Cosmology in the machine learning era

Francisco (Paco) Villaescusa-Navarro

SIMONS FOUNDATION

PRINCETON UNIVERSITY

FLATIRON INSTITUTE

MODE workshop

September 7th 2021

Take home message

Simulations are not perfect; they may never be… Do we need perfect simulations?

I want an beer

I have did the research

You was wrong

I am given a presentattion

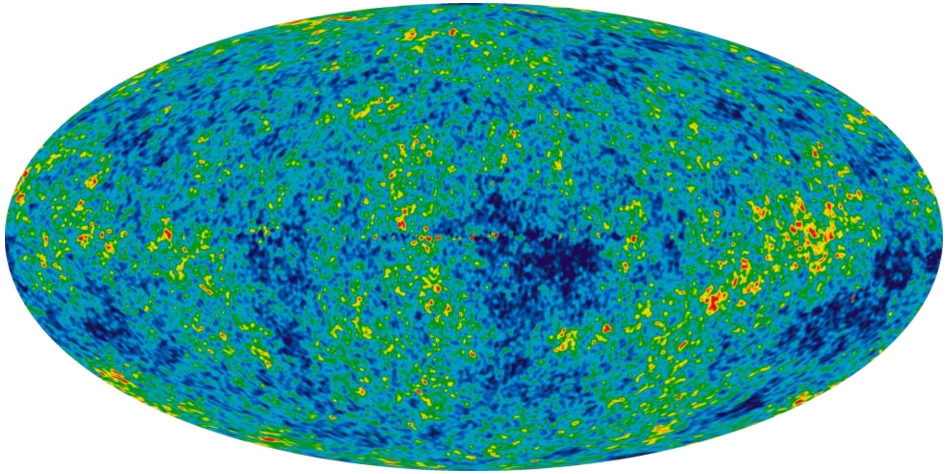
<u>To be or not to be</u>

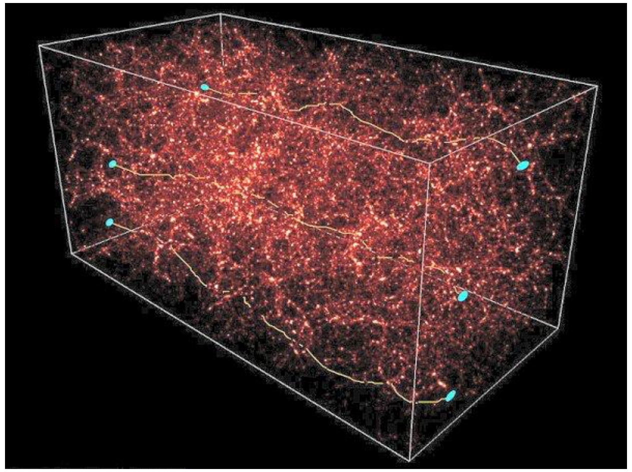
Outline

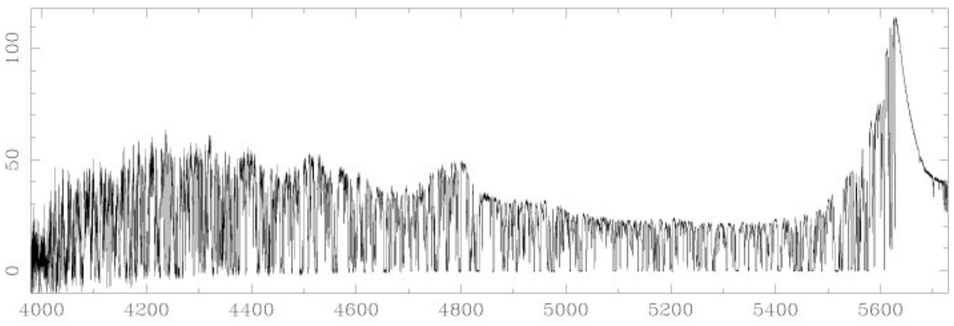
- The problem
- The potential solution
- The risks

