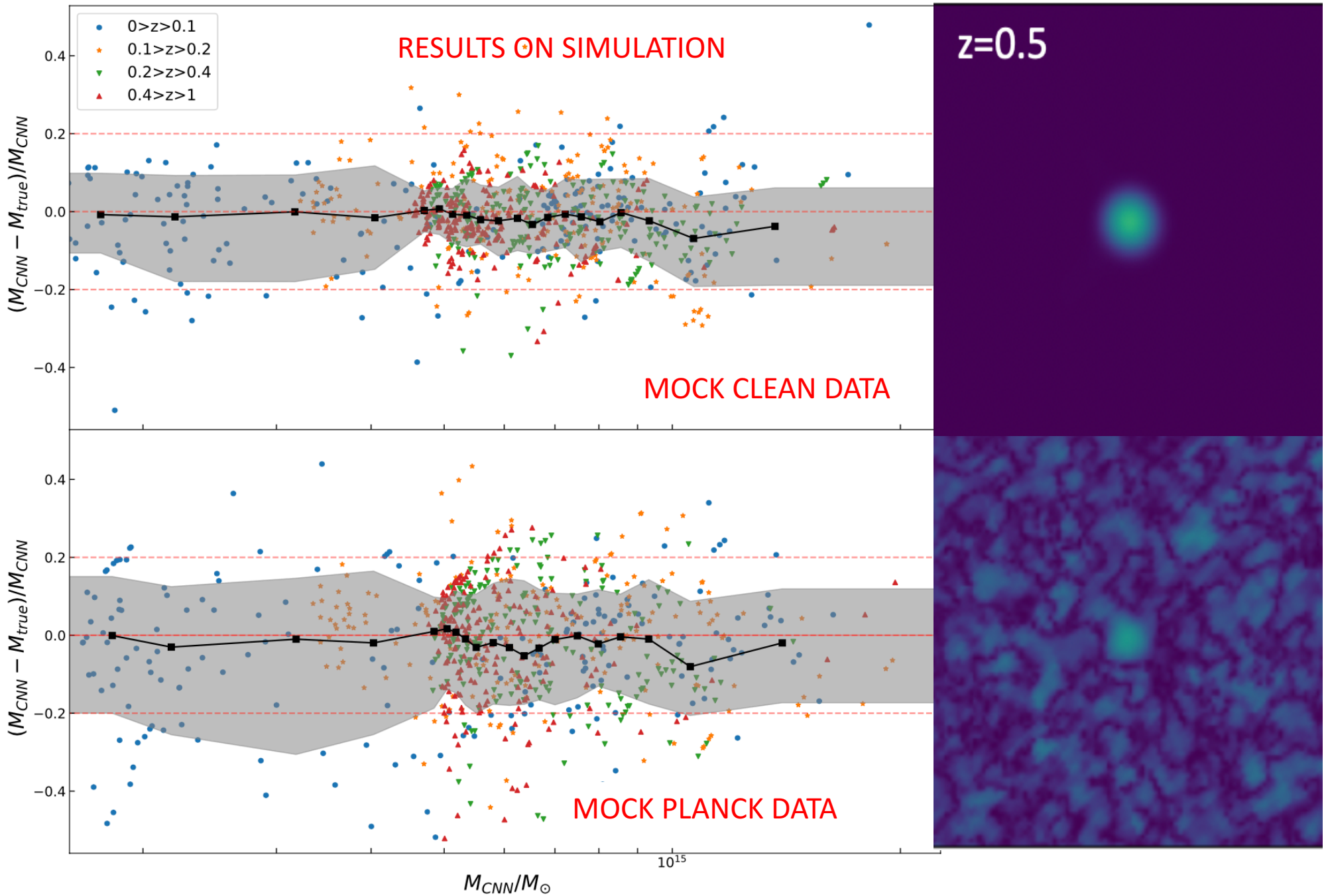


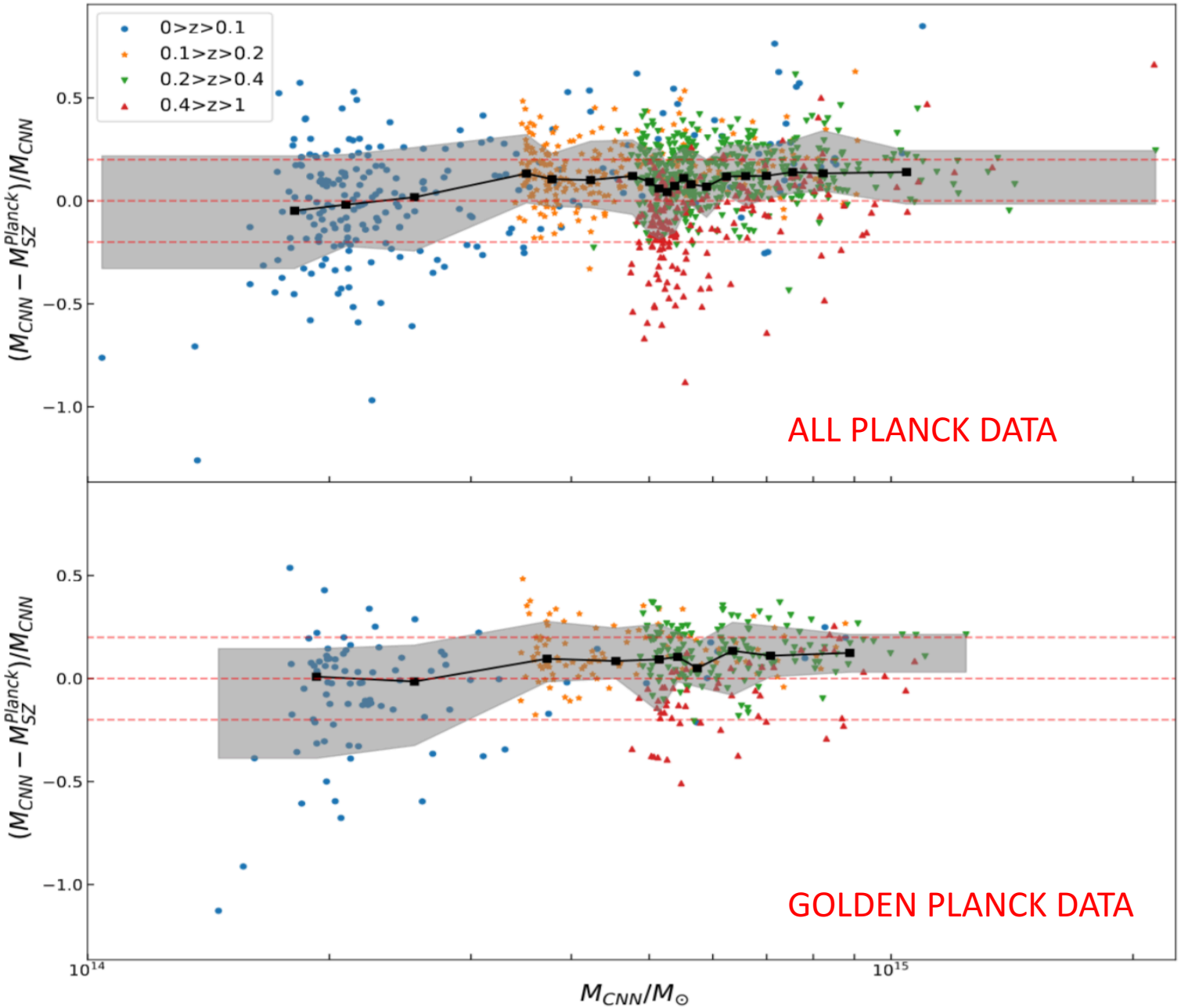


We used a version of the VGG network (Simonyan & Zisserman 2014) that has been successfully used in simulated galaxy clusters by Ntampaka et al. (2019) and Yan et al. (2020) .



RESULTS


0>z>0.1
0.1>z>0.2
0.2>z>0.4
0.4>z>1
RESULTS ON SIMULATION
MOCK CLEAN DATA
MOCK PLANCK DATA
z=0.5
$(M_{CNN} - M_{true})/M_{CNN}$
$M_{CNN}/M_{\odot}$
10^{15}
0.4
0.2
0.0
−0.2
−0.4



We found that the bias factors b

$b = (M_{CNN} - M_{planck})/M_{CNN}$

depends on the mass M_{CNN}.

GOLDEN PLANCK DATA: Free from contamination from astrophysical signal induced from the galactic plane and point sources. Reduce scatter, but compatible result.

This dependence is explained by taking into consideration that simulations don't generally agree with Planck Y_{500}-M_{500} scaling relation.

$$E(z)^{-2/3} d_A(z)^2 Y_{500} = B \left[\frac{M_{500}^{SZ}}{3 \times 10^{14} h_{70}^{-1} \, M_{\odot}} \right]^{\alpha}$$

Conclusions

- We have managed to give an estimation to the total cluster mass **without any prior assumption on the cluster dynamical state**. https://github.com/ddeandres/DeepPlanck

- We have found that $\mathbf{Y_{500}}$ **might be overestimated for low mass clusters** at low redshift in the Planck catalog due to the scaling relation used in Planck.

- However, the results on our simulated data and the results on real data are slightly different but nevertheless, **statistically compatible and they both show an overall similar trend**. The possible difference can be due to the physics modelled by the simulations.

- We tested our algorithm on the **GIZMO simulation** (different subgrid feedback modelling) of the same clusters finding ***no significant difference*** with the results presented here.

- Nevertheless, as in other areas of image processing, the success of these techniques depends on the quality and accuracy of the training set. To this end**, hydrodynamical numerical simulations are an indispensable tool** to provide the proper mock observations on which we can train CNN architectures.

THANKS, QUESTIONS?