Improvements in the exploration of the CD-EoR in the era of AI

Stockholm University

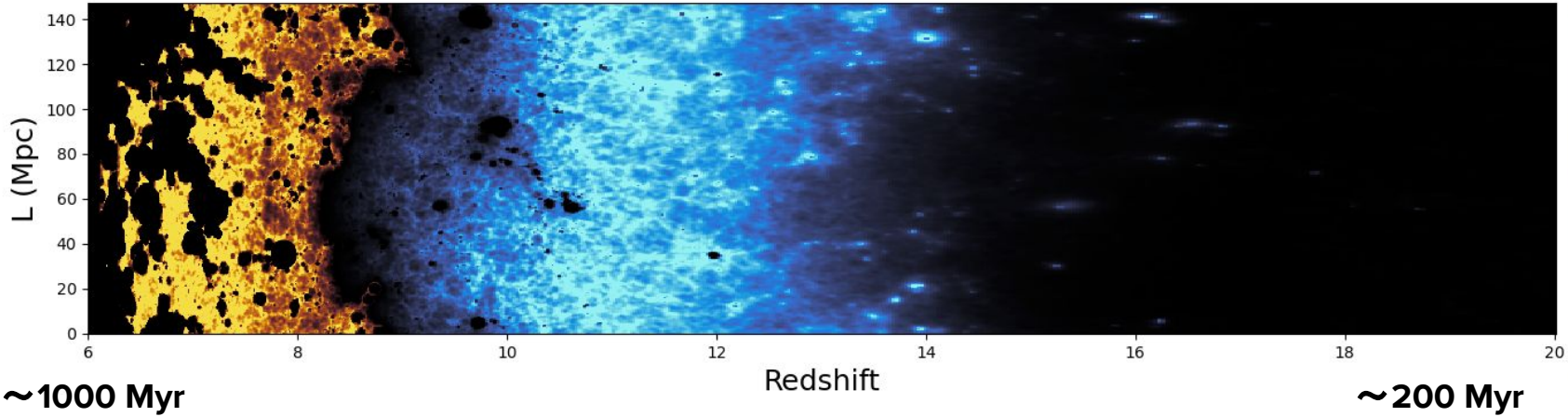
Sambit Giri

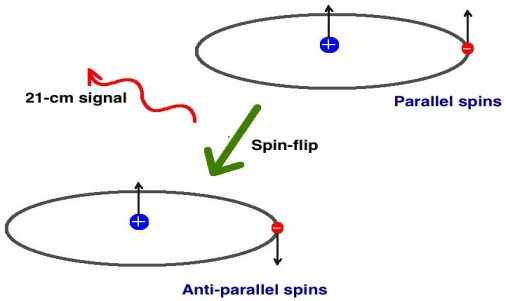
NORDITA fellow

NORDITA

18 July 2024

The 21-cm signal will probe the intergalactic medium





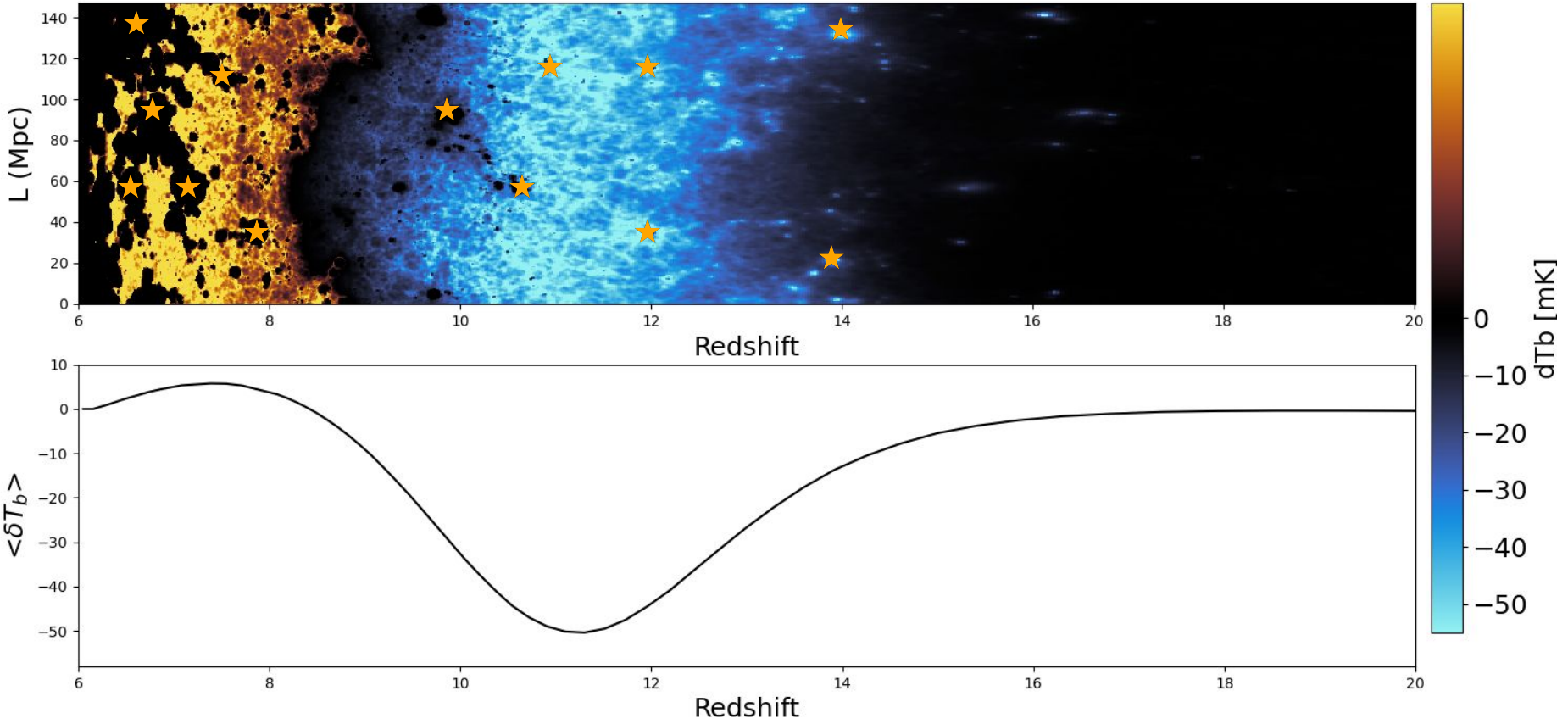
$$\delta T_{\rm b} \propto x_{\rm HI}(1+\delta)\left(1-\frac{T_{\rm CMB}}{T_{\rm S}}\right)$$