The problem

The ΛCDM model







$$\Omega_m \pm \delta\Omega_m$$

$$\Omega_b \pm \delta\Omega_b$$

$$h \pm \delta h$$

$$n_s \pm \delta n_s$$

$$\sigma_8 \pm \delta\sigma_8$$

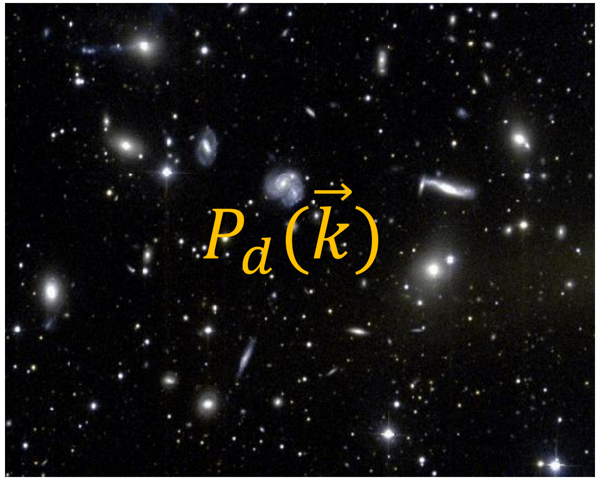
$$M_\nu \pm \delta M_\nu$$

$$w_0 \pm \delta w_0$$

$$w_a \pm \delta w_a$$

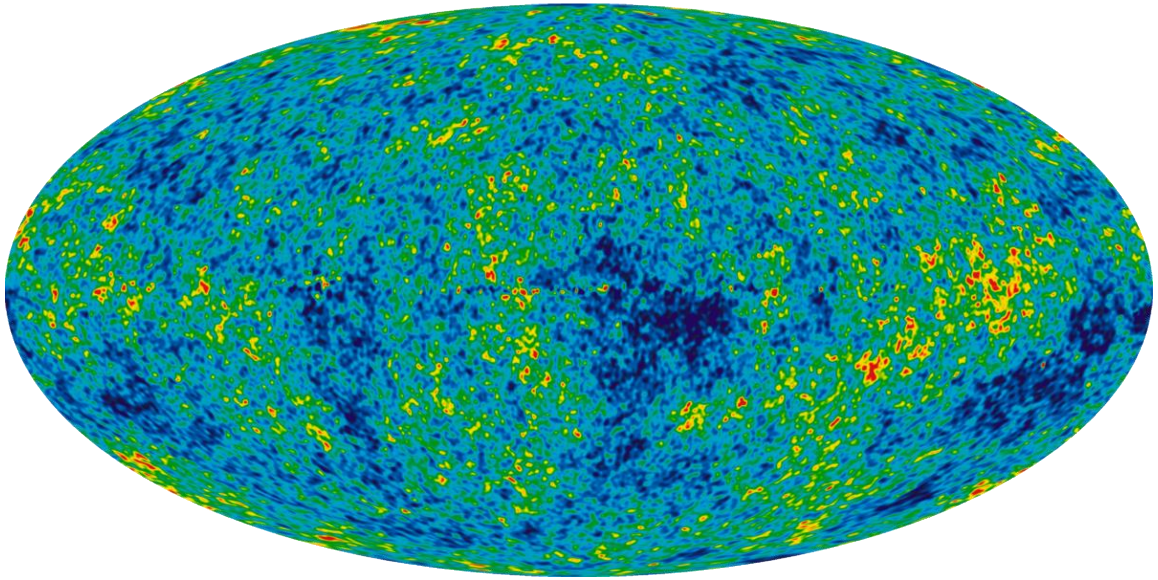
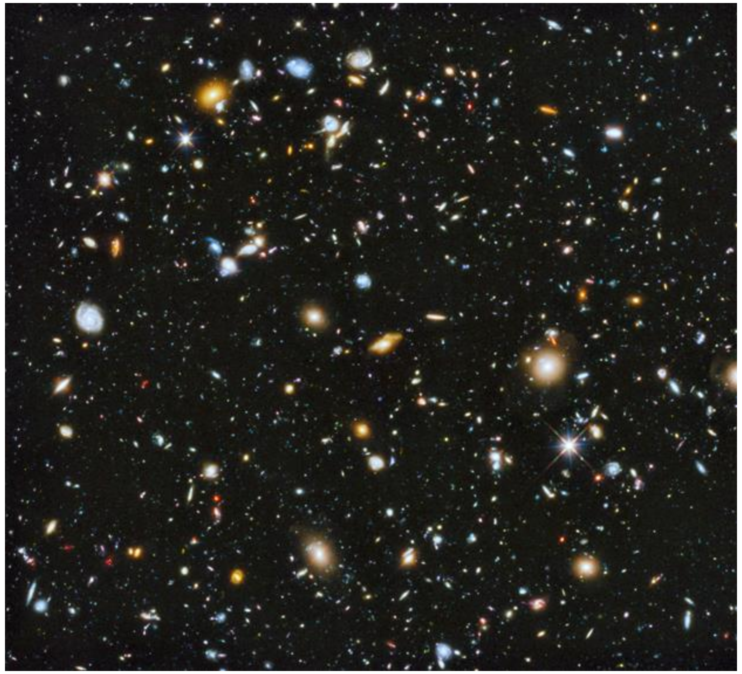
$$N_{\rm eff} \pm \delta N_{\rm eff}$$

Parameter inference

Observations	**Theory**
	$P_t(\vec{k}\|\vec{\theta})$

What summary statistics shall we use to determine $\vec{\theta}$ with the smallest error?

Parameter inference: summary statistics

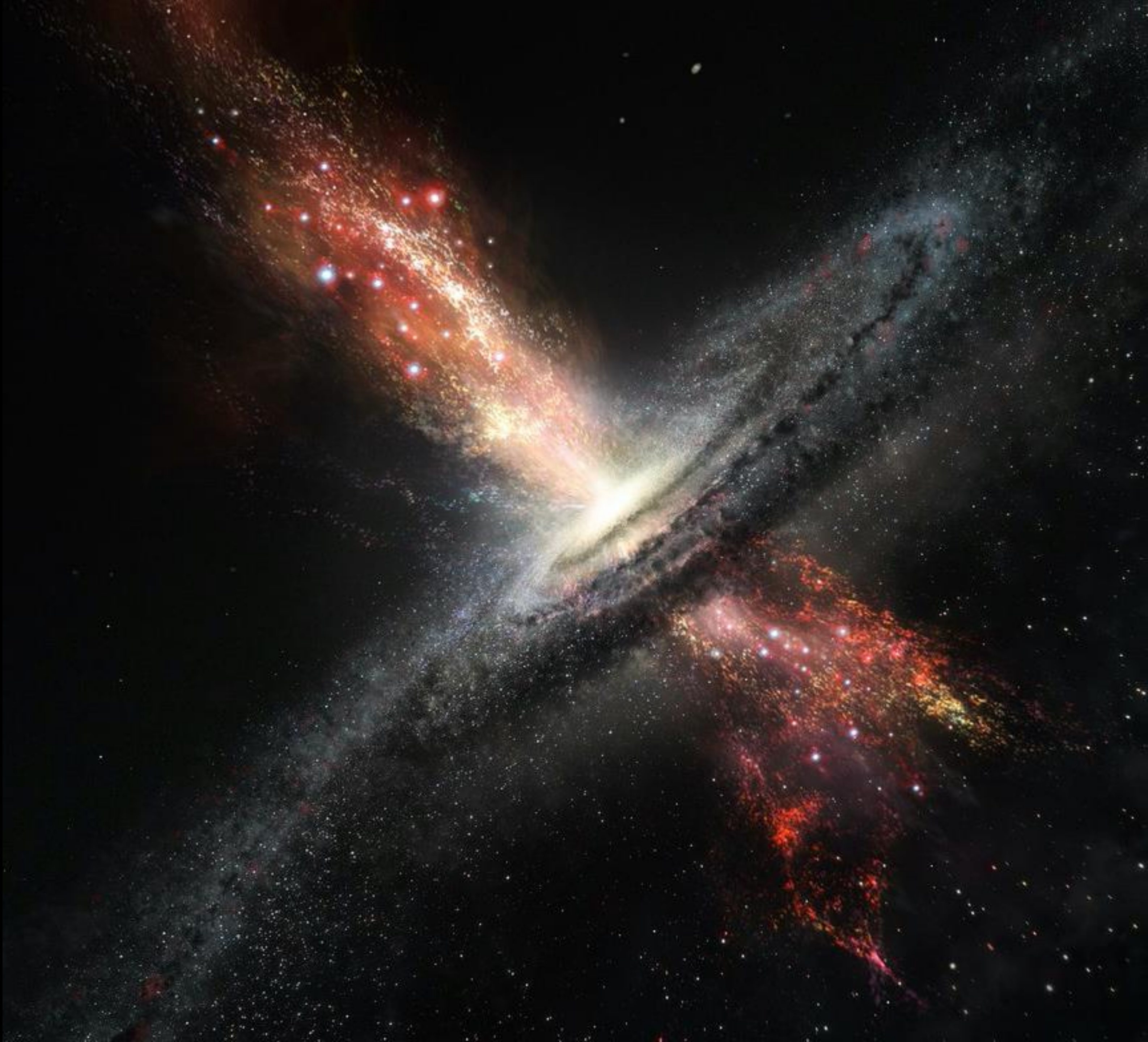
Gaussian density field	Non-Gaussian density field
	
Optimal statistic: power spectrum	Optimal statistic: ???

