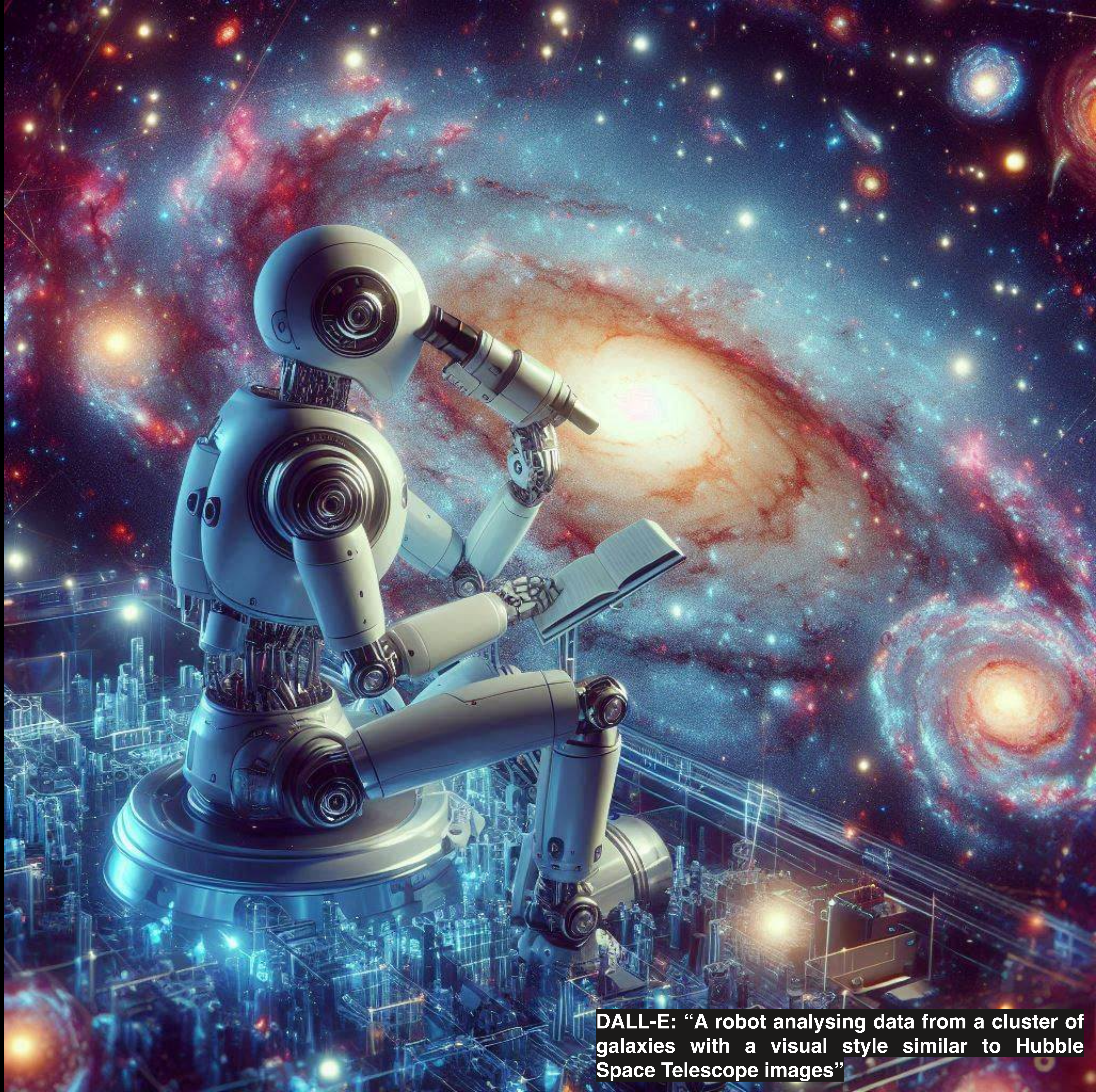
Artificial Intelligence Applications in The Three Hundred Simulation Project

Daniel de Andrés Hernández

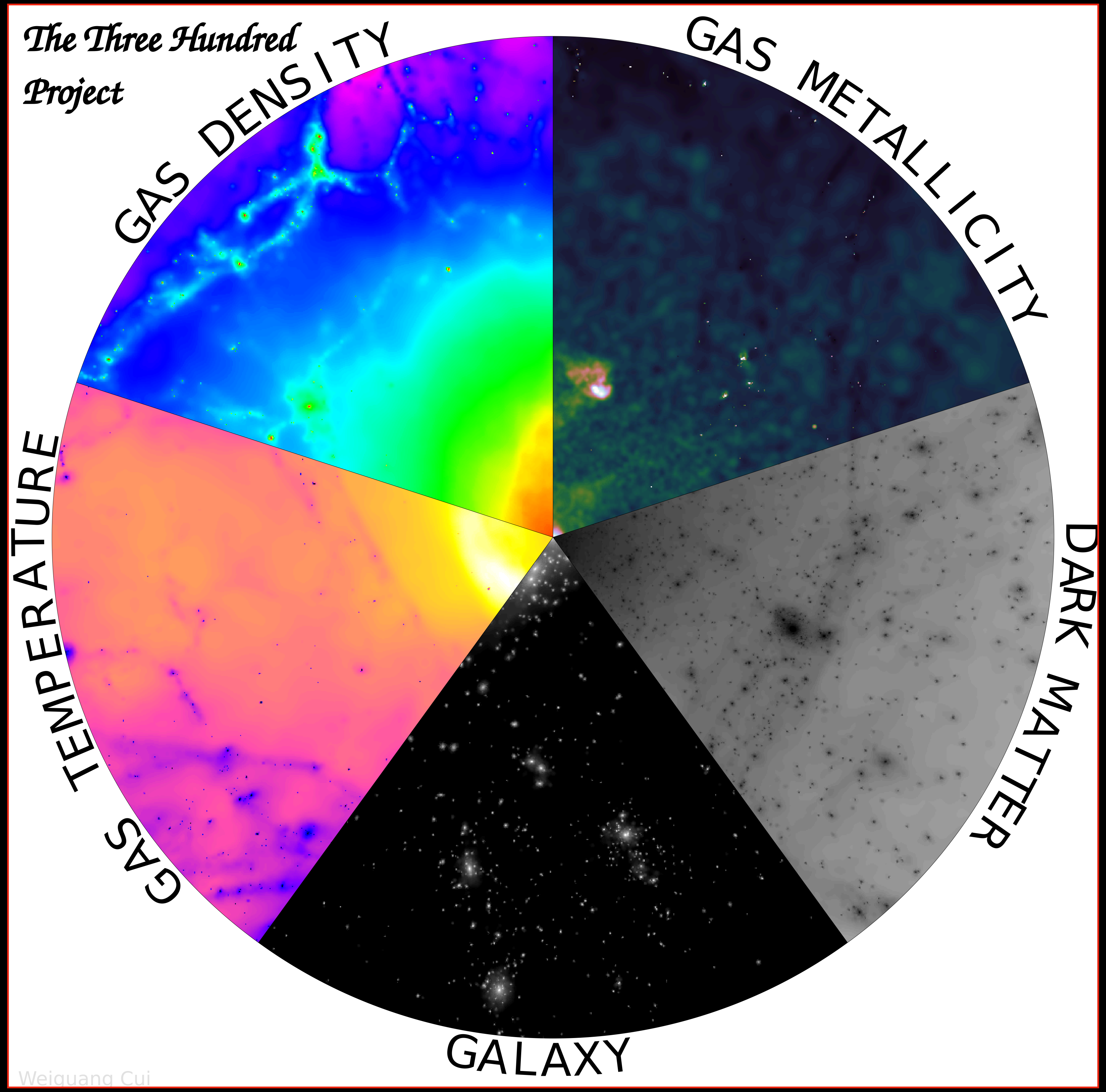
Universidad Autónoma de Madrid

AI GOES MAD²



DALL-E: "A robot analysing data from a cluster of galaxies with a visual style similar to Hubble Space Telescope images"

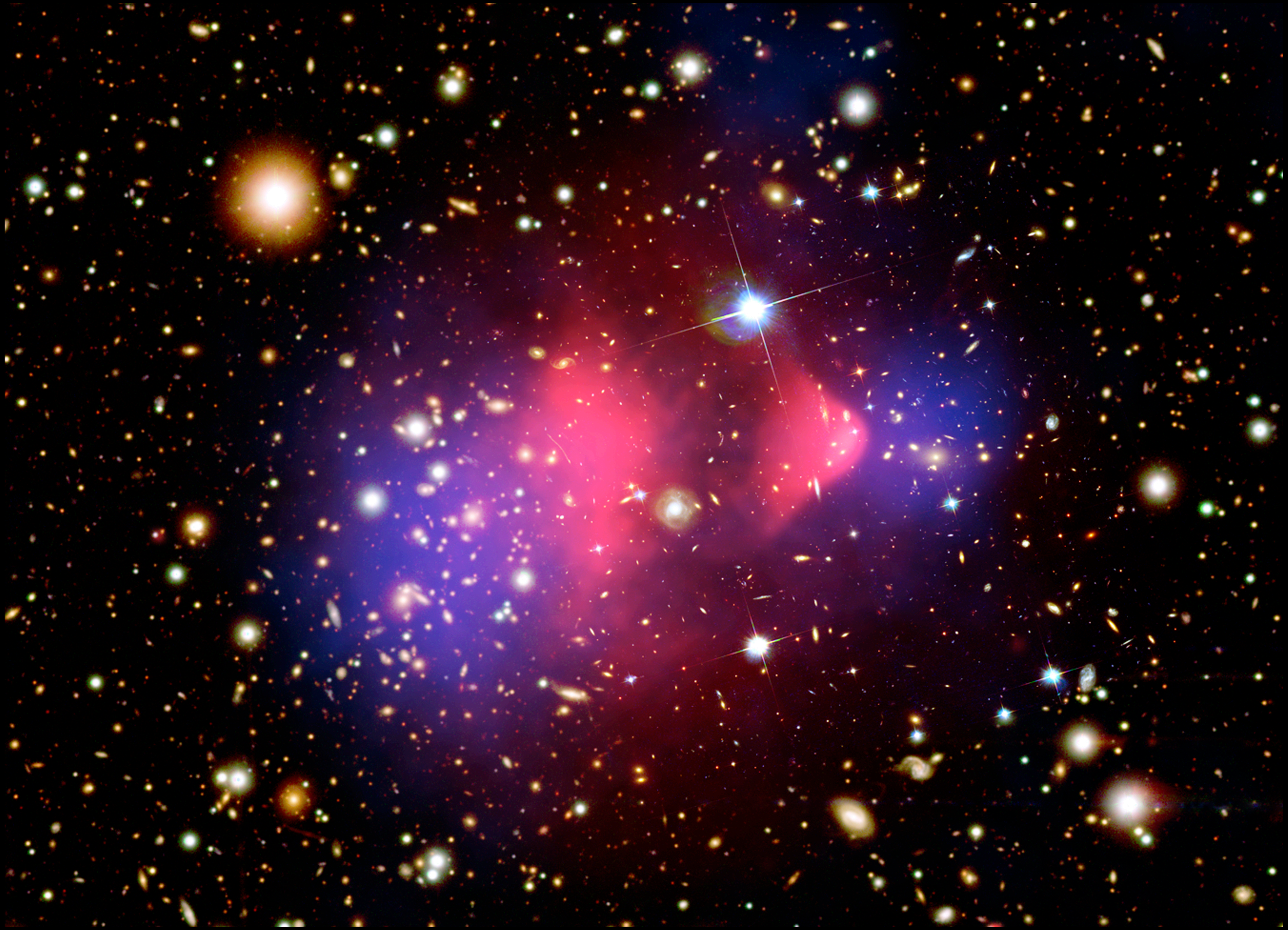
Thanks to all collaborators of The Three Hundred!



- Direct collaborators: Gustavo Yepes; Weiguang Cui; Marco De Petris; Antonio Ferragamo; Florian Ruppin; Federico De Luca; Giulia Gianfagna; Jesús Vega Ferrero; Raphaël Wicker; Ihraf Lahouli; Gianmarco Aversano; Romain Dupuis; Mahmoud Jarraya; Amélie Nef; and Félicien Schiltz.

Clusters of Galaxies

- They are the most massive gravitationally-bound structures in The Universe: $\sim 10^{15} M_{\odot}$
- Main components are: Dark matter (DM, ~80%), Intra-Cluster Medium (ICM, ~12%) and stars (~8%).
- DM cannot be directly observed, typically by its interactions of baryons or gravitational lensing.
- Stars -> optical band.
- The ICM, hot gas, -> X-ray and ->Sunyaev-Zeldovich (SZ), mm wavelengths.



The "bullet cluster",The two pink clumps correspond to the hot gas detected in X-rays, and the optical image from the Magellan and the Hubble Space Telescope shows the galaxies in orange and white. The blue area corresponds to the concentration of mass inferred by gravitational lensing effects.

Cluster of Galaxies in Cosmology and Astrophysics

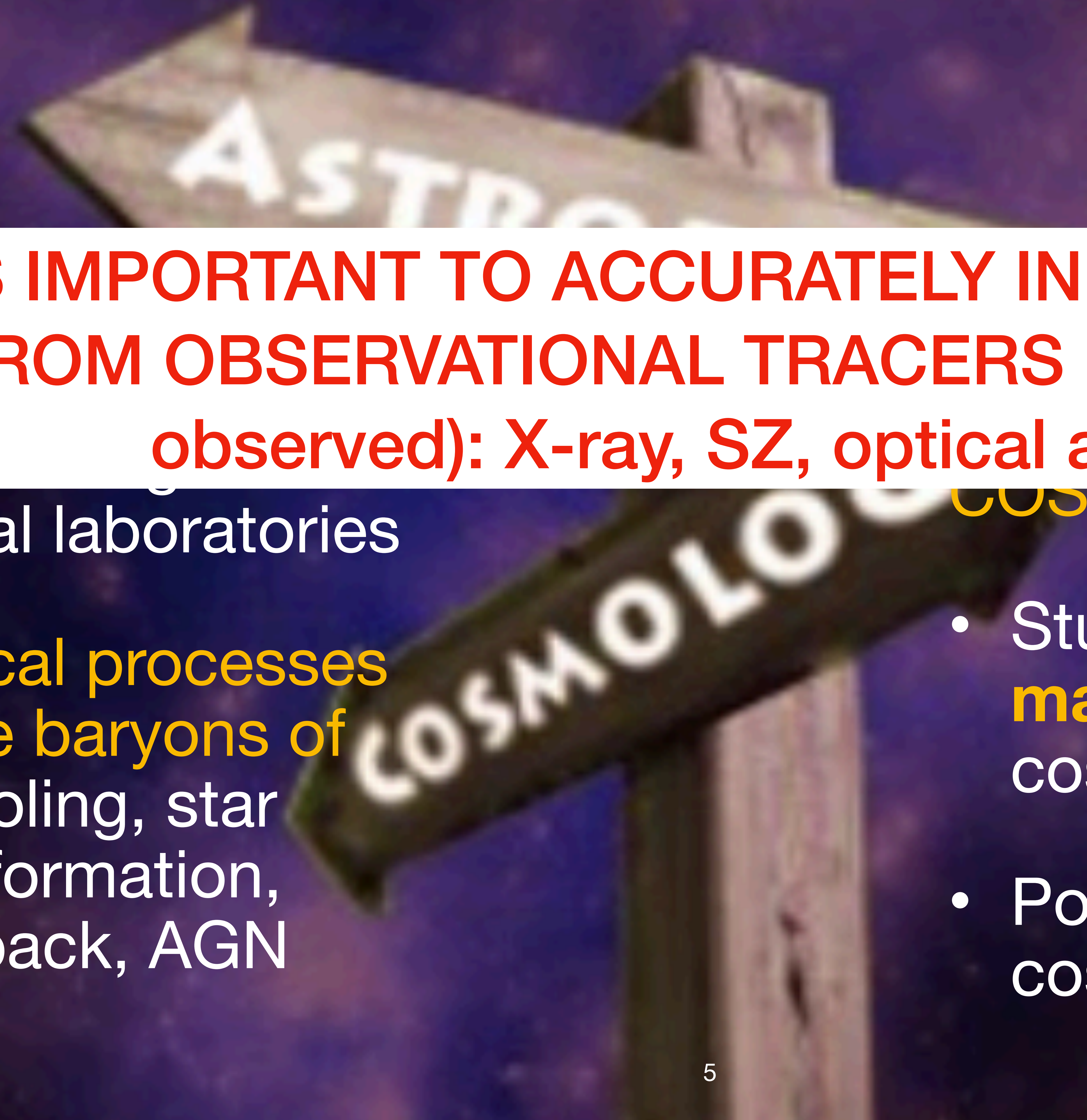
ASTROPHYSICS:

- Isolated system: giant astrophysical laboratories
- Many physical processes involving the baryons: cooling, galaxy formation, stellar feedback, AGN feedback…

COSMOLOGY:

- Study of **their abundance in mass and redshift** to test cosmological models
- Powerful tool to estimate cosmological parameters

Cluster of Galaxies in Cosmology and Astrophysics



IT IS IMPORTANT TO ACCURATELY INFER THEIR MASSES FROM OBSERVATIONAL TRACERS (DM is not directly observed): X-ray, SZ, optical and lensing.

ASTROP

- Isolate … astrophysical laboratories
- Many physical processes involving the baryons of the ICM: cooling, star and galaxy formation, stellar feedback, AGN feedback…

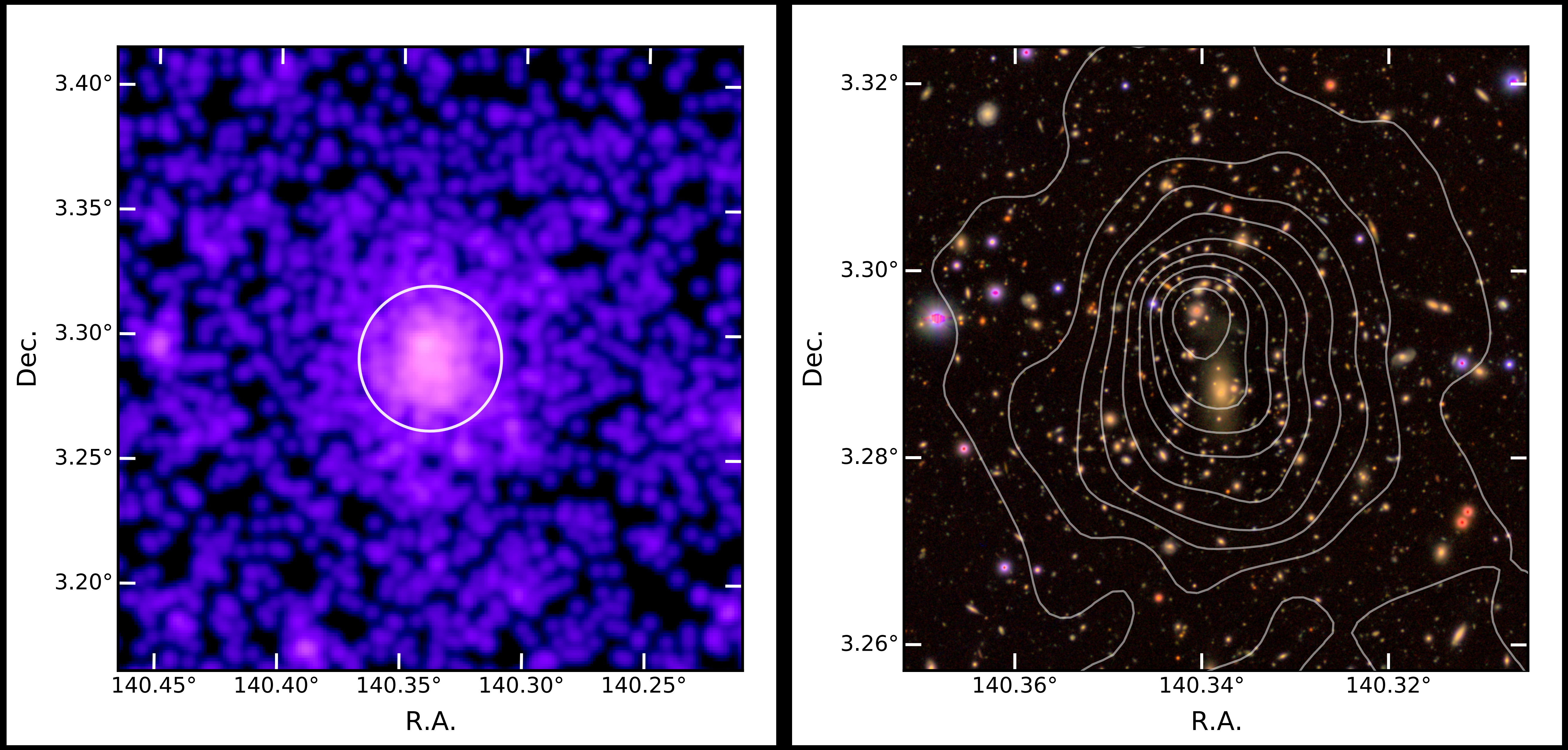
COSMOLOGY:

- Study of **their abundance in mass and redshift** to test cosmological models
- Powerful tool to estimate cosmological parameters.