Neutral hydrogen

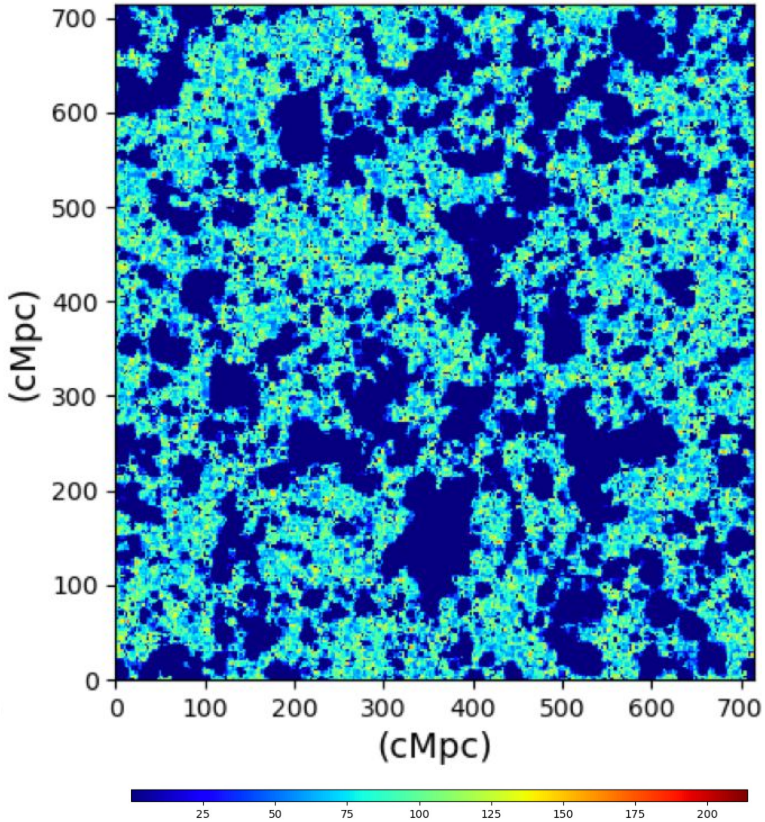
Baryon density

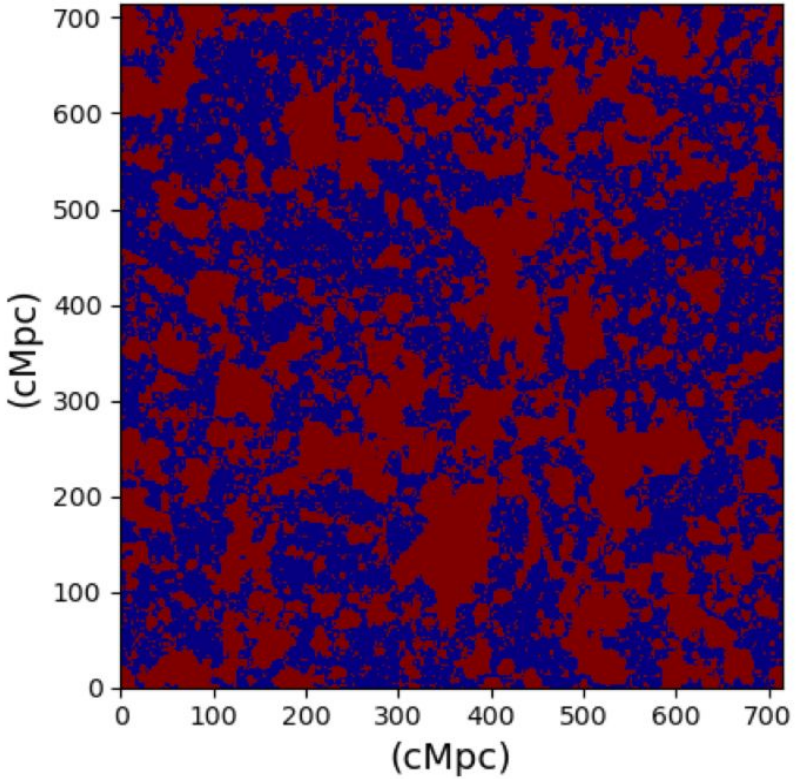
Spin temperature

Evolution of the signal is driven by early luminous sources

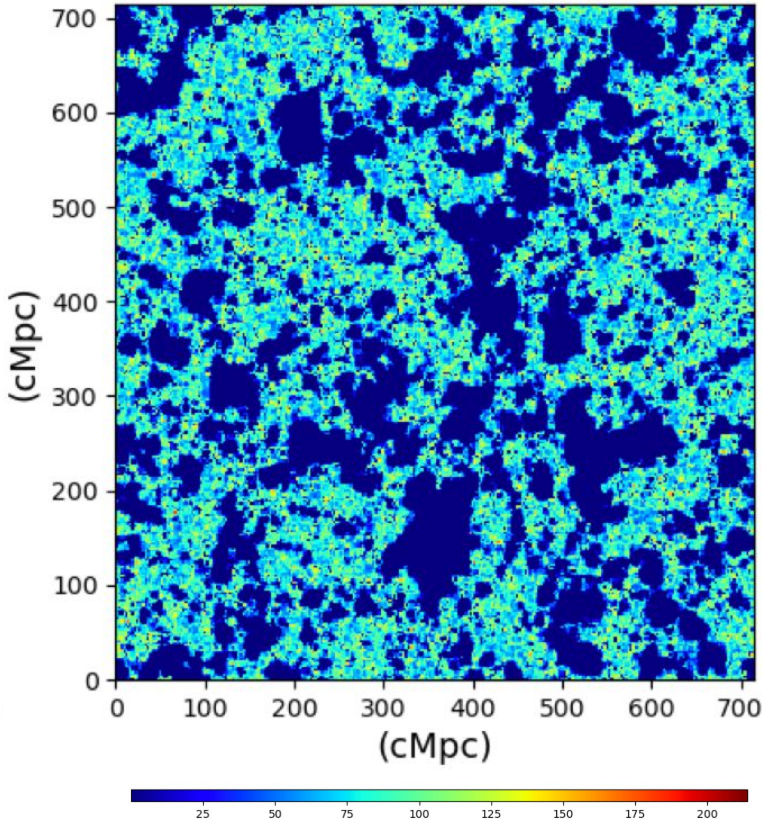


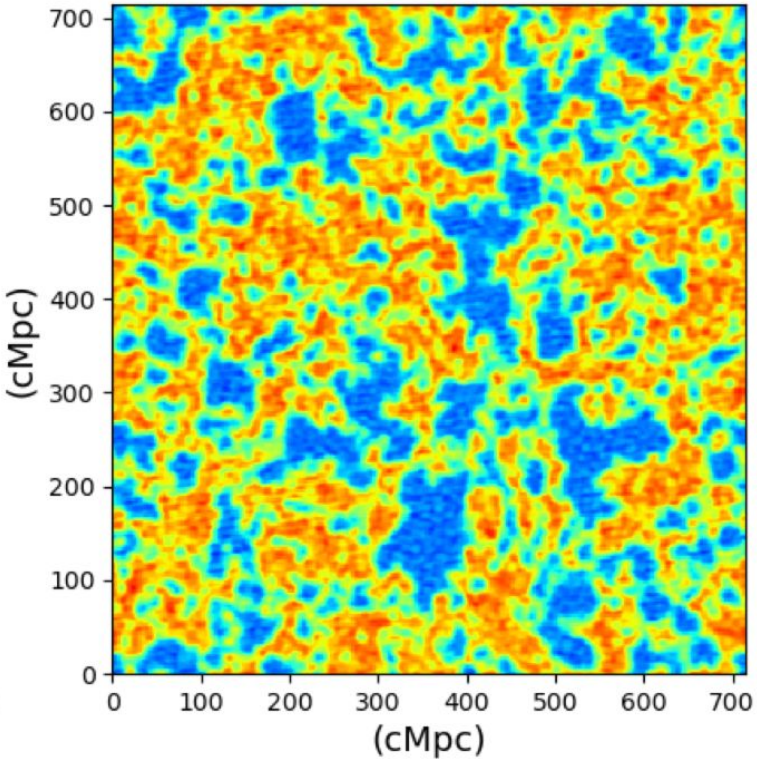
Ionized regions are imprinted onto the 21-cm signal





Simulated SKA-Low image





Instrumental effects using Tools21cm (Giri+2018b; Giri+2020)

How do we study ionized regions in SKA images?

Survey	Data per night/day	Galaxies	Cost	Scientists
DES	1 TeraB	~300 Million (all observed)	~$70M	~400
DESI	40 GigaB	~35 Million (12M observed)	~$70M	~600
Rubin-LSST	15 TeraB	~Billions	~$1.0B	~1000
Euclid	850 GigaB	~Billions	~$1.5B	~1500
SKA	1 PetaB	~Billions	~$1.3B	~1000

Big Data!

(Lahav 2023)

Low frequency component of Square Kilometre Array (SKA-Low)