The Quijote Simulations

(https://quijote-simulations.readthedocs.io)

- A set of 44,100 full N-body simulations
- More than 7,000 cosmologies in $\{\Omega_m, \Omega_b, h, n_s, \sigma_8, M_\nu, w\}$ hyperplane
- Around 10 trillion particles over a volume larger than entire observable Universe
- Catalogues with billions of halos, voids and galaxies: Molino and Gigantes datasets
- 35 Million CPU hours; 1 Petabyte of data
- Everything publicly available

Generic conclusion:
Lots of information on small scales beyond P(k)



Benefits: Lots of information

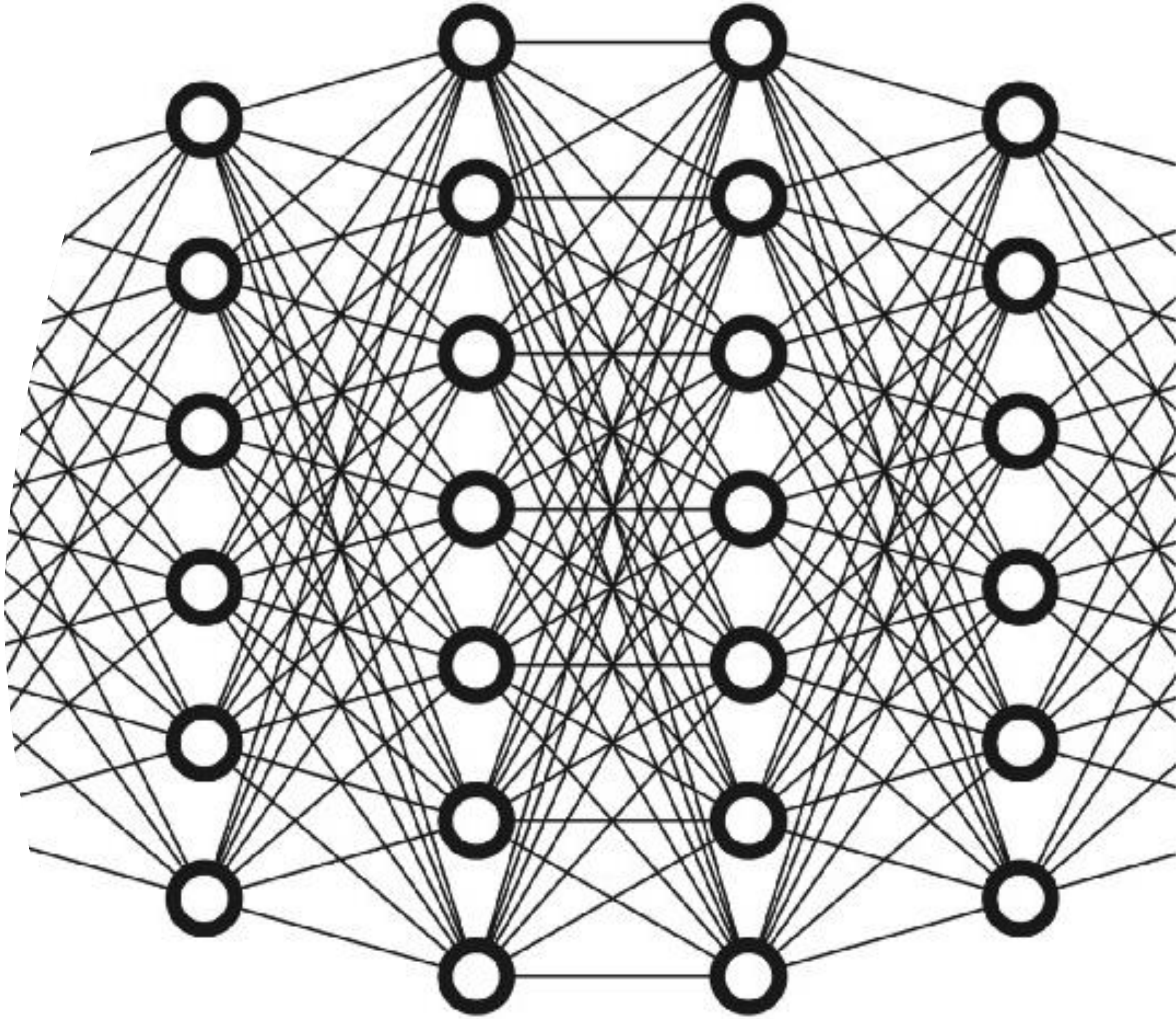
Problems: Non-linearities & baryonic effects

Summary

- We don’t know how to read the cosmological information written on the sky. We may be missing the most important part of the book

- The tools we typically use to extract information are suboptimal

The potential solution



A machine learning solution

Can we extract ALL information from the field while marginalizing over uncertain baryonic effects? YES!

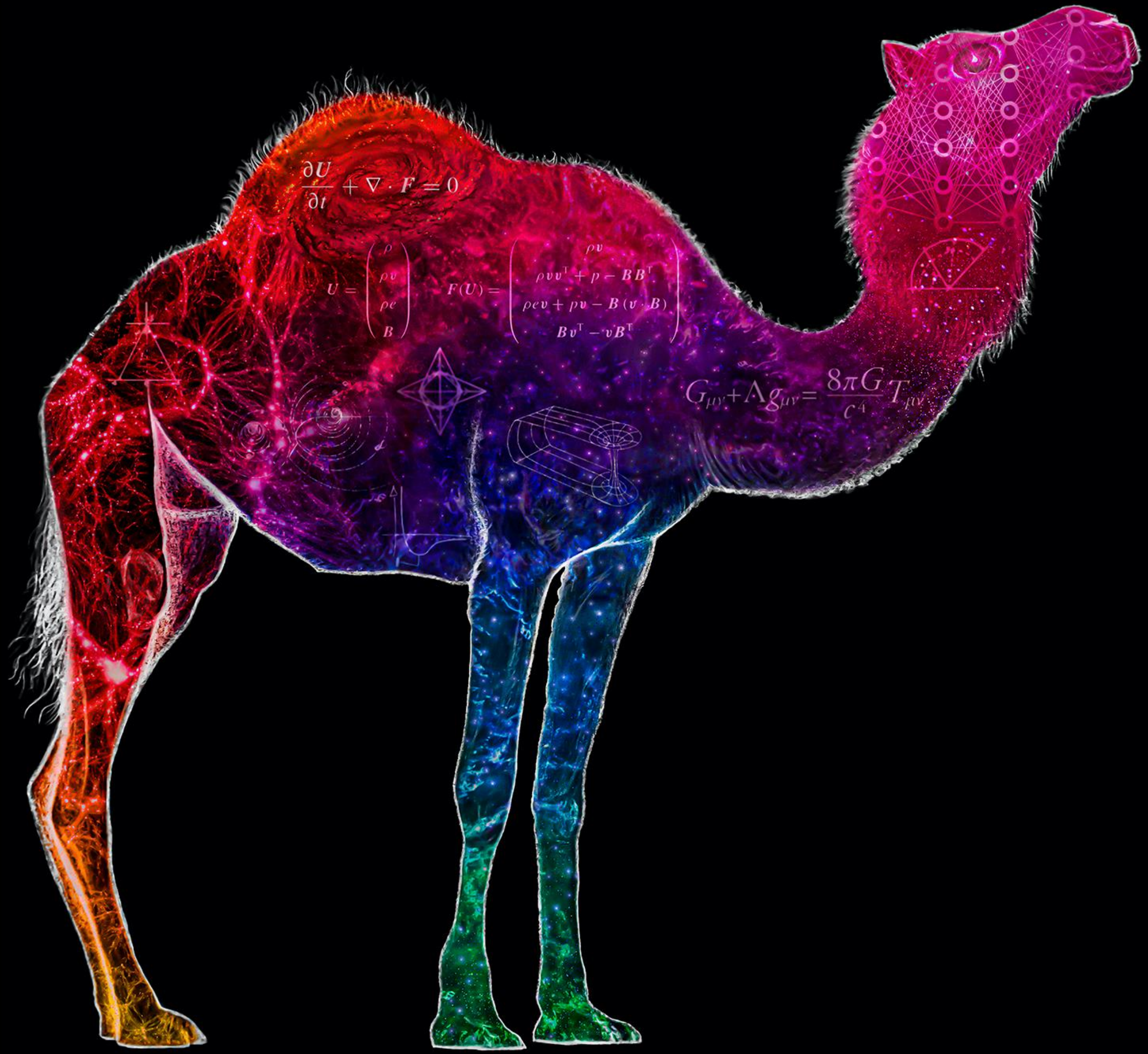
What we need?

