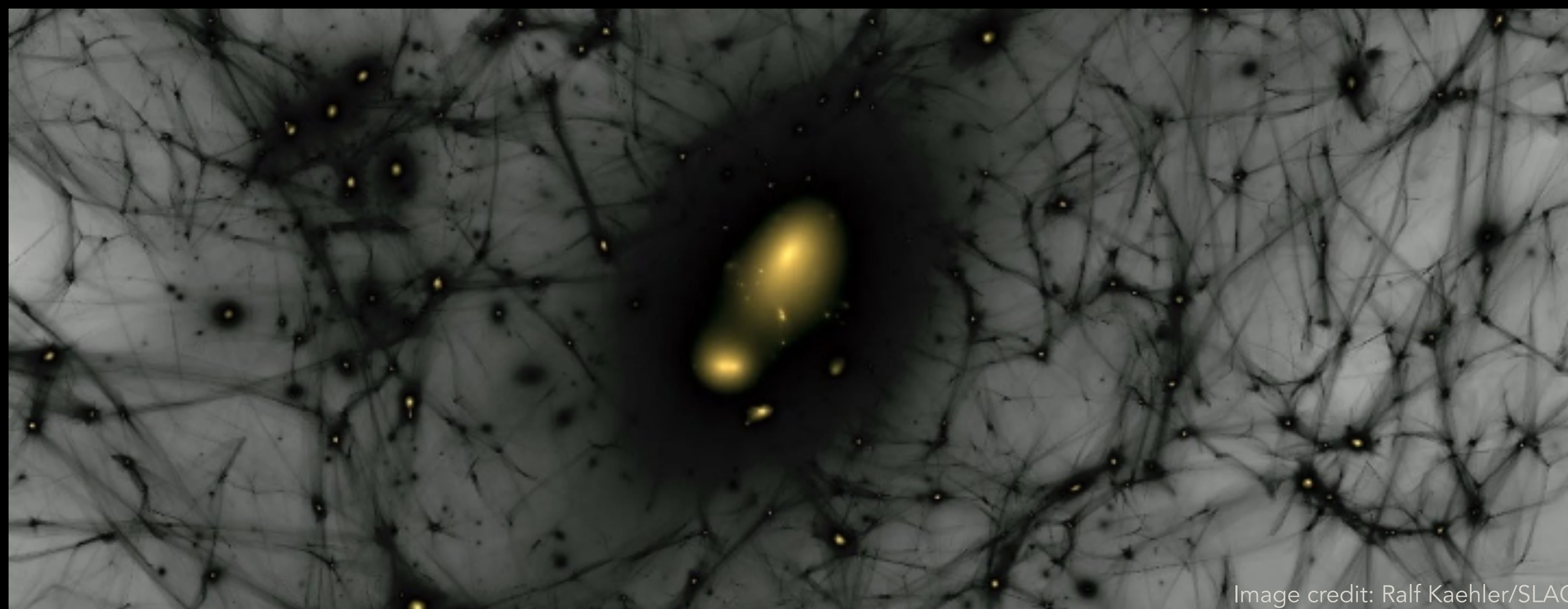
UNVEILING THE HALO-GALAXY CONNECTION WITH MACHINE LEARNING

Cosmo21, Chaniá, May 2024



Natalí de Santi
USP → Flatiron & Berkeley

Natália Rodrigues
USP → ?

Antonio Montero-Dorta
UTSM/Chile

Raul Abramo

Physics Institute, University of São Paulo

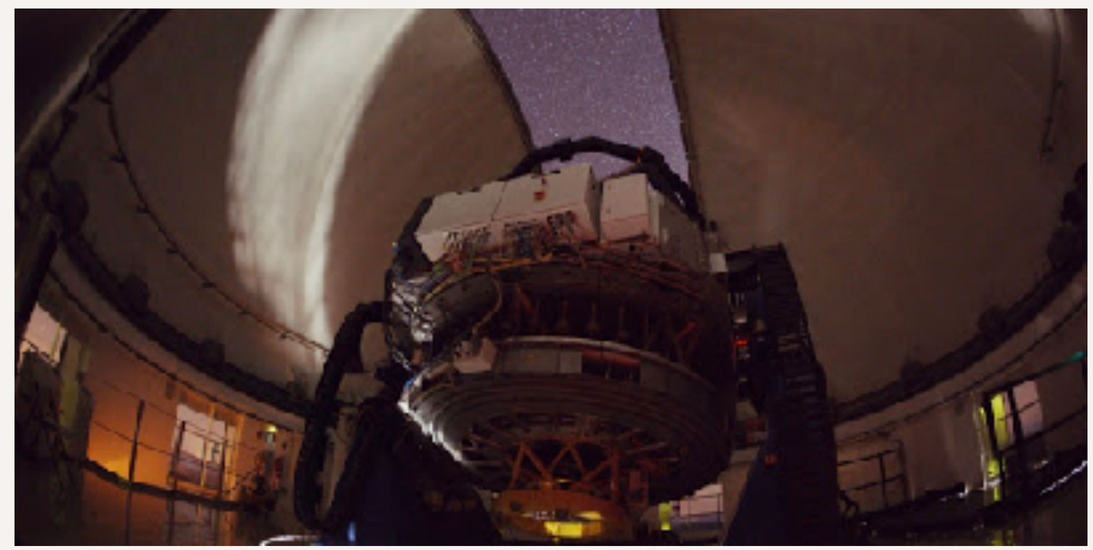
Physics Institute

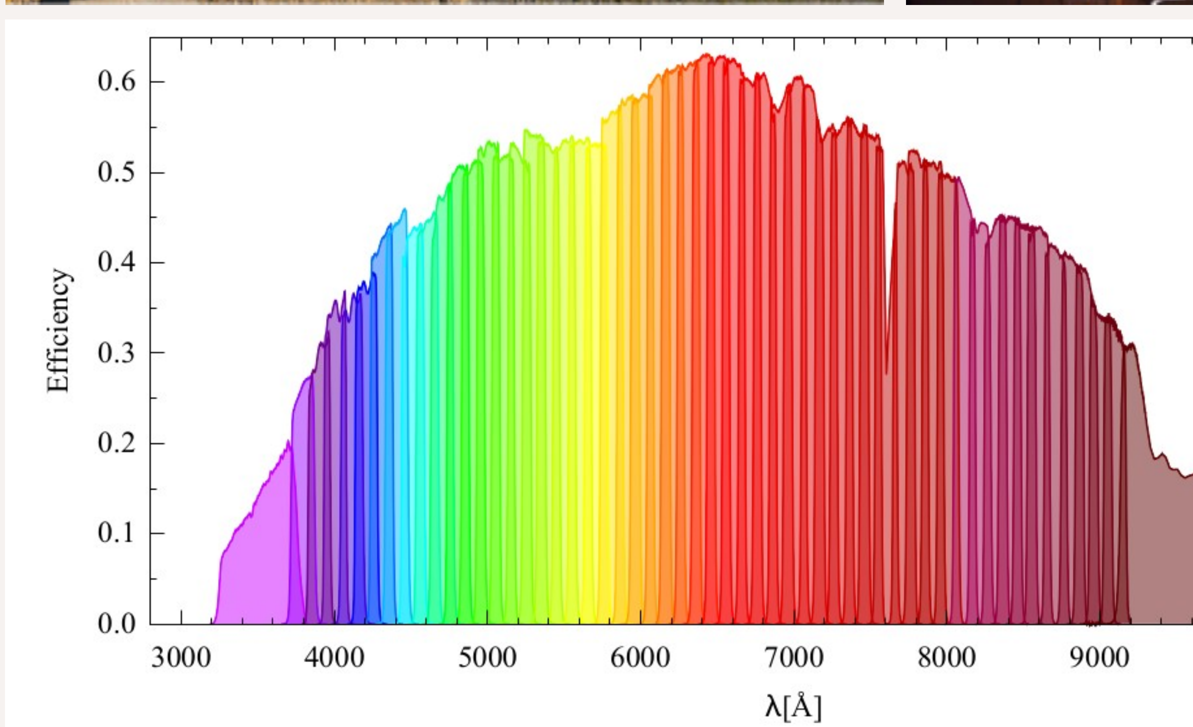
FAPESP

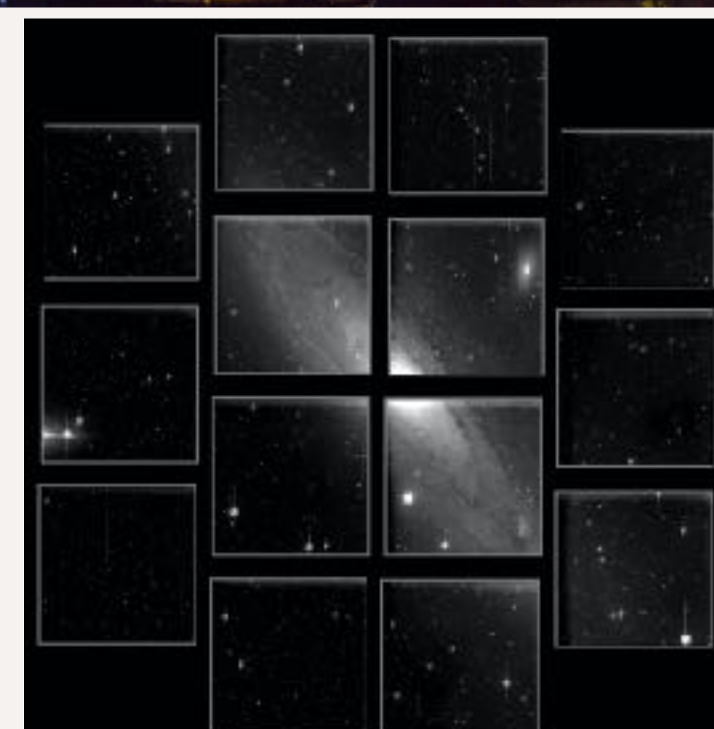
University of São Paulo

(my) Motivation: the J-PAS and WEAVE-QSO surveys

- J-PAS is a **narrow-band optical survey** conducted from a 2.5m telescope in Spain — **www.j-pas.org**
- It is **now** (!) taking **images in 56 filters** of widths ~100 Å : ~300 deg^2 already observed (→ ~800 deg^2/year)
- First **public data release** by the end of 2024