Methods to automatically identify meaningful features

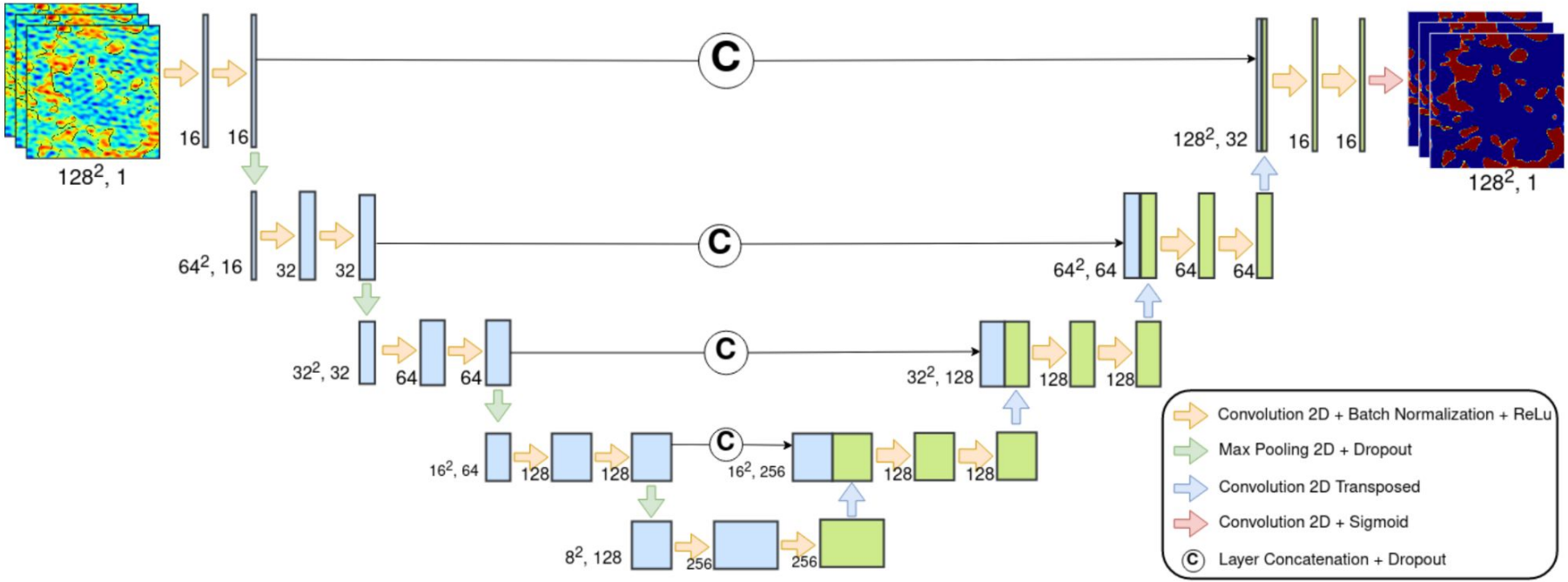
Survey	Data per night/day	Galaxies	Cost	Scientists
DES	1 TeraB	~300 Million (all observed)	~$70M	~400
DESI	40 GigaB	~35 Million (12M observed)	~$70M	~600
Rubin-LSST	15 TeraB	~Billions	~$1.0B	~1000
Euclid	850 GigaB	~Billions	~$1.5B	~1500
SKA	1 PetaB	~Billions	~$1.3B	~1000

(Lahav 2023)

- **Superpixel method**
  Image processing (**Giri**+2018b)
- **SegU-Net**
  Deep learning (Bianco, **Giri**+2021;2024)

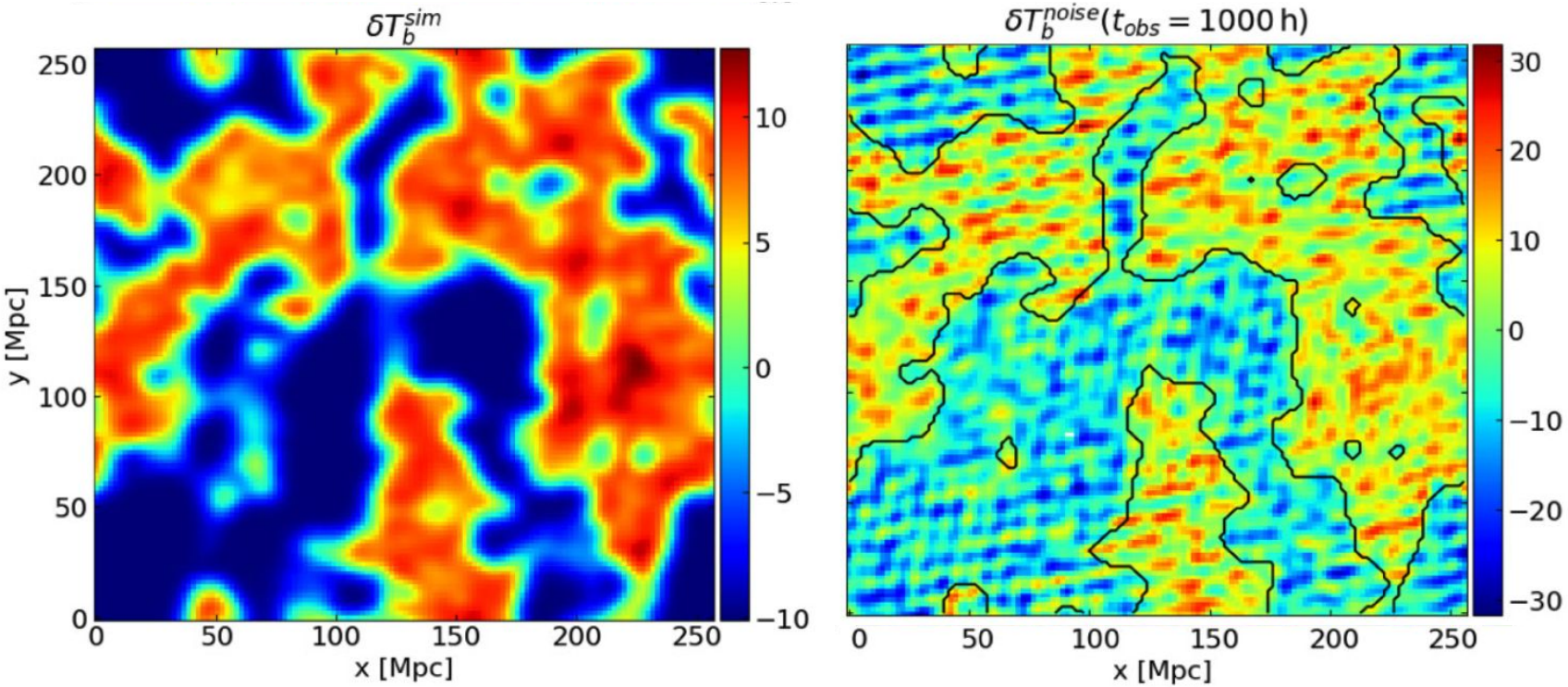
Low frequency component of Square Kilometre Array (SKA-Low)