- Many simulations with different cosmologies & astrophysics
- Train neural networks
- Check robustness of the estimators found by the networks

CAMELS

https://www.camel-simulations.org

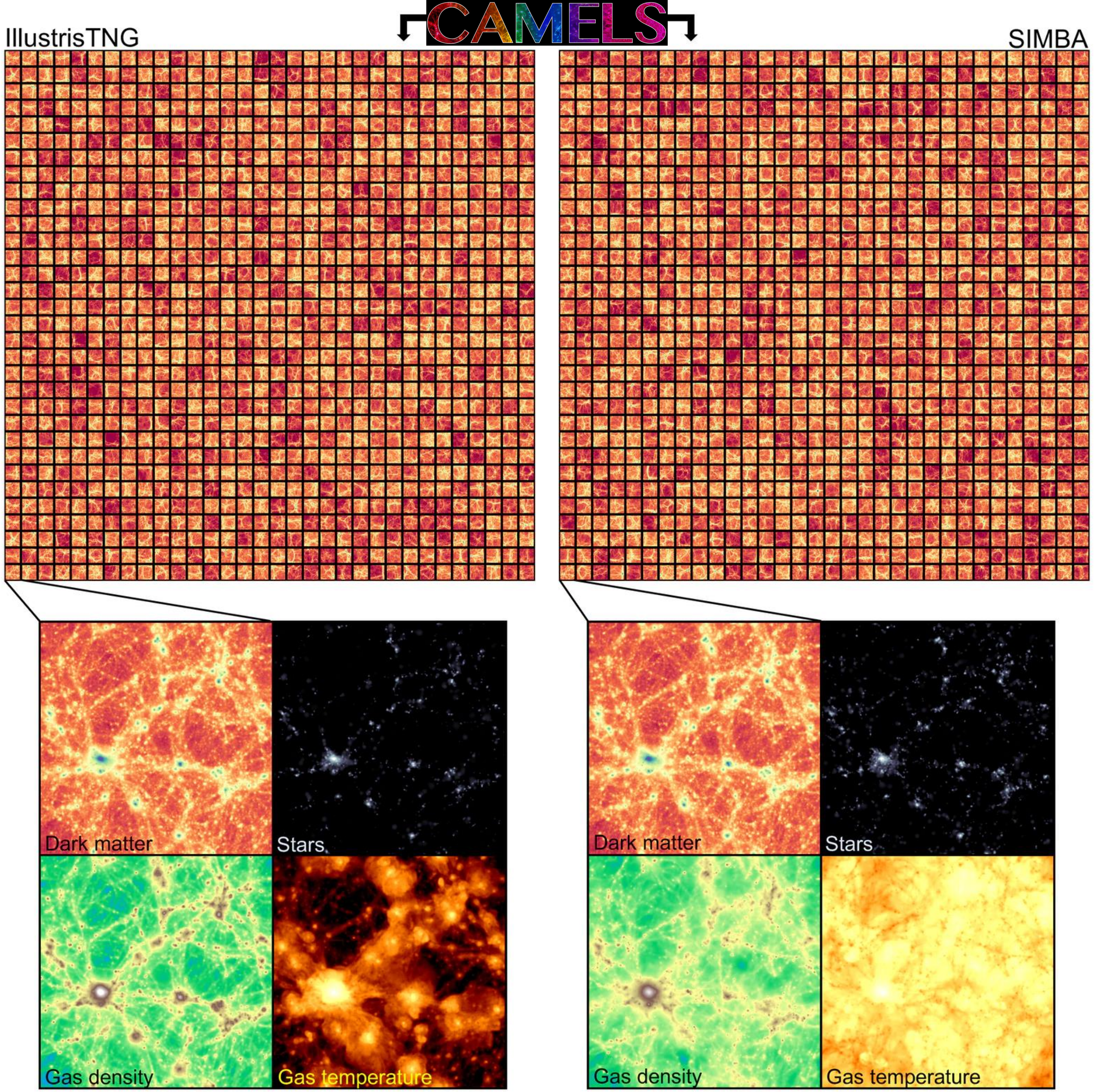
Cosmology and Astrophysics with MachinE Learning Simulations



- A suite of 4,233 simulations
- 2,049 N-body; Gadget-III
- 2,184 state-of-the-art (magneto-)hydrodynamic sims
- AREPO/IllustrisTNG + GIZMO/SIMBA
- 6 parameters: $\{\Omega_m, \sigma_8, A_{\mathrm{SN1}}, A_{\mathrm{SN2}}, A_{\mathrm{AGN1}}, A_{\mathrm{AGN2}}\}$
- More than 100 billion resolution elements over combined volume of ~$(400\ \mathrm{Mpc}/h)^3$
- More than 2,000 cosmologies & astrophysics models; more than 140,000 snapshots
- Designed for machine learning applications

1,000 different simulations with AREPO + IllustrisTNG

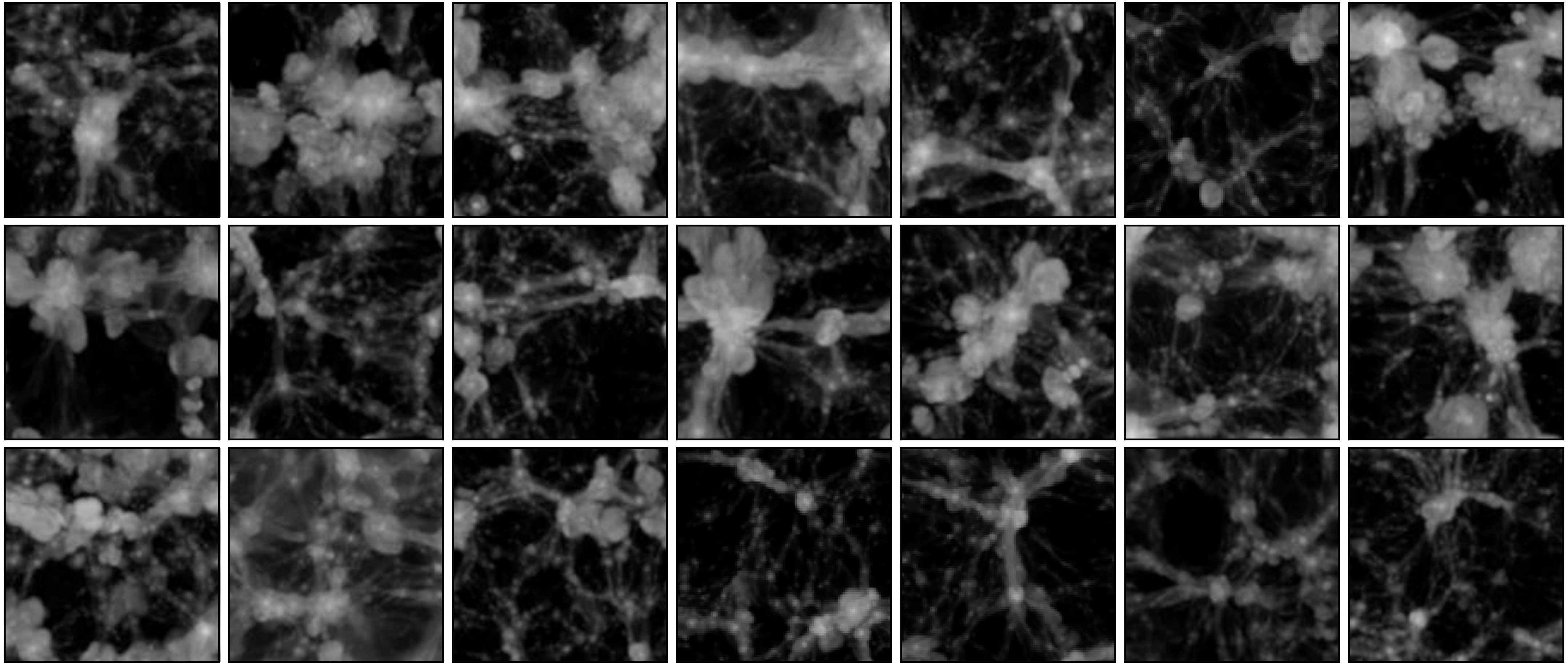
Each simulation has a different cosmology and astrophysics model