Cluster of Galaxies: mass from X-ray

- The temperature of the ICM is high, $T \sim 10^8$K , the **dominant process of radiative emission is X-ray.**
- The integrated luminosity Lx at radius R_{500} is a very important quantity and it is well correlated with mass through the **scaling relation.**
- A theoretical model is fitted for estimating the electron density profile and temperature profiles, useful for inferring the mass assuming the **hydrostatic equilibrium (HE) hypothesis.**

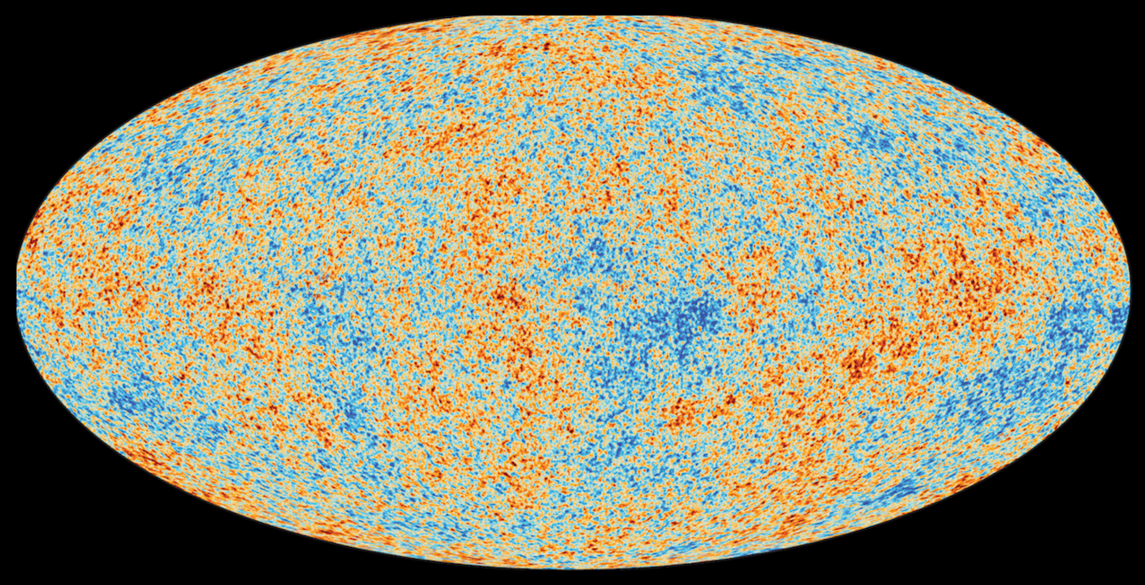
eROSITA satellite

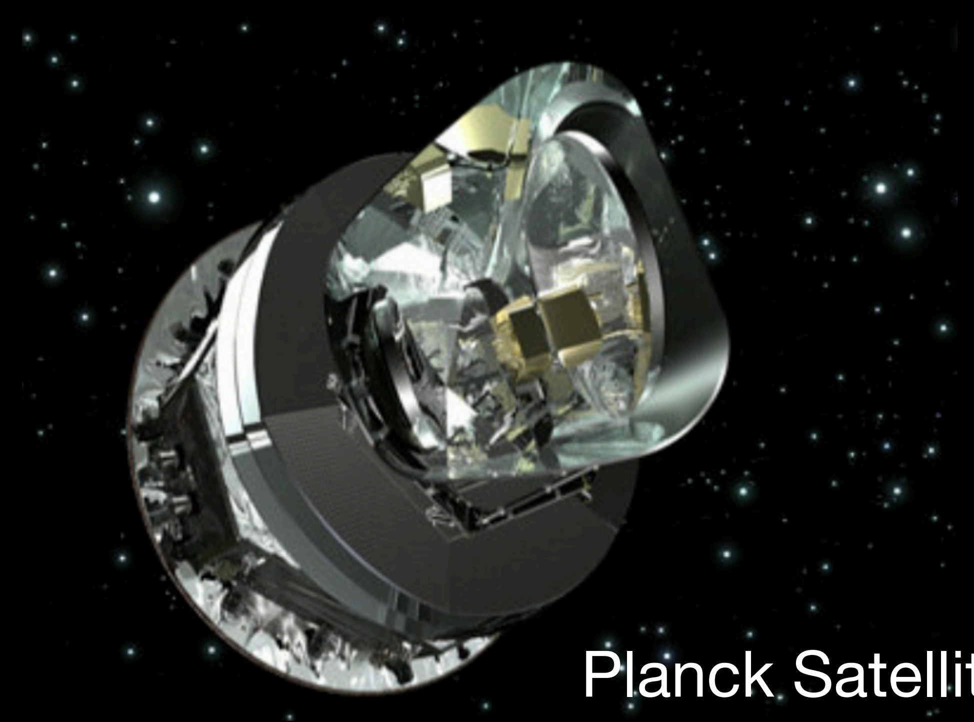


Cluster eFEDS J092121.2+031726 at redshift 0.333 (spectroscopic) soft band luminosity (0.5–2 keV) eROSITA image. Liu et al. (2022)

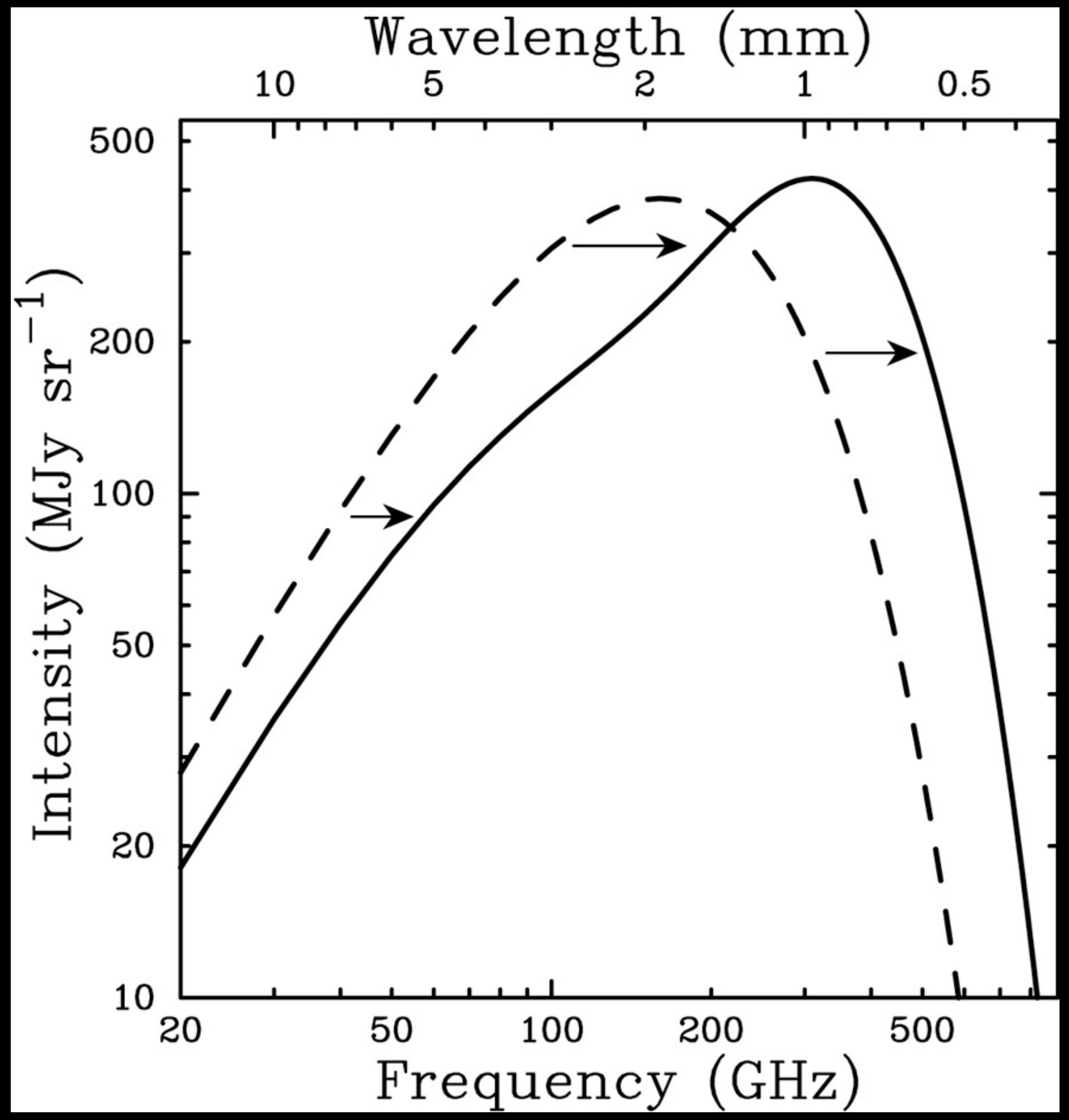
Cluster of Galaxies: mass from SZ

- The **Sunyaev-Zeldovich** effect is the inverse Compton scatter of the CMB photons with the hot electrons within the ICM. This effect is observed at mm wavelengths.
- The intensity of the SZ effect is characterised by the **Compton-y parameter map,** which is the gas pressure integrated over the l.o.s.
- The integrated y-map Y is very well related to the mass through the Y-M **scaling relation.**
- From the y-map, the pressure of the gas can be estimated, and the mass can be computed using the **HE hypothesis** .

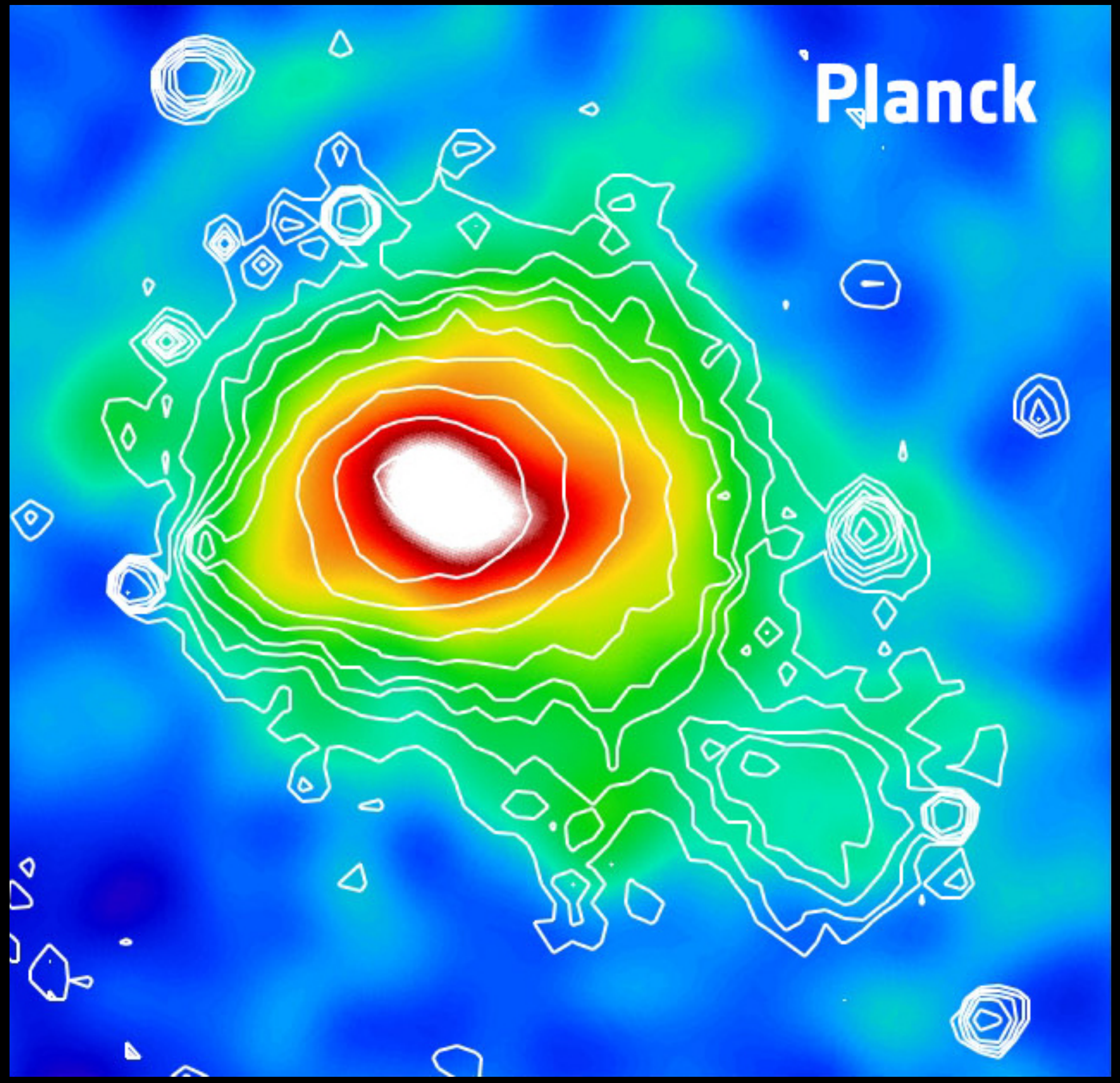




Planck Satellite



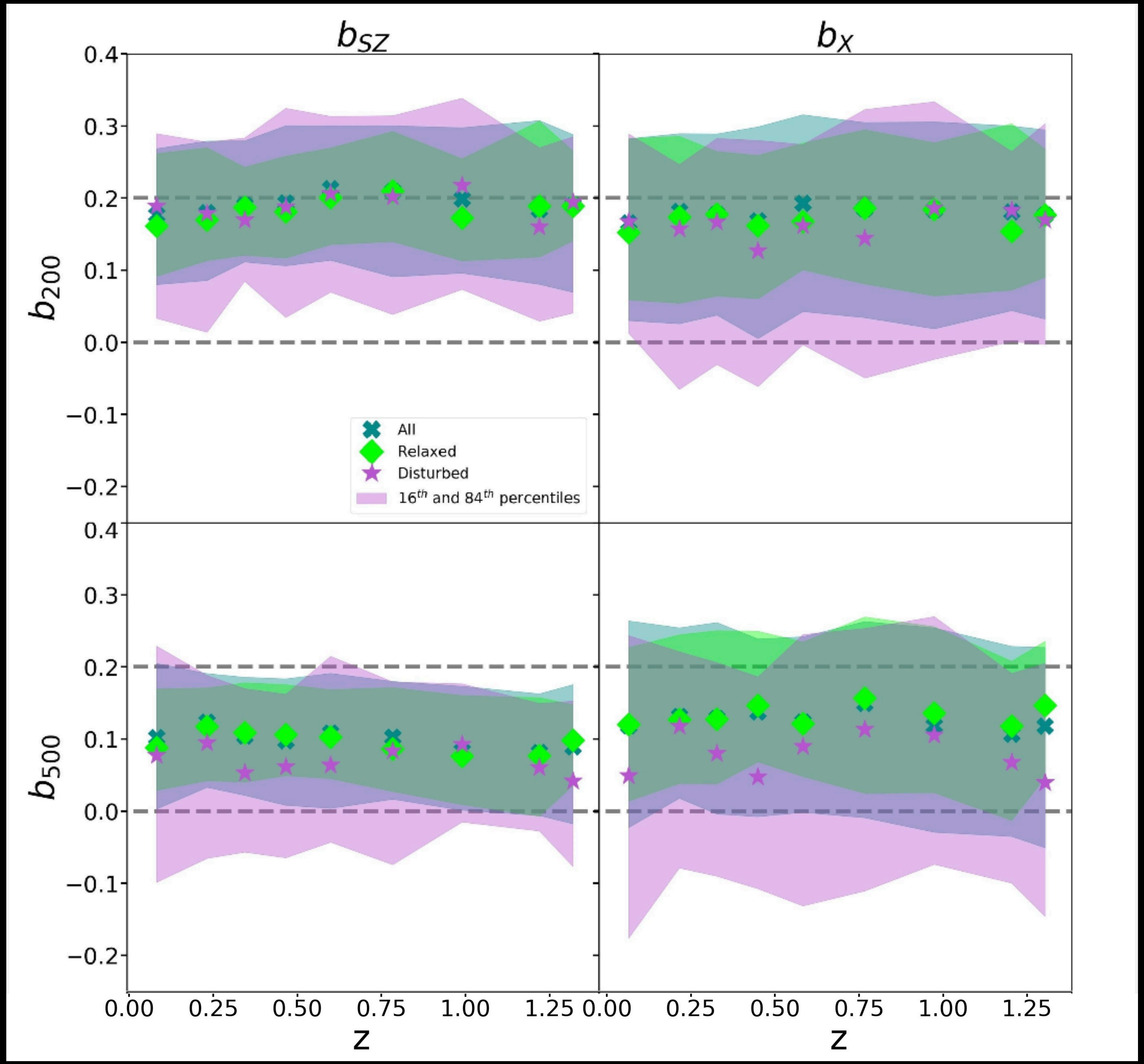
CMB power spectrum is distorted by the ICM plasma



Coma Cluster observed by the Planck Satellite

Cluster of Galaxies: HE mass bias

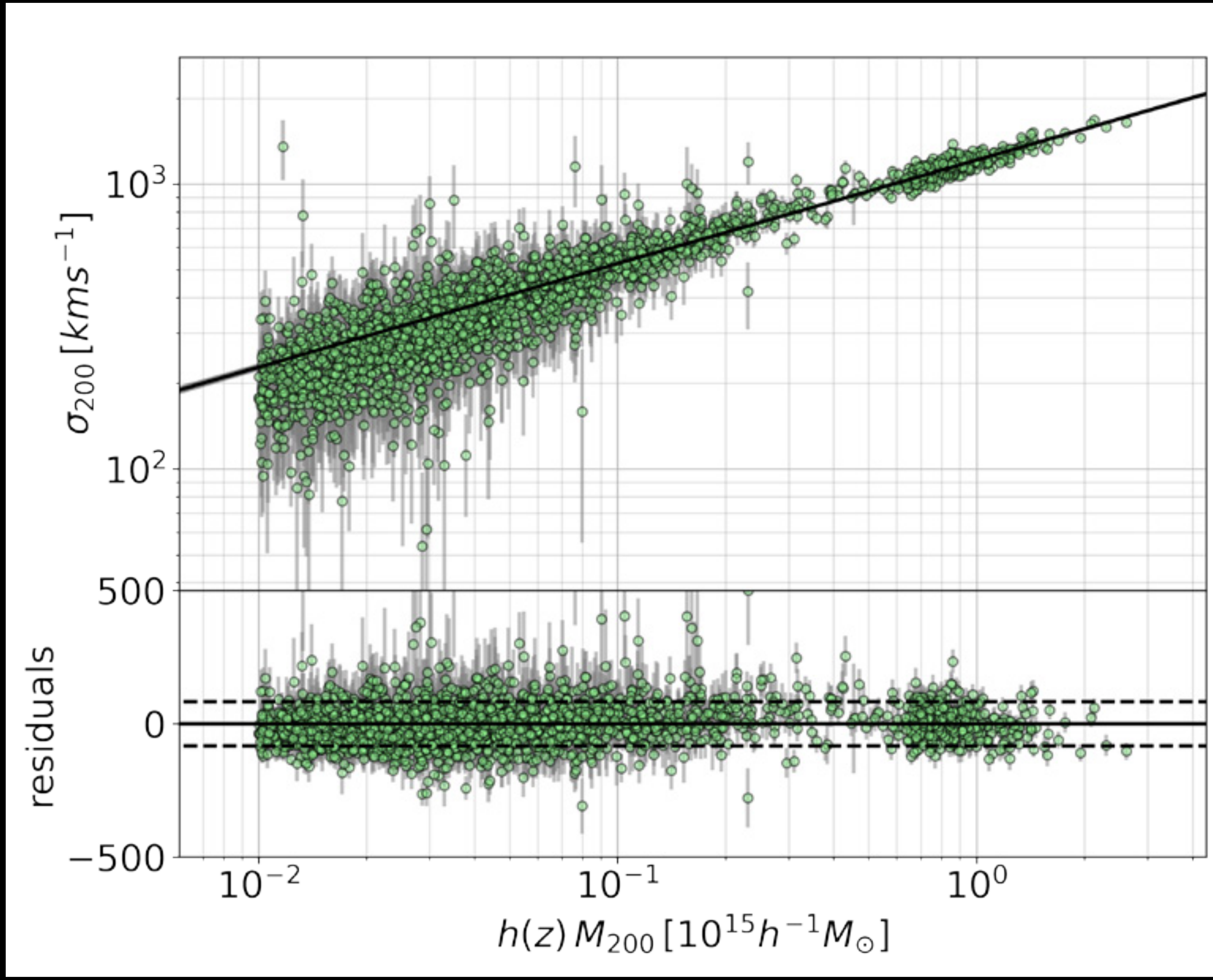
- Masses of clusters of galaxies can be estimated assuming that the gas pressure is in hydrostatic equilibrium with the gravitational potencial. **M_{HE}**
- These masses are found to be biased, defining the **bias parameter** as: $b = \dfrac{M_{\text{tot}} - M_{\text{HE}}}{M_{\text{tot}}}$.
- This parameter is calibrated typically using simulations. The median value is found to be around 10-20%.



