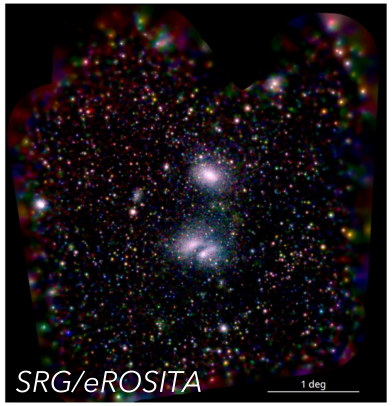
ML tools for Cosmology :
Simulation-based modeling and inference

Nicolas Cerardi
Postdoc @ Ecole Polytechnique Fédérale de Lausanne, Switzerland

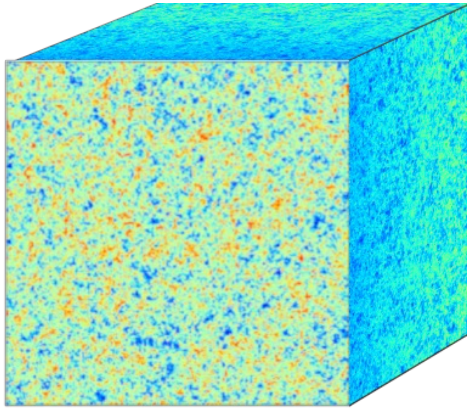
EPFL

Simulation-based methods for cosmology



Galaxy clusters
$z < 2$
X-ray detected
Catalogues

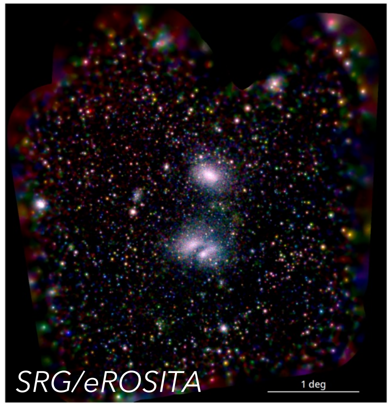
21cm signal from EoR
$7<z<9$
Low radio frequencies
Power spectrum



Cosmo & Astro parameters → Simulations → Predictions (summary stats) → Fit ← Observed data

Forward model everything (…if you can)
All known physics + instrumental effects
→ ***Simulation-based modeling***

Inference at an observable level
Intractable likelihood ?
→ ***Simulation-based inference***



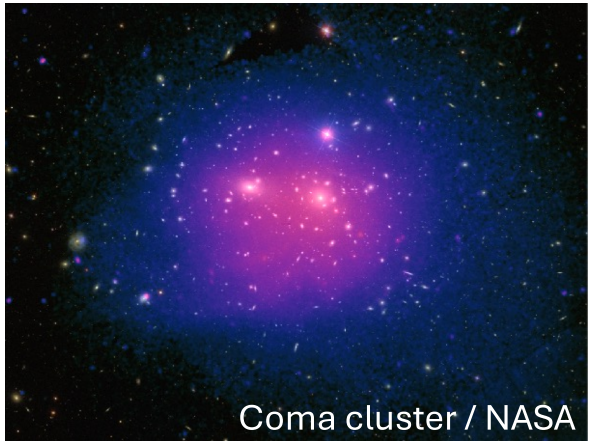
Galaxy clusters
z < 2
X-ray detected
Catalogues

21cm signal from EoR
7<z<9
Low radio frequencies
Power spectrum

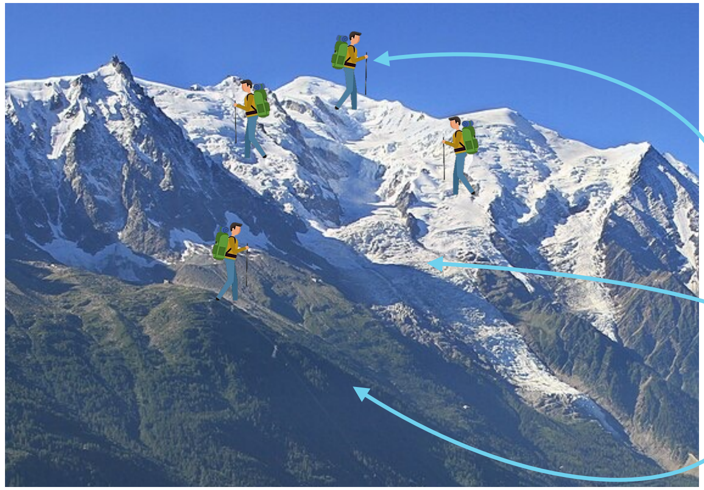
Simulation-based Cosmology with Galaxy Clusters

PhD work, under the supervision of Marguerite Pierre and François Lanusse @ CEA Paris-Saclay
Cerardi+25

Galaxy Clusters and Cosmology



=

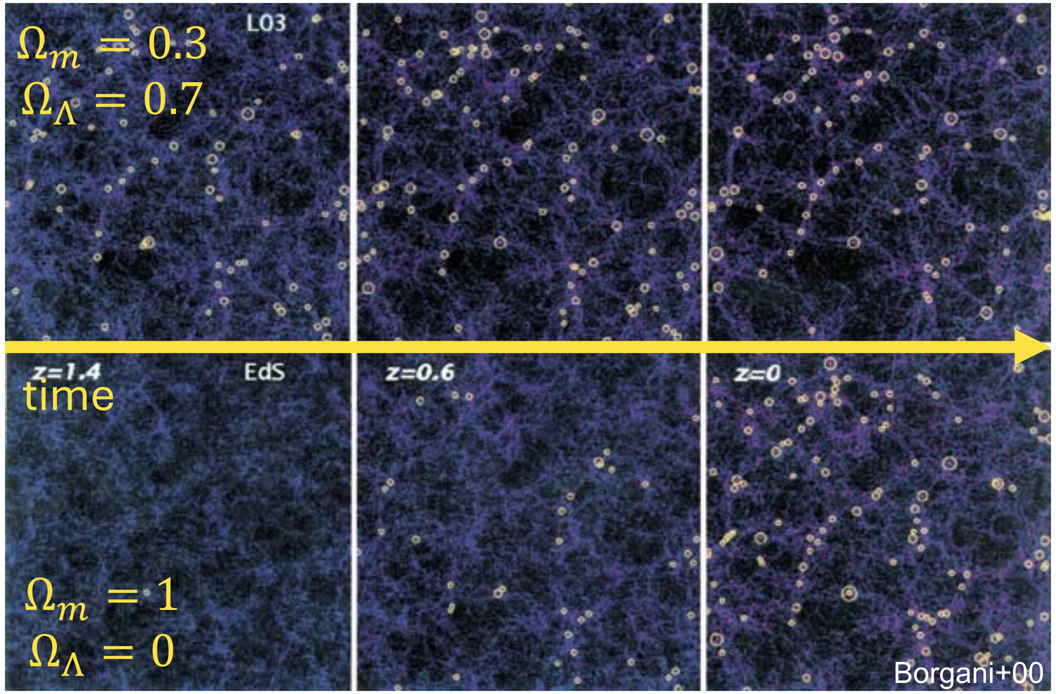


Mass composition

5% Galaxies - hikers

15 % Hot gas glaciers

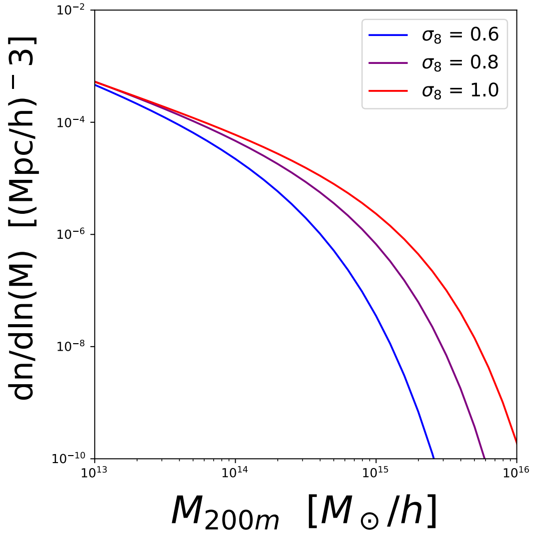
80% DM rocks



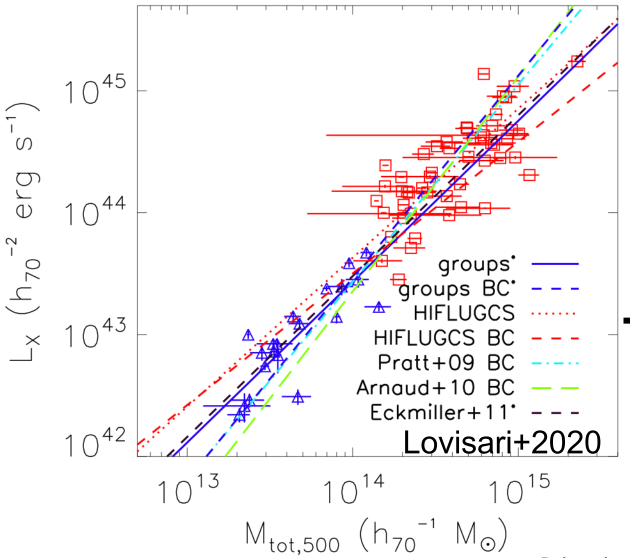
- Probe for the growth of structure and the geometry of the universe
- Population studies: abundancy, angular correlation…
- Standard candels: gas fraction…