SegU-Net: deep learning segmentation model



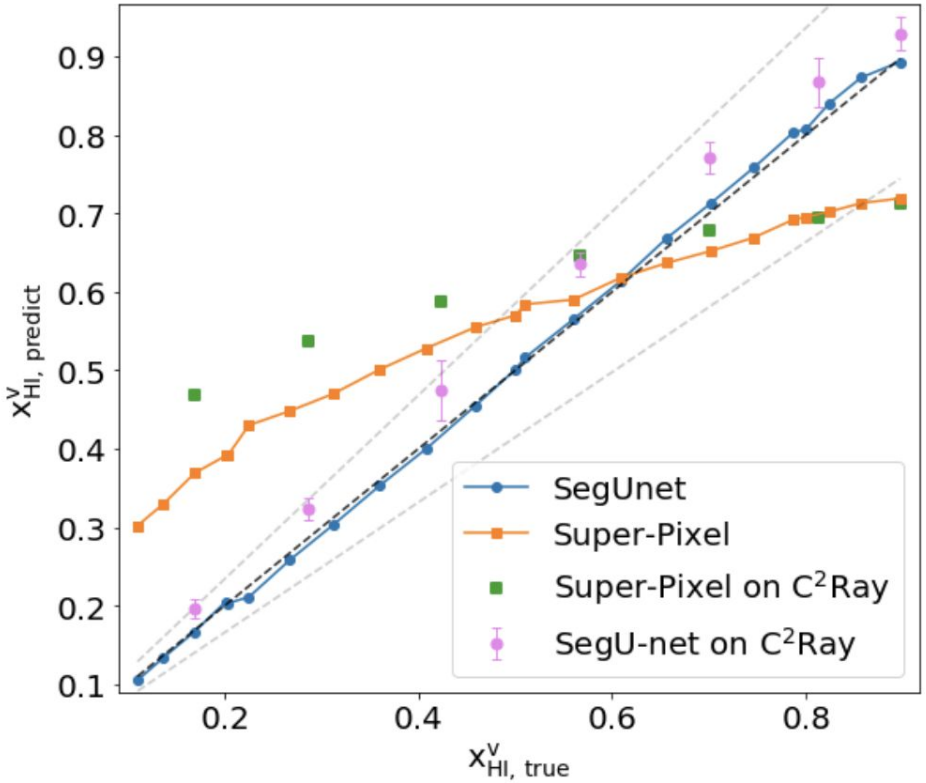
(Bianco,Giri+2021;2024)

SegU-Net identifying ionized regions



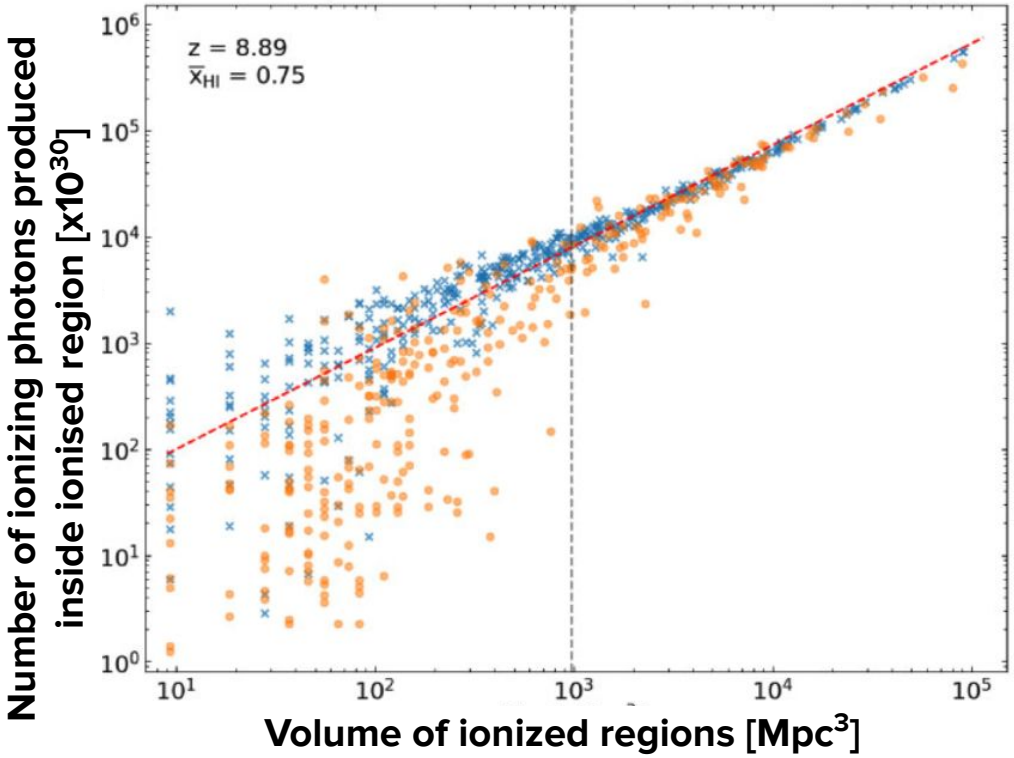
(Bianco,Giri+2021;2024)

Estimate the mean neutral fraction



(Bianco,Giri+2021)