HE mass bias as a function of redshift for SZ (left) and X-ray (right) for The300 simulation. Gianfagna et al. (2023)

Cluster of Galaxies: Mass by galaxies

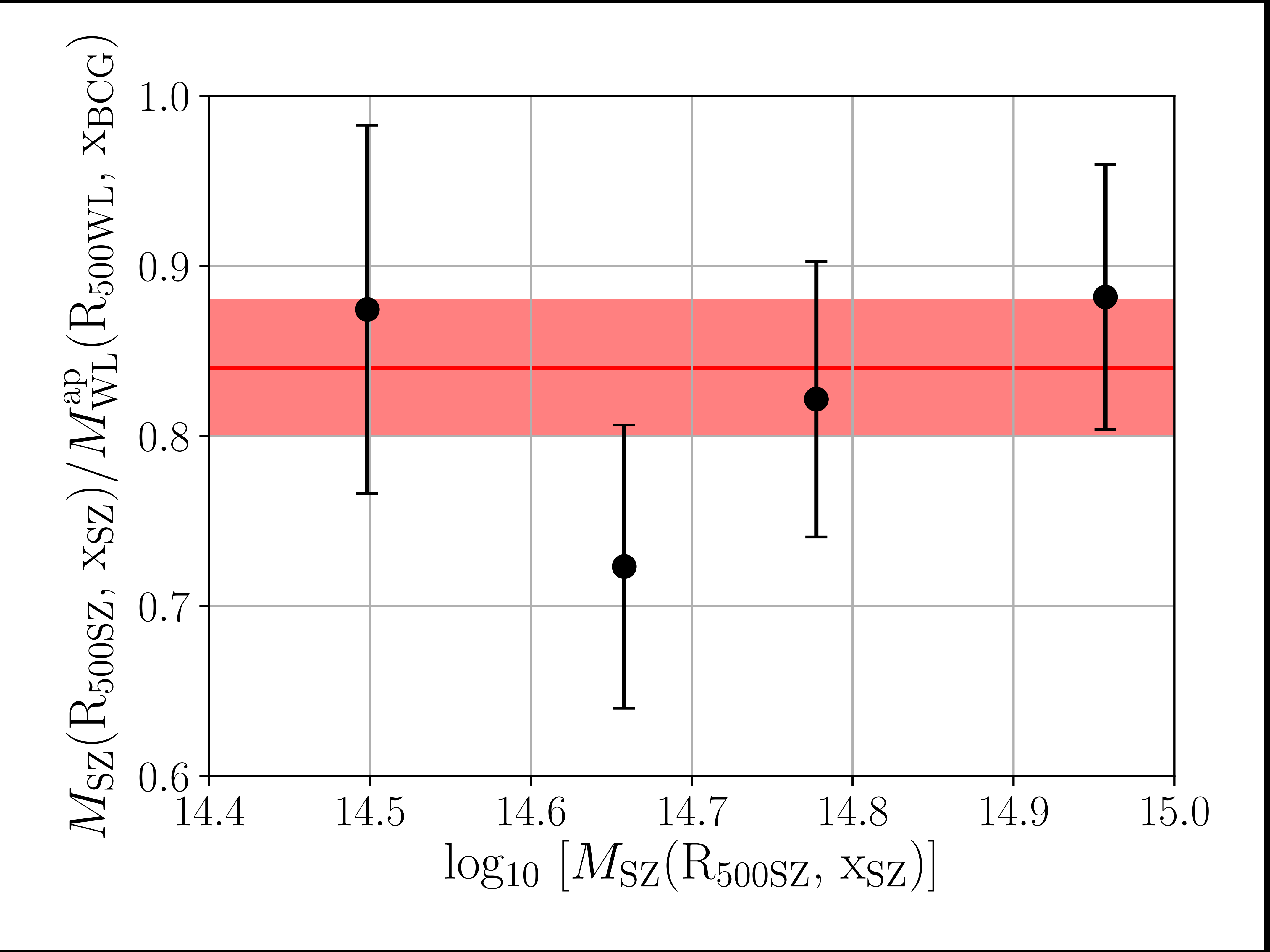
- Masses can also be estimated by measuring the velocity dispersion of the galaxies.
- The mass is typically estimated by considering the σ-M scaling relation.



Sub-halo velocity dispersion against halo mass at z=0. (Ferragamo+2022)

Cluster of Galaxies: Mass by weak gravitational lensing

- The massive galaxy cluster introduces a weak distortion (shear) of the light from background galaxies, which scales with mass.
- The mass is typically computed by theoretically modelling the relation between the shear and mass.
- Simulations show that these masses are almost unbiased, so other mass proxies are typically calibrated with the lensing mass.



SZ mass assuming HE divided by the WL mass.
The results show a biased of 1-b=0.84.
Herbonnet+2019

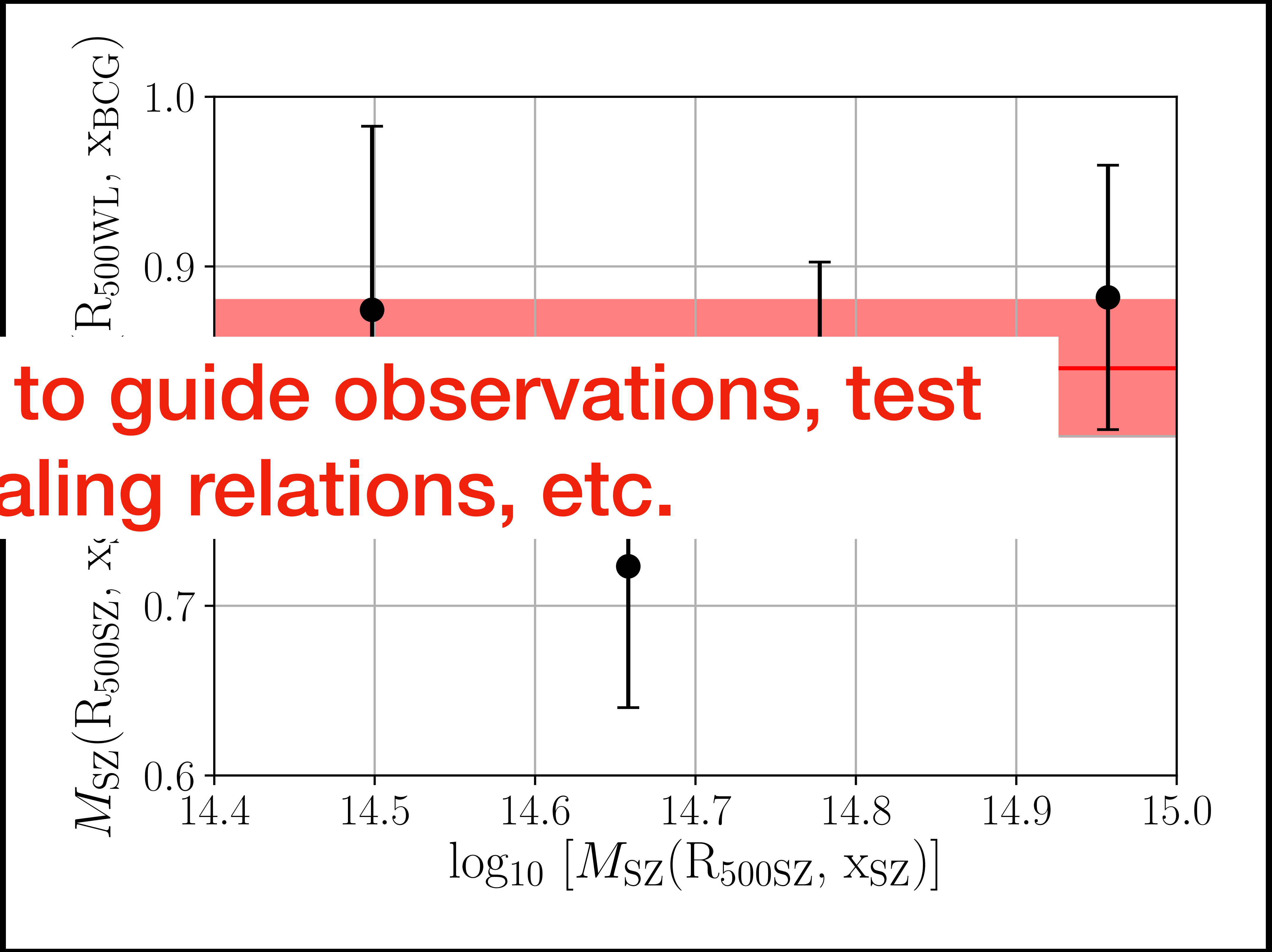
Cluster of Galaxies: Mass by weak gravitational lensing

- The massive galaxy cluster introduces a weak distortion (shear) of the light from background galaxies, which scales with

Simulations are powerful tools to guide observations, test pipelines, calibrate scaling relations, etc.

- The theoretically modelling the relation between the shear and mass.

- Simulations show that these masses are almost unbiased, so other mass proxies are typically calibrated with the lensing mass.

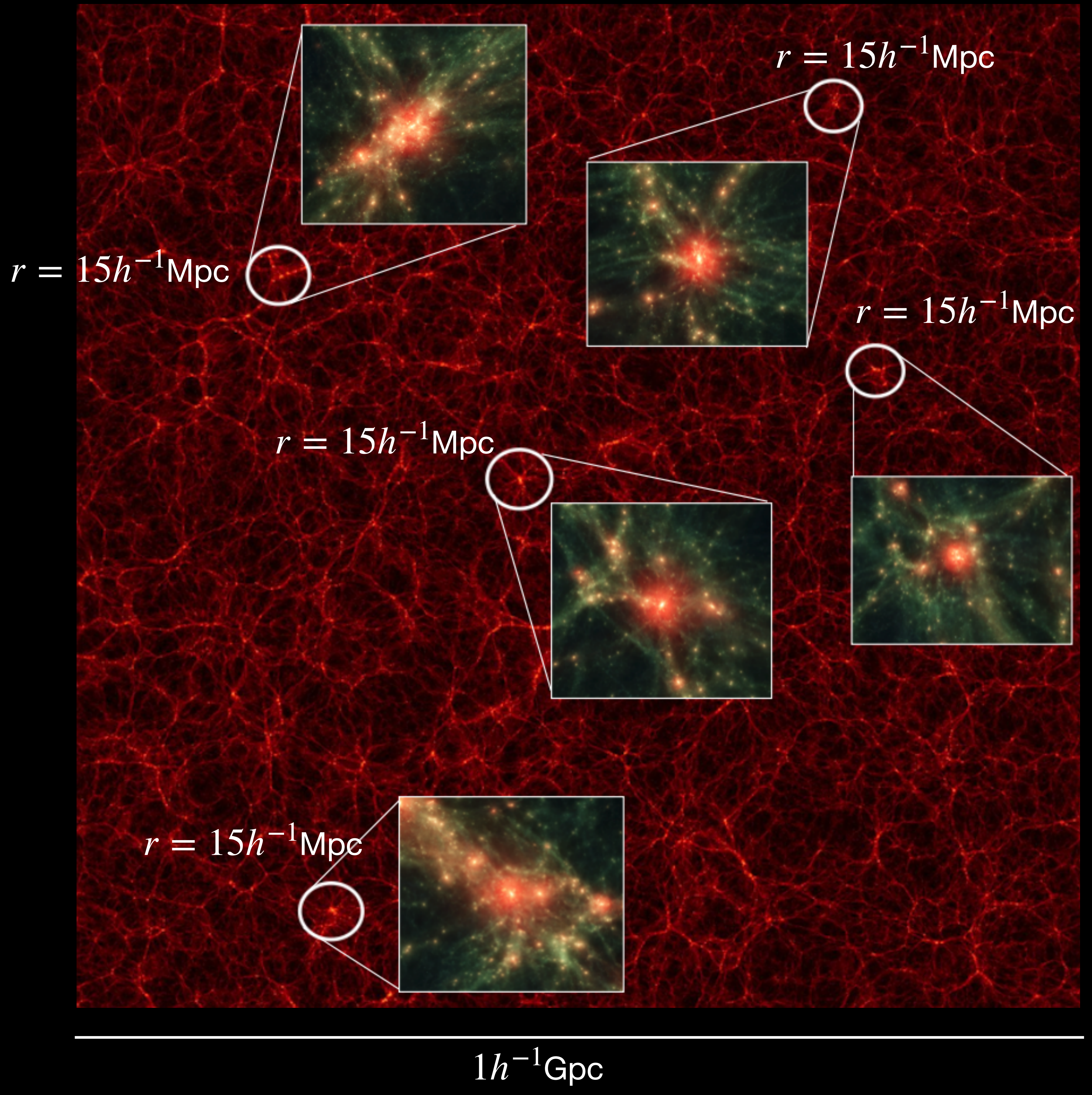


SZ mass assuming HE divided by the WL mass.
The results show a biased of 1-b=0.84.
Herbonnet+2019

Cosmological simulations: The Three Hundred Project

- Galaxy clusters are important for cosmology and astrophysics-> large volumes+small scales
- Big cosmological simulations of volume $(1h^{-1}\text{Gpc})^3$ include only Dark Matter, e.g., Multidark Planck (MDPL2) simulation.
- Hydrodynamical simulations are smaller and typically lack statistics of massive galaxy clusters ($\sim 10^{15} M_{\odot}$).
- The solution is to run zoom-in simulations. High resolution + hydrodynamics only in the region of interest.
- The Three Hundred (The300) project is a set of 324 zoom-in hydrodynamical simulations centred at the most massive clusters at z=0 of the MDPL2 simulation-> 324 spheres of $r = 15h^{-1}$Mpc.

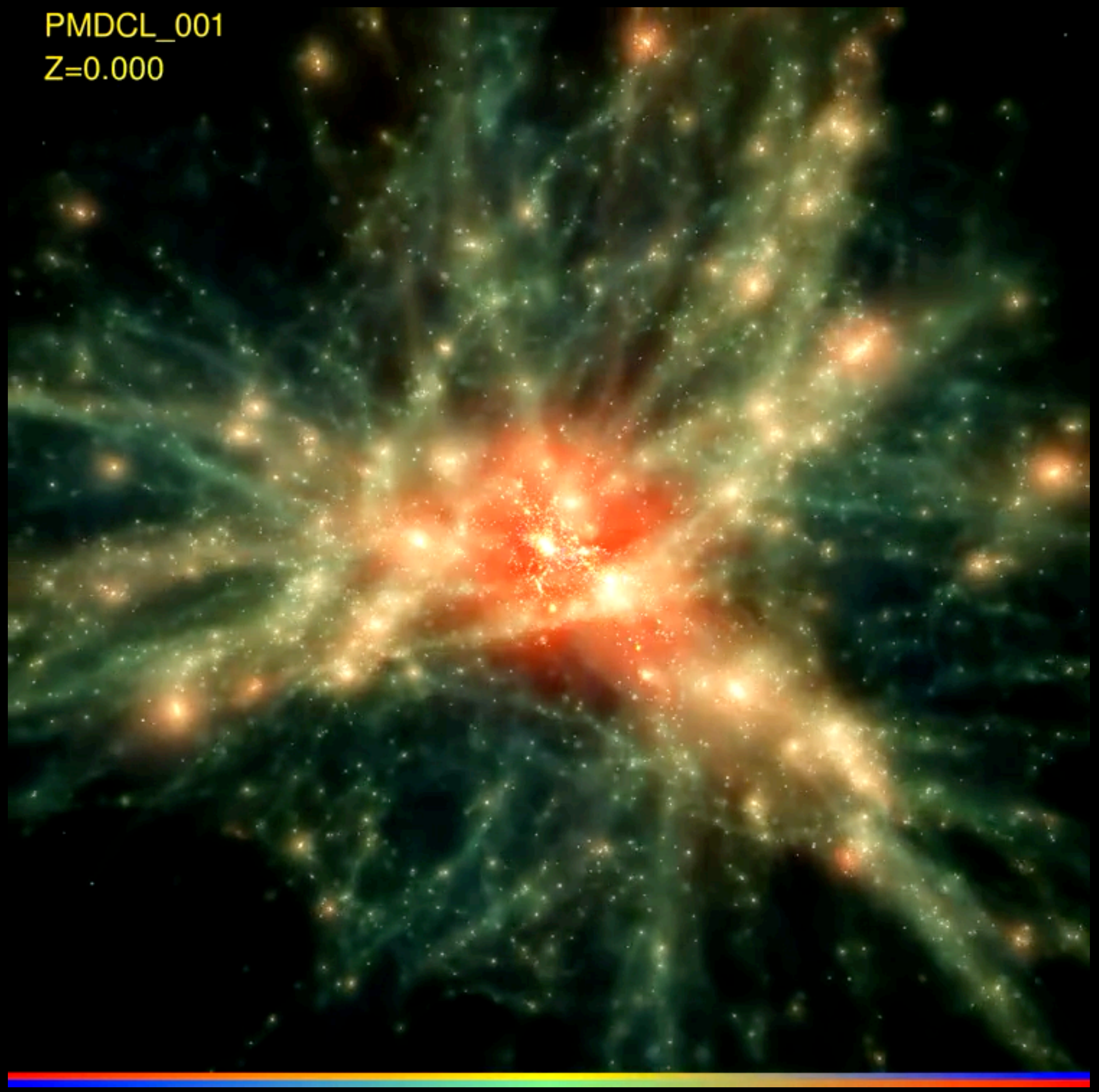
The Three Hundred Project



12

Cosmological simulations: The Three Hundred Project

- The simulations were run within ΛCDM cosmology and the parameters are consistent with the Planck collaboration.
- $m_{DM} = 12.7 \times 10^8 h^{-1} M_\odot$,
  $m_{\text{gas}} = 2.23 \times 10^8 h^{-1} M_\odot$.
- Gravity and hydrodynamics implemented at the particle level.
- The rest of the processes are developed as analytical prescriptions known as "subgrid physics", such as stellar feedback and AGN feedback.



Cosmological simulations: The Three Hundred Project

The300 runs

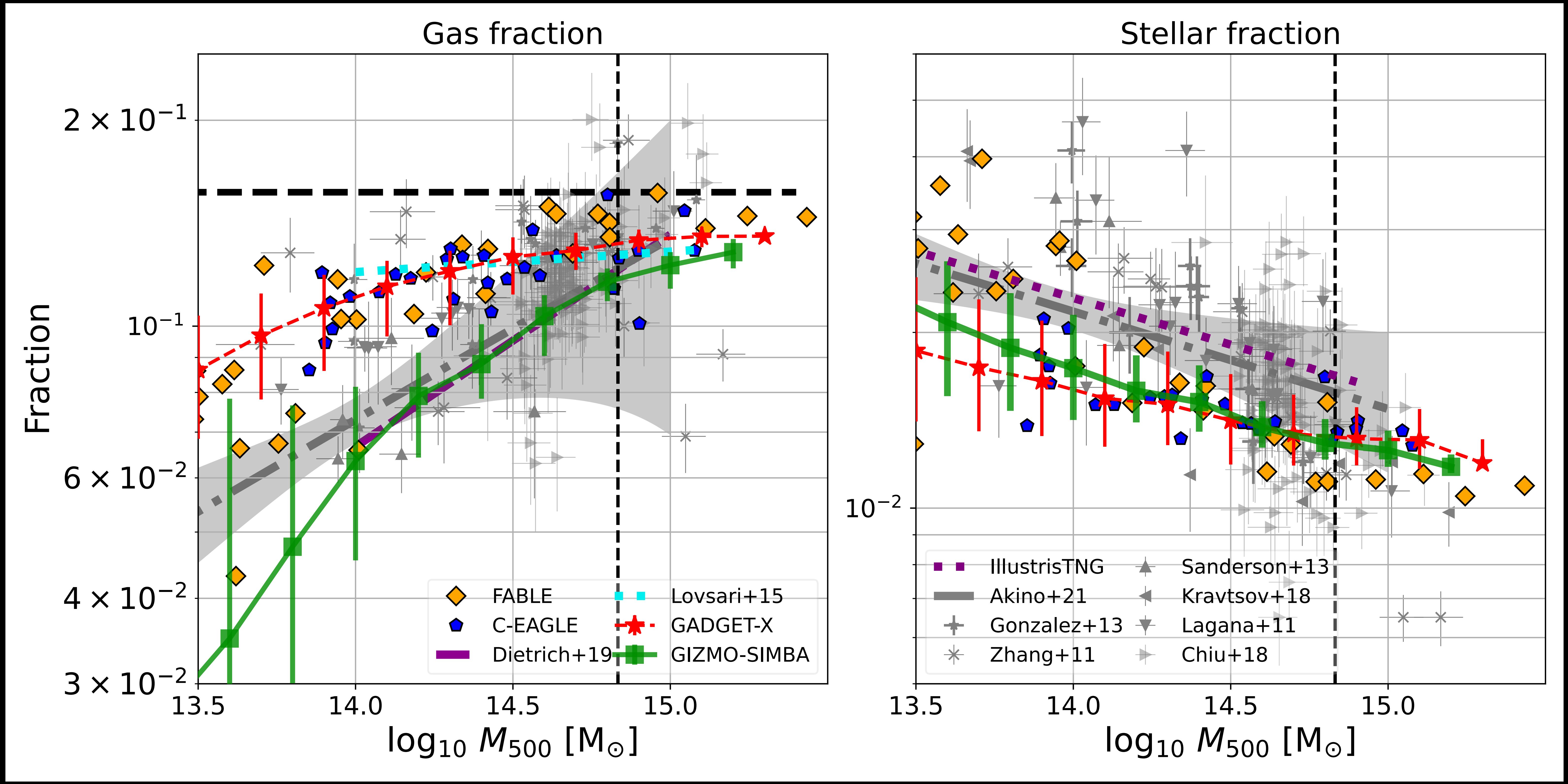
MUSIC run: SPH, but no black holes

GADGET-X run: SPH with AGN feedback

GIZMO-SIMBA run: MFM solver and more efficient AGN feedback.