Cosmology with clusters

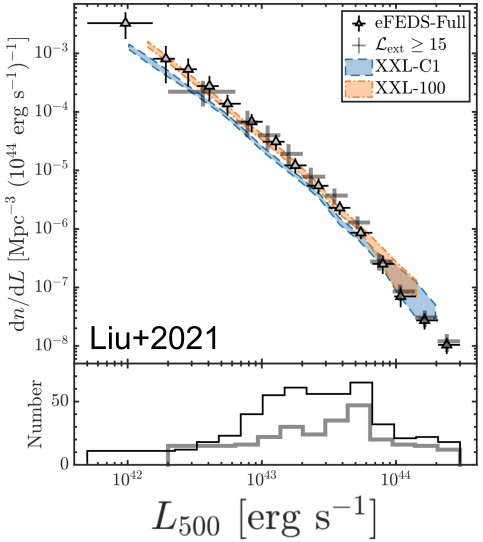
HMF
Fitted on simulations



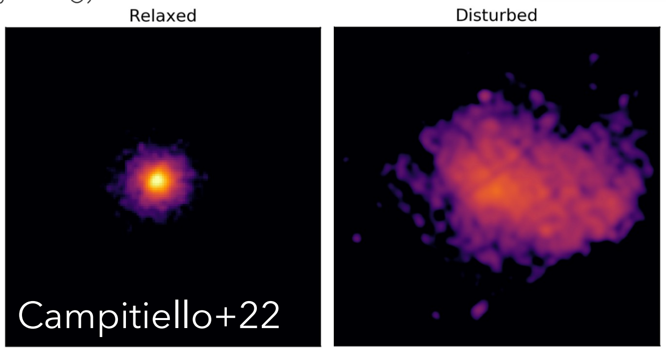
Scaling relations
Fitted on observed samples



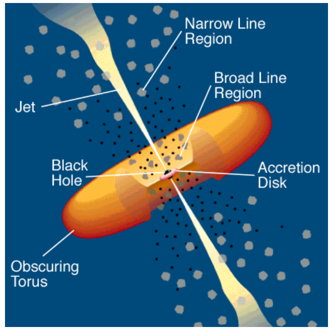
Sample of observed clusters
Certain level of purity and completeness



- Scaling relations have a theoretical explanation and prediction
- Empirical fit : $\log X = \alpha + \beta \log M + \gamma \log E(z) \quad \pm \sigma.$
- Nuisance parameters do not directly trace cluster physics, and are degenerate with the cosmology.



Cluster dynamics



AGN & SN feedback

From explicit to implicit inference

Explicit

- Poisson likelihood:

$$\ln p(x \mid \theta) = \left(\sum_i x_i \ln N_i(\theta)\right) - N_{tot}(\theta)$$

- Predicted number counts:

$$N(\theta) = \int p(S \mid \mathcal{O}) p(\mathcal{O} \mid M, z, \theta) \frac{\mathrm{d}N(\theta)}{\mathrm{d}M\mathrm{d}z\mathrm{d}V} \mathrm{d}V$$

Selection function

Scaling relations

HMF

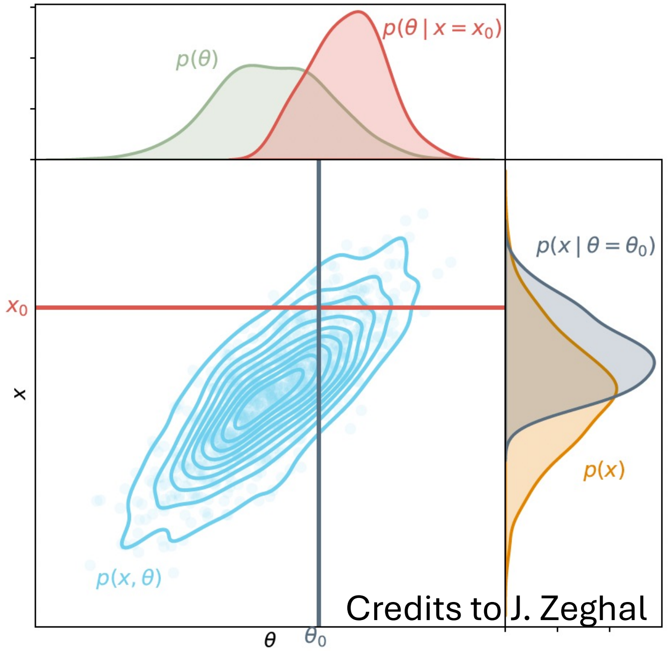
Implicit / Likelihood-free / Simulation-based

- Sample the **joint distribution**:

$$\theta_i \sim p(\theta), \;\; x_i \sim p(x \mid \theta = \theta_i)$$

- Train a **density estimator**:

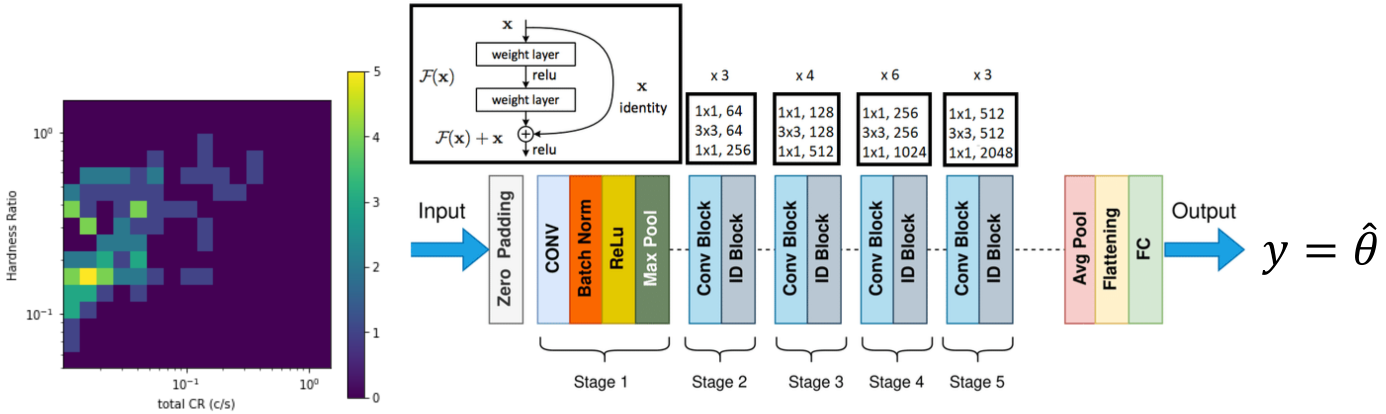
$$p(\theta \mid x = x_0) \propto q_\varphi(\theta \mid x = x_0)$$