Estimate number of galaxies inside ionized regions

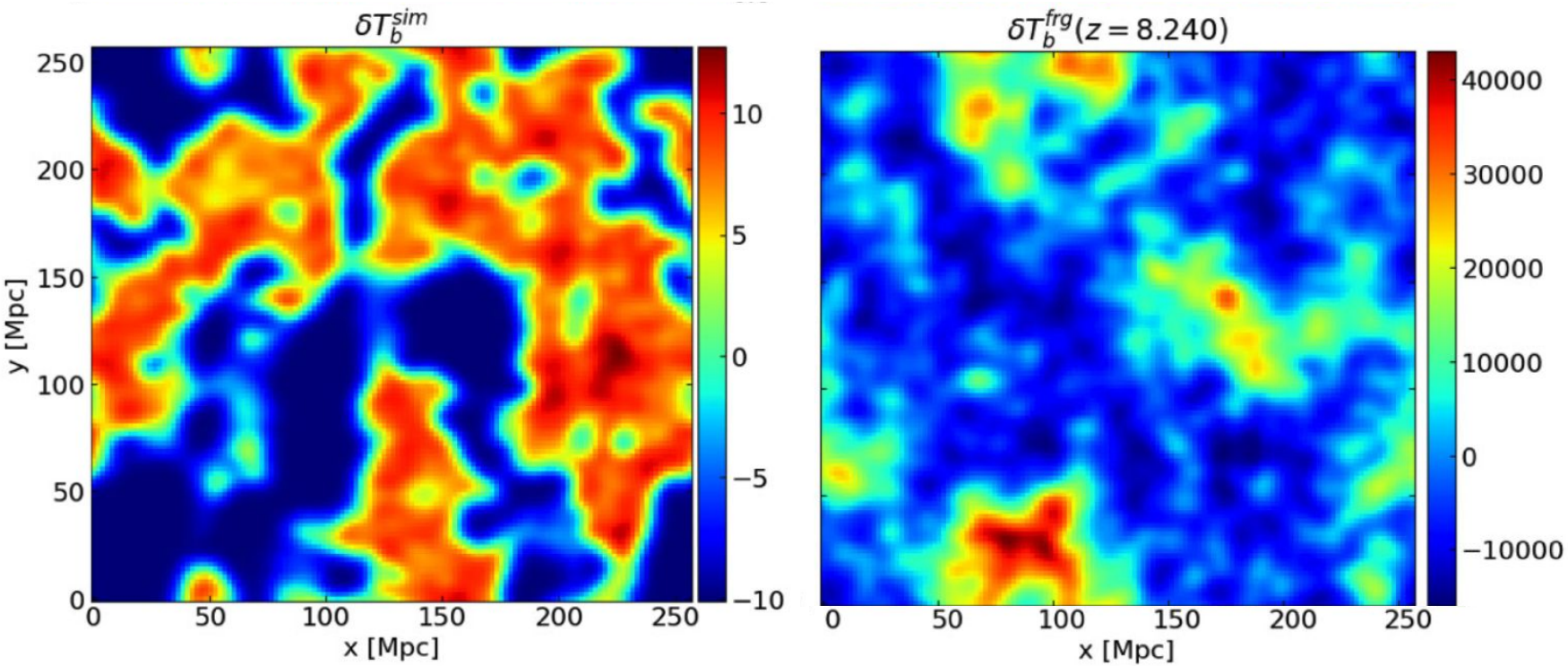


$$M_{\rm star,tot} \approx 8\times10^{10} M_\odot \left(\frac{V_{\rm ion}}{10^3 {\rm cMpc}^3}\right)\left(\frac{\langle f_{\rm esc}\rangle}{0.1}\right)^{-1}$$

(Zackrisson,…,Giri+2020)

(Bianco,Giri+2024)

21-cm signal contaminated with Foreground



(Bianco,Giri+2021;2024)

Science Data Challenge 3

Foregrounds

Scoring

Teams

Leaderboard

Rules

Data

Test Data

Our 'Foregrounds' challenge asks participants to remove obscuring sources of emission which prevent analysis of the underlying hydrogen-21cm signal from the Epoch of Reionisation. This foreground emission stems from both Galactic and extragalactic sources, both of which have previously observed, and unobserved components.

Given the lack of a model for the finer structure of Galactic emission at SKA-LOW frequencies, the removal of Galactic emission from the dataset represents a significant challenge. By similar reasoning, source confusion from previously unknown extragalactic sources, especially at the coarser resolution at metre-wavelengths, complicates the matter further.

From our synthetic datasets, participants are asked to extract the cylindrically-averaged power spectrum of the EoR signal, clean from foregrounds contamination.

To assess resulting submissions, our scoring ('figure-of-merit') algorithms will take resulting cylindrical power spectra, and return a score. Ancillary analytical data products can be assessed, however, the cylindrical power spectra results will be the only ones that affect the 'leaderboard'. Given that this challenge runs over 6 months, we expect this to be an iterative process and encourage competing teams to make full use of the computing facilities assigned to them.

SKA SDC3 Overview Challenges HPC Partners Registration Reproducibility Badges Forum FAQs

Science Data Challenge 3

Foregrounds

Astrophysicist

Michele Bianco,[1,2,3]★ Sambit. K. Giri,[4,5] Rohit Sharma,[6] Tianyue Chen,[1] Shreyam Parth Krishna,[1]
Chris Finlay[7], Viraj Nistane[7], Philipp Denzel,[6] Massimo De Sanctis,[8] Hatem Ghorbel[8]

[1] *Laboratoire d'Astrophysique, Ecole Polytechnique Federale de Lausanne (EPFL), Observatoire de Sauverny, Versoix*
[2] *Astronomy Centre, Department of Physics & Astronomy, Pevensey III Building, University of Sussex, Falmer, Brighton* m
[3] *Institute for Particle Physics and Astrophysics, ETH Zurich, Wolfgang-Pauli-Str 27, 8093 Zurich, Switzerland*
[4] *Nordita, KTH Royal Institute of Technology and Stockholm University, Hannes Alfvens väg 12, SE-106 91 Stockholm, Sweden*
[5] *Institute for Computational Science, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland*
[6] *Fachhochschule Nordwestschweiz, Bahnhofstrasse 6, 5210 Windisch, Switzerland*
[7] *Département de Physique Théorique and Center for Astroparticle Physics, Université de Geneve, 24 quai Ernest Ansermet, 1211 Geneve 4, Switzerland*
[8] *Data Analytics Group, Haute Ecole Arc Ingnierie HES-SO, University of Applied Sciences Western Switzerland*