1,000 different simulations With GIZMO + SIMBA

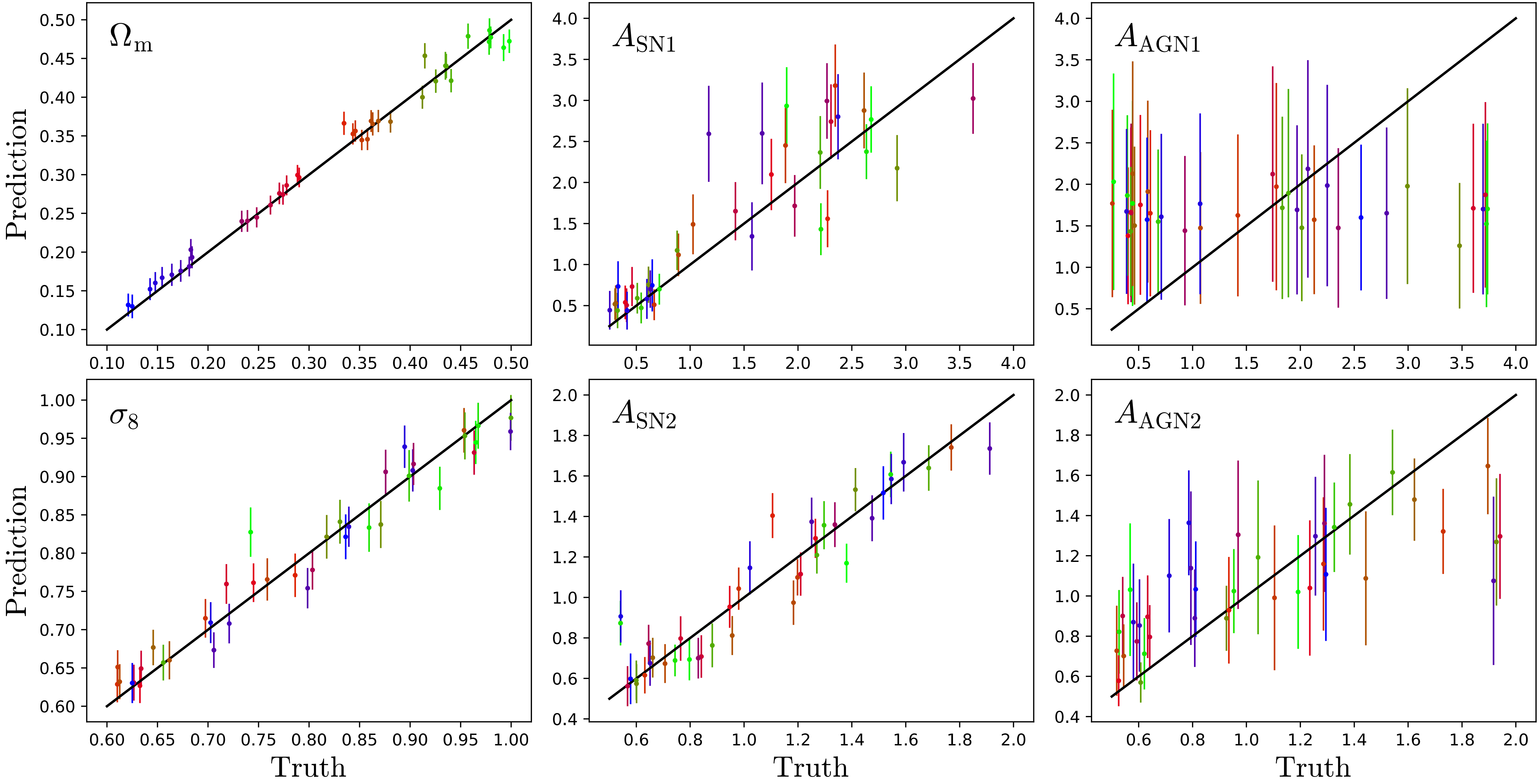
Each simulation has a different cosmology and astrophysics model