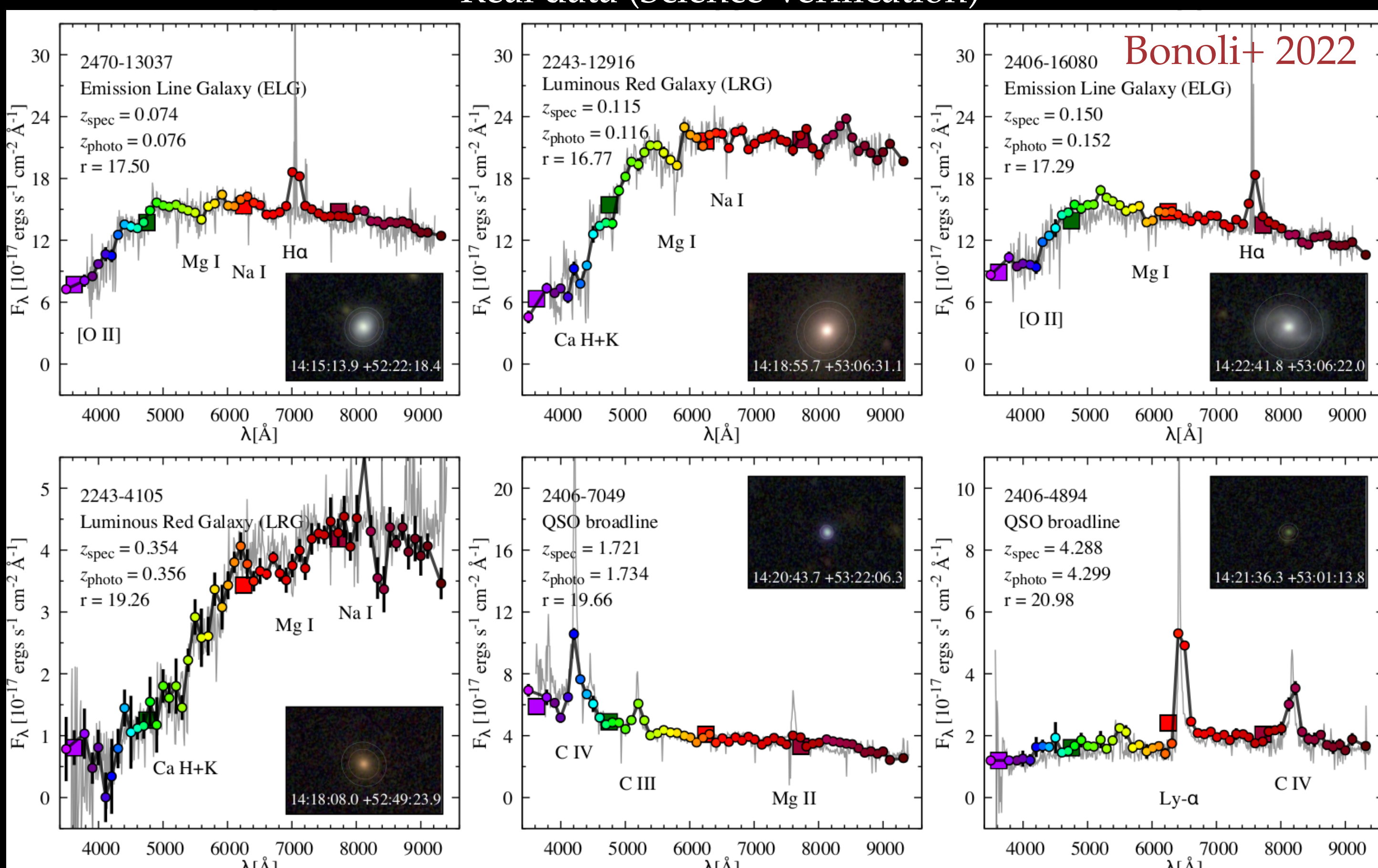
Bonoli+ 2022

N. Rodrigues, C. Queiroz
J-PAS & WEAVE-QSO
Quasar ID team

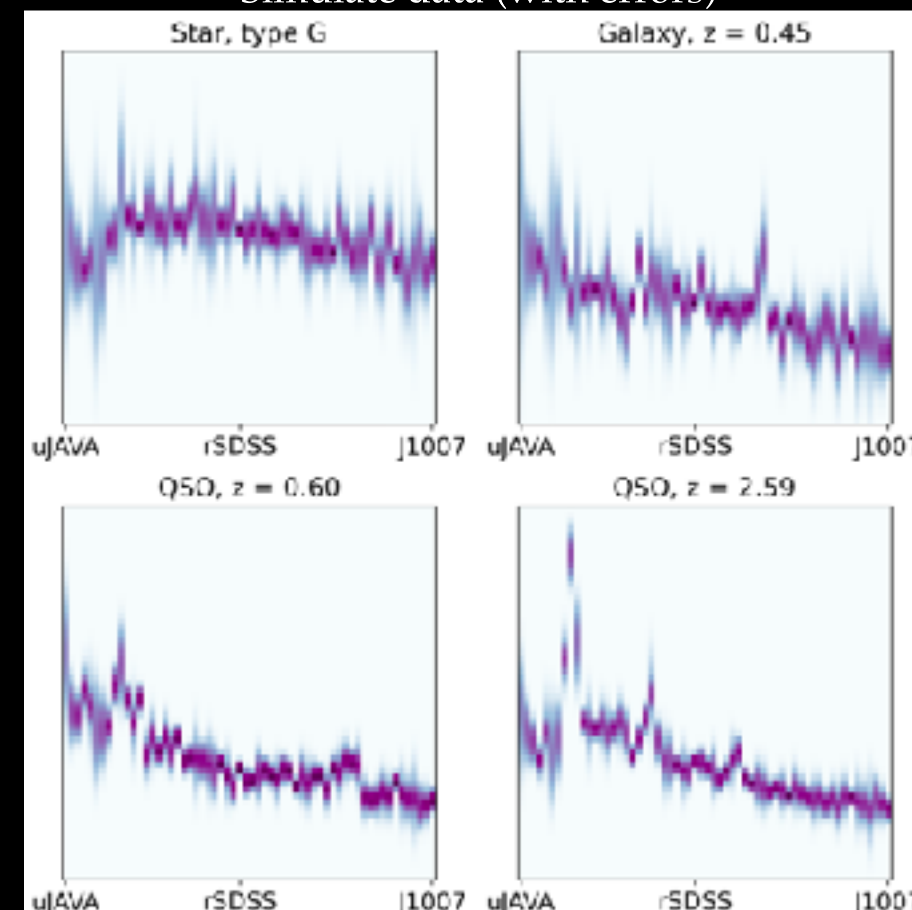
10^9 J-PAS sources: ML Classification

- ~ 5×10^4 objects/deg^2 with low-resolution spectra (*pseudo-spectra*)
- 10^4 galaxies/deg^2 with photo-zs $\sigma_z < 1\%$ + galaxy properties ⇒ superb **multi-tracer optical survey**
- ~ 200 QSOs/deg^2, of which ~75 at z>2.2
- J-PAS high-z QSOs will be observed by WEAVE (Ly-α forest w/ ~30% denser sampling comp. w/ DESI)

Real data (Science Verification)



Simulate data (with errors)

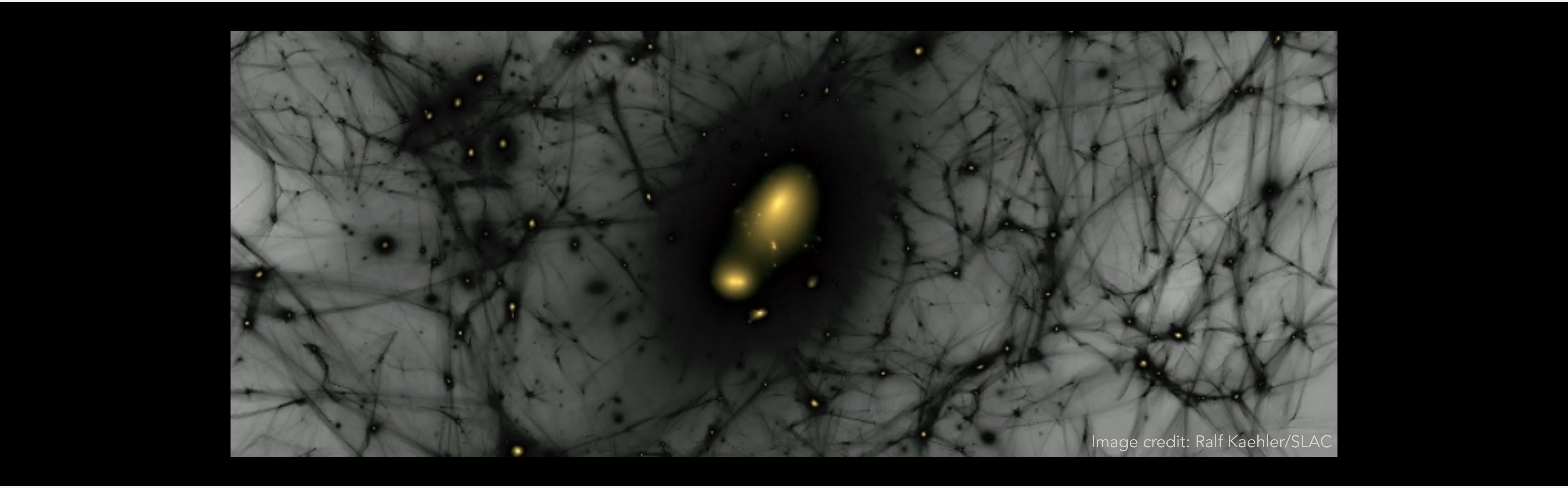


- We use ML to find those "needles in the haystack" (high-z QSOs are ~0.01% of our sources)
- For the training set, we **forward-simulated our data**, from LF and SDSS spectra down to J-PAS data-based flux/magnitudes with real uncertainties (Queiroz+ 2022) → CNNs & other ML methods (Rodrigues+ 2023, Pérez-Ràfols+ 2023)

From halos to galaxies using ML

+ Bia Tucci, Celeste Artale

- LSS relies heavily on **numerical simulations** in order to capture physical mechanisms on a wide **range of scales**



- N-body DM sims are complex enough, but actual **tracers** are even more so (baryonic physics, environments etc.)
- However… N-body + hydro sims are **very expensive**
- The precise relationships between **halos** and **galaxies** can be quite intricated (SAMs, SHAMs)
- Machine learning can **predict** with high accuracy how **central galaxies** form **inside halos**, depending on their **properties** and **environment.** We study these relations with the help of the **IllustrisTNG 300** hydro simulation.

IllustrisTNG: Springel+2018 ; Pillepich+2018 ; color bimodality: Nelson+2018 ; comp. w/. HSC galaxies Eisert+2023;

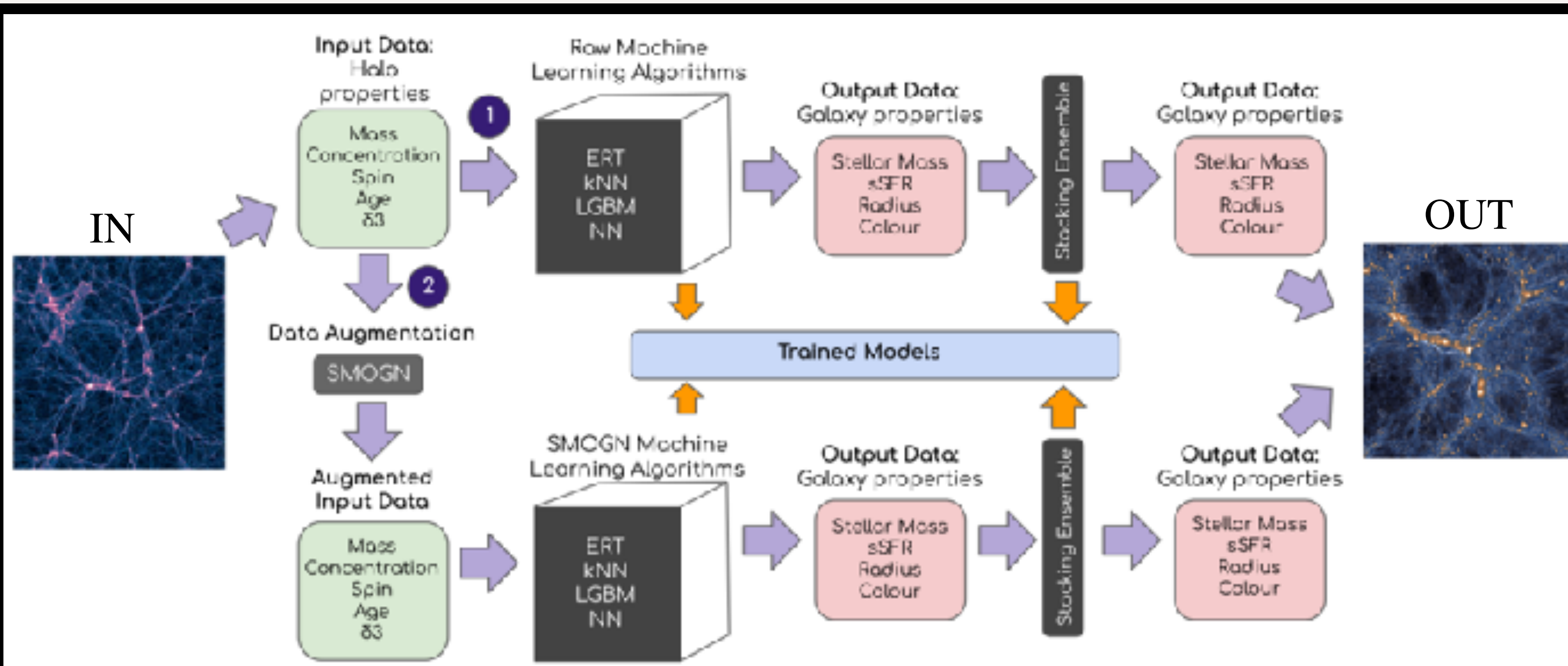
From halos to galaxies using ML: regression

+ Bia Tucci, Celeste Artale

- The detailed relations between tracers and halos are key to model **tracer bias** with **high accuracy**: e.g., assembly bias/secondary bias parameters (e.g., Lin+2016; Zehavi+ 2018; Montero-Dorta+ 2020; Wu+ 2024)

- Simple ML methods (e.g., NNs) can be trained to infer non-parametric relations between continuous variables in some input space (halo properties), and continuous variables in the output space (galaxy properties)

- In de Santi+ 2022 we used IllustrisTNG 300 to predict **central galaxy properties** from their **host halo properties** (can also be done with merger trees — Chuang+ 2024)