Data Scientists

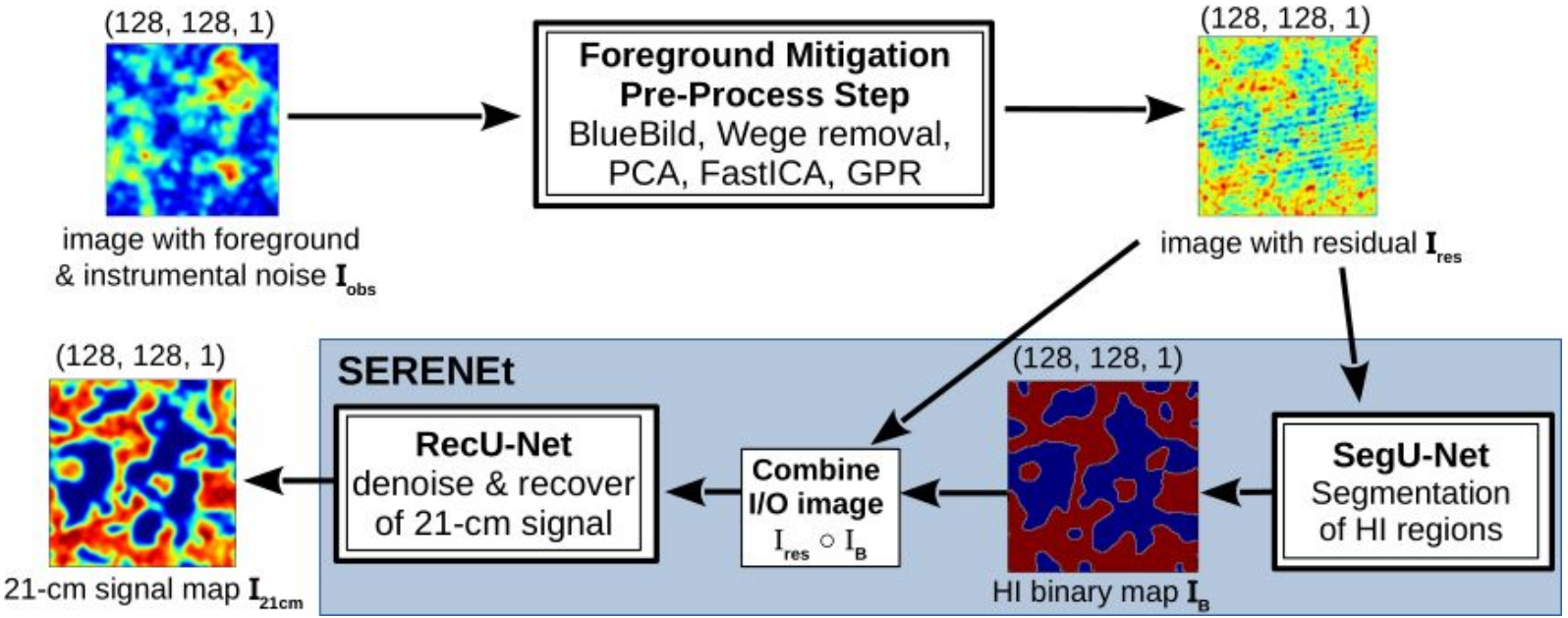
Data

Test Data

the EoR signal, clean from foregrounds contamination.

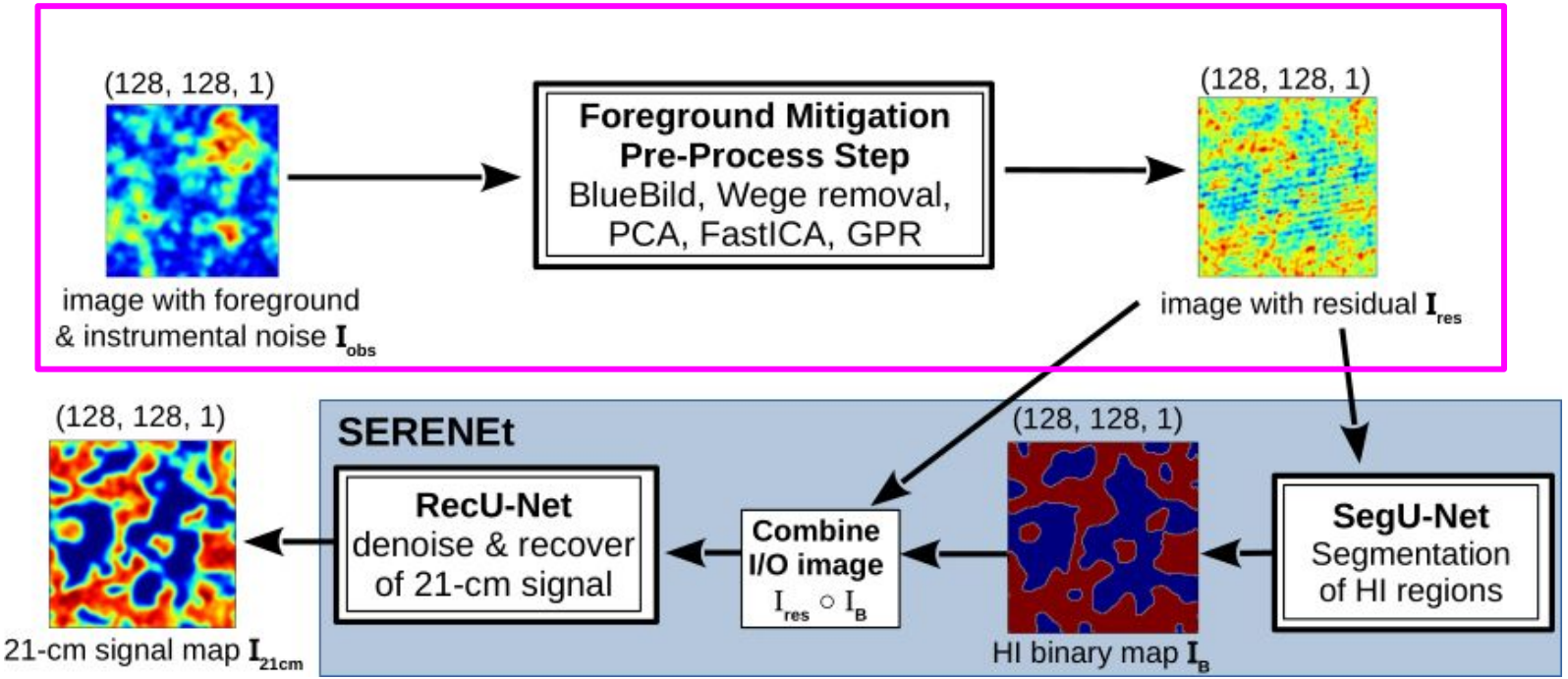
To assess resulting submissions, our scoring ('figure-of-merit') algorithms will take resulting cylindrical power spectra, and return a score. Ancillary analytical data products can be assessed, however, the cylindrical power spectra results will be the only ones that affect the 'leaderboard'. Given that this challenge runs over 6 months, we expect this to be an iterative process and encourage competing teams to make full use of the computing facilities assigned to them.