Example I: Gas temperature



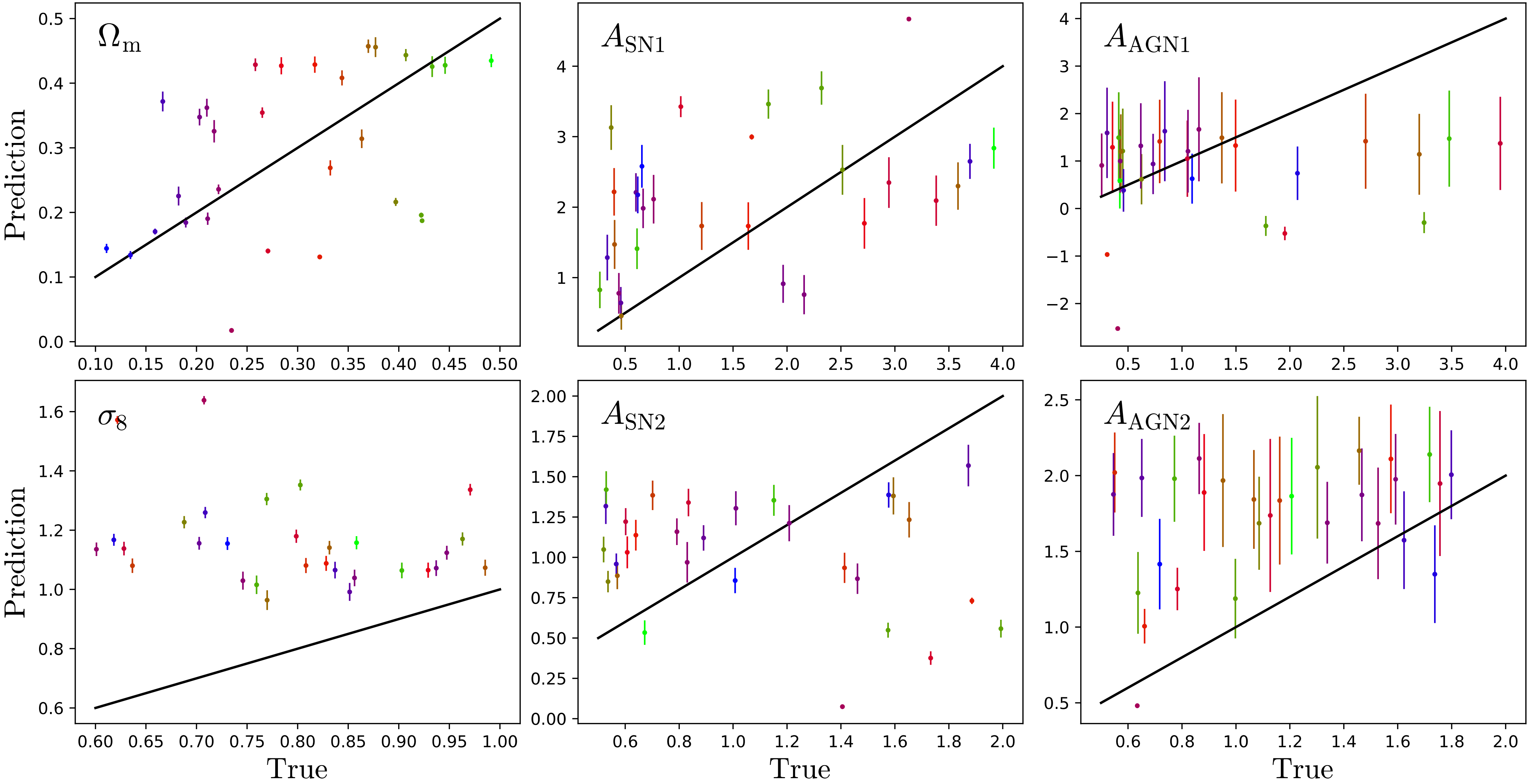
Every map has 256×256 pixels, covers an area of 25×25 $(h^{-1}\mathrm{Mpc})^2$, and has a different cosmology & astrophysics. 15,000 images in total.