Credits to J. Zeghal

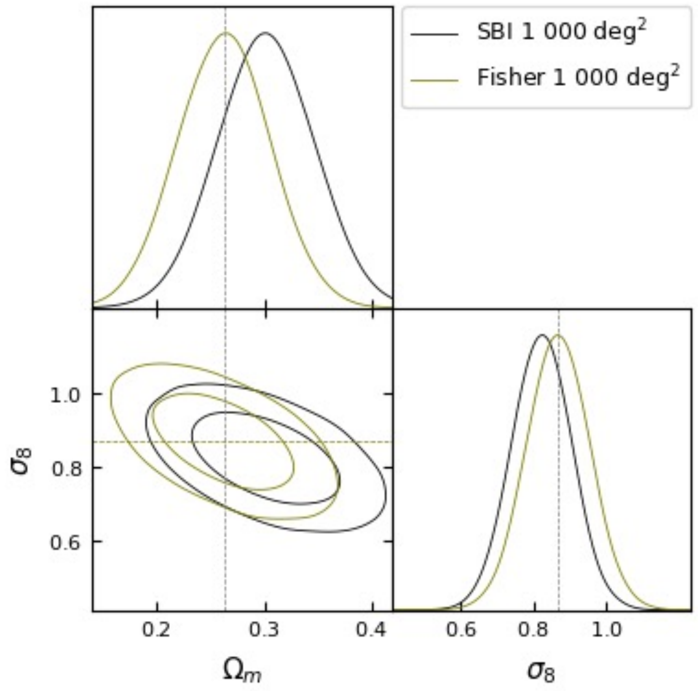
Cluster cosmology with SBI

Kosiba, Cerardi+24:
SBI on cluster cosmology, still using scaling relations in the modeling



1. Compress the summary statistics before inference

$$y = f(x)$$

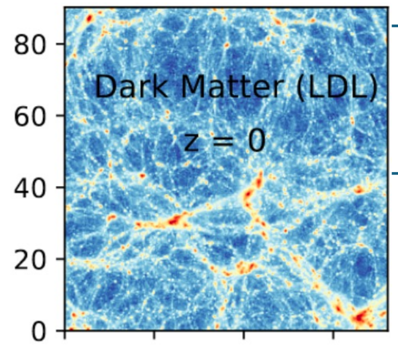


2. Neural posterior estimation & comparison with Fisher analysis

$$p(\theta \mid y = y_0) \propto q_\varphi(\theta \mid y = y_0)$$

Fast simulation-based modelling

Lagrangian Deep Learning (LDL, Dai+21): Baryon pasting on DMO simulations



Dai+21

$\mathbf{x}$

$f(\delta) = (1+\delta)^{\gamma}$
Potential

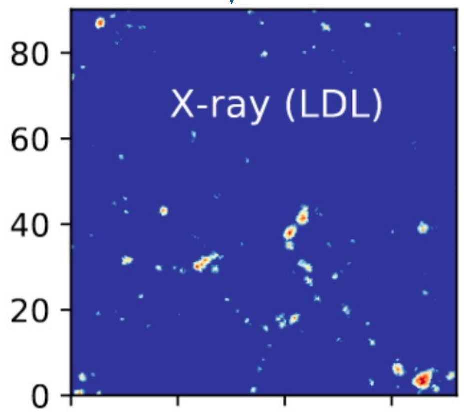
Corrected Potential (Fourier Space)

$\mathcal{S}_{\varphi}(k)$
1D Fourier filter

$\mathrm{d}\mathbf{x} = \alpha \nabla \mathcal{F}^{-1}(\mathcal{S}_{\varphi}(k) f(\delta))$
Displacement field

$\delta'(\mathbf{x} + \mathrm{d}\mathbf{x})$
Updated field

$F(\mathbf{x}) = \mathrm{ReLU}(b_1(1+\delta')^{\mu} + b_0)$
Non-linear activation



Advantages

- Adopts physical principles.
- Lighter model than deep generative models.

Extension : conditioning

$$\Omega_m, \sigma_8, A_{feedback} \xrightarrow{NN} \gamma, \varphi, \alpha, b_0, b_1, \mu$$

Training simulations

CAMELS dataset

- Thousands of simulated volumes.
- HD Codes : **IllustrisTNG** / SIMBA / Astrid / Magneticum.
- Fiducial simulations: 27x(50 Mpc/h)^3
- Varied simulations: 500x(25 Mpc/h)^3

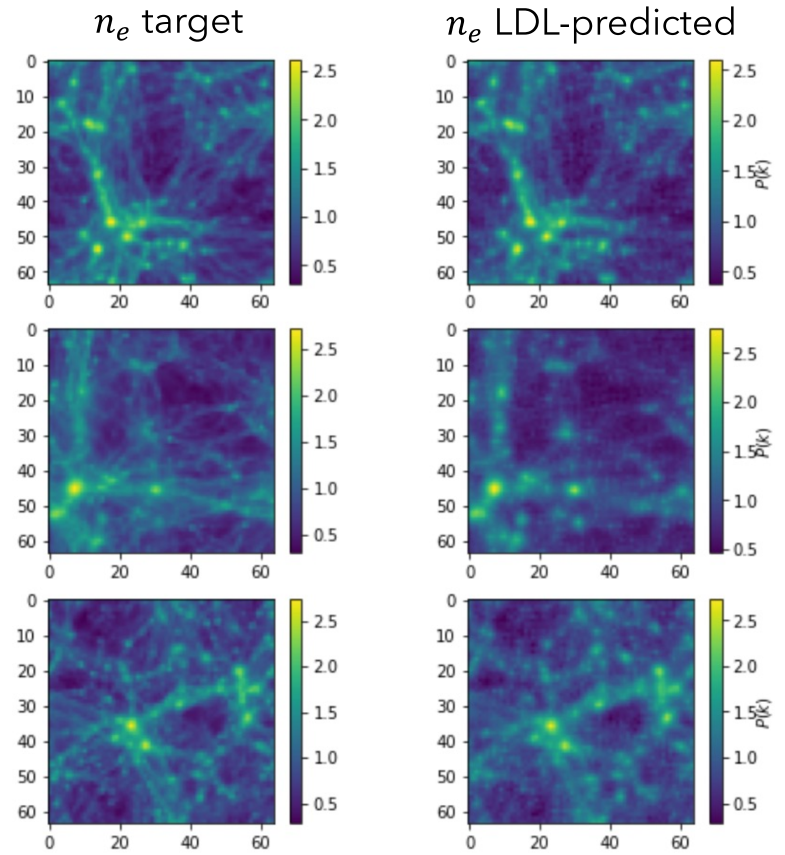
  Cosmology : Ω_m , σ_8

  SN : A_{SN1}, A_{SN2}, energy and speed of galactic winds.