DM-only 3K (same as hydro), 7K y 15K particle resolution.

Galaxy SAM: SAG, Galactic and SAGE for DM 3K (Knebe, A. +2017) . SAG and SAGE for 7k y 15k (Gómez, J.+2024).



The baryon fractions within R_{500}: gas fractions on the left-hand side panel and stellar fractions on the right-hand side panel at $z = 0$. The AGN feedback mechanism is very efficient in the GIZMO-SIMBA simulation blowing gas well outside the virial radius. Cui et al. (2022)

Cosmological simulations: The Three Hundred Project

The300 runs

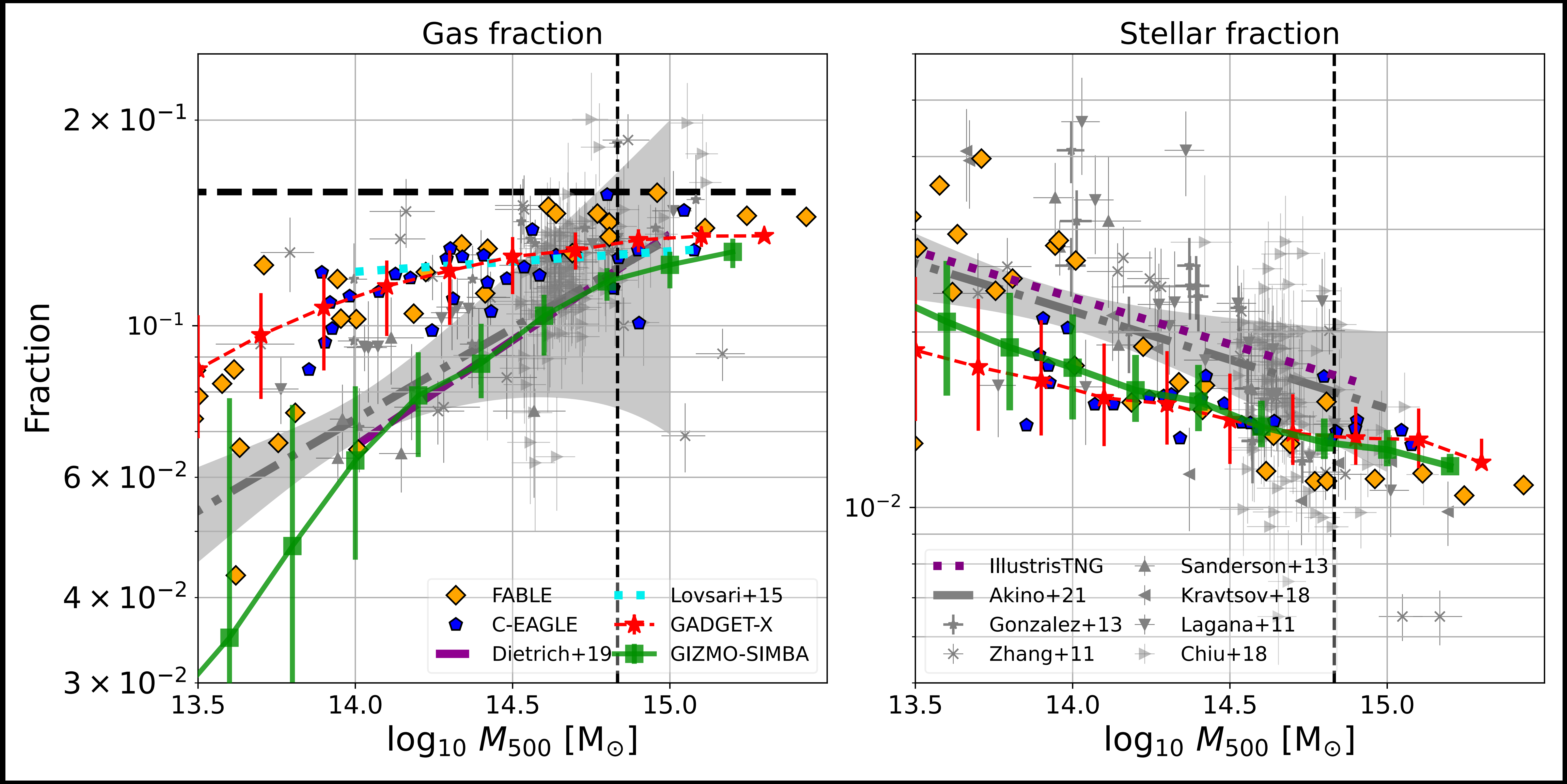
MUSIC run: SPH, but no black holes

GADGET-X run: SPH with AGN feedback

GIZMO-SIMBA run: MFM solver and more efficient AGN feedback.

DM-only 3K (same as hydro), 7K y 15K particle resolution.

Galaxy SAM: SAG, Galactic and SAGE for DM 3K (Knebe, A. +2017) . SAG and SAGE for 7k y 15k (Gómez, J.+2024).

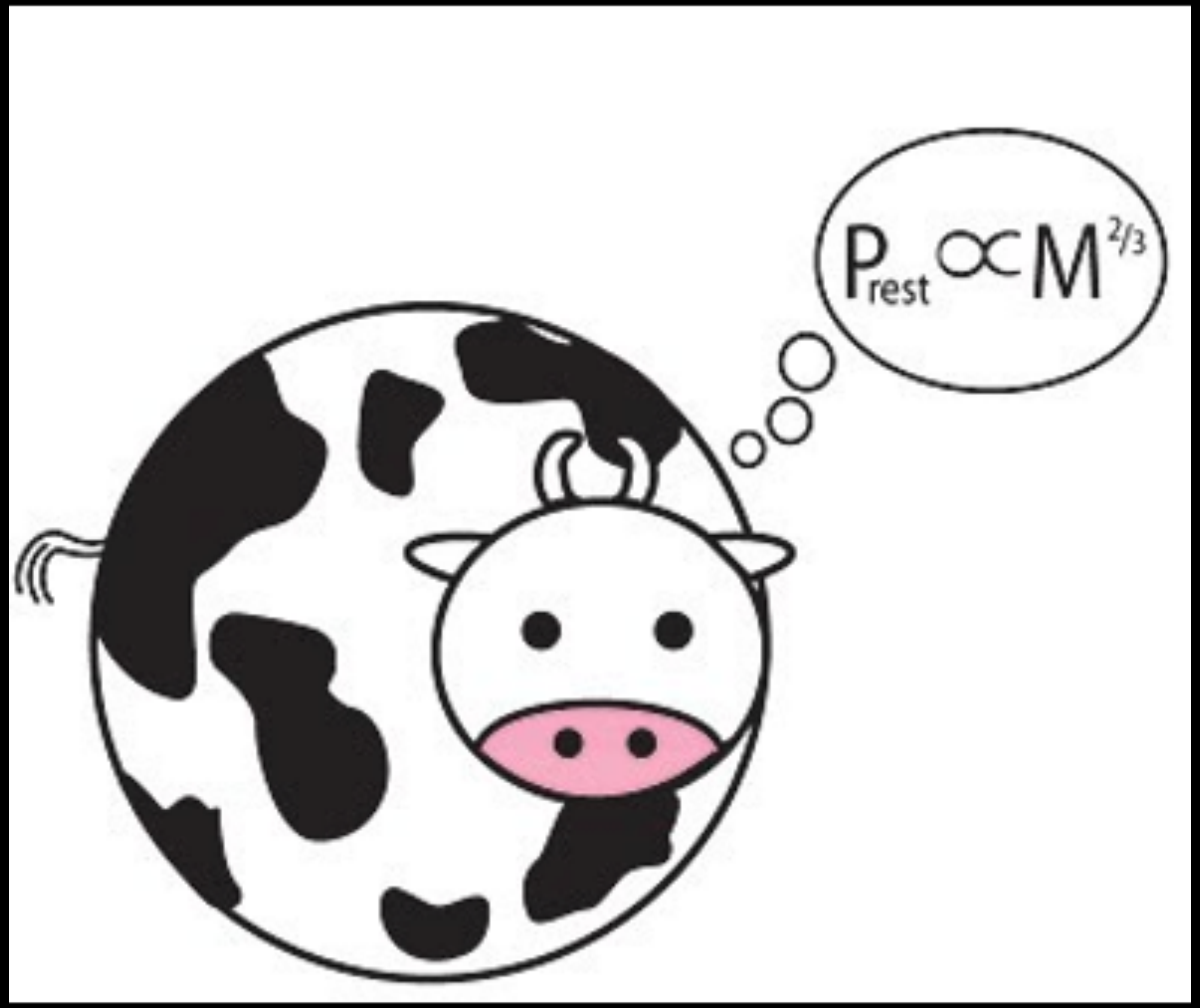


The baryon fractions within R_{500}: gas fractions on the left-hand side panel and stellar fractions on the right-hand side panel at z = 0. The AGN feedback mechanism is very efficient in the GIZMO-SIMBA simulation blowing gas well outside the virial radius. Cui et al. (2022)

Cosmological simulations: The Three Hundred Project

Importance of ML

"Traditional methods" to infer masses use The300 and assume symmetries of The ICM that lead to a bias result,
Gianfagna+2023



Machine learning

ML methods use The300 data to learn directly the underlying relation between mass and observables.

The main limitation is the physics implemented in the simulations.

In general, ML allows to address problems in a different way, o problems that were intractable.

nature astronomy

Article

https://doi.org/10.1038/s41550-022-01784-y

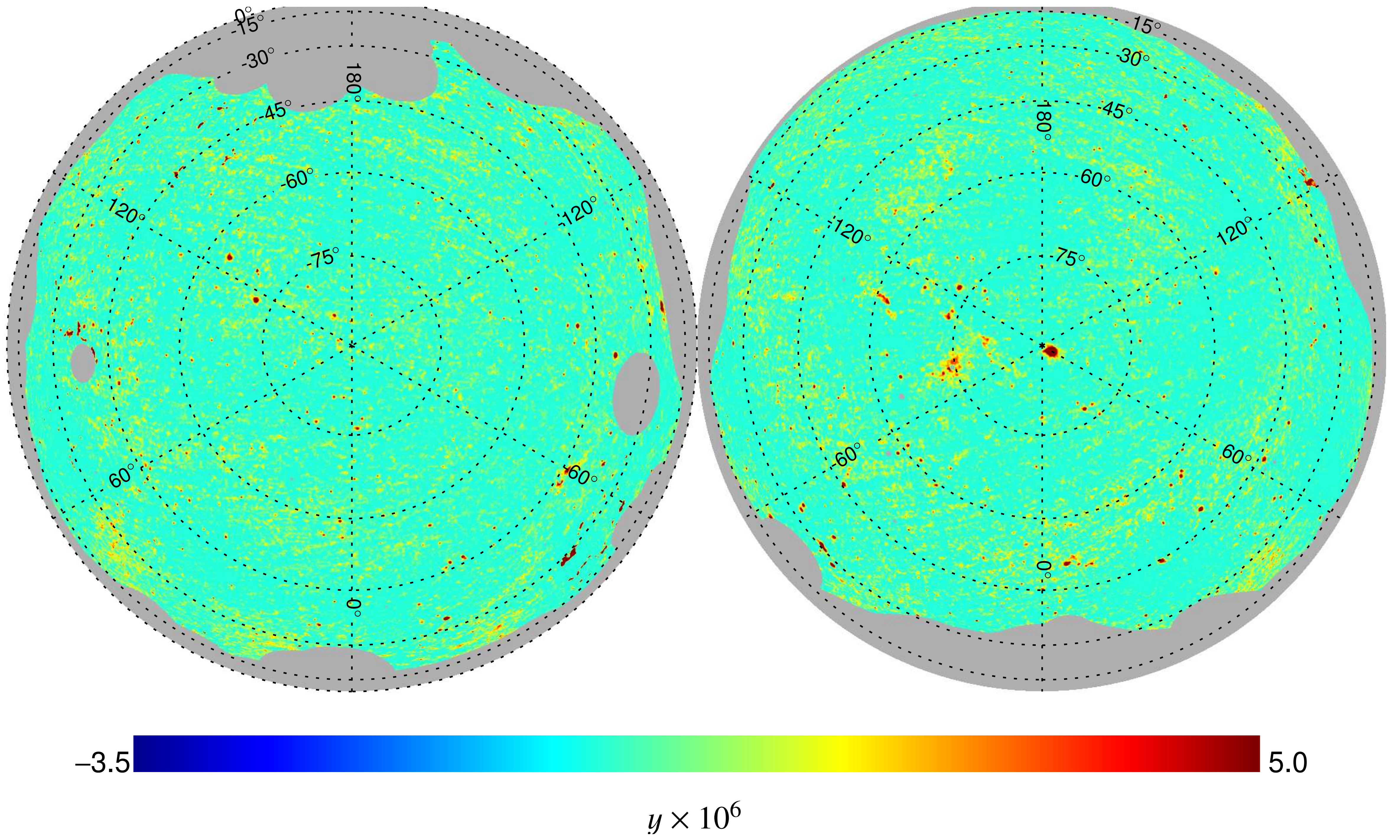
A deep learning approach to infer galaxy cluster masses from Planck Compton-*y* parameter maps

De Andres et al. 2022

- The Planck collaboration provides the Compton-y parameter (SZ) map of the full sky, which is a map of the thermal SZ effect. See Planck 2015 results

- A "blind search for galaxy clusters" creates the **PSZ2 catalogue**, with 1653 detections, of which at least 1203 are confirmed clusters with external datasets. For this work, we only considered the objects with measured redshift, a total of **1094 cluster with redshift z<1**.

- The SZ effect maps are widely studied, mainly because from simulations it is known that the integrated Compton-y parameter is a very valuable mass proxy. Therefore, the masses M_{500} of all these clusters were estimated from scaling relations.



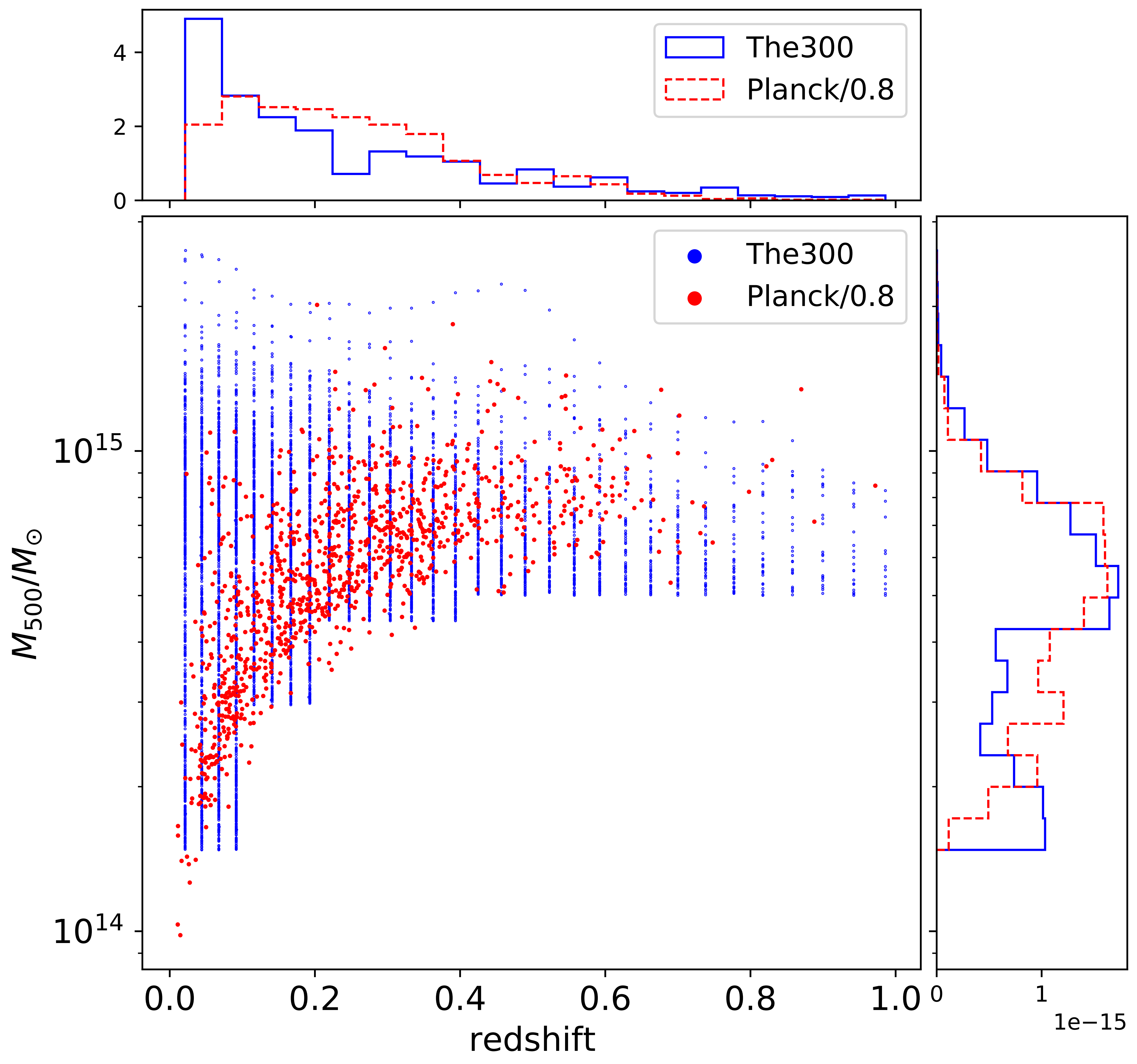
Planck all-sky Compton parameter map for MILCA orthographic projection. See Planck 2015 results.