Likelihood-free inference: gas temperature

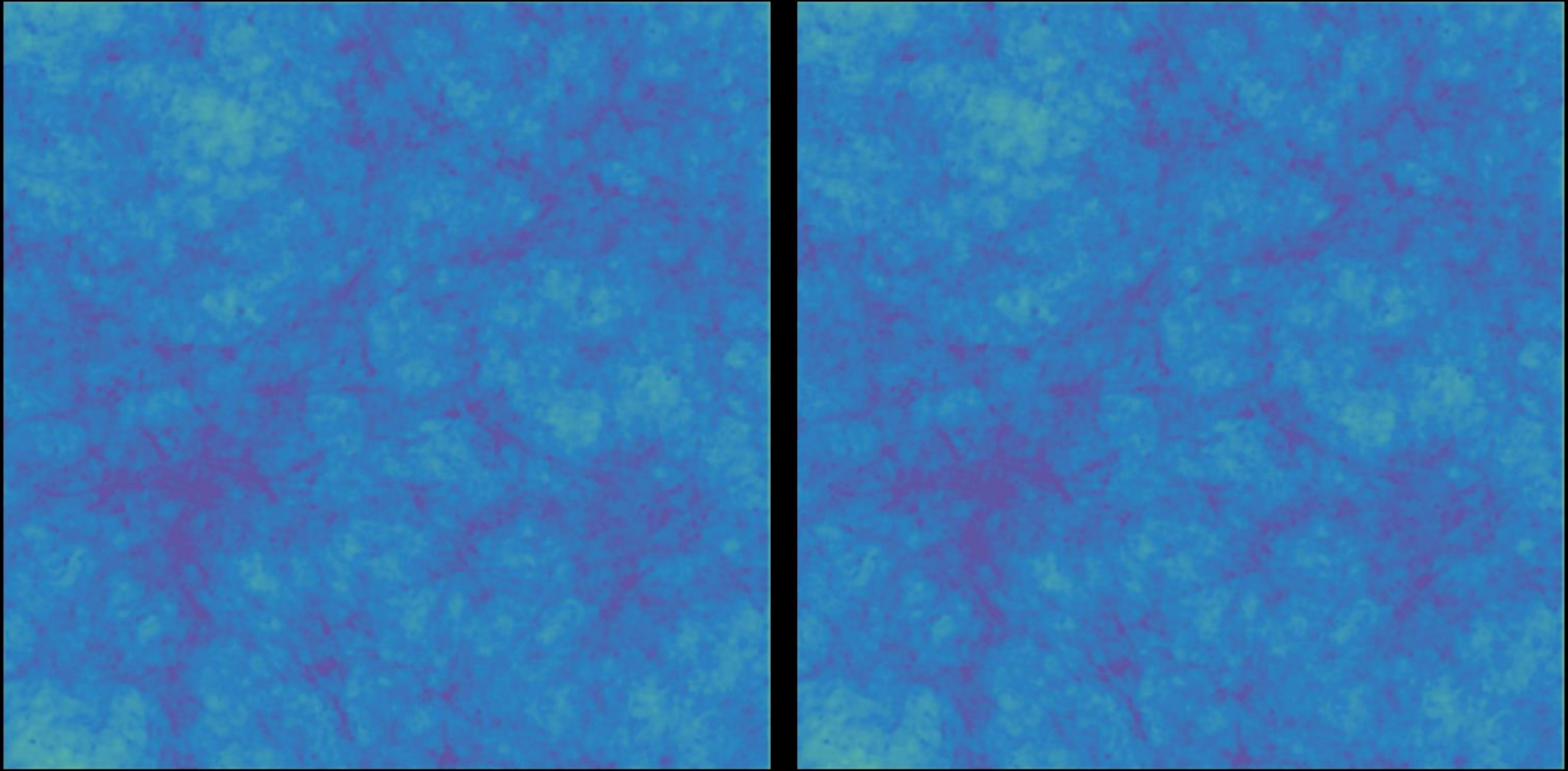


Robustness: gas temperature

Network trained on IllustrisTNG and tested on SIMBA



IllustrisTNG **Dark matter density** SIMBA



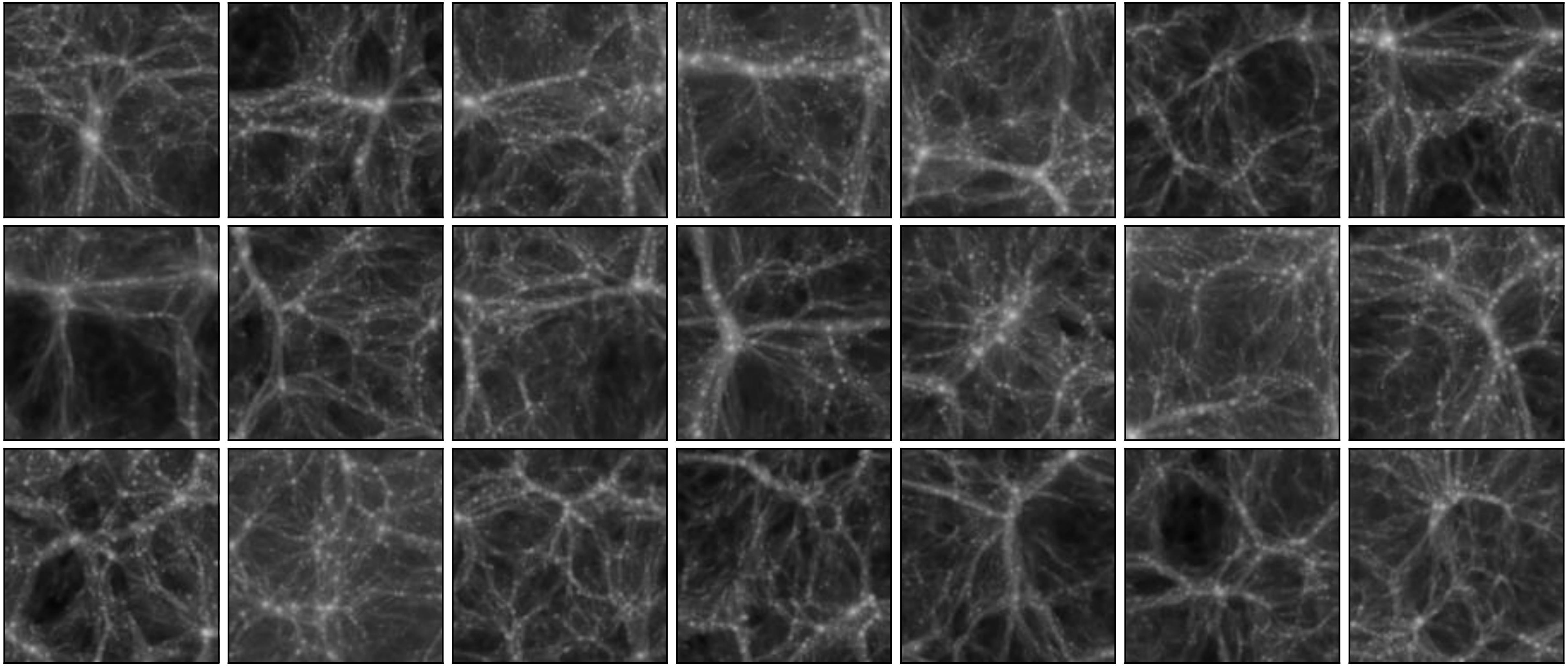
Summary

- We don’t know how to read the cosmological information written on the sky. We may be missing the most important part of the book

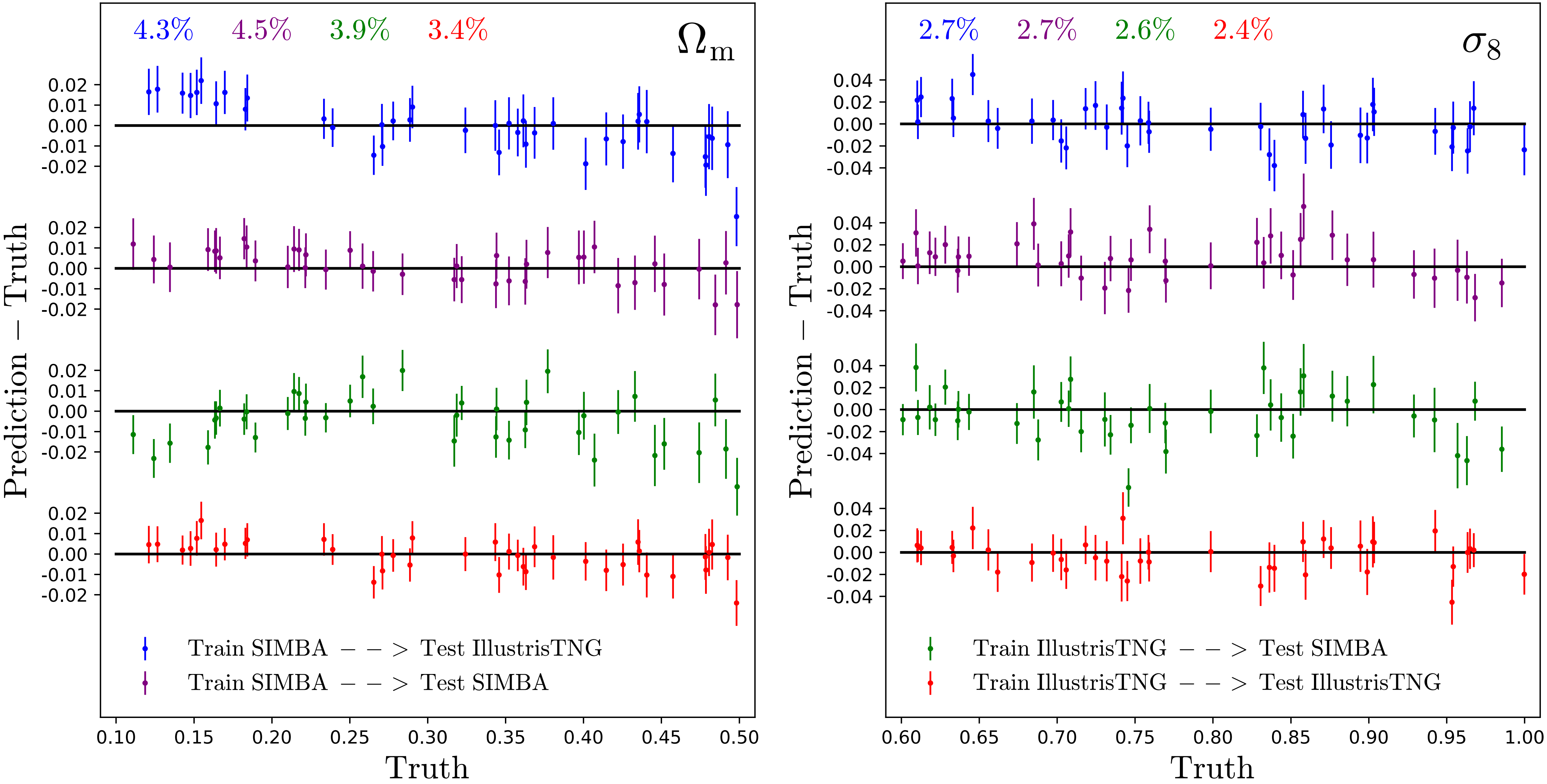
- The tools we typically use to extract information are suboptimal