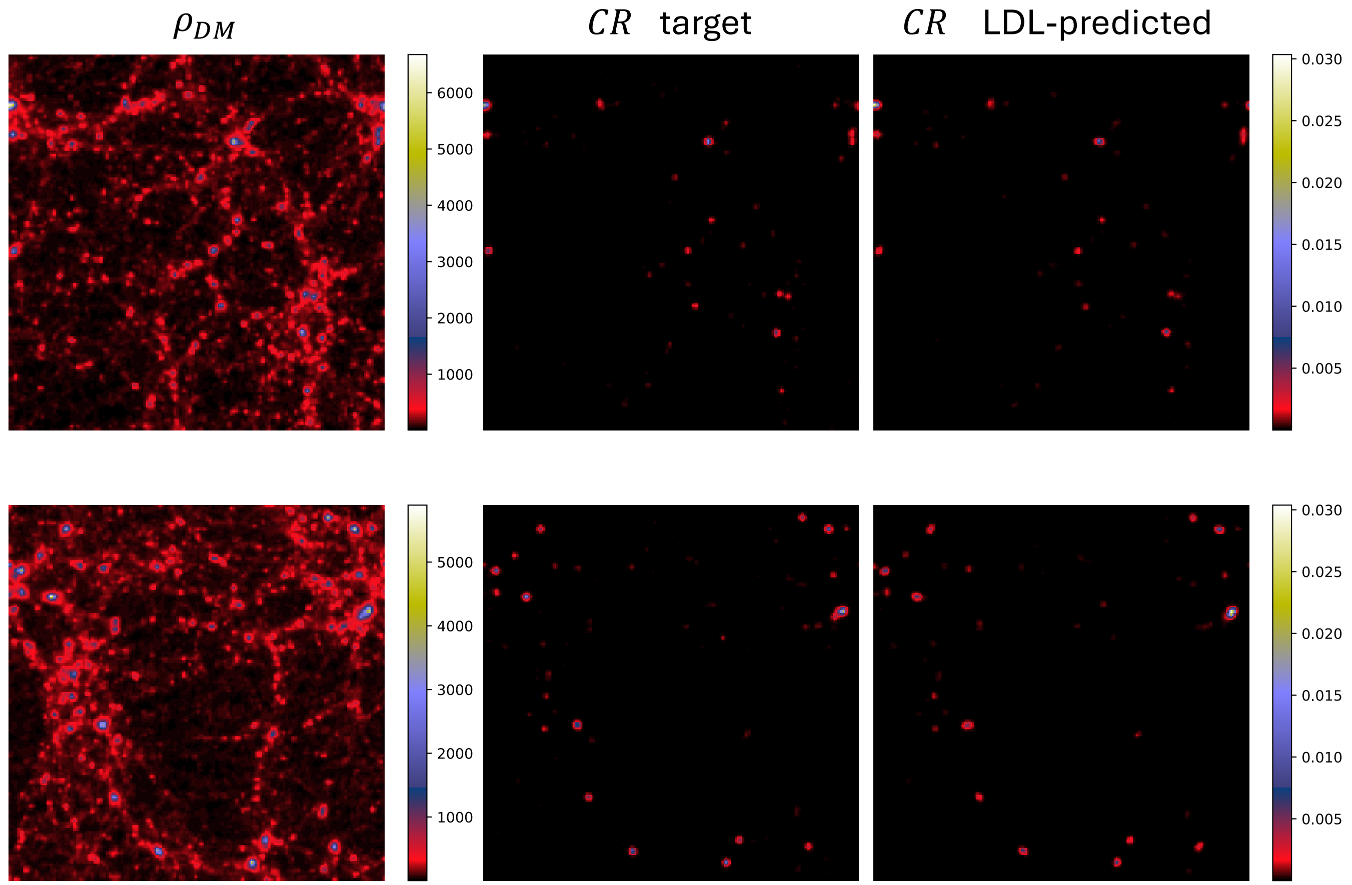
  AGN : $A_{AGN1}, A_{AGN2},$ power and burstiness of kinetic mode / low accretion rate.

Emulated fields

Baryonic properties, CAMELS/LH25



X-ray emission, CAMELS/CV50



End-to-end pipeline

Compiled on GPUs: ~30s to make a mock X-ray map!

ICs, cosmological parameters → DMO fields → Baryonic properties n_e et T → X-ray mocks → Cluster catalogue → X-ray observable diagram → Posteriors

JaxPM

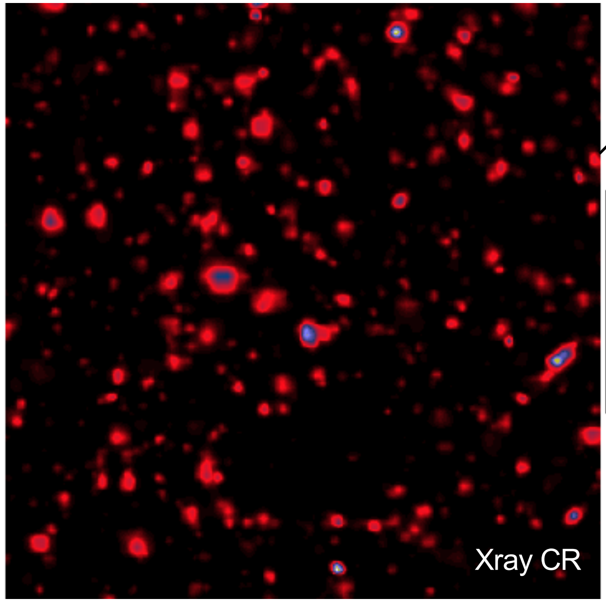
LDL

Lightcones

Detection

Histogram

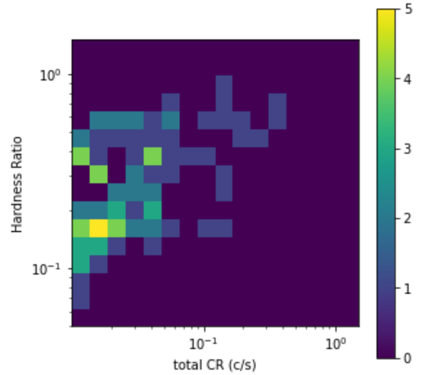
Inference



$\Omega_m = 0.47$
$\sigma_8 = 0.63$
$A_{SN1} = 0.93$
$A_{AGN1} = 0.32$
$A_{SN2} = 0.52$
$A_{AGN2} = 0.54$

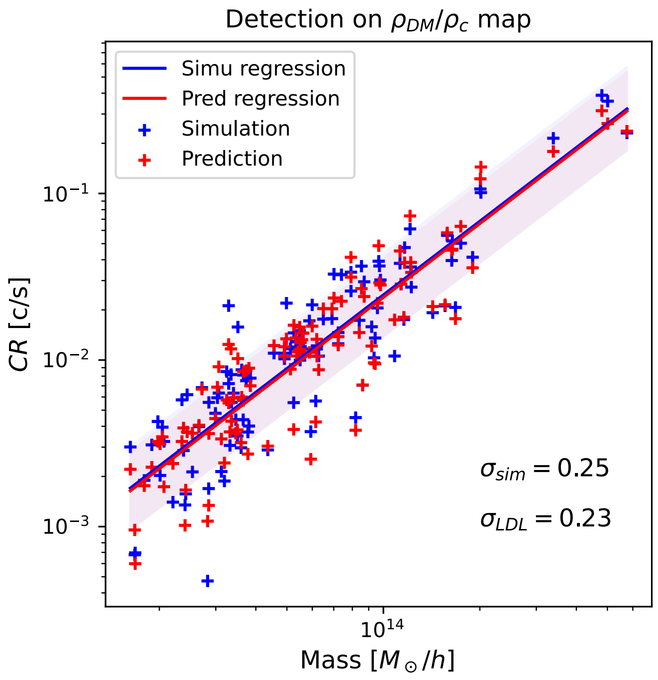
Simplified detection

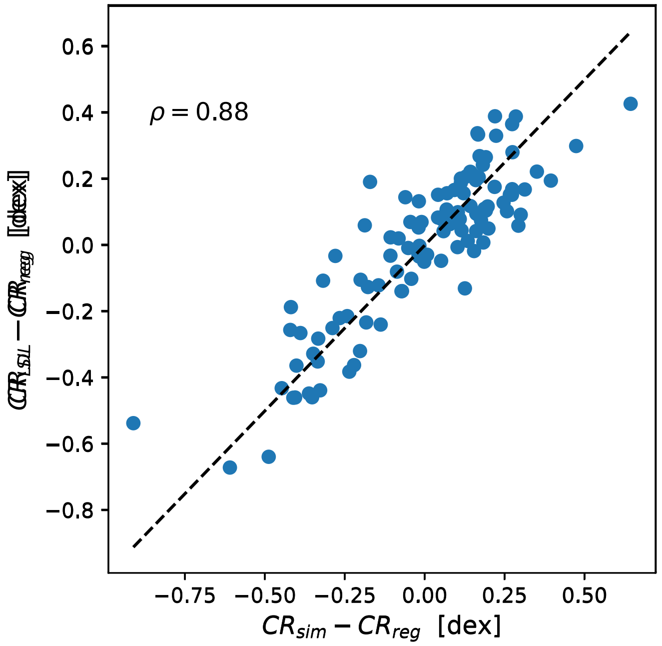
- Simplifications: no AGNs, no galactic absorption
- Detection SExtractor-like on unnoised CR maps
- Measures of clusters' CR and HR



Emulated scaling relations

- Reproduction of CR-M relation from the fiducial model @ z=0.21.
- Correlated deviations : LDL benefits from the 3D information on each halo environments.