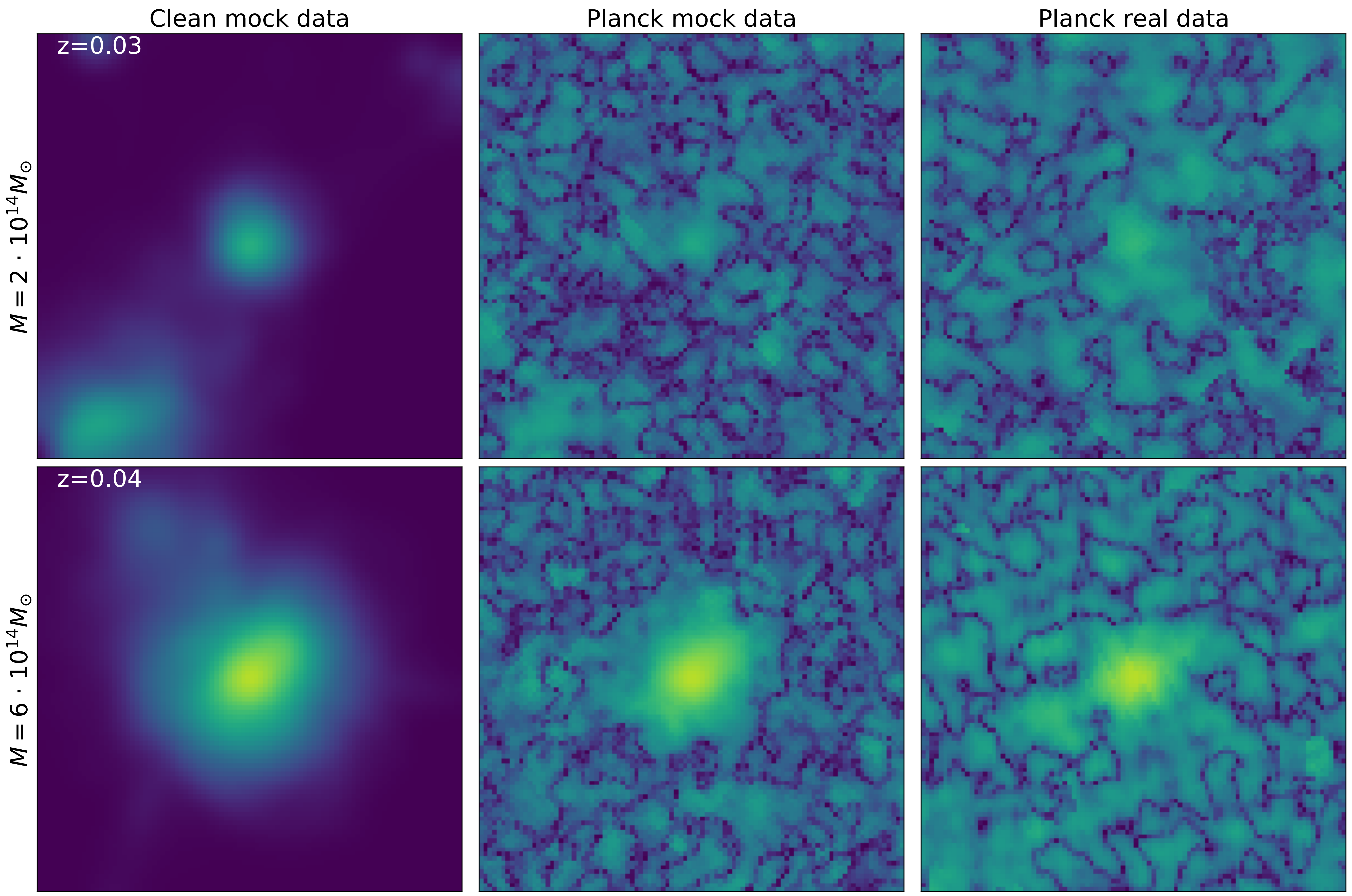
nature astronomy

Article https://doi.org/10.1038/s41550-022-01784-y

A deep learning approach to infer galaxy cluster masses from Planck Compton-y parameter maps

Datasets





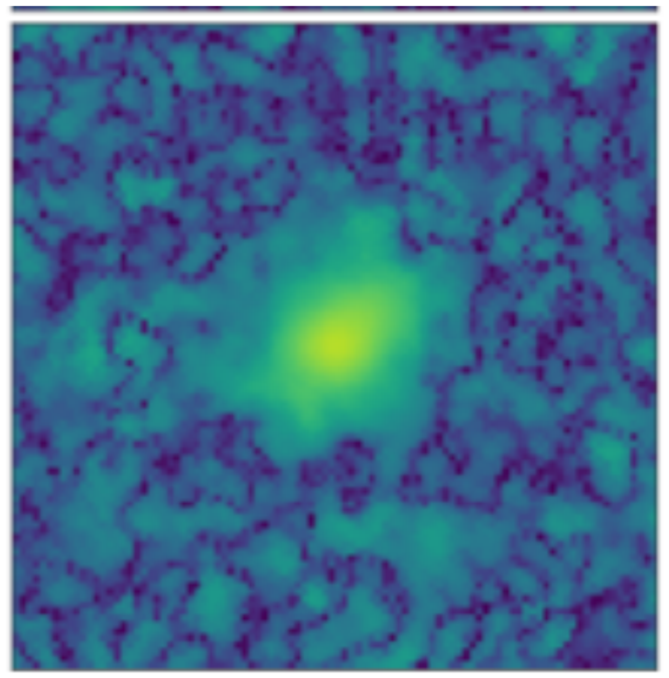
nature astronomy

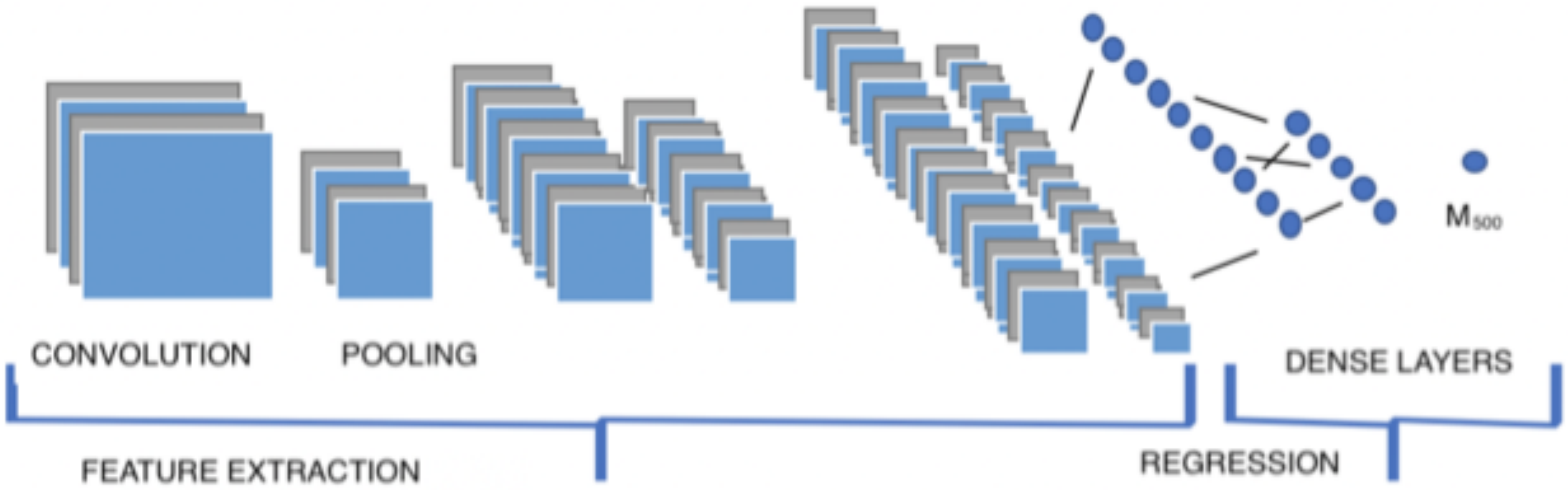
Article https://doi.org/10.1038/s41550-022-01784-y

A deep learning approach to infer galaxy cluster masses from Planck Compton-*y* parameter maps

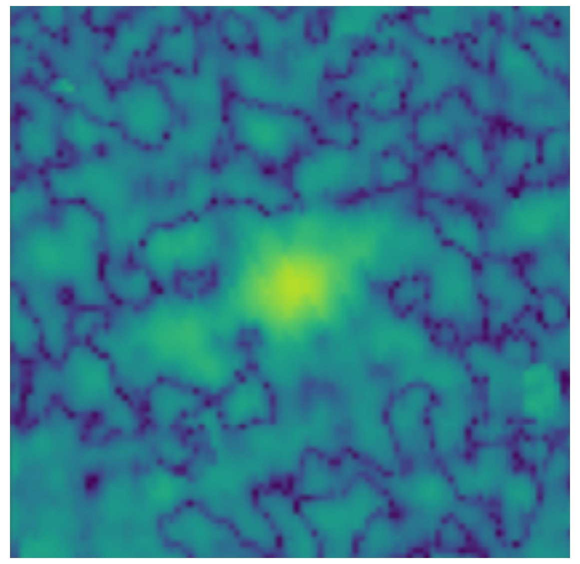
Model

Train with simulated data





Predict with real data



nature astronomy

Article https://doi.org/10.1038/s41550-022-01784-y

A deep learning approach to infer galaxy cluster masses from Planck Compton-y parameter maps

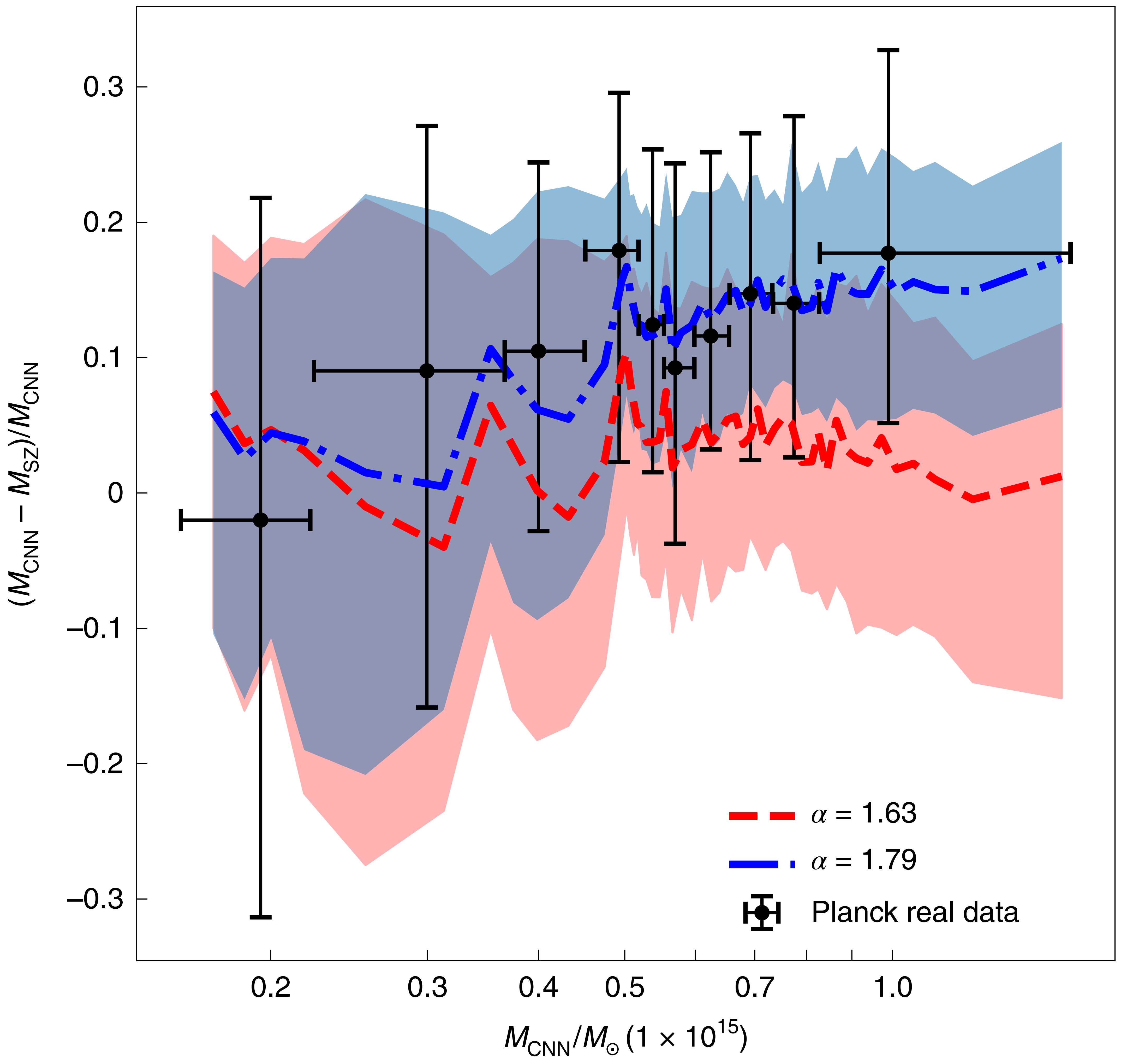
Main result: Understanding the mass bias

- We perform a simple mass inference using the Y-M scaling relation:

$$E(z)^{-2/3}\left[\frac{D_A^2(z)Y}{10^{-4}\,\mathrm{Mpc}^2}\right] = B\left[\frac{h}{70}\right]^{-2+\alpha}\left[\frac{M_{SZ}}{6\times10^{14}\,M_\odot}\right]^{\alpha},$$

- We derive Y from the original dataset (clean and high resolution). The mass M_{SZ} is computed with two slopes: $\alpha = 1.63$ (red, The300) and $\alpha = 1.79$ (blue, Planck). The **black** points corresponds to our **CNN estimates** (previous figure).

- The blue line follows the Planck data while the red line is roughly flat. Therefore, a possible explanation lies in the assumed Y-M scaling relations.



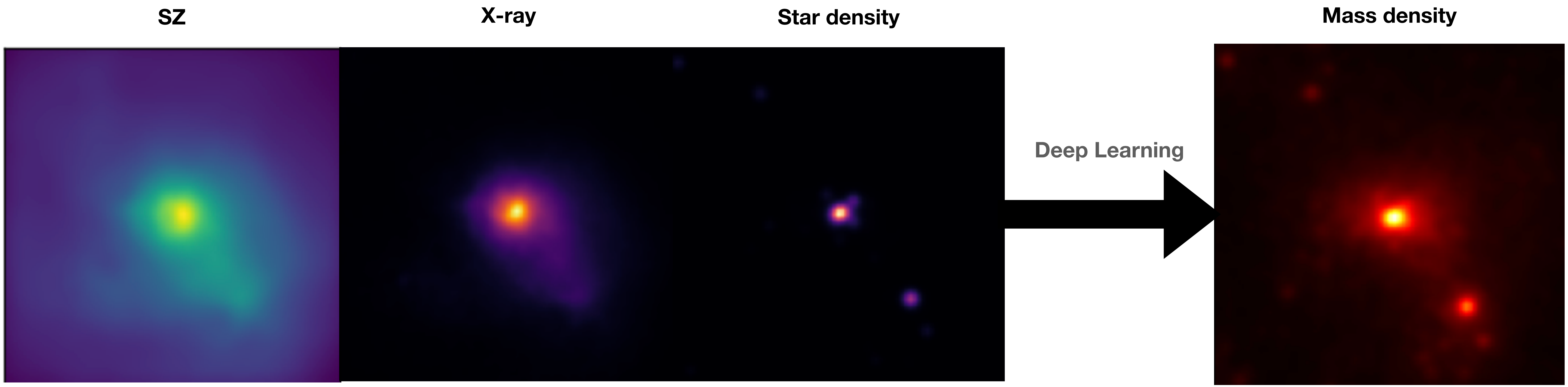
Verifying the bias causes with Y-M relation.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **528,** 1517–1530 (2024)
Advance Access publication 2024 January 9
https://doi.org/10.1093/mnras/stae071

The three hundred project: mapping the matter distribution in galaxy clusters via deep learning from multiview simulated observations

De Andres et al, 2024



- Observations of galaxy clusters are **numerous in X-ray, SZ and optical**. For instance, New X-ray mission **eROSITA** recently presented a catalog of **12,247** clusters observed in X-ray.

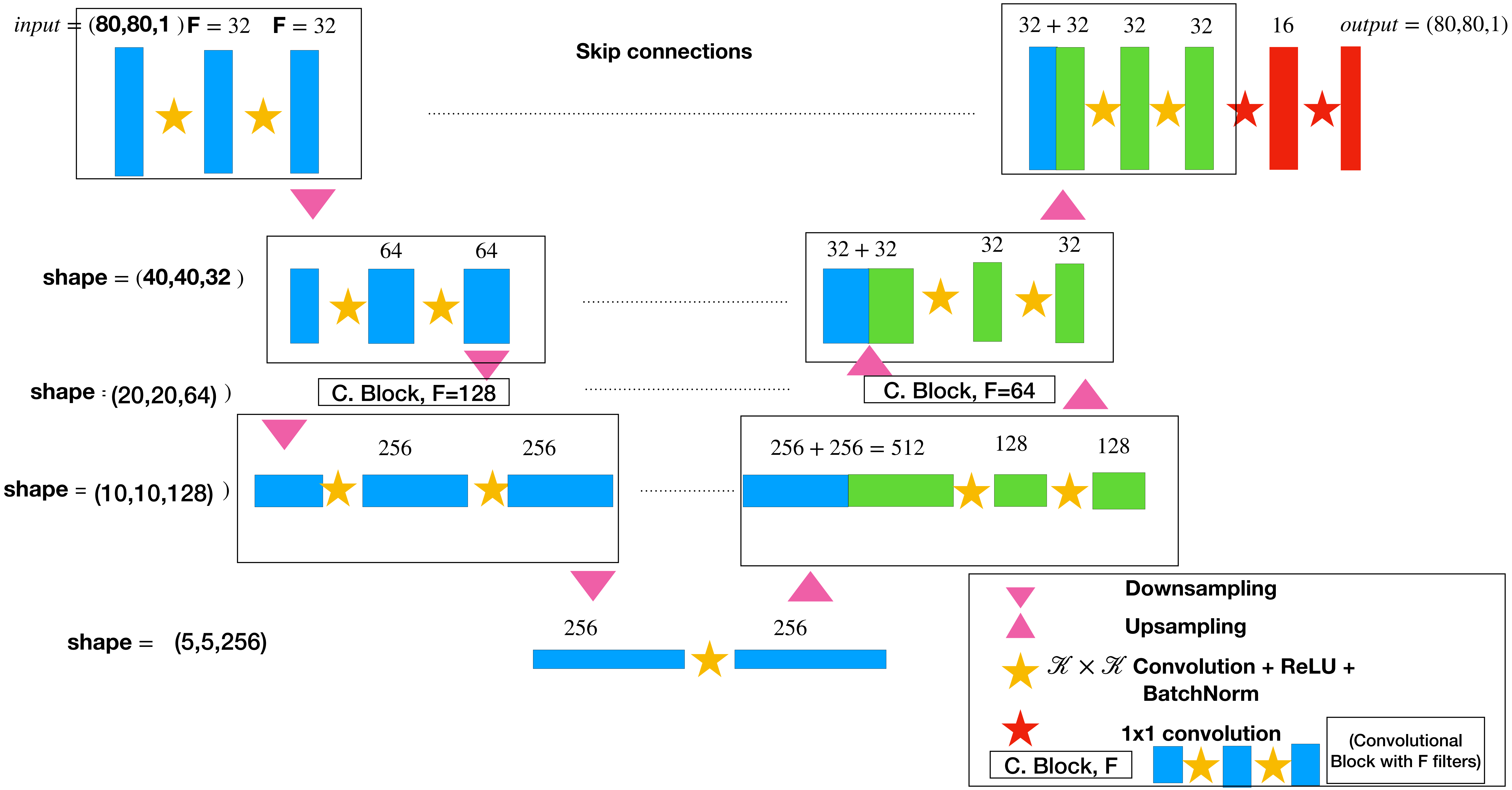
- With our method, the overall matter distribution is directly inferred from the ICM observations and stars, corresponding to a **more accesible approach, complementing the lensing methods.**

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS 528, 1517–1530 (2024) https://doi.org/10.1093/mnras/stae071
Advance Access publication 2024 January 9

Model

The three hundred project: mapping the matter distribution in galaxy clusters via deep learning from multiview simulated observations

- The U-Net model is considered the standard for image-to-image translation. It was introduced in biomedical imaging.
- We test the MAE loss function and the conditional WGAN model. The MAE loss function was as good as the conditional GAN, so for simplicity we considered MAE as the best lost function.