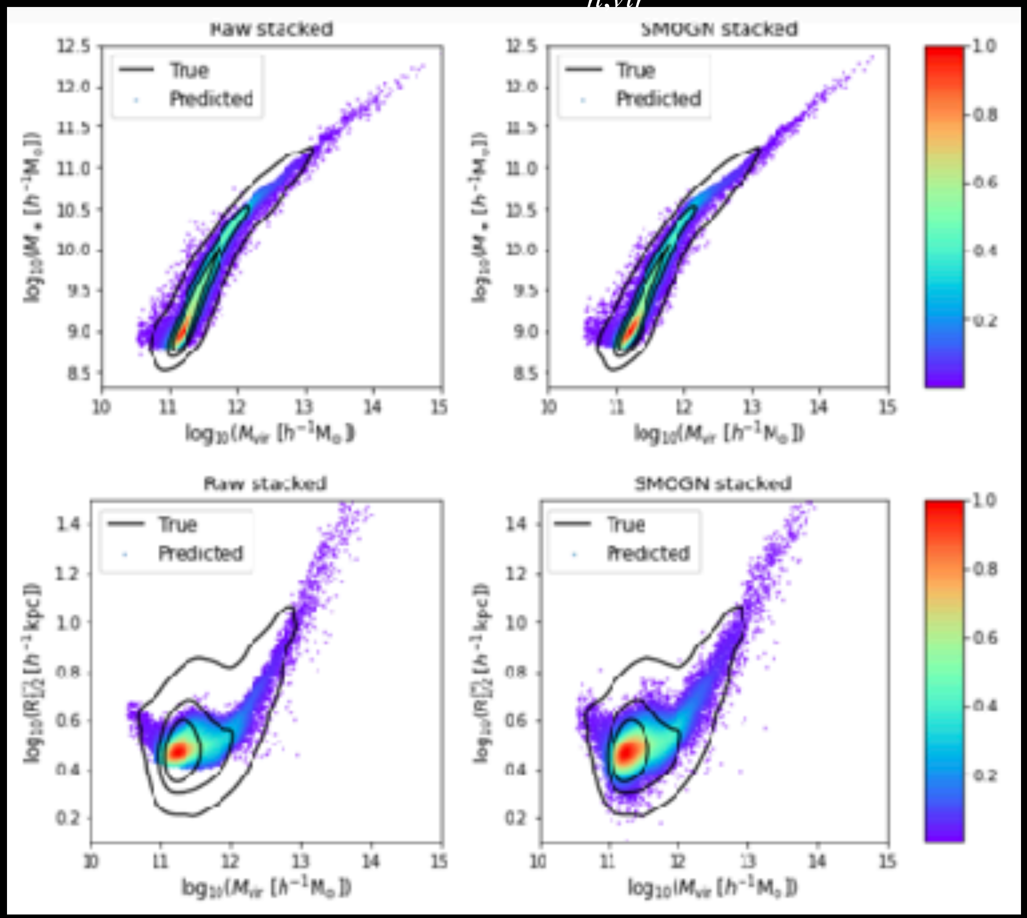
- In order to capture common as well as rare instances (e.g., high mass halos and galaxies) we used a **data augmentation** technique tailored for **imbalanced regression** problems (SMOGN - **S**ynthetic **M**inority **O**ver-Sampling technique for regression with **G**aussian **N**oise — Branco 2017 ; github.com/nickkunz/smogn)

From halos to galaxies using ML: regression

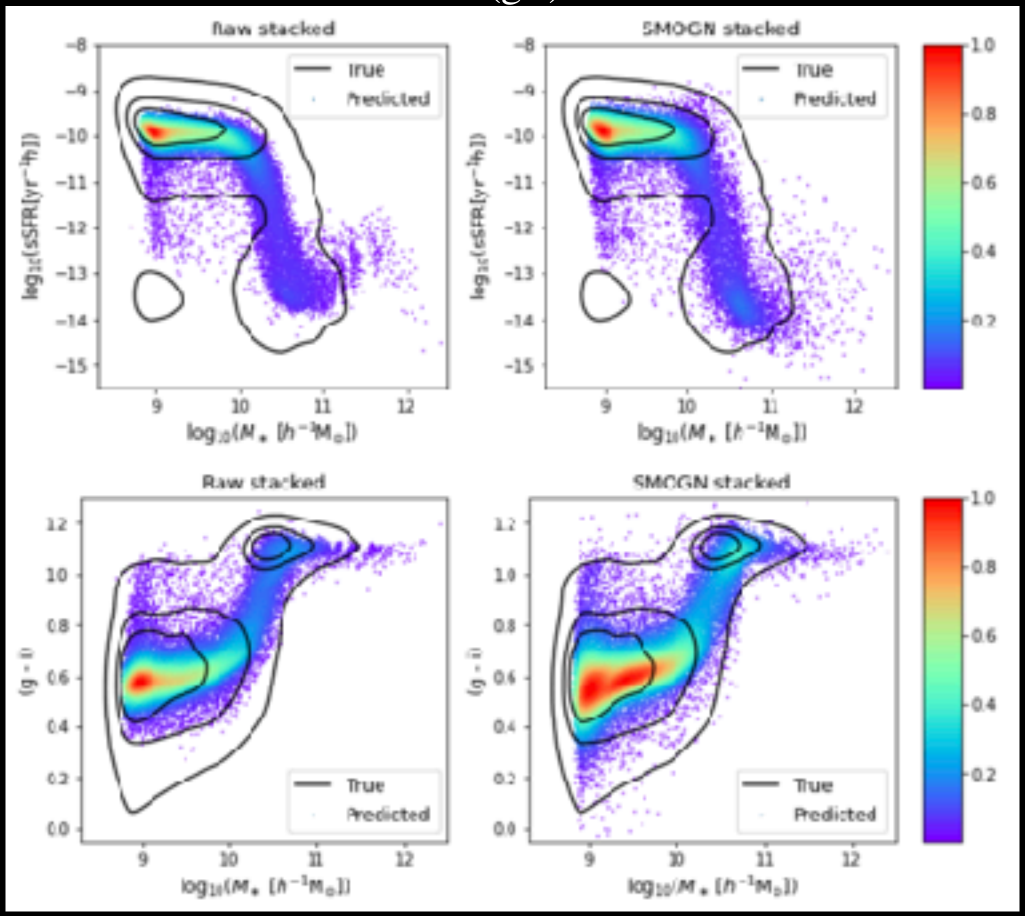
+ Bia Tucci, Celeste Artale

- Predictions for galaxy properties from halo properties at $z = 0$ (de Santi+ 2022)

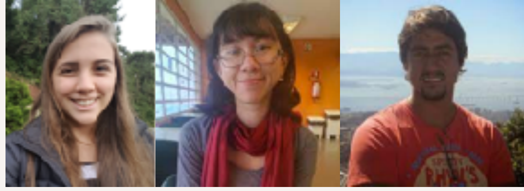
M_* & R vs. $M_{h,vir}$



sSFR & (g-i) vs. M_*

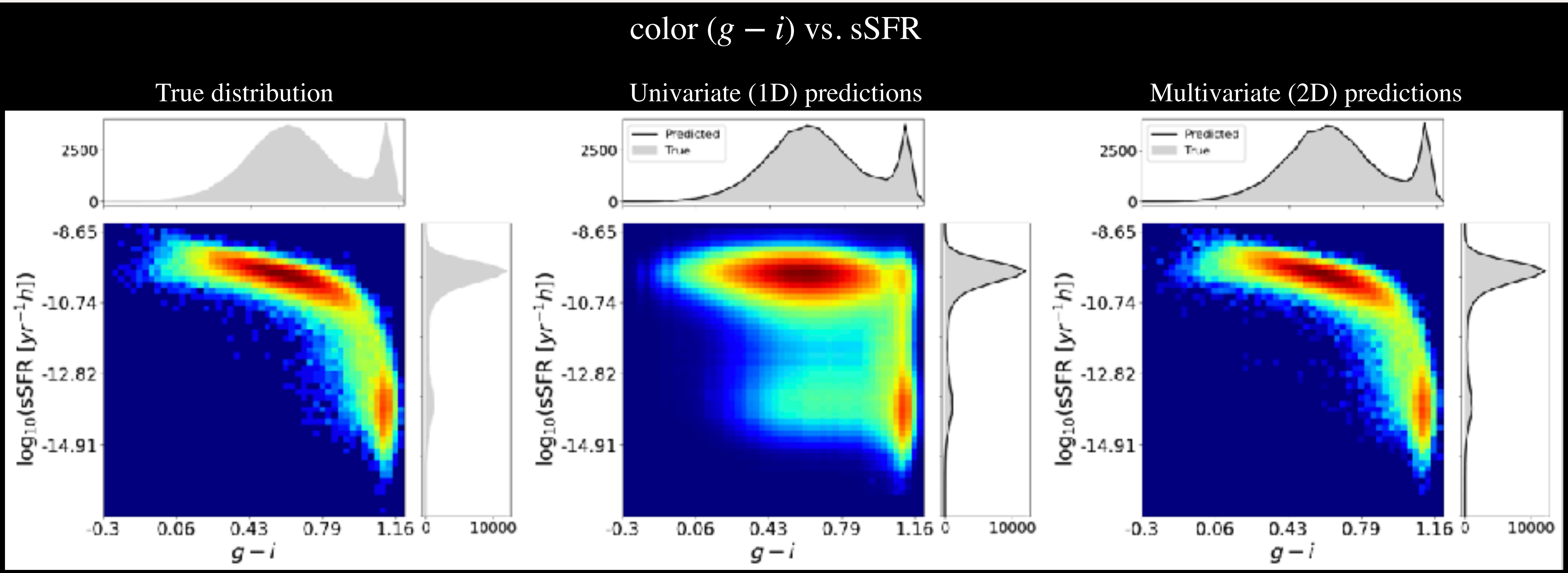


- The ML regression was able to reproduce the overall distribution of galaxy properties, **but**… there were some snags:
  - **Peaks** of the distribution were being **over-predicted** — and the **tails**, **under-predicted**
  - Each galaxy property was **trained independently** from the others
  - Method is **deterministic**



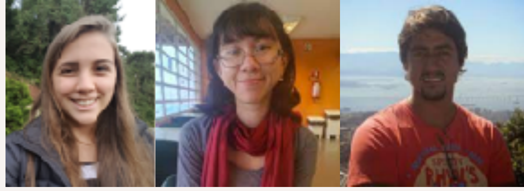
From halos to galaxies using ML: classification

- Galaxy properties are **correlated** — and, to some level, **stochastic**
- In *Rodrigues et al. 2023* we predicted the **joint distributions** of central galaxy properties (see also Alsing+2024)
- We initially achieved this by **splitting** the N-dimensional parameter space of galaxy properties into **discrete classes** ("cells"), and then applying an NN classifier which, given a halo, assign a **score** (~probability) to each class.



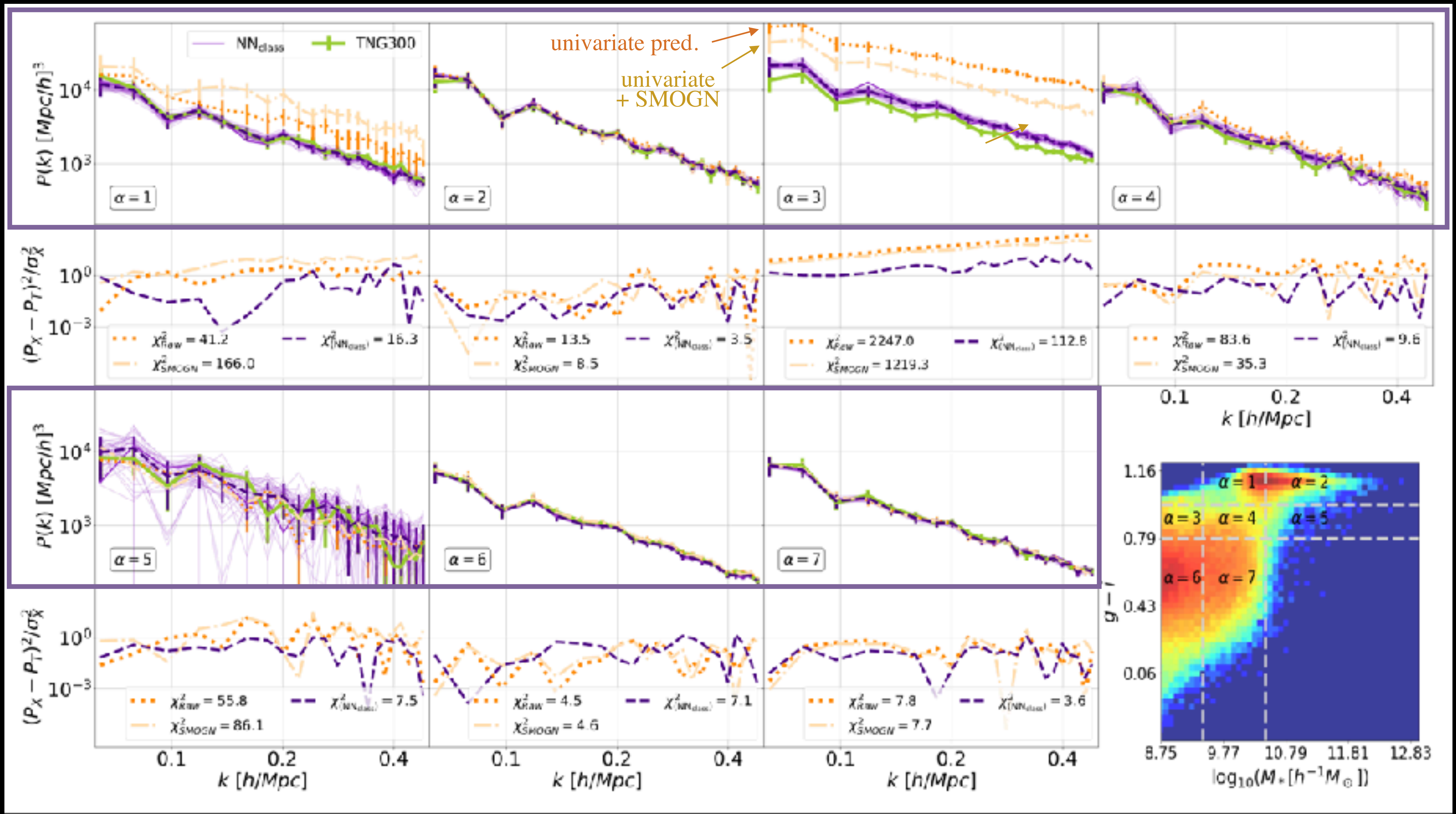
N. Rodrigues et al., MNRAS 2023

From halos to galaxies using ML: classification

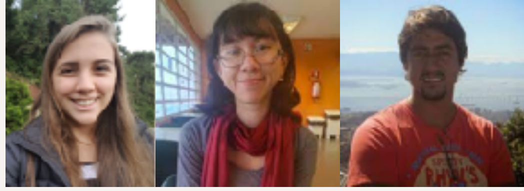


- We can now **split** galaxies into several **different populations of tracers**, each one with **very well-defined biases** — reproducing with high accuracy the **clustering** of those galaxies.