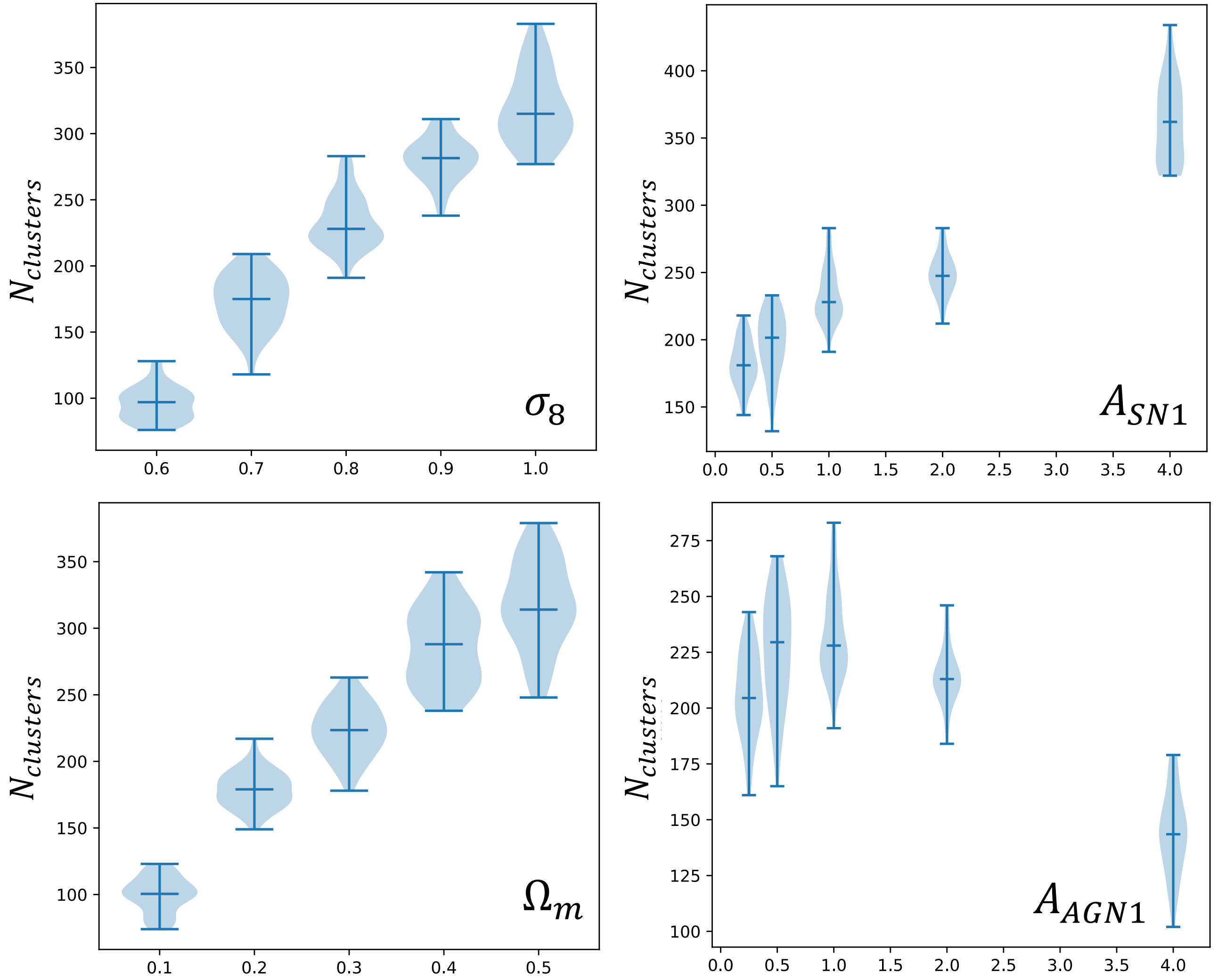


- Too few clusters in CAMELS/LH to conduct the same test

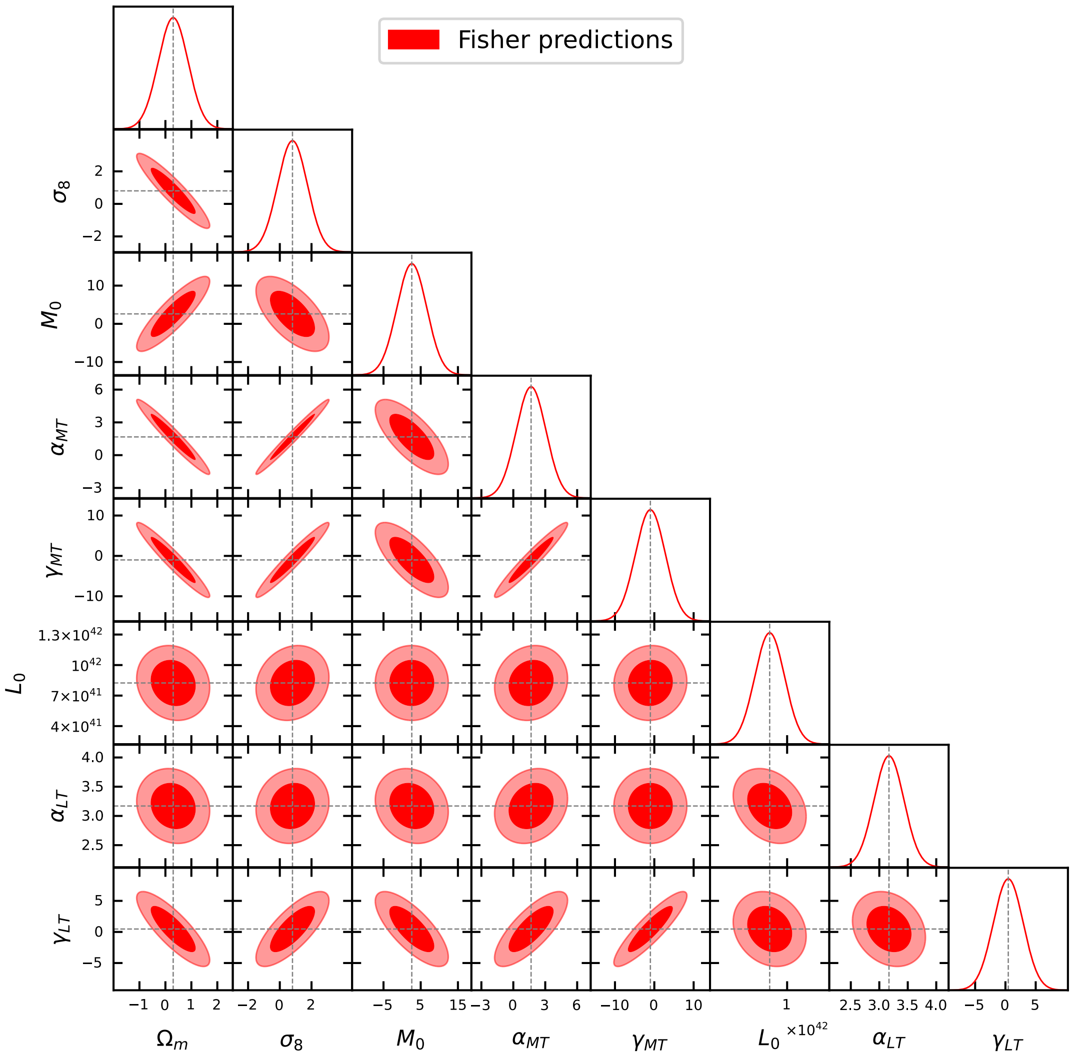
Pipeline sensitivity

Individual variation of each parameter

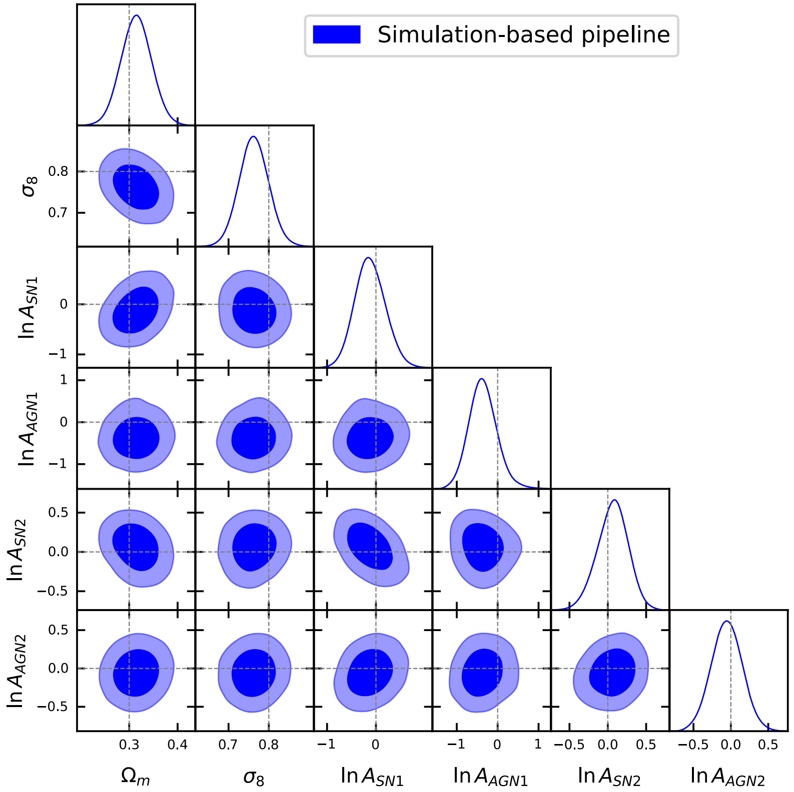
- 48 surveys of 50 deg^2 for each parameter value
- Strong sensitivity to cosmological parameters (full pipeline)
- Strong response to SN retroaction but weak to AGN parameters (extended LDL)