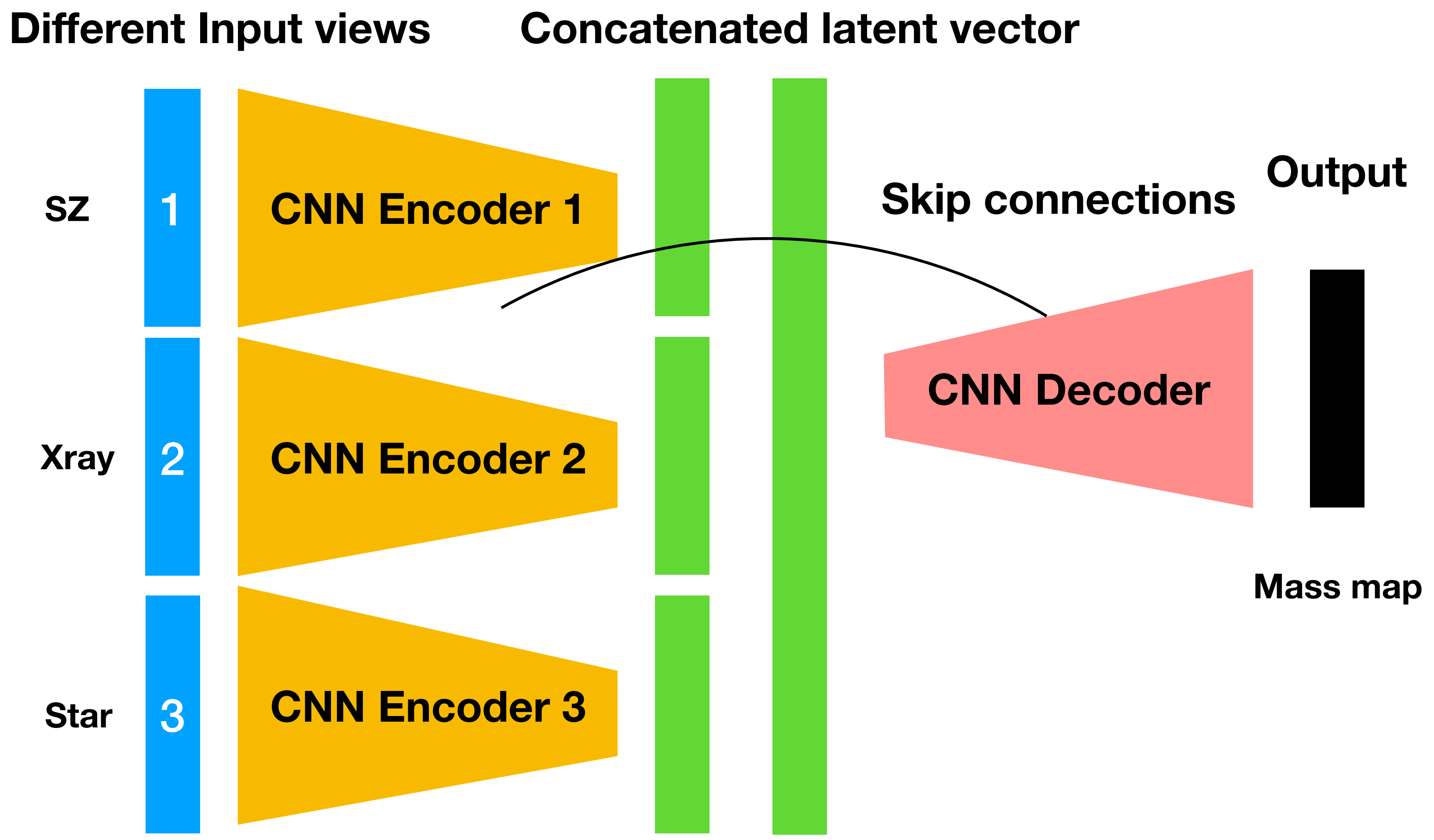
Model

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **528,** 1517–1530 (2024) https://doi.org/10.1093/mnras/stae071
Advance Access publication 2024 January 9

The three hundred project: mapping the matter distribution in galaxy clusters via deep learning from multiview simulated observations

- We **trained 5 models** in our work **depending on the input**, each a variation of the U-Net architecture.
- We have **three single-input models** star, SZ, and X-ray. For instance, the mass inferred from only star maps.
- Two **multi-input models**: one encoder or three encoders. These efficiently combine star, SZ and X-ray.
- At the end we have **5 different inferred mass maps** to compare with regarding of the input values or model. This is discussed in the results section.

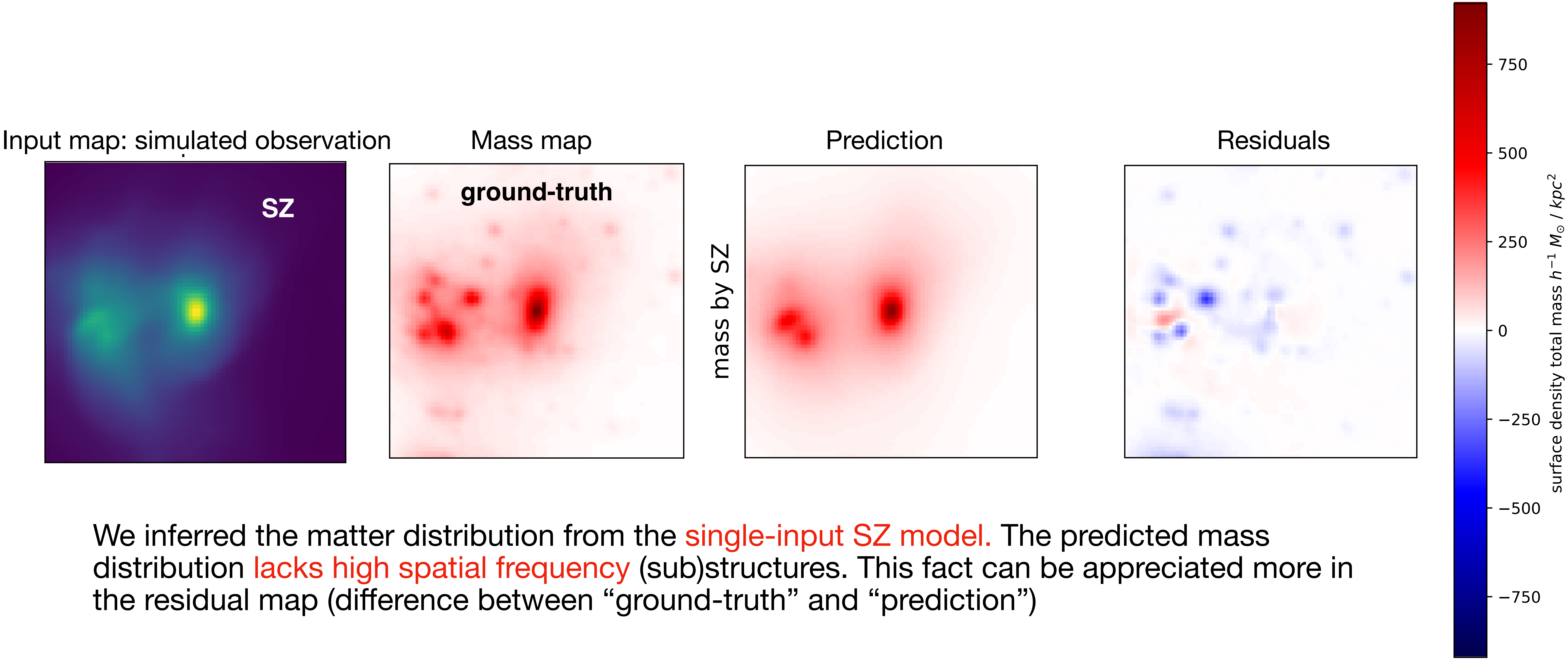
Model name	input maps	number of encoders
star	star	1
SZ	SZ	1
X-ray	X-ray	1
multi-1	star, SZ and X-ray	1
multi-3	star, SZ and X-ray	3



Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **528,** 1517–1530 (2024)
Advance Access publication 2024 January 9
https://doi.org/10.1093/mnras/stae071

The three hundred project: mapping the matter distribution in galaxy clusters via deep learning from multiview simulated observations

Results: SZ

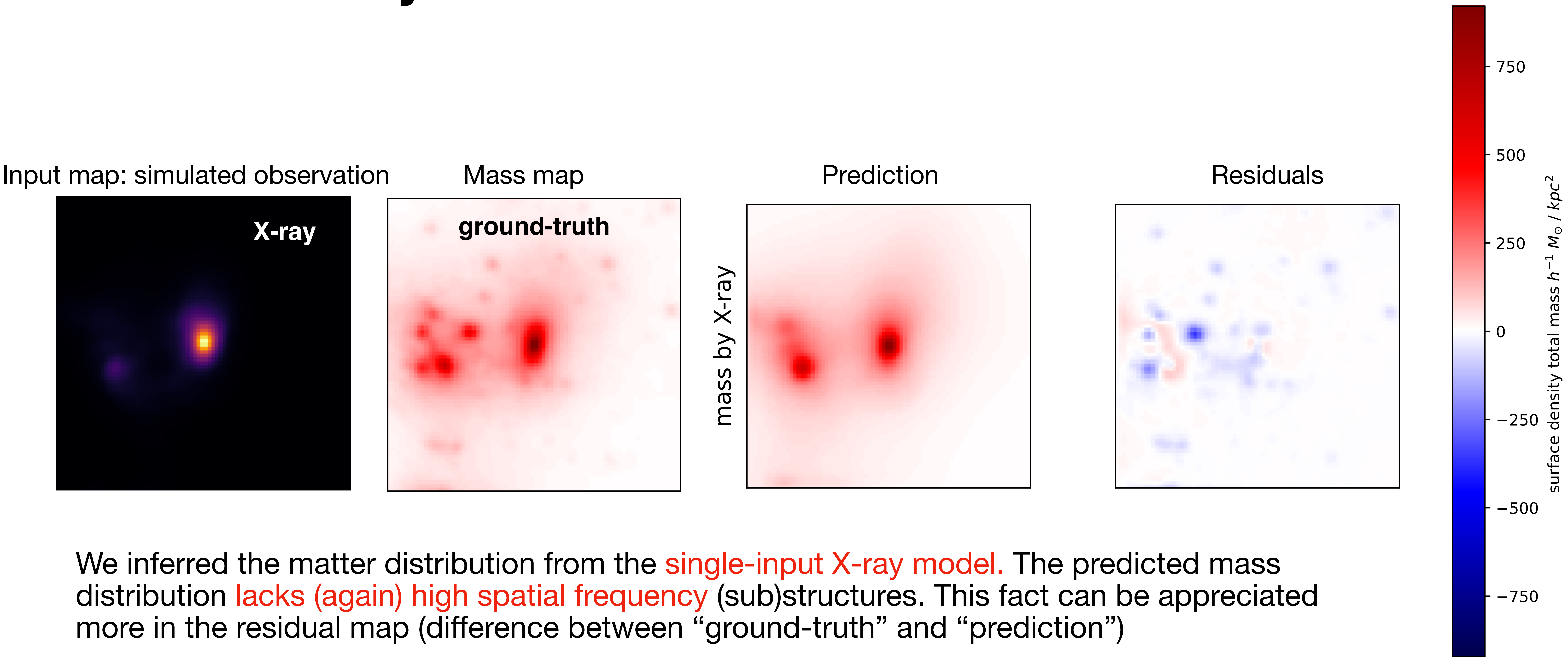


We inferred the matter distribution from the single-input SZ model. The predicted mass distribution lacks high spatial frequency (sub)structures. This fact can be appreciated more in the residual map (difference between “ground-truth” and “prediction”)


Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **528,** 1517–1530 (2024) https://doi.org/10.1093/mnras/stae071
Advance Access publication 2024 January 9

The three hundred project: mapping the matter distribution in galaxy clusters via deep learning from multiview simulated ol

Results: X-ray

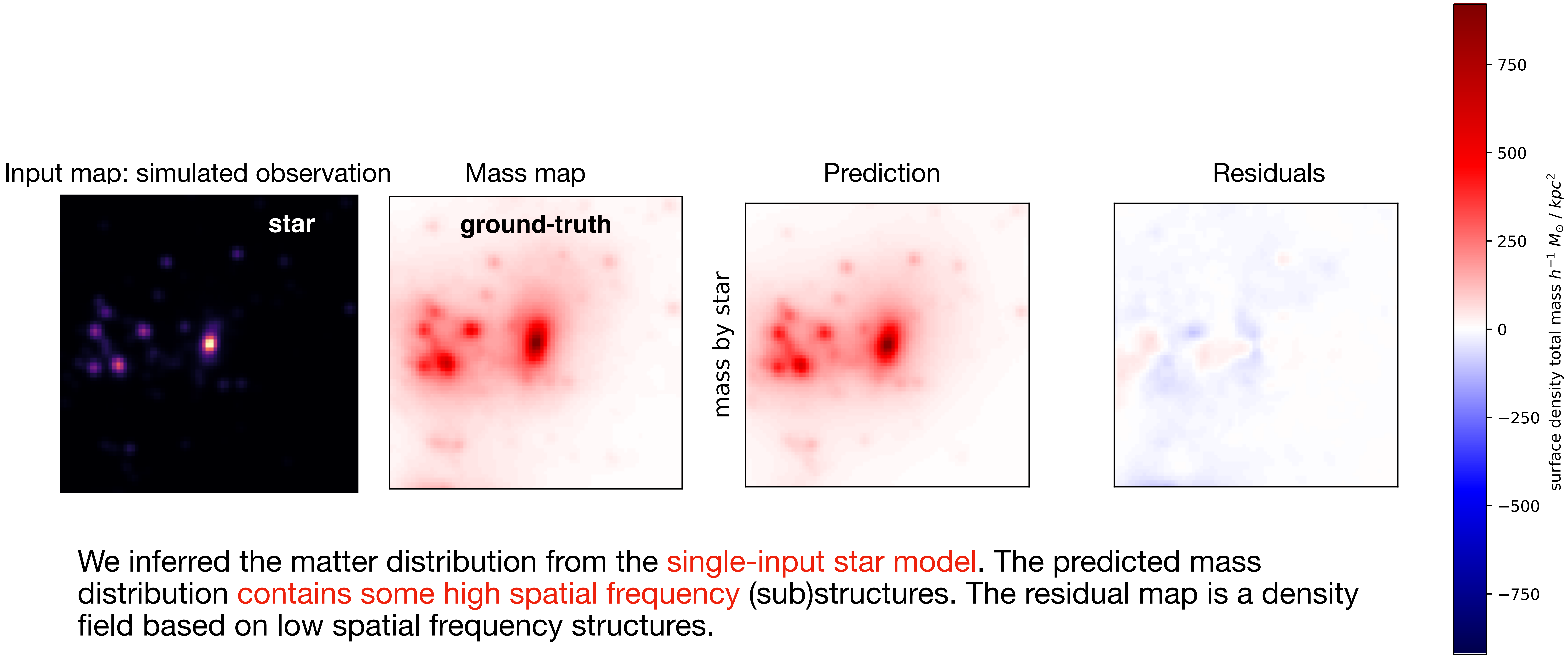


We inferred the matter distribution from the single-input X-ray model. The predicted mass distribution lacks (again) high spatial frequency (sub)structures. This fact can be appreciated more in the residual map (difference between “ground-truth” and “prediction”)

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **528,** 1517–1530 (2024) https://doi.org/10.1093/mnras/stae071
Advance Access publication 2024 January 9

Results: star

The three hundred project: mapping the matter distribution in galaxy clusters via deep learning from multiview simulated observations



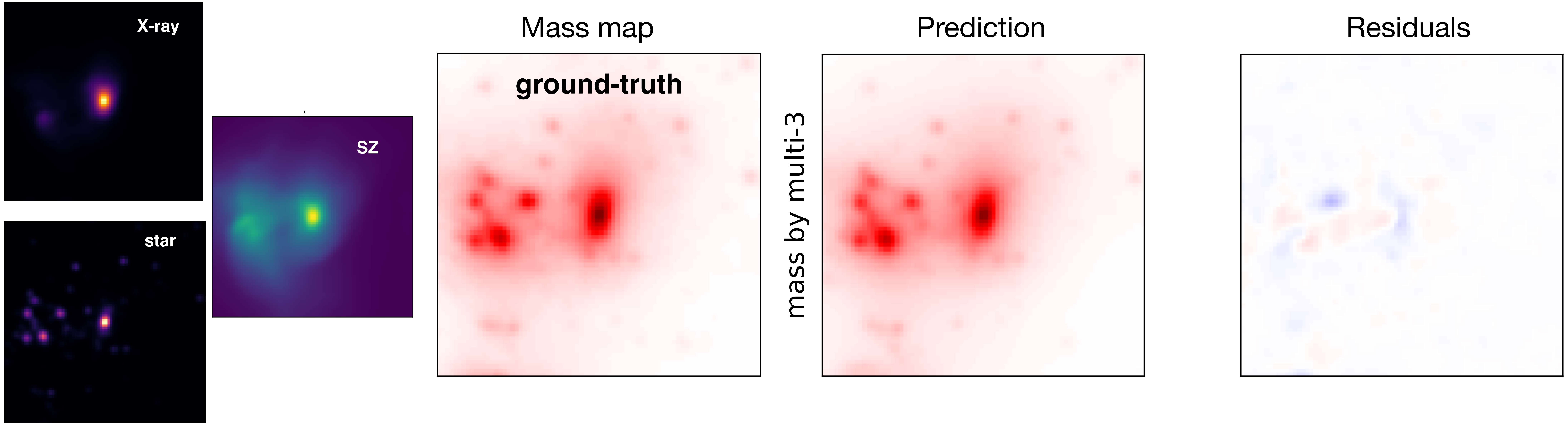
We inferred the matter distribution from the single-input star model. The predicted mass distribution contains some high spatial frequency (sub)structures. The residual map is a density field based on low spatial frequency structures.

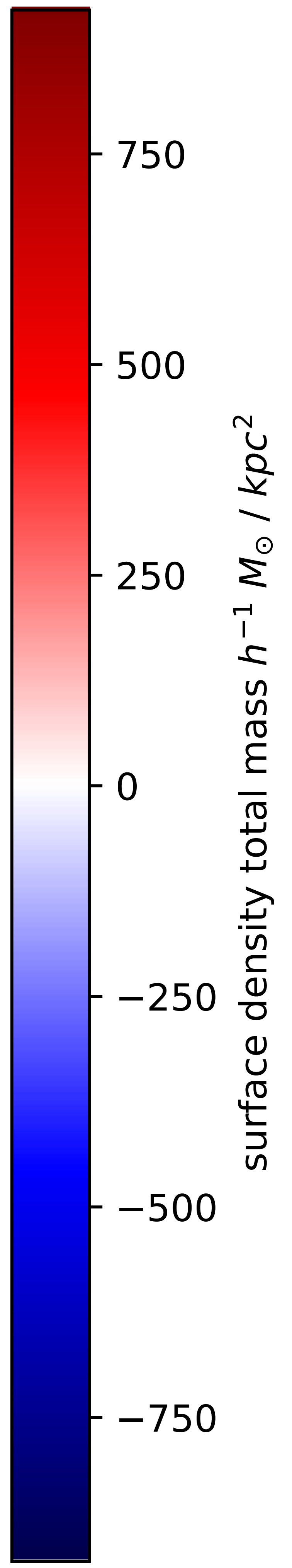
Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **528,** 1517–1530 (2024) https://doi.org/10.1093/mnras/stae071
Advance Access publication 2024 January 9

The three hundred project: mapping the matter distribution in galaxy clusters via deep learning from multiview simulated ob

Results: multi-3

Input map: simulated observation





We inferred the matter distribution from the multi-3 model: SZ, X-ray and stars. The predicted mass distribution contains both high spatial frequency (sub)structures and small spatial frequency structures. The overall scatter is reduced by a factor of 1/2.

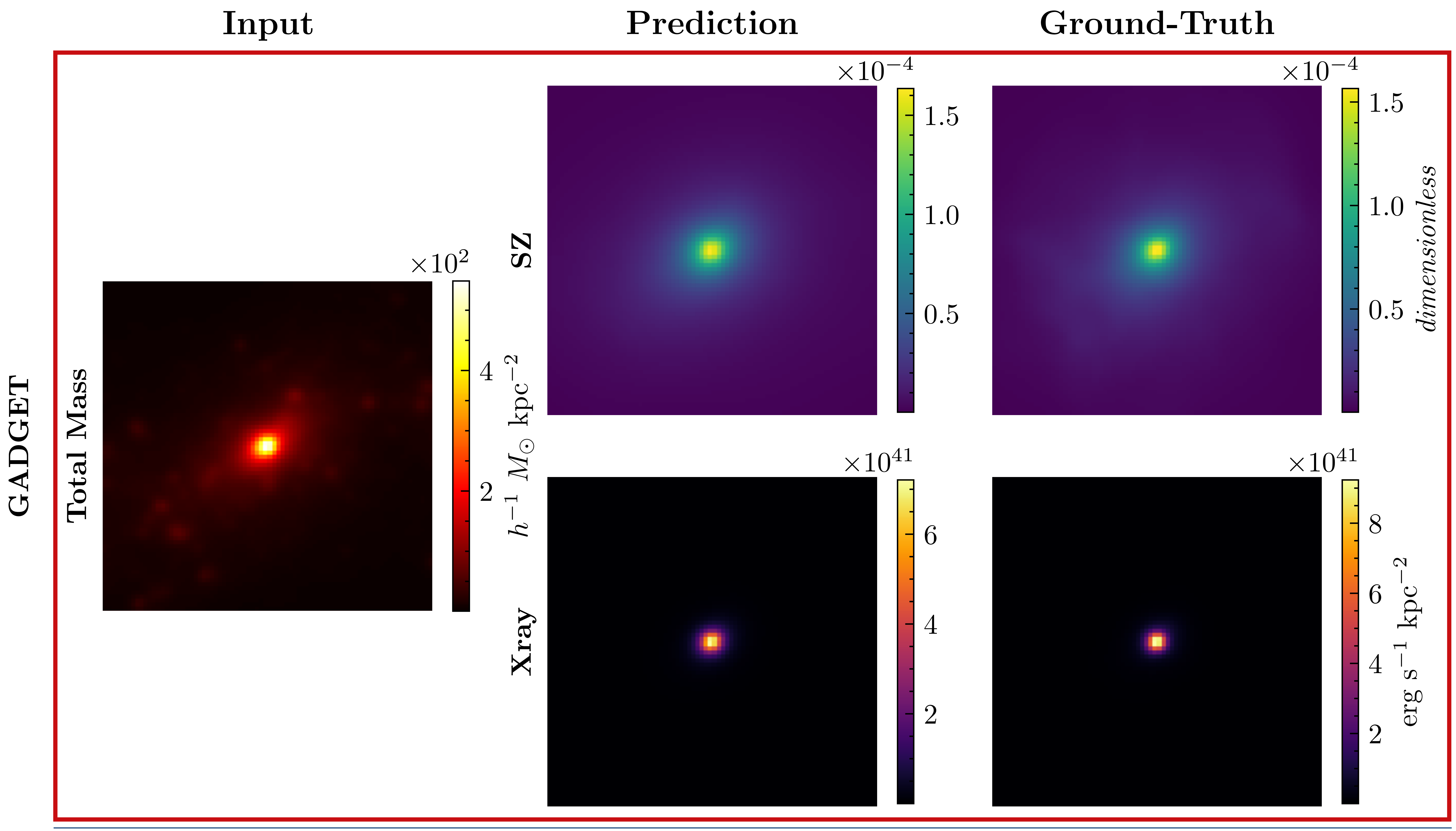
RASTI 000, 1–23 (2024) Preprint 8 October 2024 Compiled using rasti LaTeX style file v3.0

Deep Learning generated observations of galaxy clusters from dark-matter-only simulations

Submitted last week, preprint version. comments are welcome.