- **Joint estimation** is critical to correctly reproduce **tracer bias**



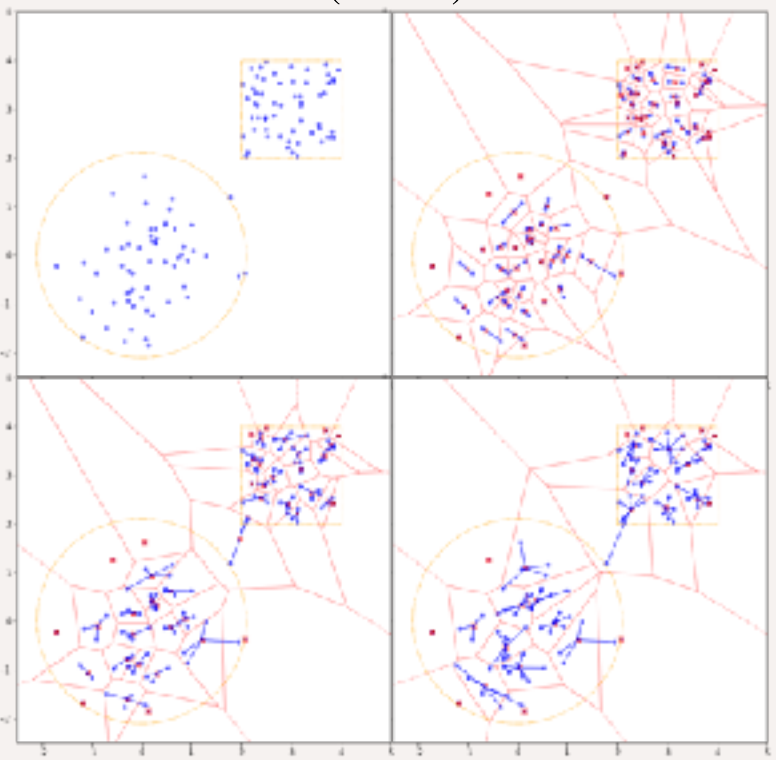
N. Rodrigues et al., MNRAS 2023

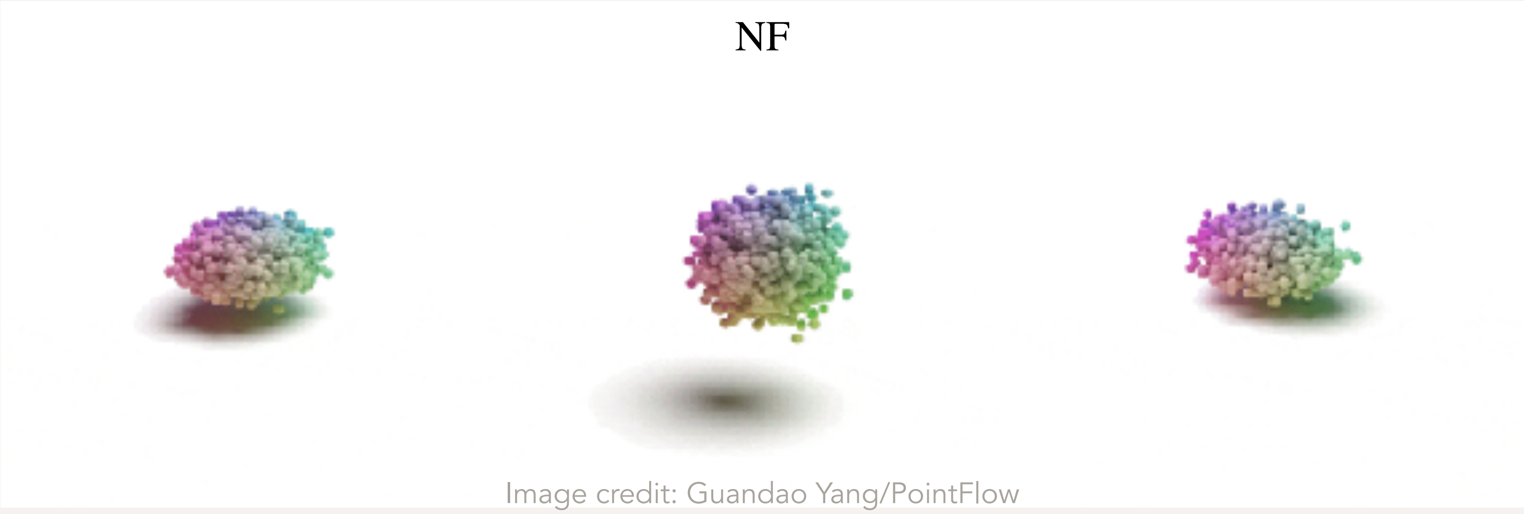


The halo-galaxy joint distribution

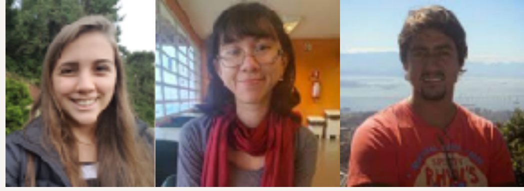
- However…
  - regular tiling/grids are highly inefficient in **higher dimensions**
  - **tails** of the distribution still not ideally represented
  - some **metrics** are still… "meh"…
- So, let's try a couple of different approaches:
  - **Hierarchical Voronoi algorithm** to define classes in high-dimensional spaces (for interpretability)
  - Targets are treated as samples from **Gaussian** distributions, with expectations and variances predicted by the NN (PyTorch GaussianNLLLoss)
  - **Normalizing flows** — for sampling, conditioning and density evaluations. Our flows were learned using conditional spline autoregressive; a single flow was sufficient for our 5 halo + 5 galaxy parameters.

Hierarchical Voronoir Allocation scheme (HiVAl)



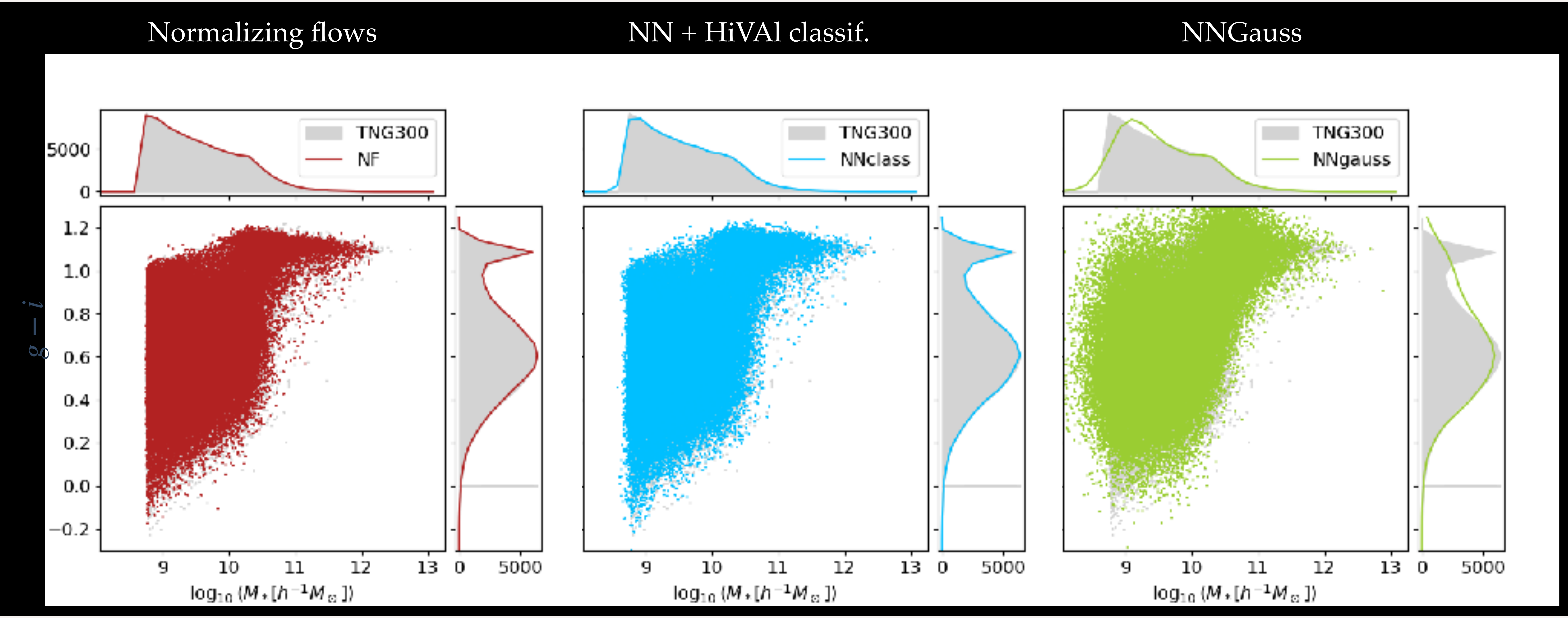


N. Rodrigues, N. de Santi, RA, A. Montero-Dorta, 2024 (to appear)

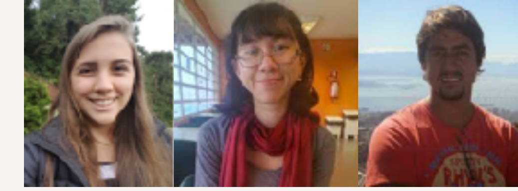


The halo-galaxy joint distribution

- Some results: stellar mass *vs*. color (all halo masses)