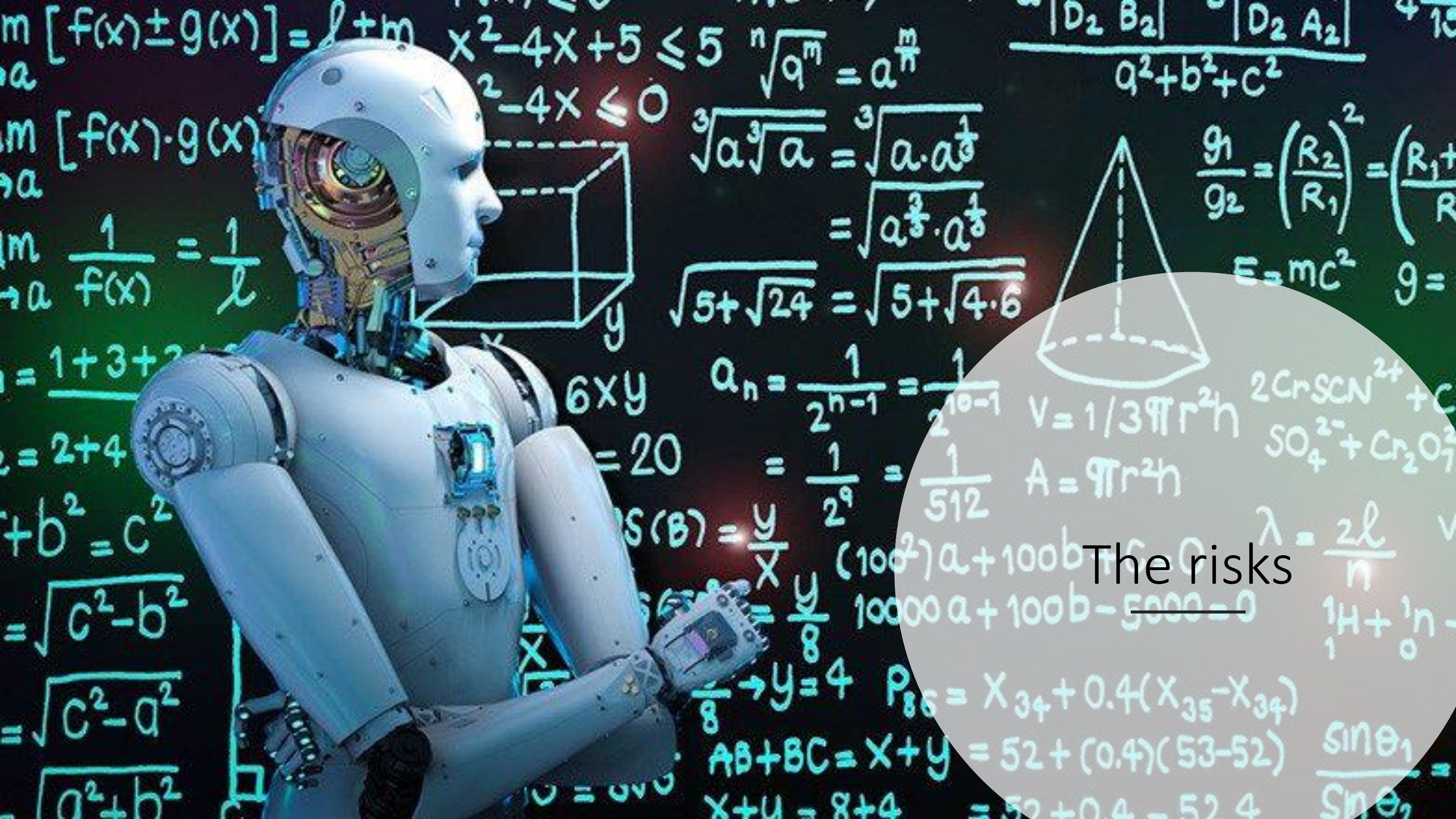
- Neural networks can find the optimal estimator to extract every single bit of cosmological information while marginalizing over uncertain astrophysical processes

Example II: Total matter



Robustness: total matter





The risks

The risks

Our simulations may never be perfect…
Can we train a perfect translator with imperfect sentences?

Do we need our simulations to overlap with reality?

How can we be sure it is not learning some artifacts/biased introduced by us?

How can we identify new physics in this formalism?

Conclusions

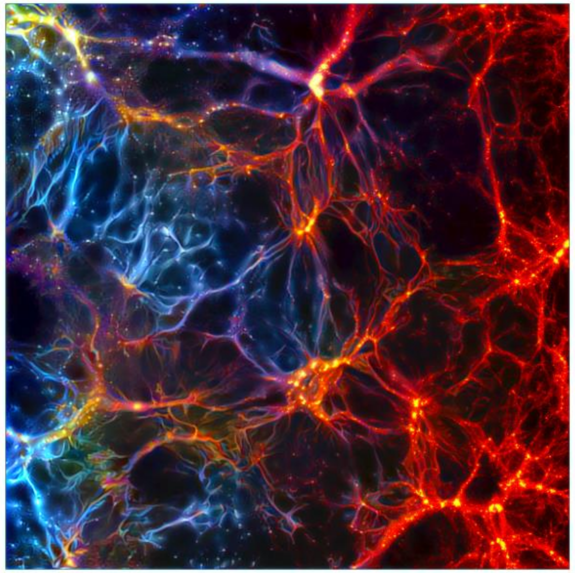
If we could simulate the Universe we could potentially learn <u>everything</u> about it.

How good should our simulations be to do this?

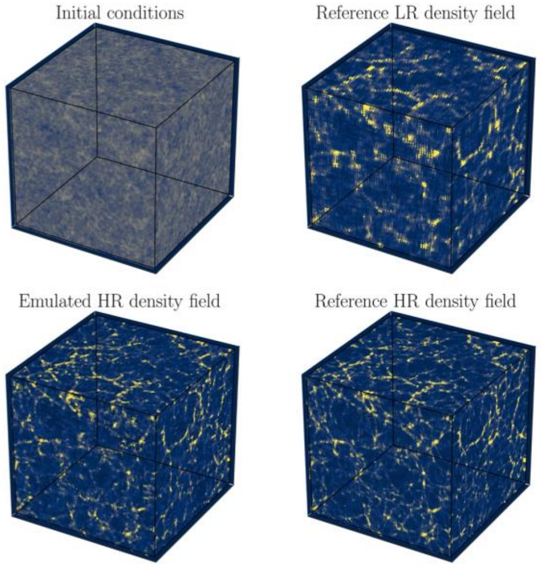
How many different kinds of simulations do we need to find a robust estimator that marginalize over subgrid physics, numerical effects, bugs…etc?

Vision/Dream

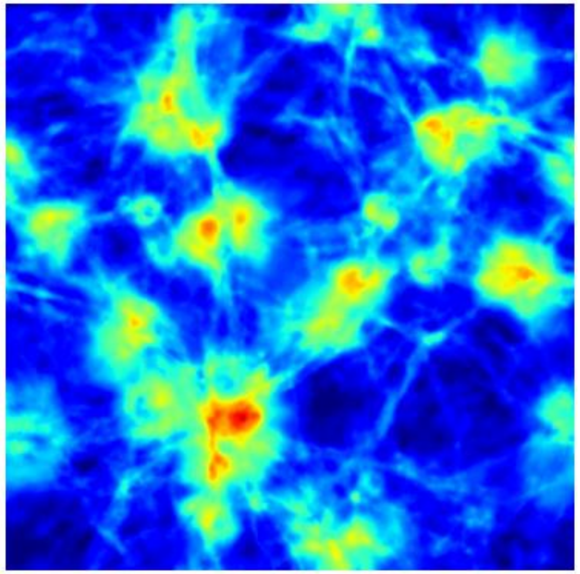
Quijote
Thousands of cosmologies



Super Resolution
$10^9\ h^{-1} M_\odot$



CAMELS
Thousands of astrophysics models



Likelihood-free inference
Extract all information. Marginalize over baryonic effects