Andrés Caro[1,2]★, Daniel de Andres[1,2]†, Weiguang Cui[1,2,3]‡, Gustavo Yepes[1,2], Marco De Petris[4], Antonio Ferragamo[4], Félicien Schiltz[5] and Amélie Nef[5]

- Apply the U-Net model from mass to observations to generate **"fast" hydrodynamical simulations.**



RASTI 000, 1–23 (2024) Preprint 8 October 2024 Compiled using rasti LATEX style file v3.0

Deep Learning generated observations of galaxy clusters from dark-matter-only simulations

Andrés Caro[1,2]⋆, Daniel de Andres[1,2]†, Weiguang Cui[1,2,3]‡, Gustavo Yepes[1,2], Marco De Petris[4], Antonio Ferragamo[4], Félicien Schiltz[5] and Amélie Nef[5]

Model Notation	GIZMO dataset	GADGET dataset	Observable
U-Net GIZMO+GADGET	✓	✓	SZ & X-ray
U-Net GIZMO	✓	-	SZ & X-ray
U-Net GADGET	-	✓	SZ & X-ray

Train **3 models with different hydro** simulations:

- GADGET-X
- GIZMO-SIMBA
- GADGET-X+GIZMO-SIMBA

Test with **DM-only** simulations GADGET and MUSIC-DM

RASTI 000, 1–23 (2024) Preprint 8 October 2024 Compiled using rasti LaTeX style file v3.0

Deep Learning generated observations of galaxy clusters from dark-matter-only simulations

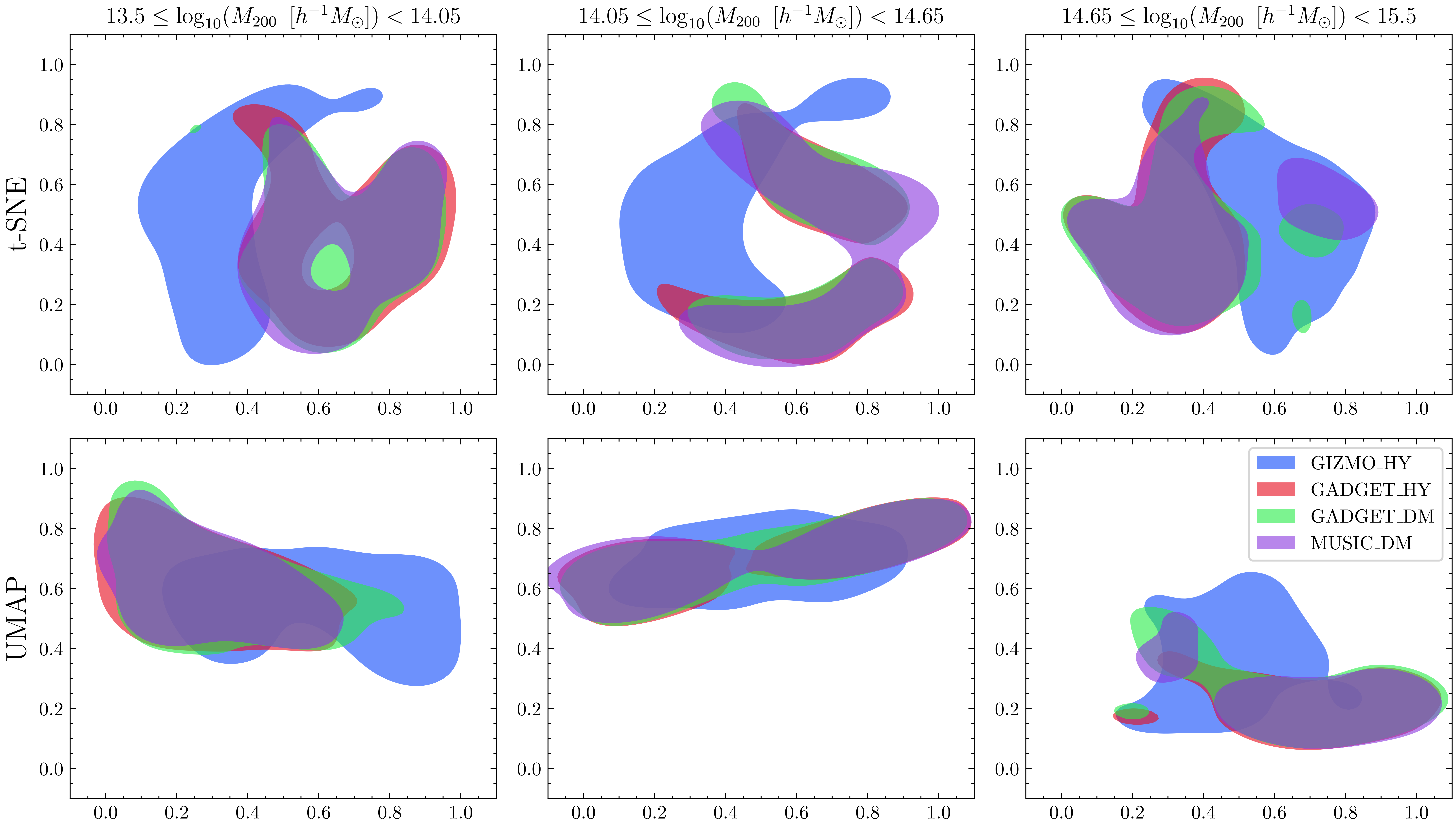
Andrés Caro[1,2]★, Daniel de Andres[1,2]†, Weiguang Cui[1,2,3]‡, Gustavo Yepes[1,2], Marco De Petris[4], Antonio Ferragamo[4], Félicien Schiltz[5] and Amélie Nef[5]

Model Notation	GIZMO dataset	GADGET dataset	Observable
U-Net GIZMO+GADGET	✓	✓	SZ & X-ray
U-Net GIZMO	✓	-	SZ & X-ray
U-Net GADGET	-	✓	SZ & X-ray

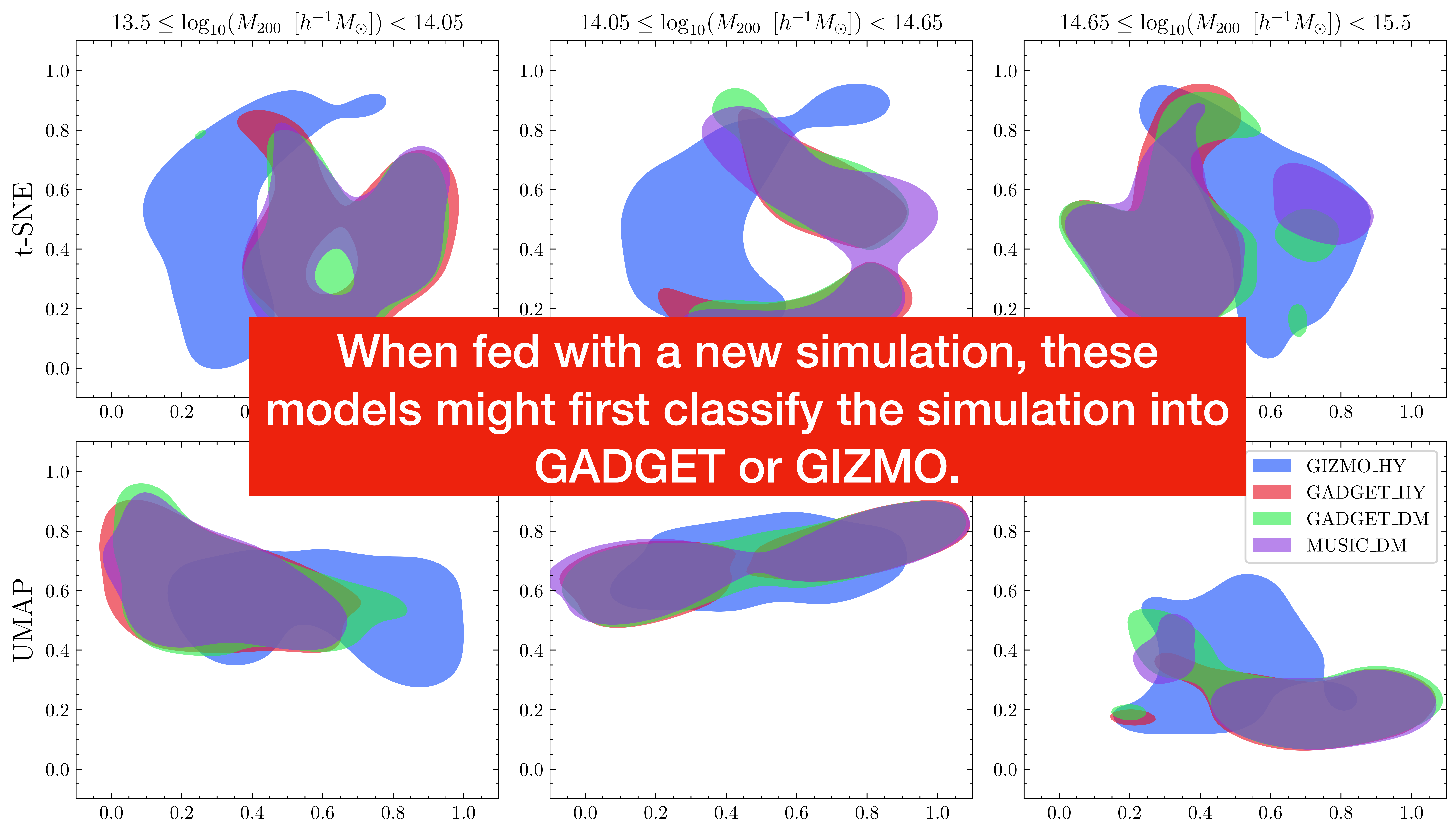
- GADGET-X+GIZMO-SIMBA

Test with **DM-only** simulations GADGET and MUSIC-DM

Generated maps follow the pixel distribution of GADGET, ignoring GIZMO-SIMBA, why?



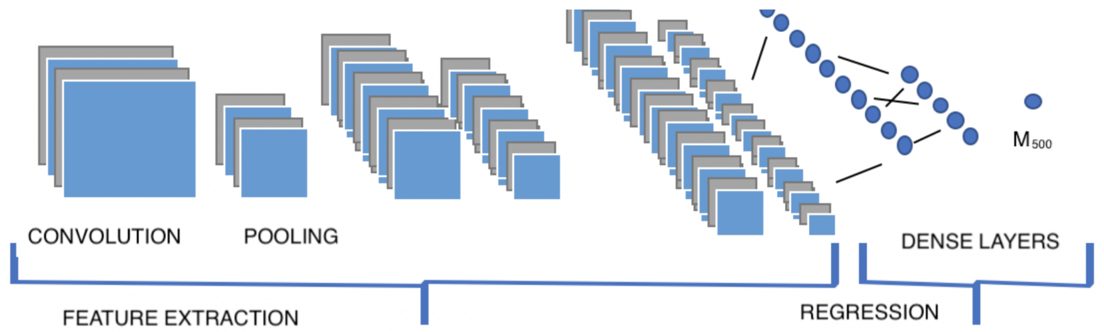
$13.5 \leq \log_{10}(M_{200}\ [h^{-1}M_\odot]) < 14.05$
$14.05 \leq \log_{10}(M_{200}\ [h^{-1}M_\odot]) < 14.65$
$14.65 \leq \log_{10}(M_{200}\ [h^{-1}M_\odot]) < 15.5$
t-SNE
UMAP
GIZMO_HY
GADGET_HY
GADGET_DM
MUSIC_DM
0.0
0.2
0.4
0.6
0.8
1.0

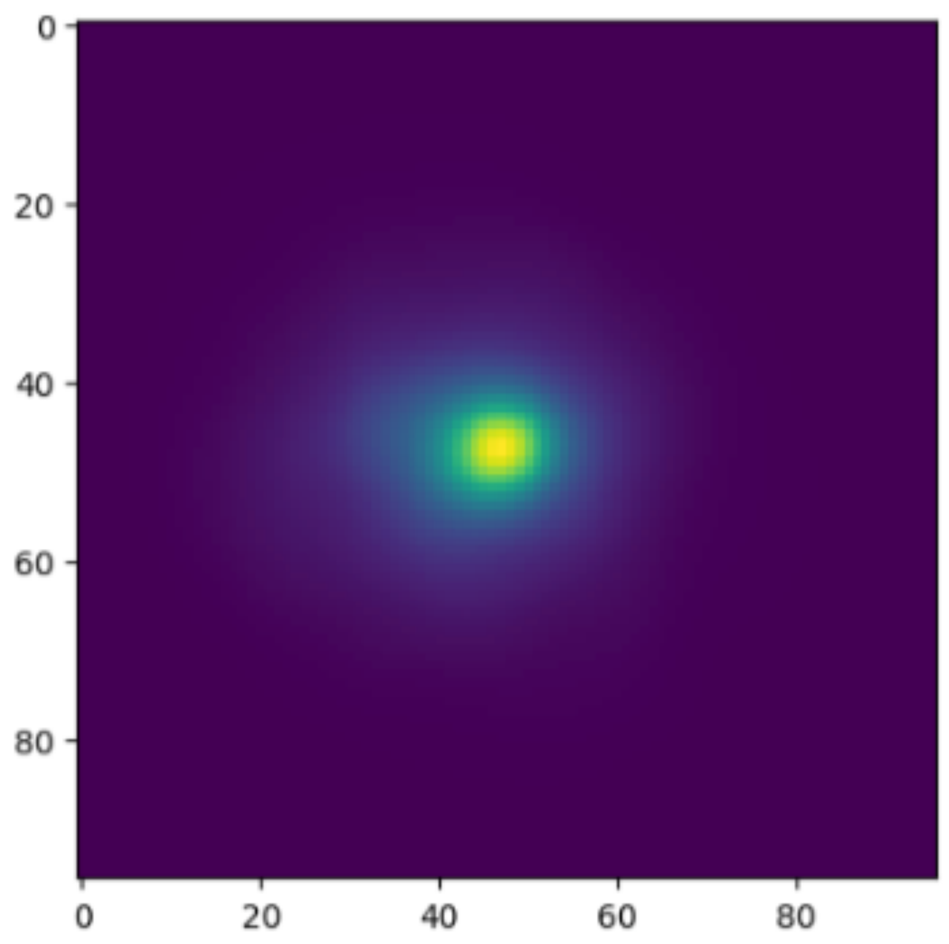


When fed with a new simulation, these models might first classify the simulation into GADGET or GIZMO.

Work in progress: **Domain adaptation methods**

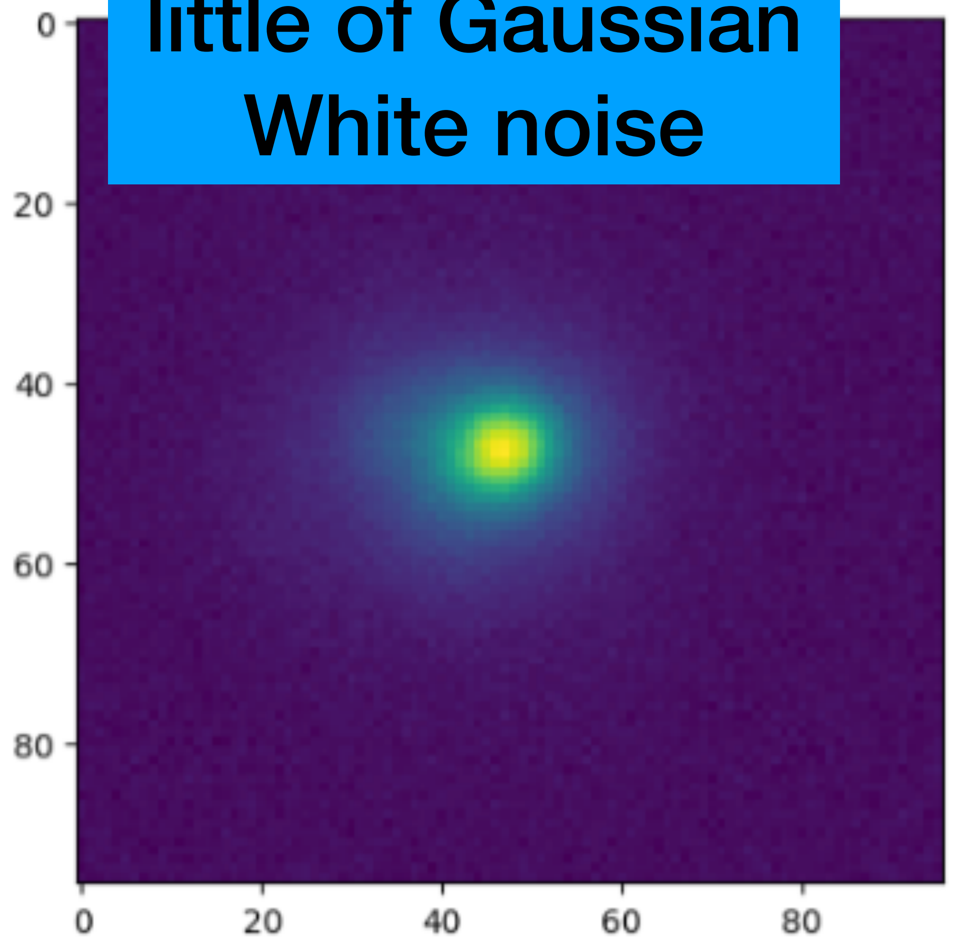
- De Andres et al 2022 masses of galaxy clusters are inferred form SZ maps