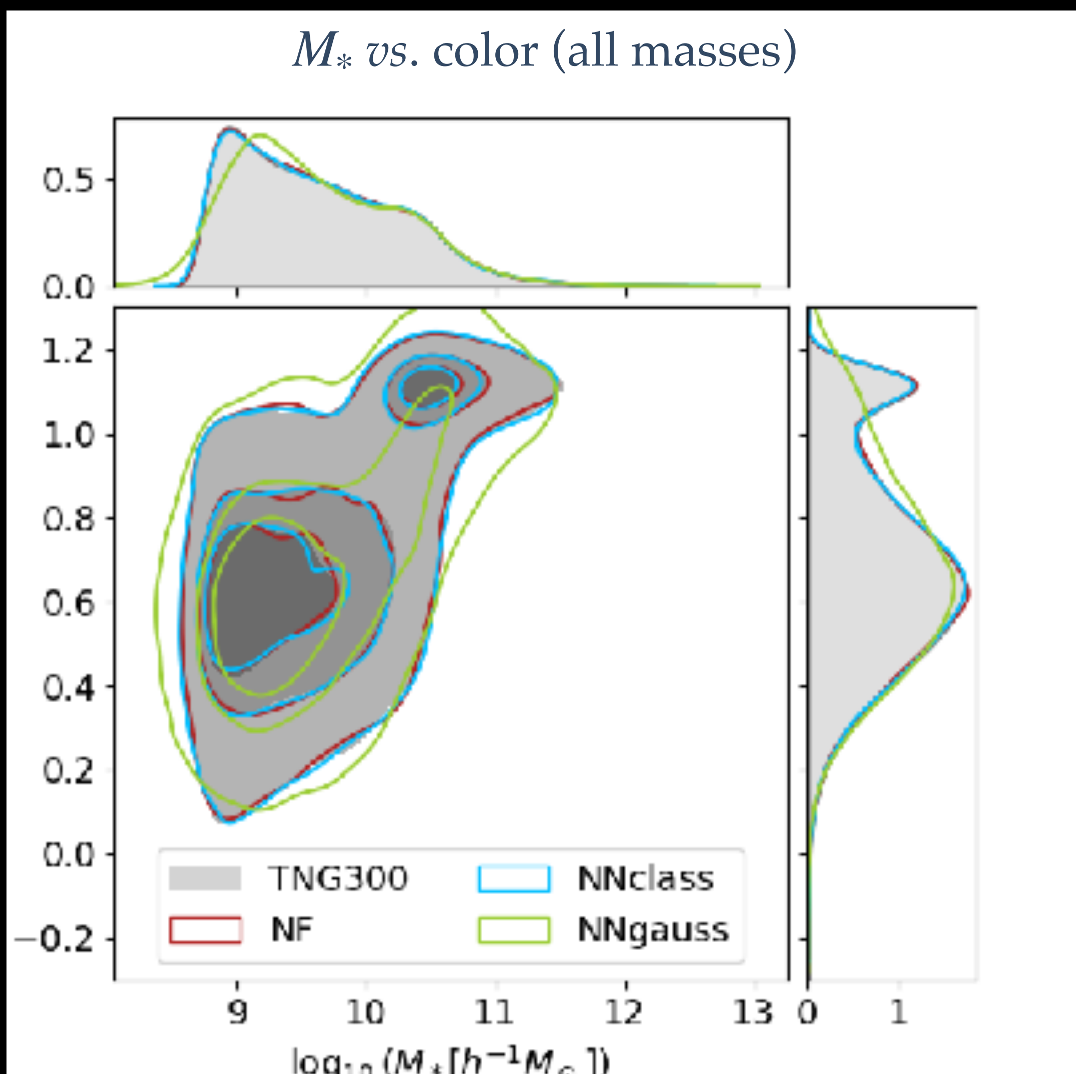


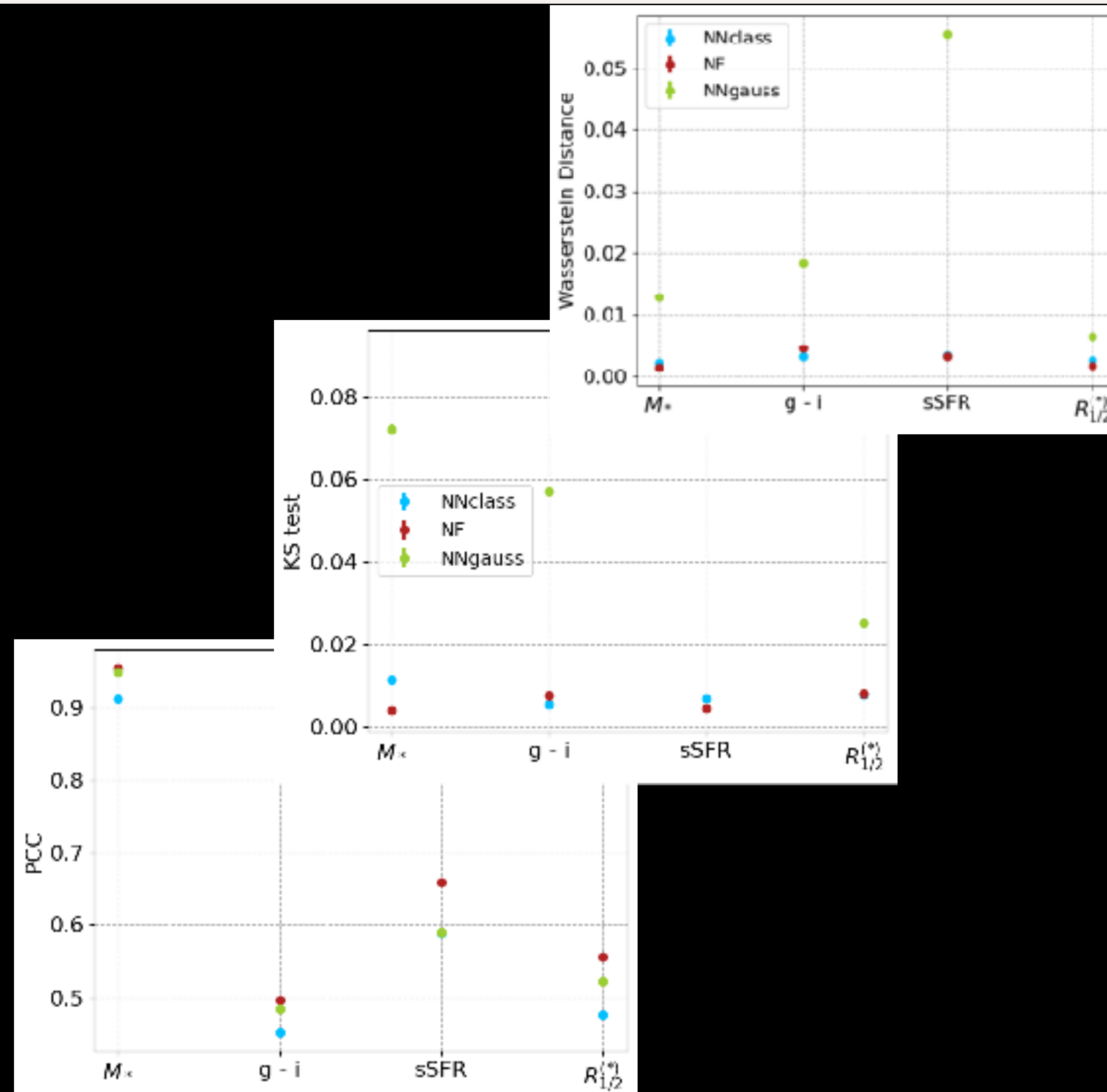
N. Rodrigues, N. de Santi, RA, A. Montero-Dorta, 2024 (to appear)



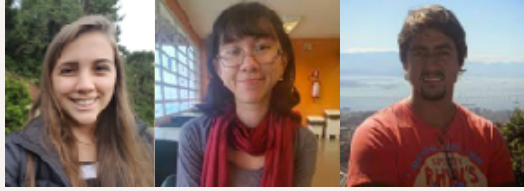
The halo-galaxy joint distribution

- Overall distributions *vs.* predictability



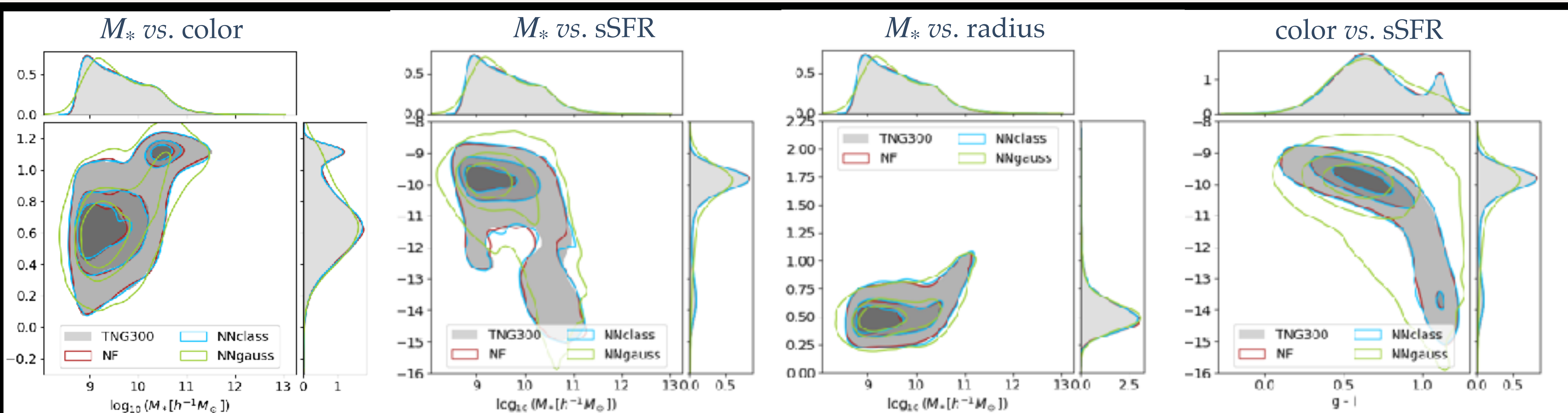


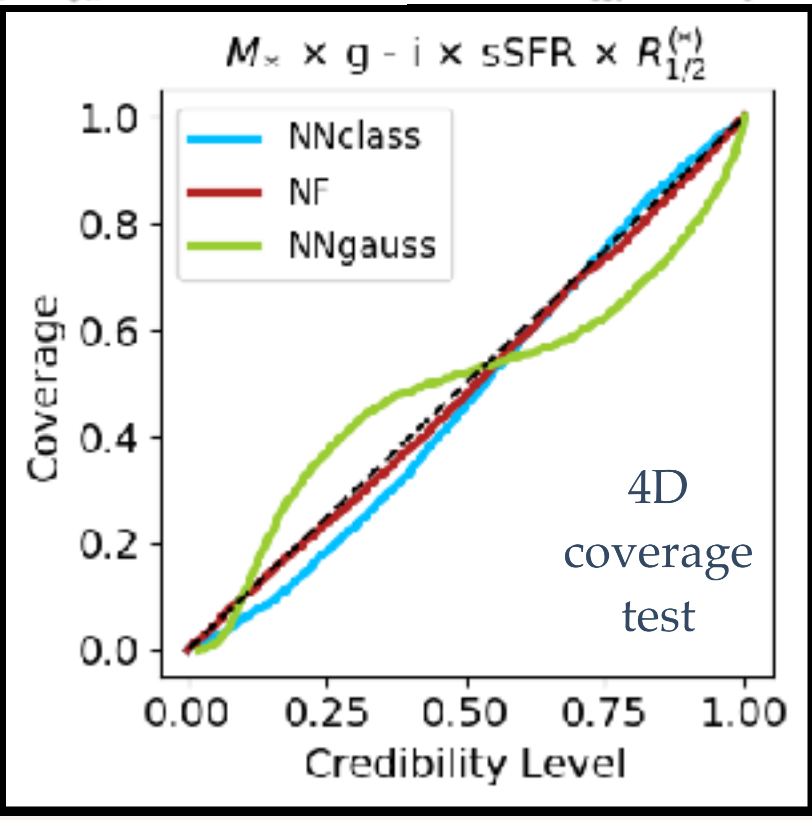
N. Rodrigues, N. de Santi, RA, A. Montero-Dorta, 2024 (to appear)



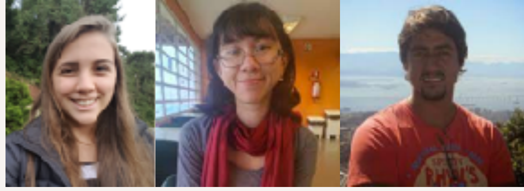
The halo-galaxy joint distribution

- Some results: 2D distributions (all halo masses)