Explicit vs Implicit modelling

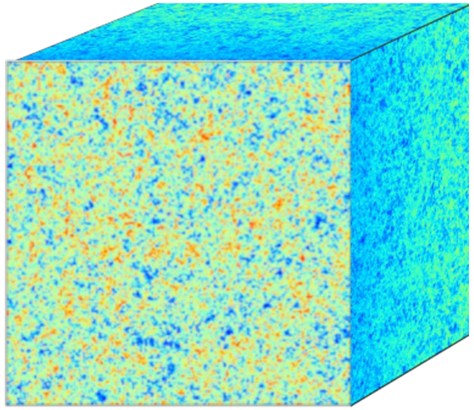


- Full posteriors for explicit model (empirical scaling relations) and simulation-based model.



Galaxy clusters
z < 2
X-ray detected
Catalogues

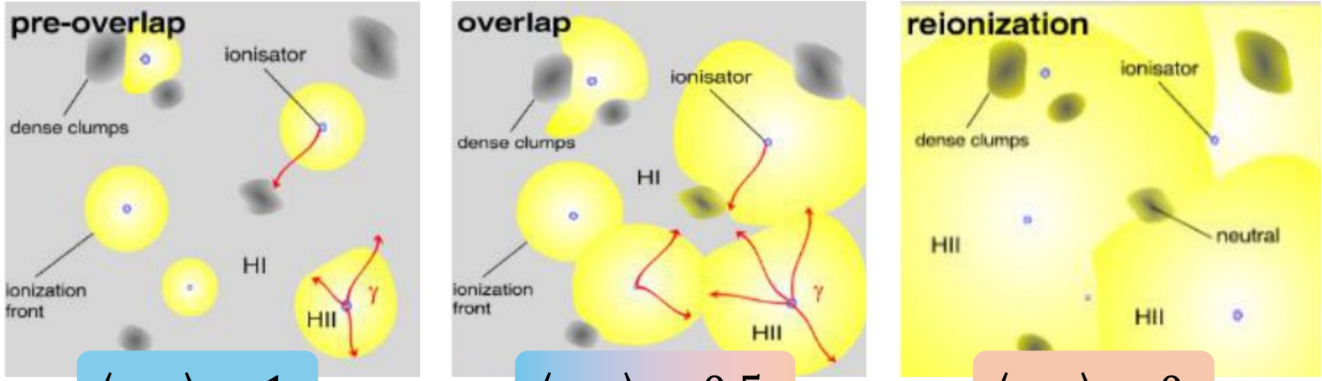
21cm signal from EoR
7<z<9
Low radio frequencies
Power spectrum



Epoch of Reionization and neutral fraction inference

With the SEarCH team: Michele Bianco, Sambit Giri, Massimo de Santis, Emmanuel de Salis, Davide Piras, Philipp Denzel…
De Salis+25, accepted at EPIA

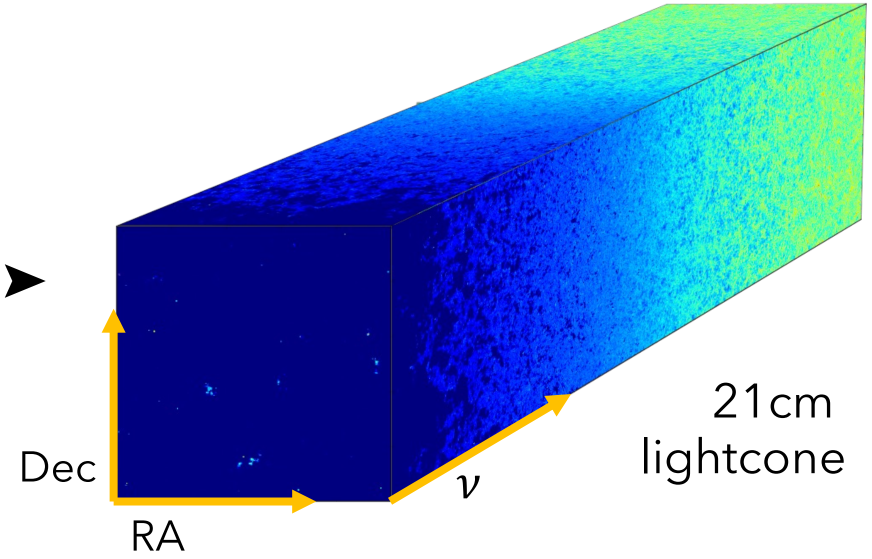
Epoch of Reionization in nutshell



$\langle x_{HI} \rangle \approx 1$ $\langle x_{HI} \rangle \approx 0.5$ $\langle x_{HI} \rangle \approx 0$