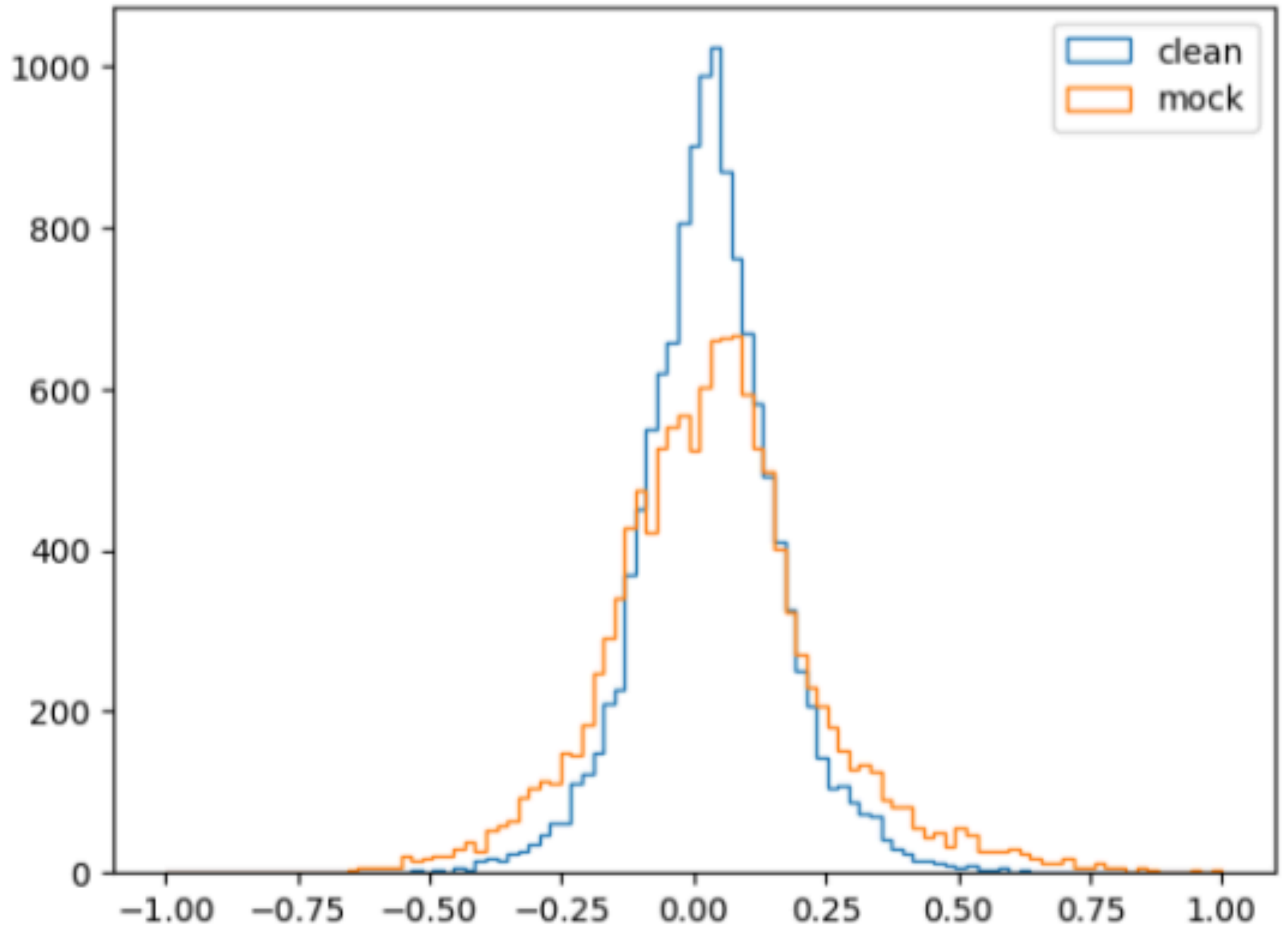


Train with

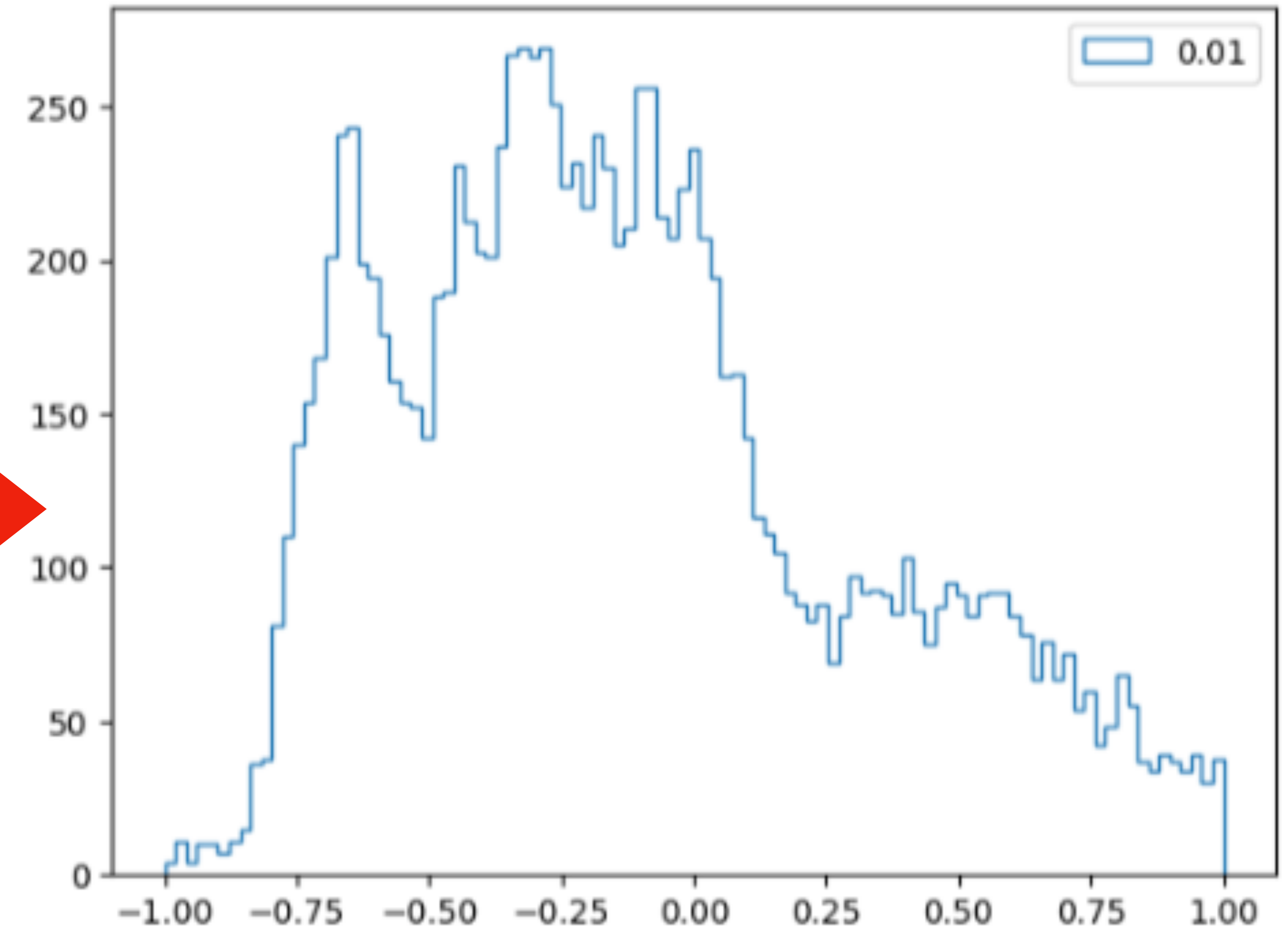
After adding a little of Gaussian White noise



Test with



Expected results for the bias distribution

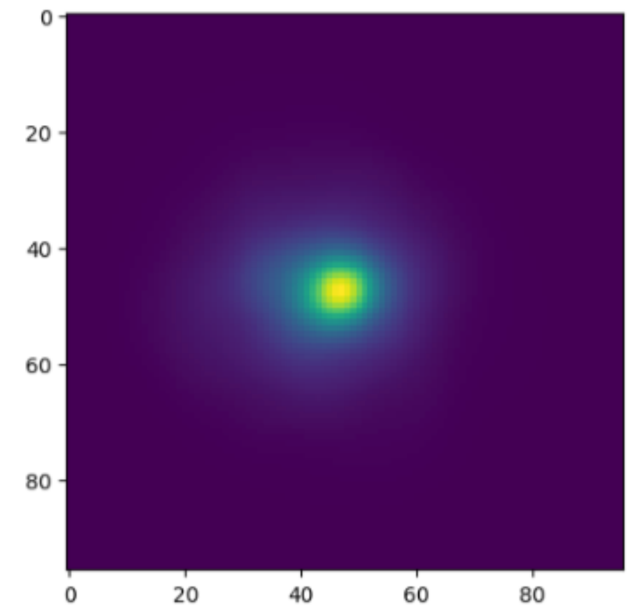


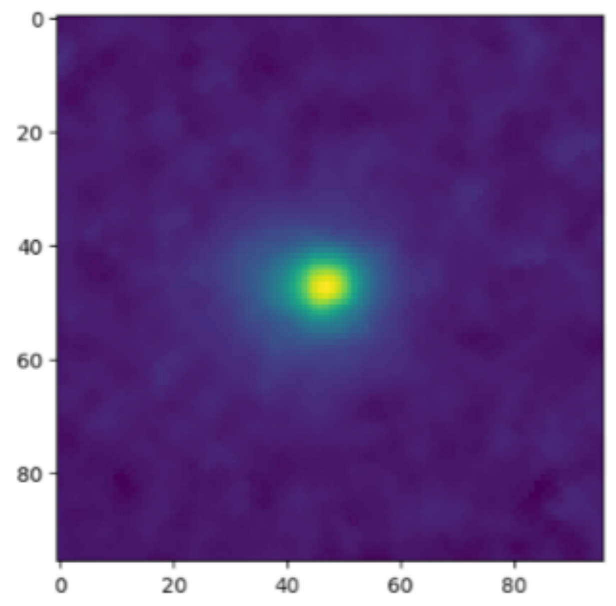
$(M_{CNN} - M_{true})/M_{CNN}$

Bias distribution is destroyed

Work in progress: **Domain adaptation methods**

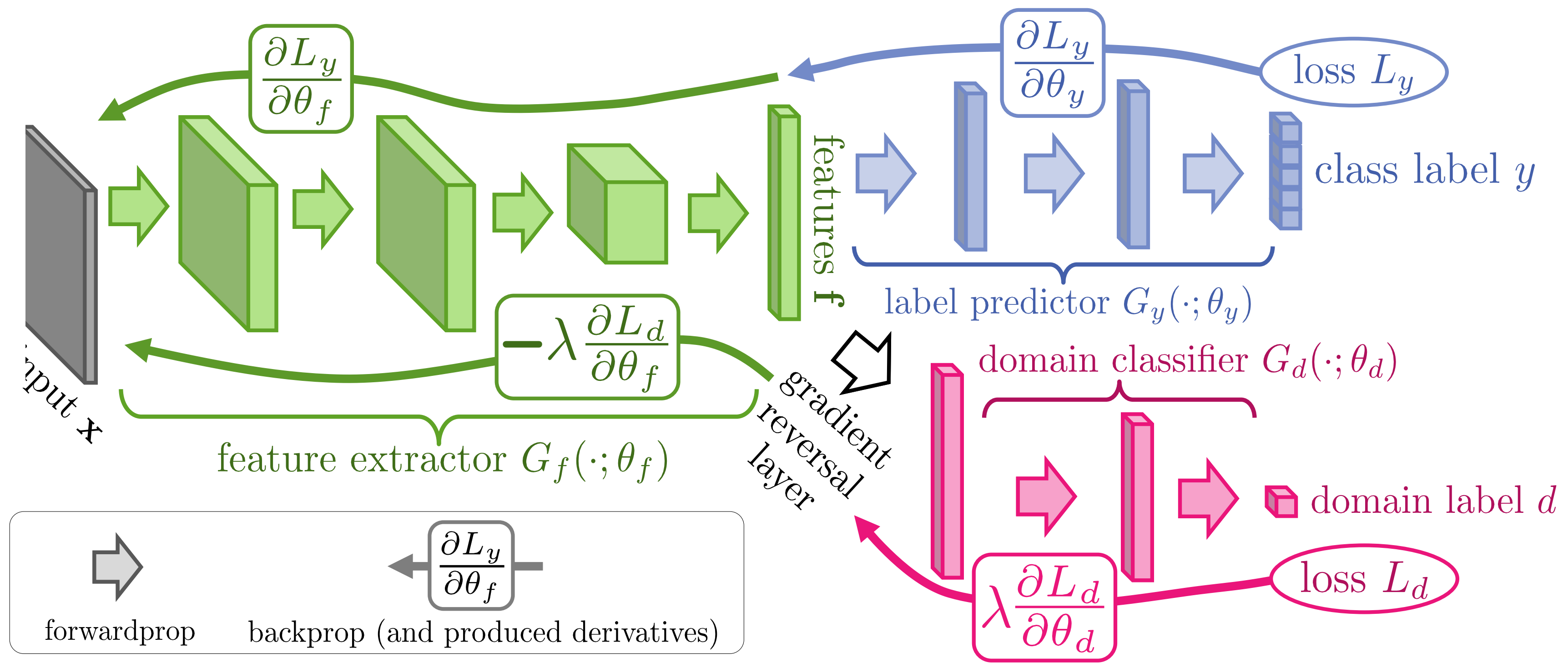
Simulation





Reality

- DANN model

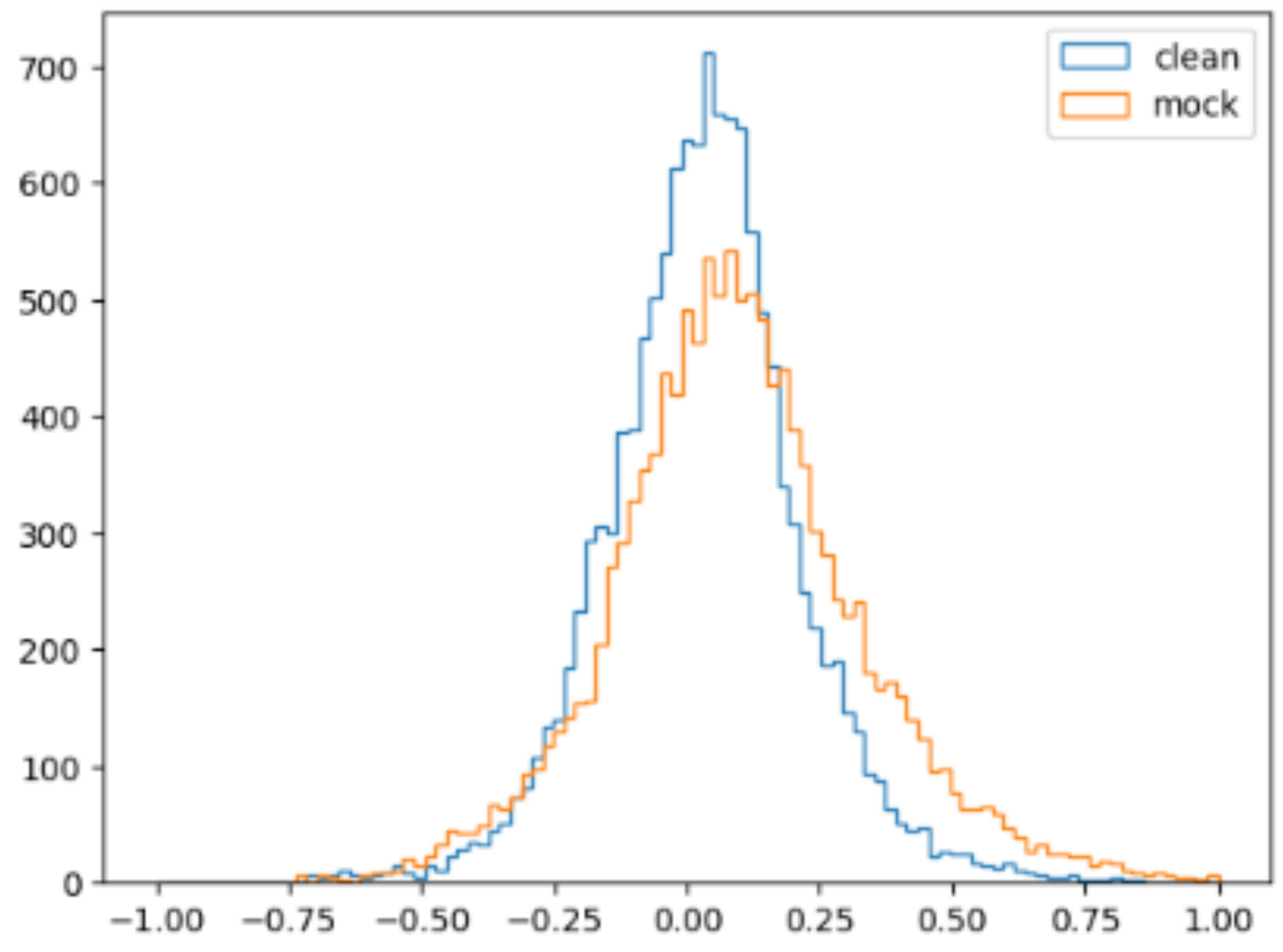


No need to make realistic simulations any more

- **Find common features in the latent space:**
- **Simulation=signal**
- **Real universe= signal+instrumental effects**

- **Ignore instrumental effects.**

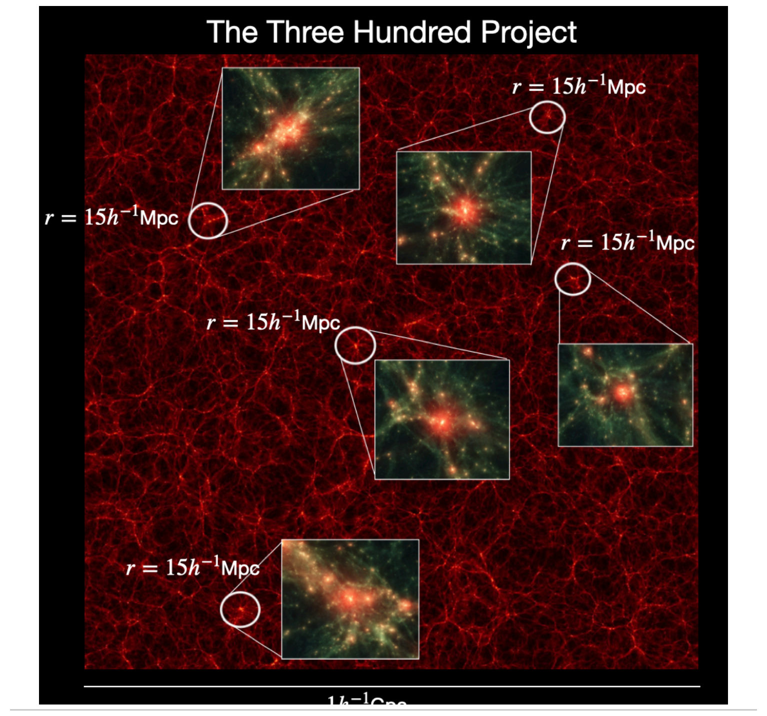
First tests very promising. Bias in **agreement with previous results,** in which instrumental effects are **known.**



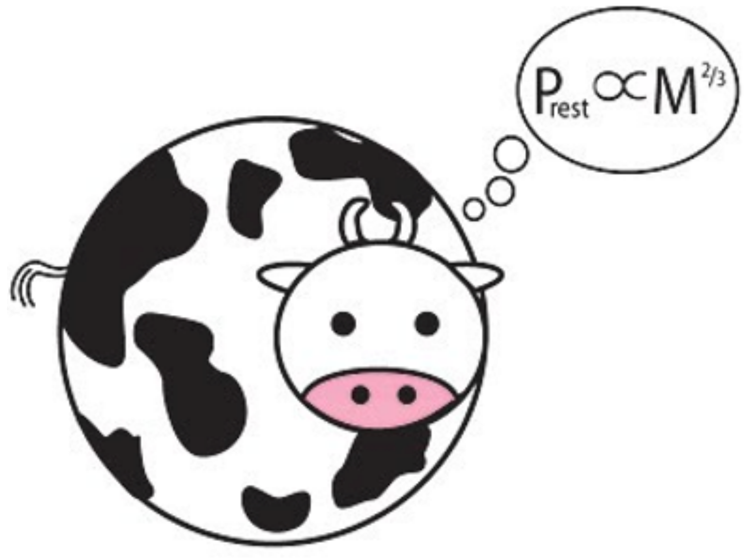
$(M_{CNN} - M_{true})/M_{CNN}$

- Summary:
- New method. **Zero knowledge of the instrumental effects.**
- Theoretically is **learning to use common properties**, therefore taking the signal only.
- It could be used for **learning invariant representations across multiple simulations.**

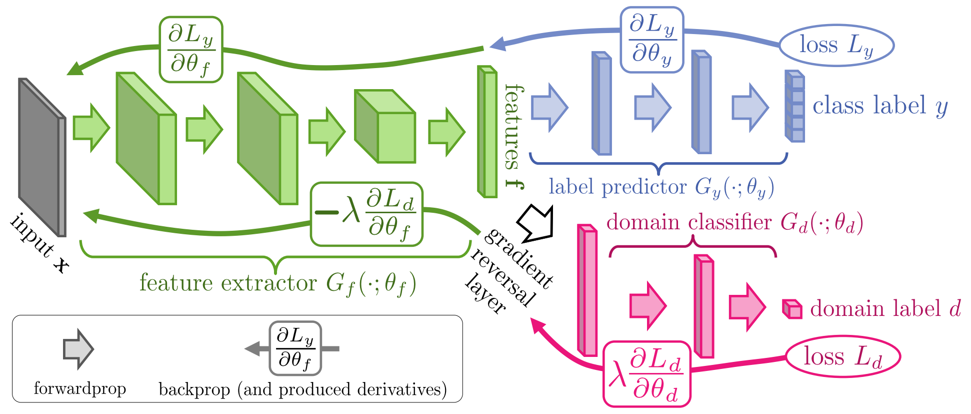
Summary



- **The Three Hundred Project**: Cosmological hydrodynamical zoom-in simulations with **good statistics of massive galaxy clusters** $\sim 10^{15} M_{\odot}$ and **different baryonic physics** models.



- Perfect database for **training** deep learning models that go beyond classical methods.



- The challenge and our **objective** is to apply models which are **trained with simulations to real data**. **Domain Adaptation** techniques address this problem.