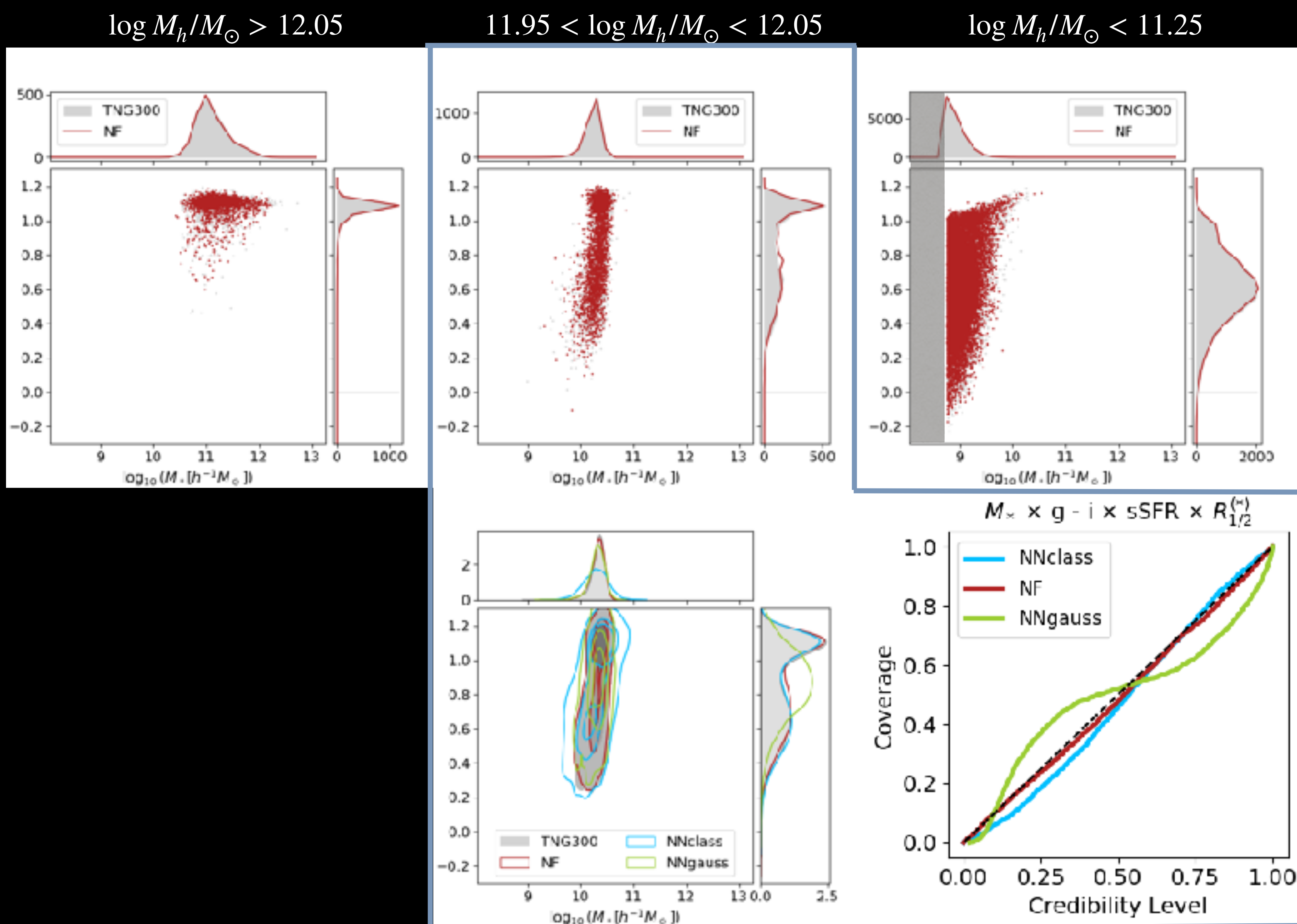


N. Rodrigues, N. de Santi, RA, A. Montero-Dorta, 2024 (to appear)

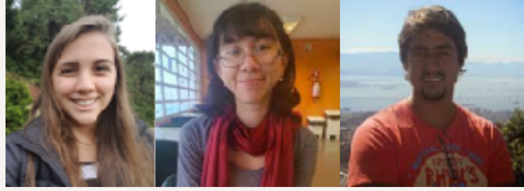


The halo-galaxy joint distribution

- Some results: stellar mass *vs*. color for **different halo masses**



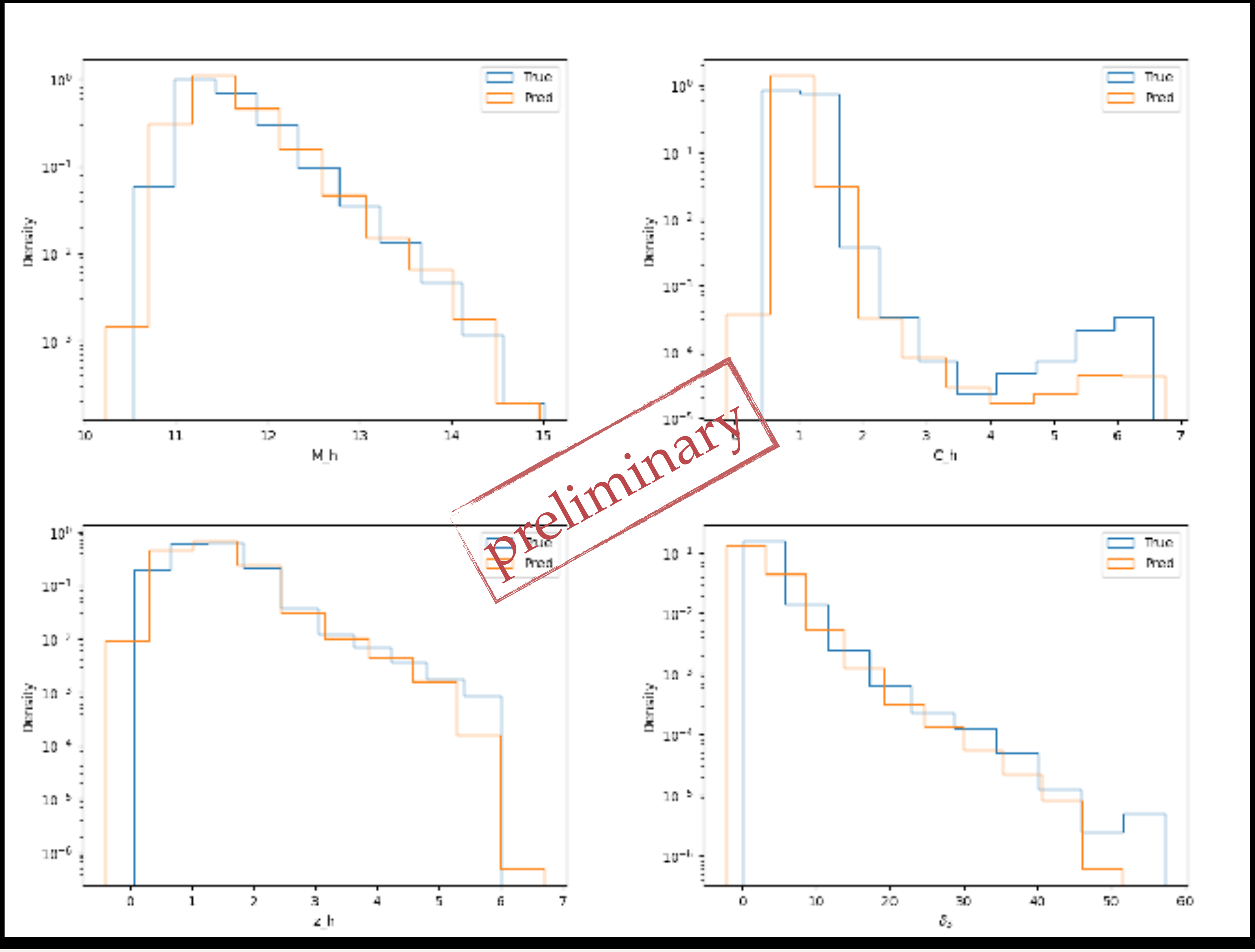
N. Rodrigues, N. de Santi, RA, A. Montero-Dorta, 2024 (to appear)



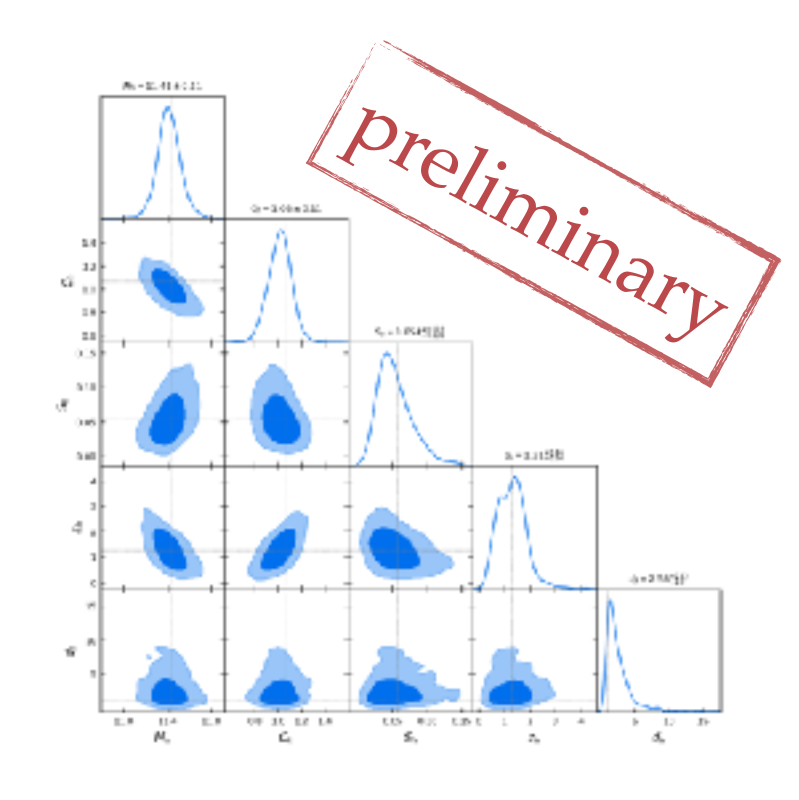
The halo-galaxy joint distribution

- Conditional distributions: not only halo → galaxy , but can do also **galaxy → halo**

halo population correctly reproduced

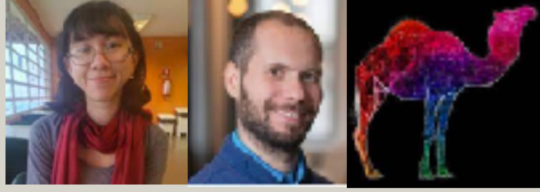


individual halo predictions



N. Rodrigues, N. de Santi, RA, A. Montero-Dorta, 2024 (to appear)

Extra stuff:
Cosmology directly from galaxy groups/clusters

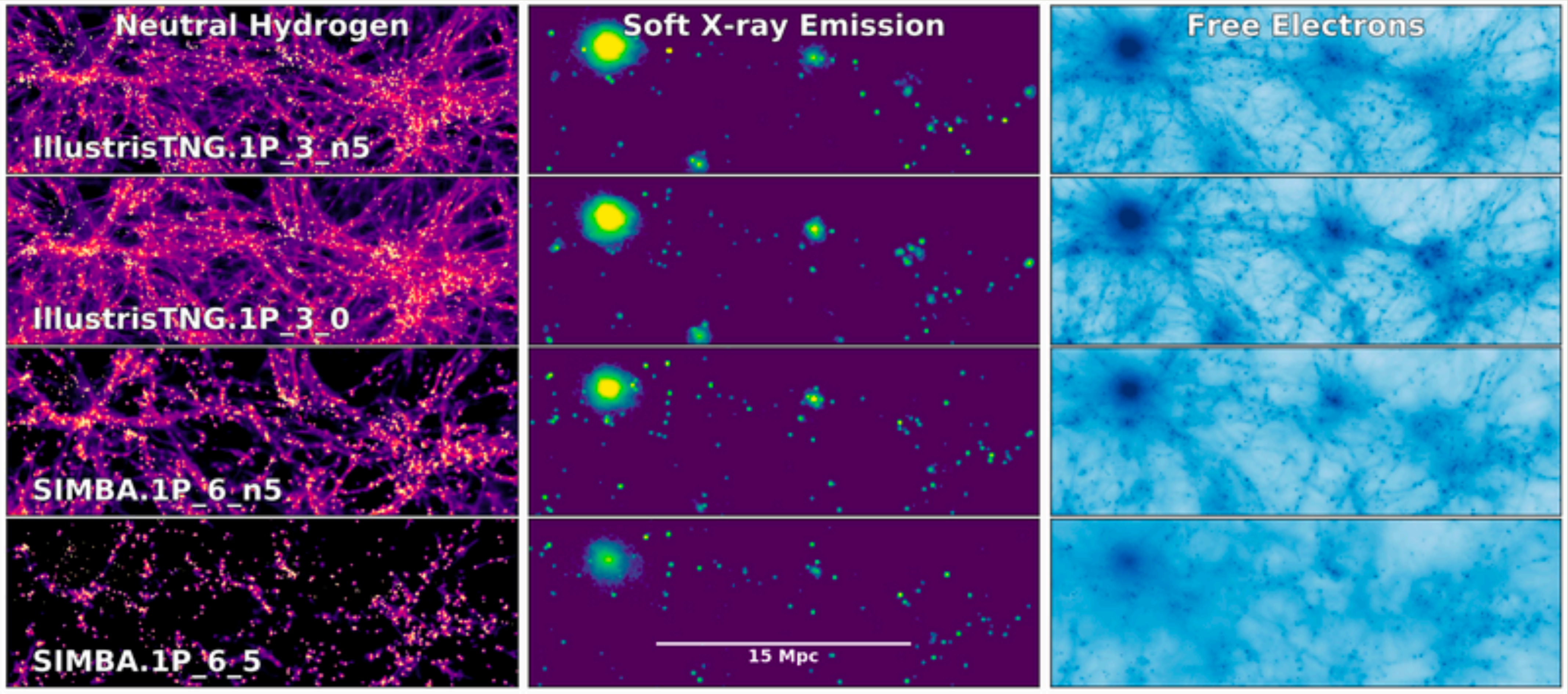


N. de Santi, F. Villaescusa-Navarro
& CAMELs team

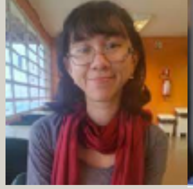
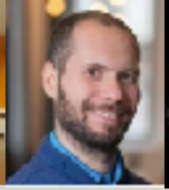
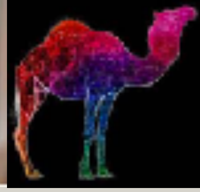
- Computing power has increased dramatically, such that we are now able to run thousands of **hydro simulations** (dark matter + baryons + radiative transfer/feedback), over **huge volumes.**

- We are also able to explore different initial conditions, as well as the differences between codes and between models of SN and AGN feedback — see, e.g., CAMELS (**C**osmology and **A**strophysics with **M**achin**E** **L**earning **S**imulations, Villaescusa-Navarro et al. 2021).