time →

➤

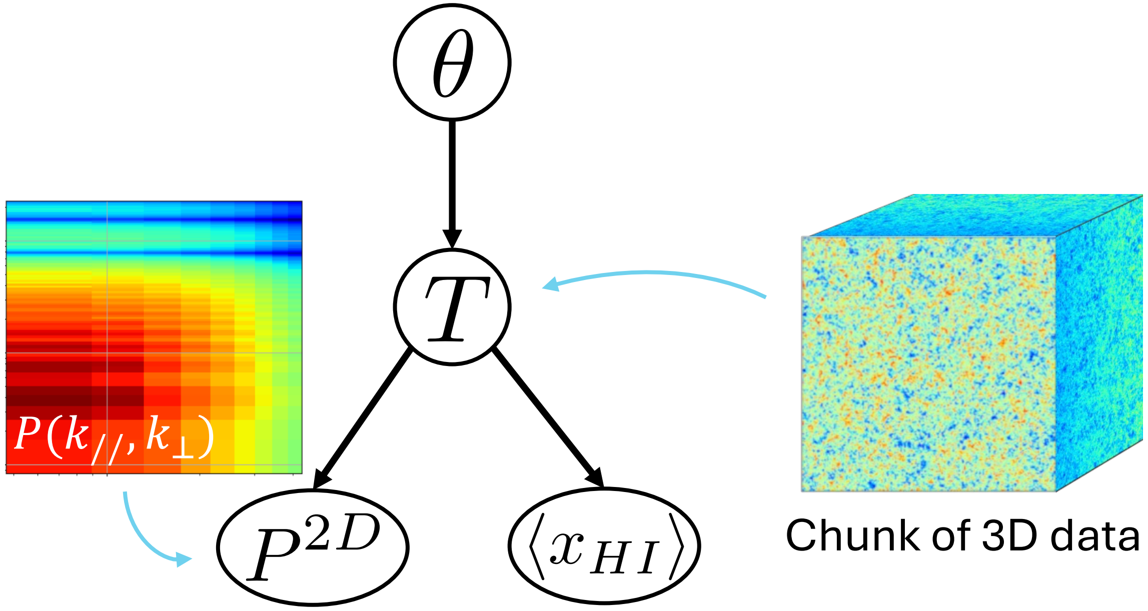


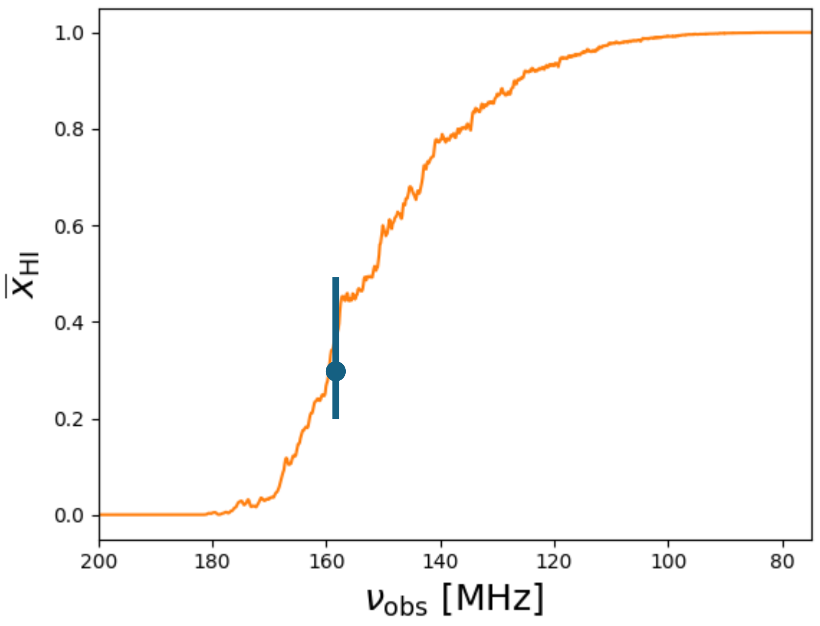
$$\delta T_b(\theta,z) \propto [1+\delta_b(\theta,z)]\, x_{HI}(\theta,z)$$

SKA Data Challenge: Inference

- Generate 21cm lightcones with 21cmFAST (Mesinger+11, Murray+20)
- 6 free nuisance astrophysical parameters

  $\theta = (F_\star, \alpha_\star, F_{esc}, \alpha_{esc}, M_{turn}, T_{star})$
- No foregrounds ⚠️
- Instrumental noise ✅





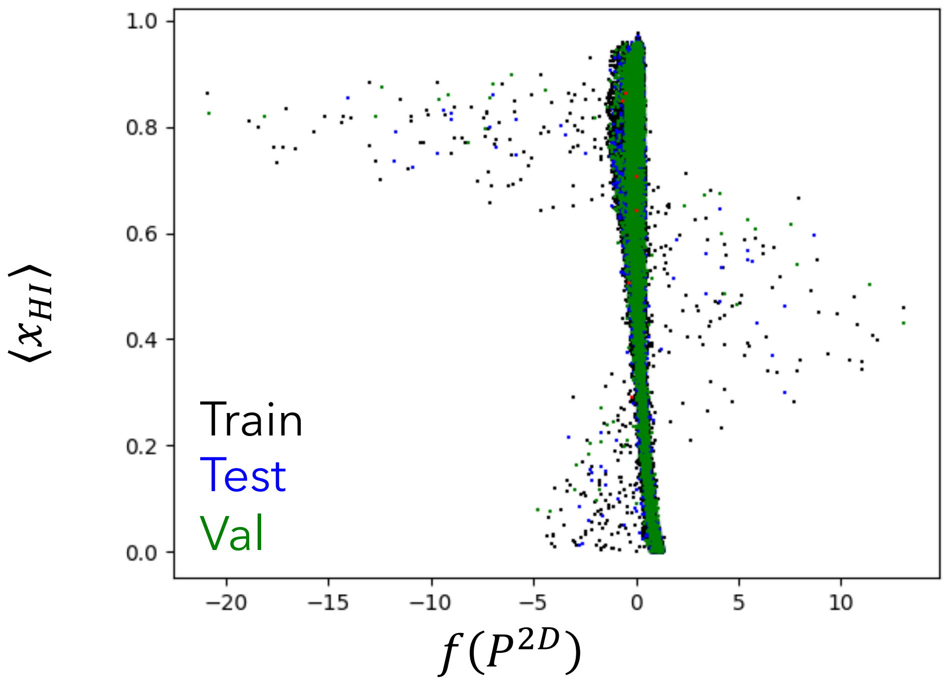
Goal : constrain average neutral fraction

$$p(\langle x_{HI} \rangle \mid P(k_\parallel, k_\perp))$$

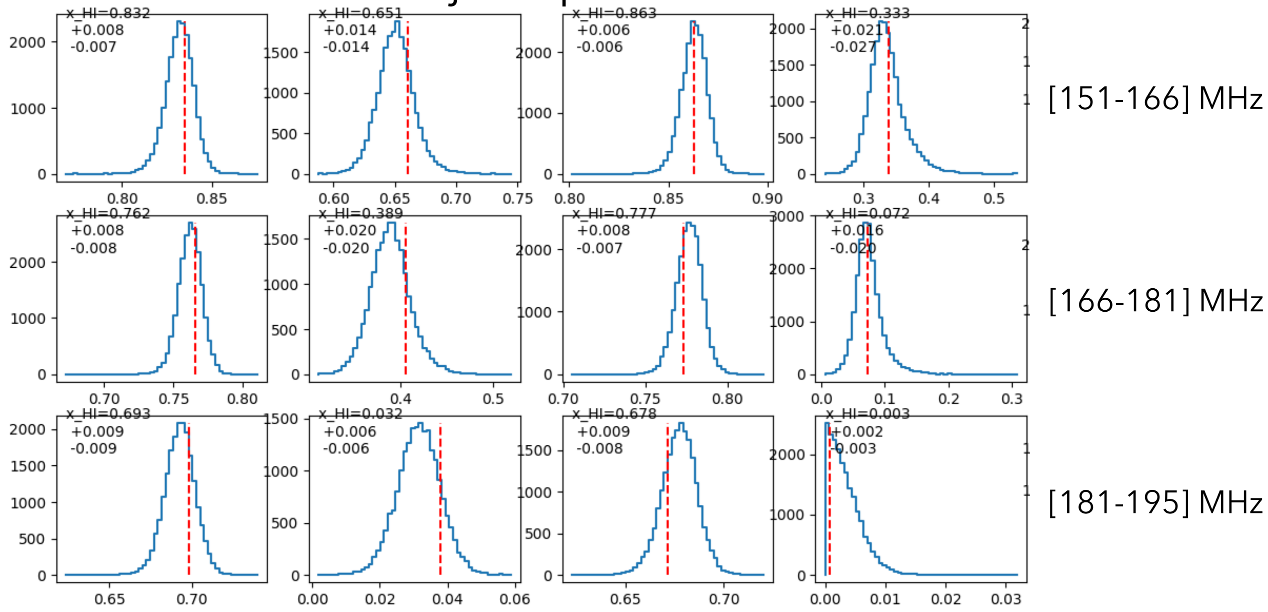
Inference methods

- No regression anymore
- Task of the ResNet : maximize mutual information between $f(P^{2D})$ and $\langle x_{HI} \rangle$ (VMIM, Serdega+20)
- Estimate joint posterior on the bands [151-166], [166-181], [181-195] MHz