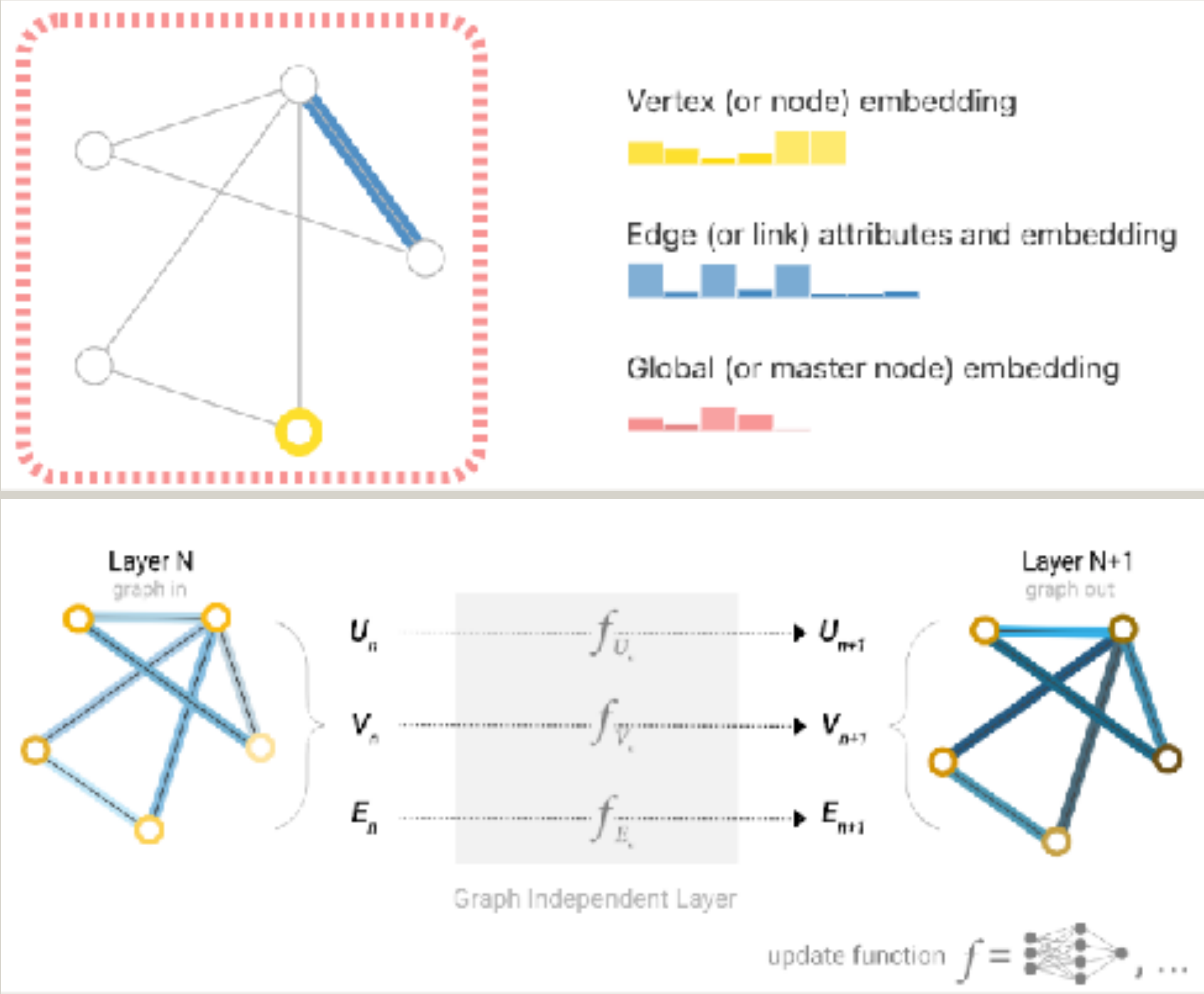
- If the simulations are anything like the "real thing", then we can use ML to help find patterns and to reproduce mechanisms which may be too complex to tackle in a parametric way (e.g., SAM/SHAM)

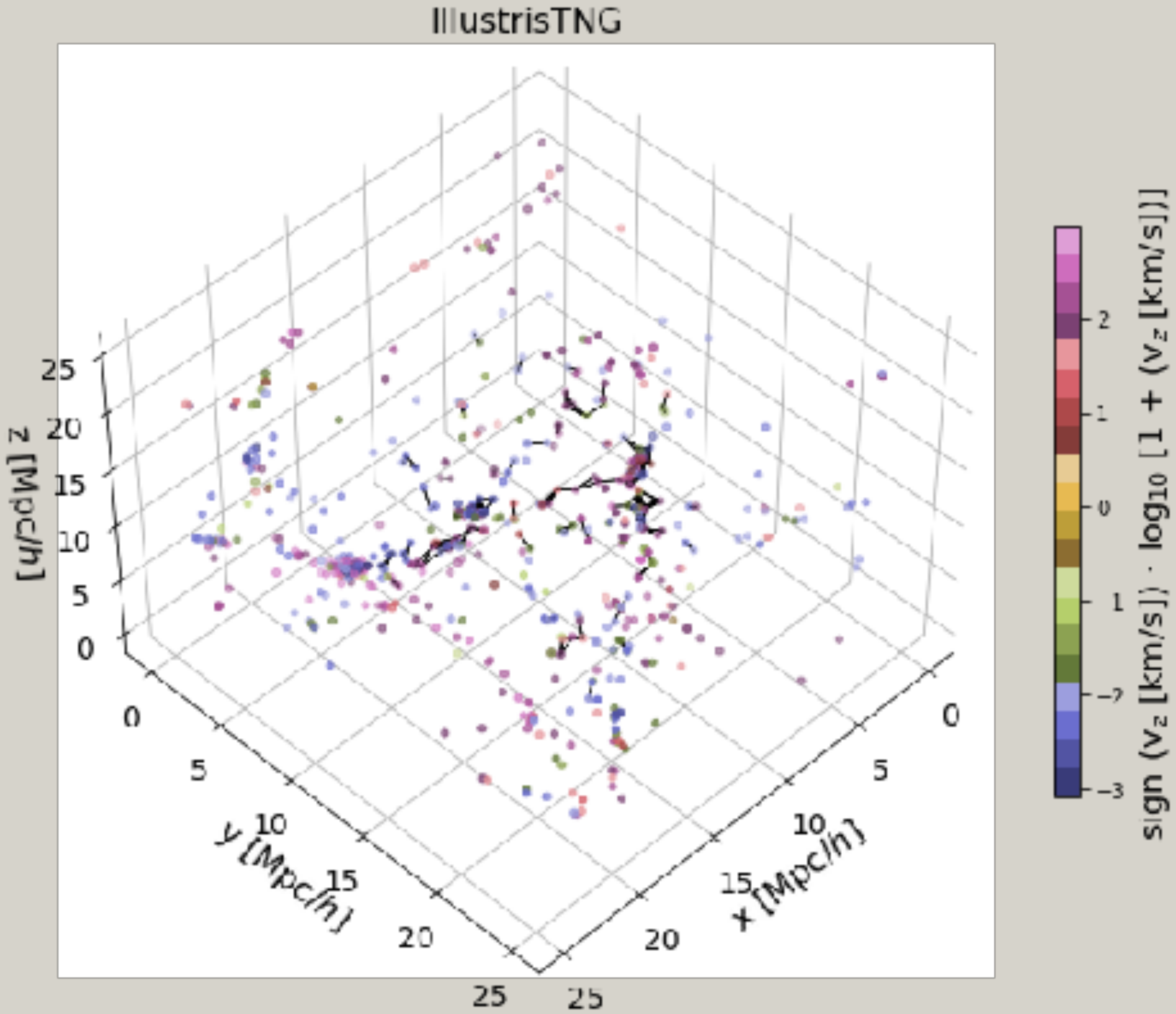


Extra stuff:
Cosmology directly from galaxy groups/clusters

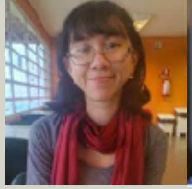
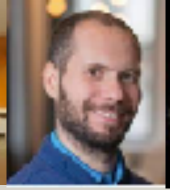
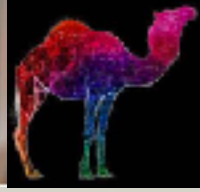




- In Natalí de Santi+ 2023 we showed how, just on the basis of **galaxies** and their **immediate environments** (scales <~ 3 Mpc), it is possible to **infer the matter density Ω_m** (see also Wu, Jespersen & Wechsler 2024)
- This was achieved by means of GNNs (**Graph Neural Networks** — see, e.g., https://distill.pub/2021/gnn-intro/)