1. Learn optimal compressions



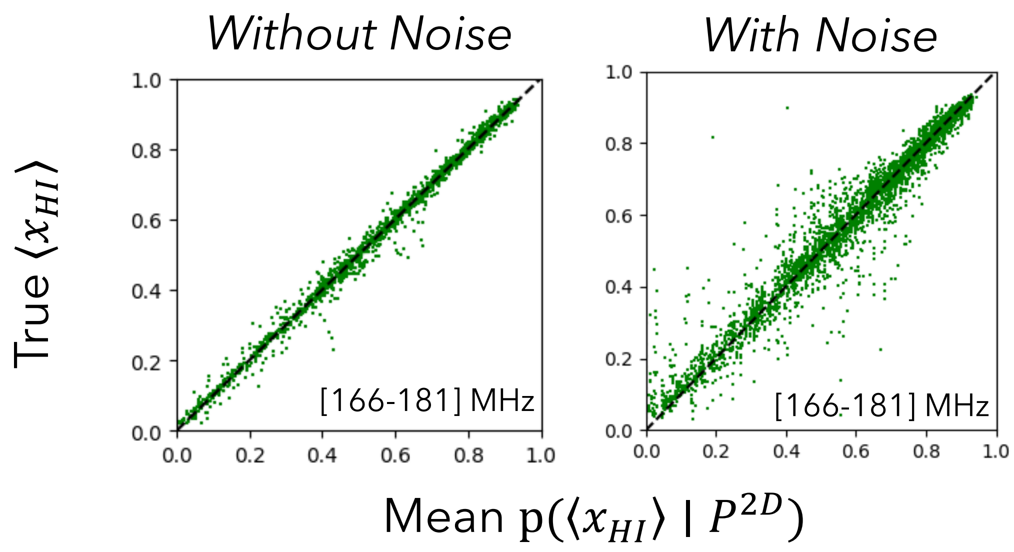
2. Learn joint posterior

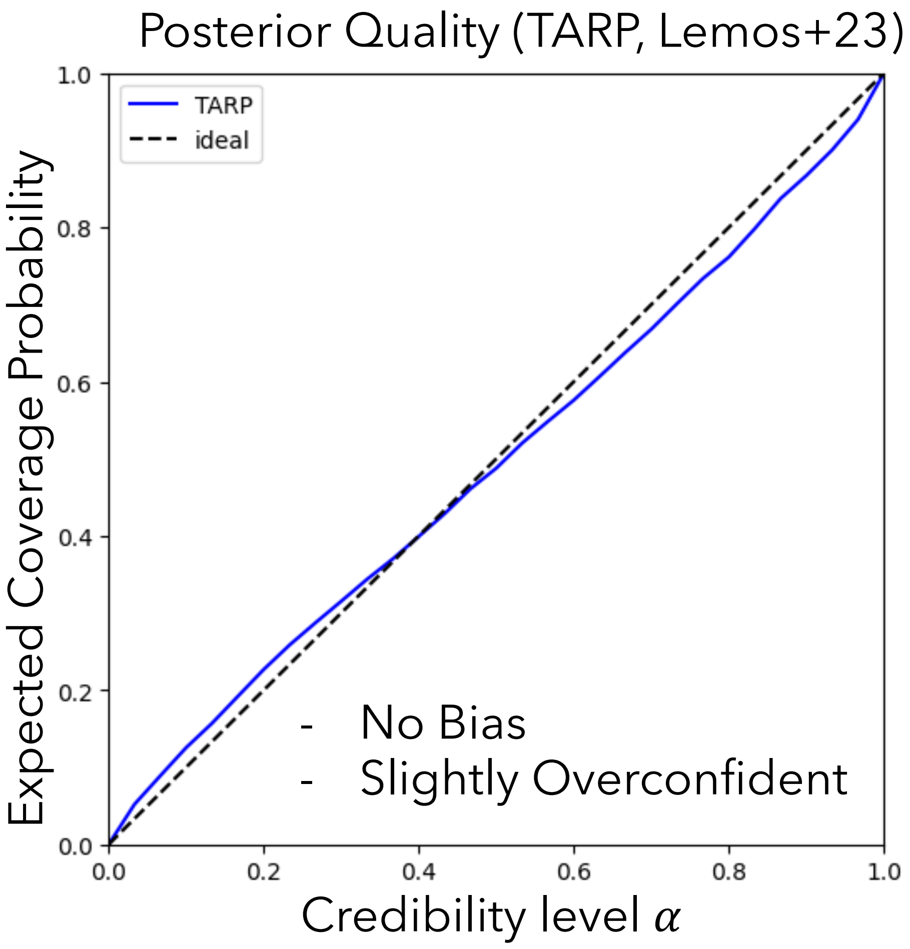


Posterior samples, Truth

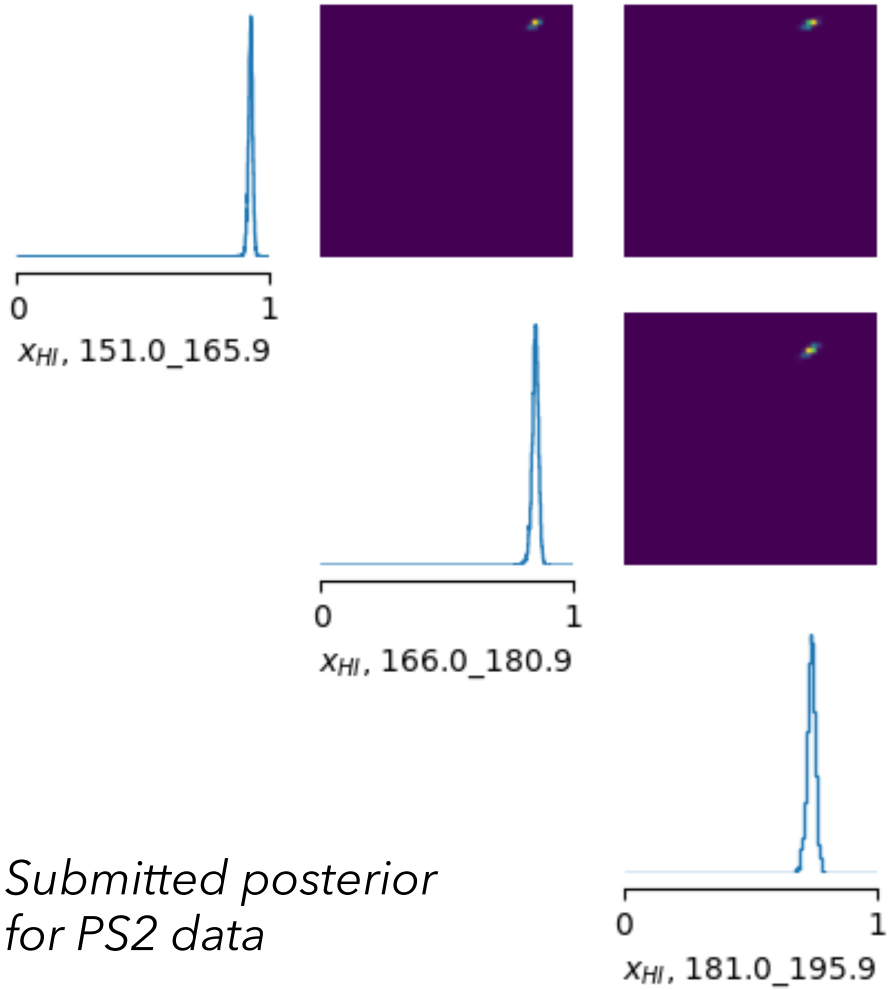
Inference Results

- Compare posterior mean vs truth





Data Challenge results



Submitted posterior for PS2 data

Data Challenge preliminary rankings

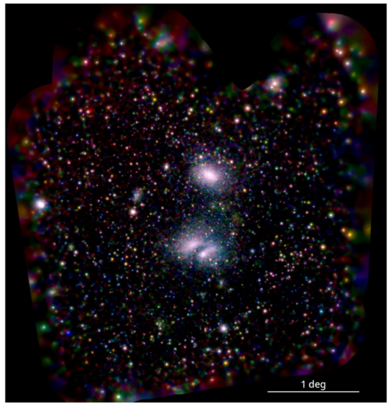
Top 10 scoring teams:

Top teams scoring over 0.1

1. Cantabrigians
2. Akashanga
3. Mordern SEarCH
4. Traditional SEarCH
5. ToSKA-model selection
6. ToSKA Explicit likelihood
7. YEYE

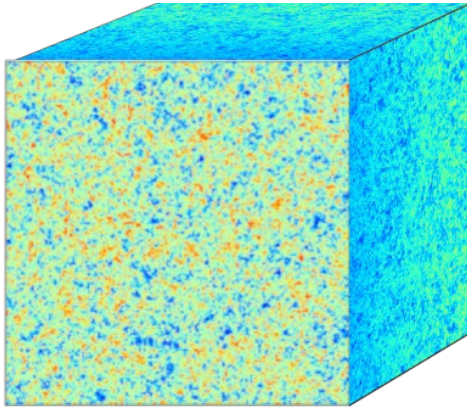
Out of 27 teams

Conclusions



Galaxy clusters
z < 2
X-ray detected
Catalogues

21cm signal from EoR
7<z<9
Low radio frequencies
Power spectrum



- ML to speed up a simulation pipeline
- SBI approach to infer cosmo + astrophysical parameters
- Method relies on accurate MHD cosmological simulations…

- Applied a SBI framework for neutral fraction inference
- The data model seems to play a crucial role
- Needs to include foreground residuals