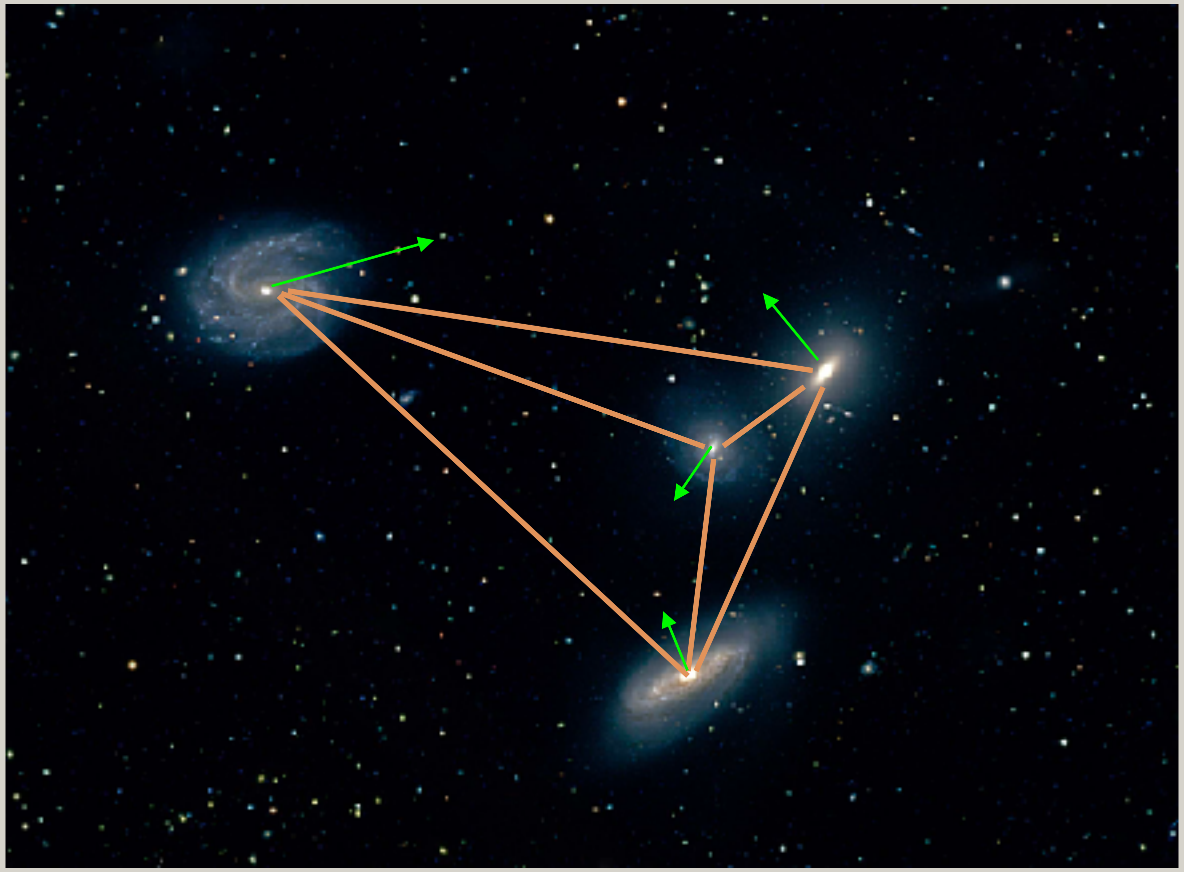


N. de Santi et al., ApJ 2023

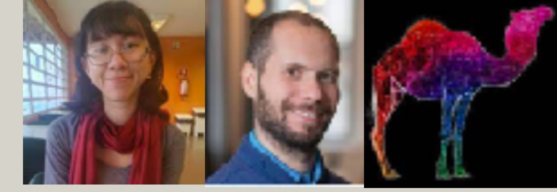




Extra stuff:
Cosmology directly from galaxy groups/clusters

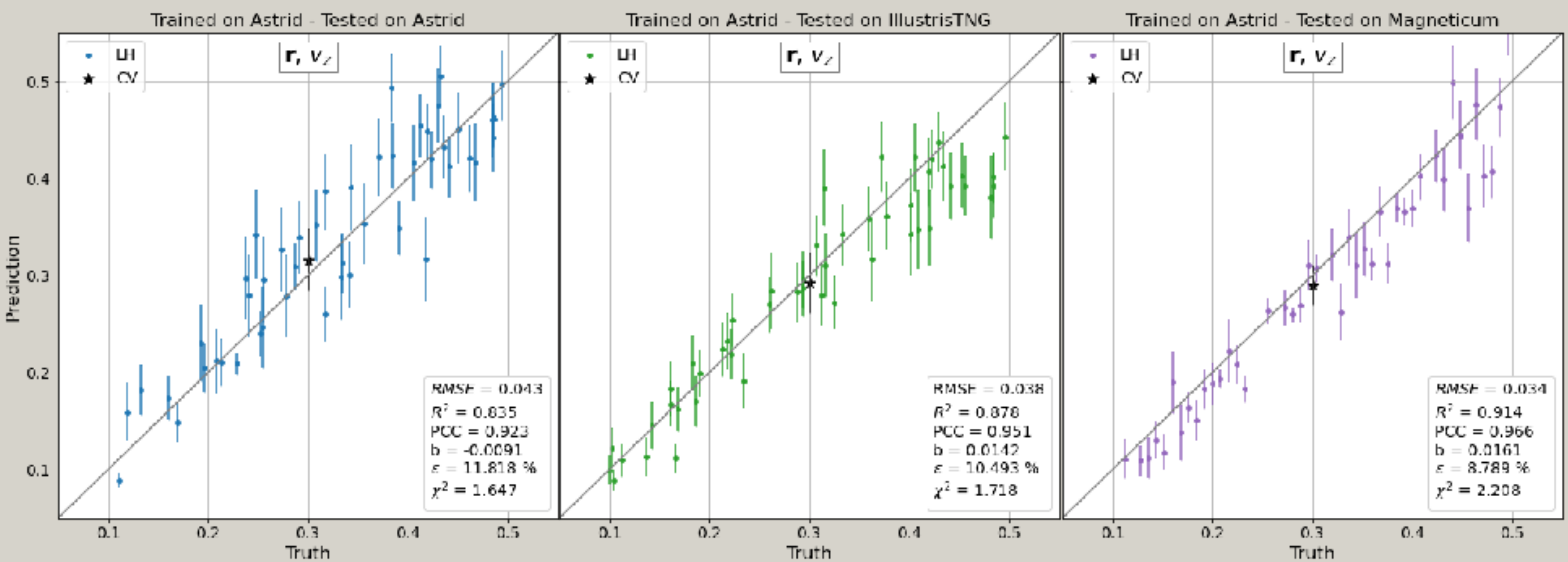
- Natalí used galaxies with $M_\star \gtrsim 1.3 \times 10^8\ M_\odot$, and an $r_{link} = 1.25\, h^{-1}\,\mathrm{Mpc}$
- The properties of the graphs are mainly:
  - the number of galaxies (global feature of the **graph** — a few to several dozen)
  - the positions $\{x, y, z\}$ and the velocities $\{v_x, v_y, v_z\}$ of the individual galaxies (**vertex** properties)
  - the distances between all pairs (**edge** properties)



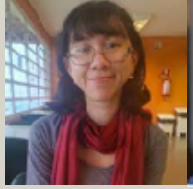
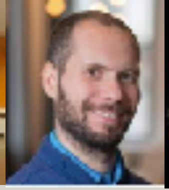
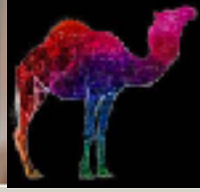
Extra stuff:
Cosmology directly from galaxy groups/clusters



- Natalí **trained the GNNs in one set** of simulations (e.g., Illustris) and **tested it in another set** (e.g., Astrid), in order to see if the method was able to generalize its results.
- We find that the models are **robust** with respect to changes in (i) the astrophysics (feedback models), (ii) the subgrid physics, and (iii) the subhalo/galaxy finder (de Santi+ 2023).
- Natalí tested those models on **thousands of simulations** that cover a **vast region in parameter space** – variations in **5 cosmological** as well as many **astrophysical parameters**.
- We were finally able to **recover Ω_m** across different simulations with a **~10% uncertainty**!

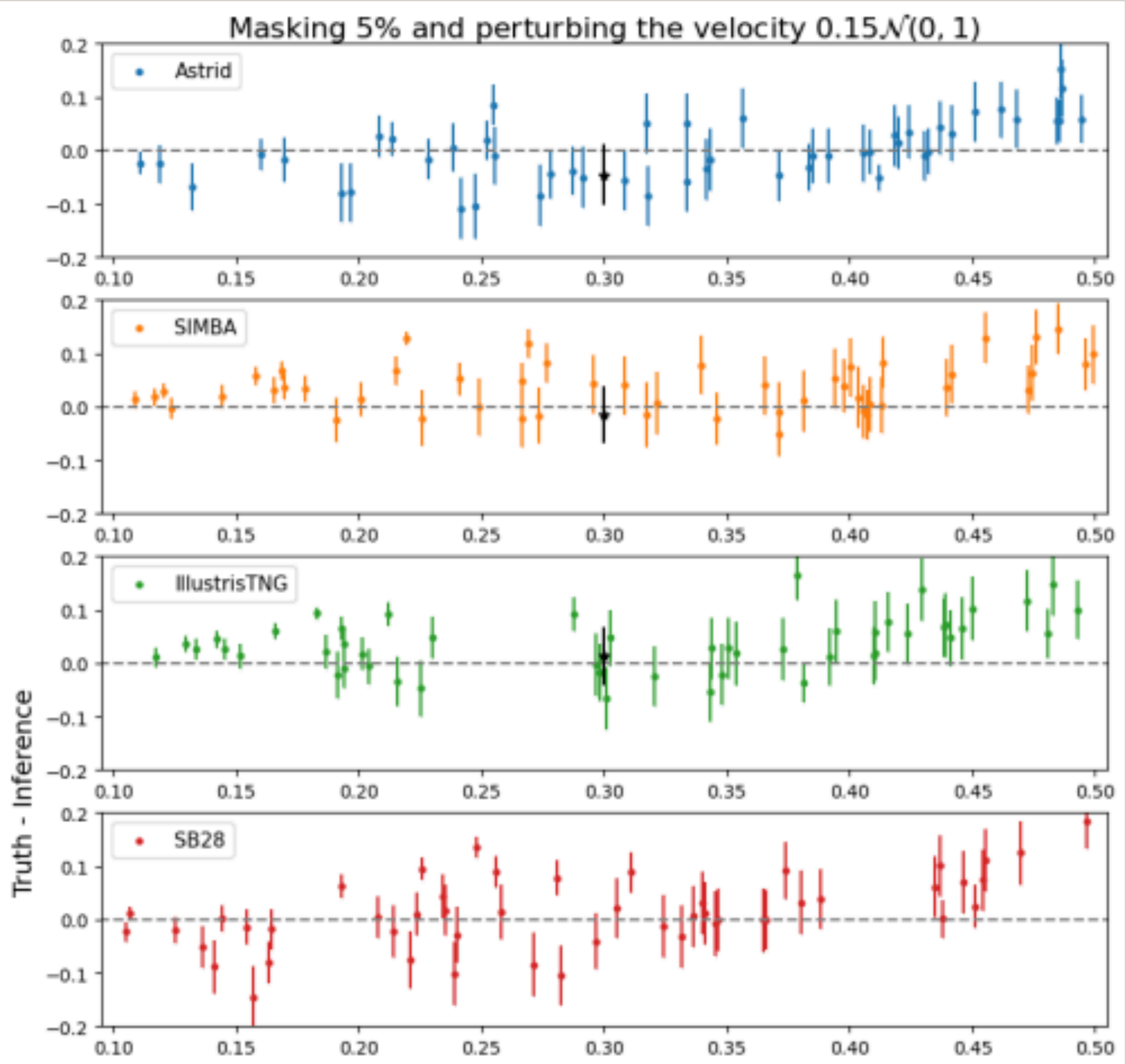


Extra stuff:
Cosmology directly from galaxy groups/clusters

- In de Santi+ 2023b we showed that these results are robust to **real-life features** such as:
  - Redshift-space effects: **only $\perp$ input positions** $\{x, y\}$ **and radial/peculiar velocity** v_z

    $\{x, y, z\,;\ v_x, v_y, v_z\}$
  - Masks & color cuts
  - Errors in the positions and velocities/redshifts
- We are **still** able to **recover** Ω_m across different simulations **with a ~15% uncertainty**



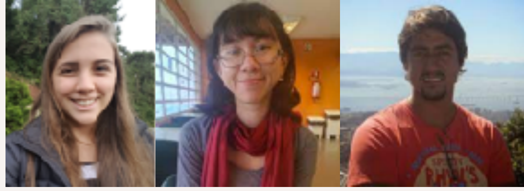
What's next?

- Centrals → + satellites & sub-halos
- Robustness across different sims / sub-grid models (Illustris / SIMBA / Astrid)
- Stochasticity
- Extension to mocks
- From galaxies back to halos

Let us know what you think…

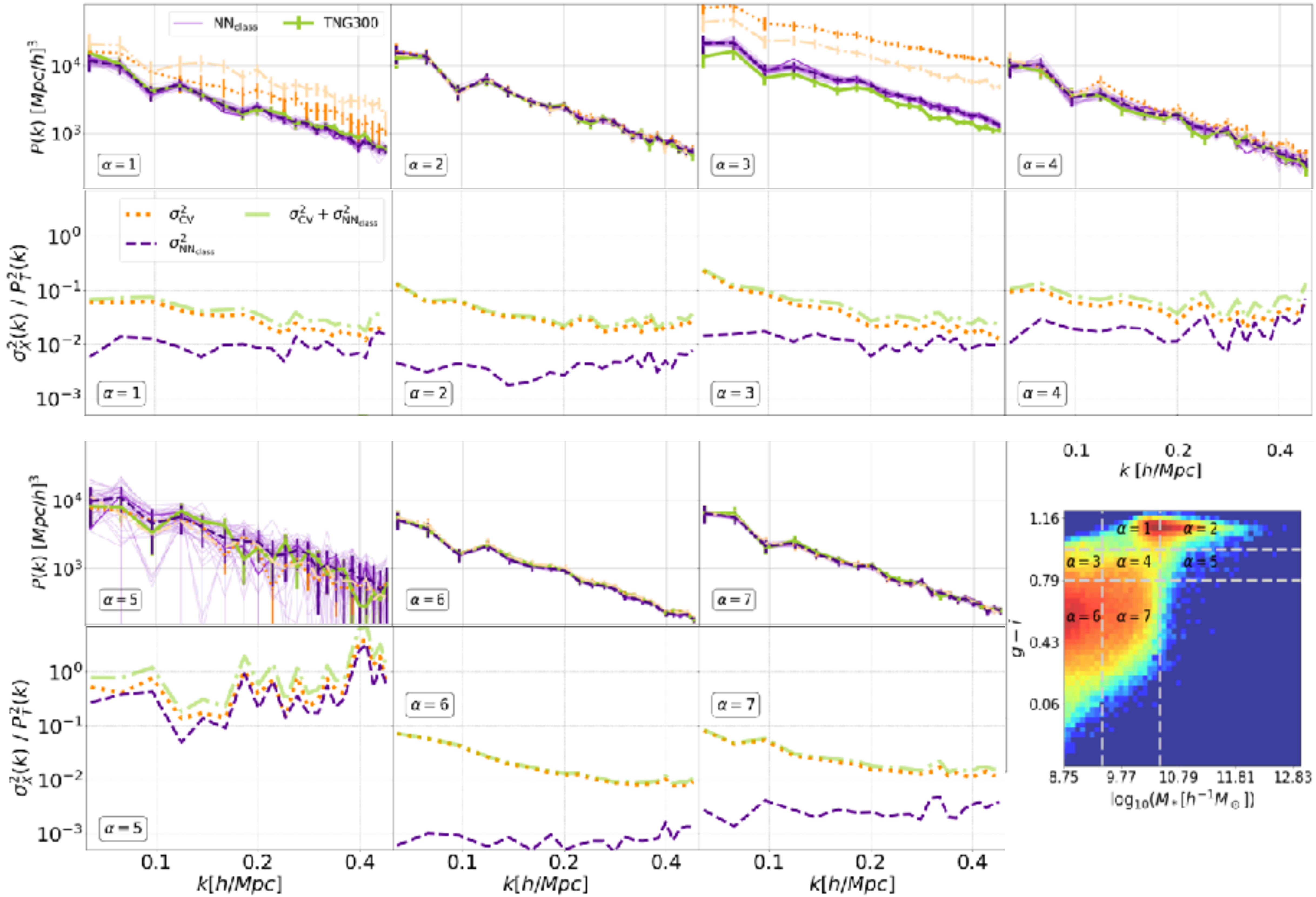
raulabramo@usp.br

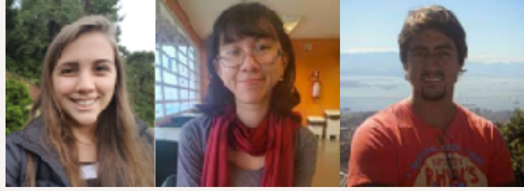
Thanks!



Extra slides

- Contribution of **stochasticity** in galaxy properties to **variance in clustering** (Rodrigues+ 2023)





Extra slides

- Permutation feature importance (of halo properties)