SERENet foreground mitigation pipeline



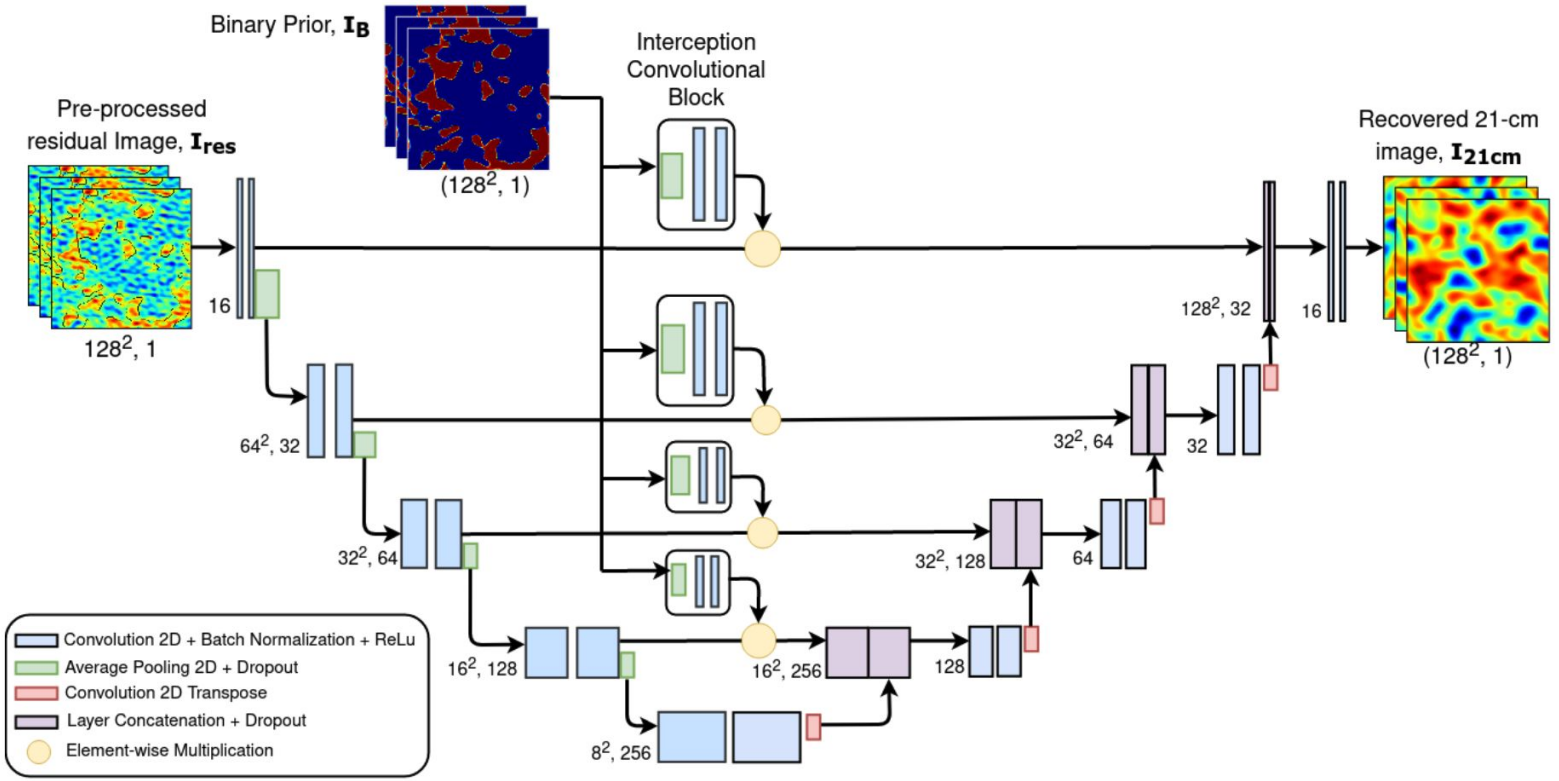
(Bianco,Giri+in prep)

SERENet: Pre-processing with PCA



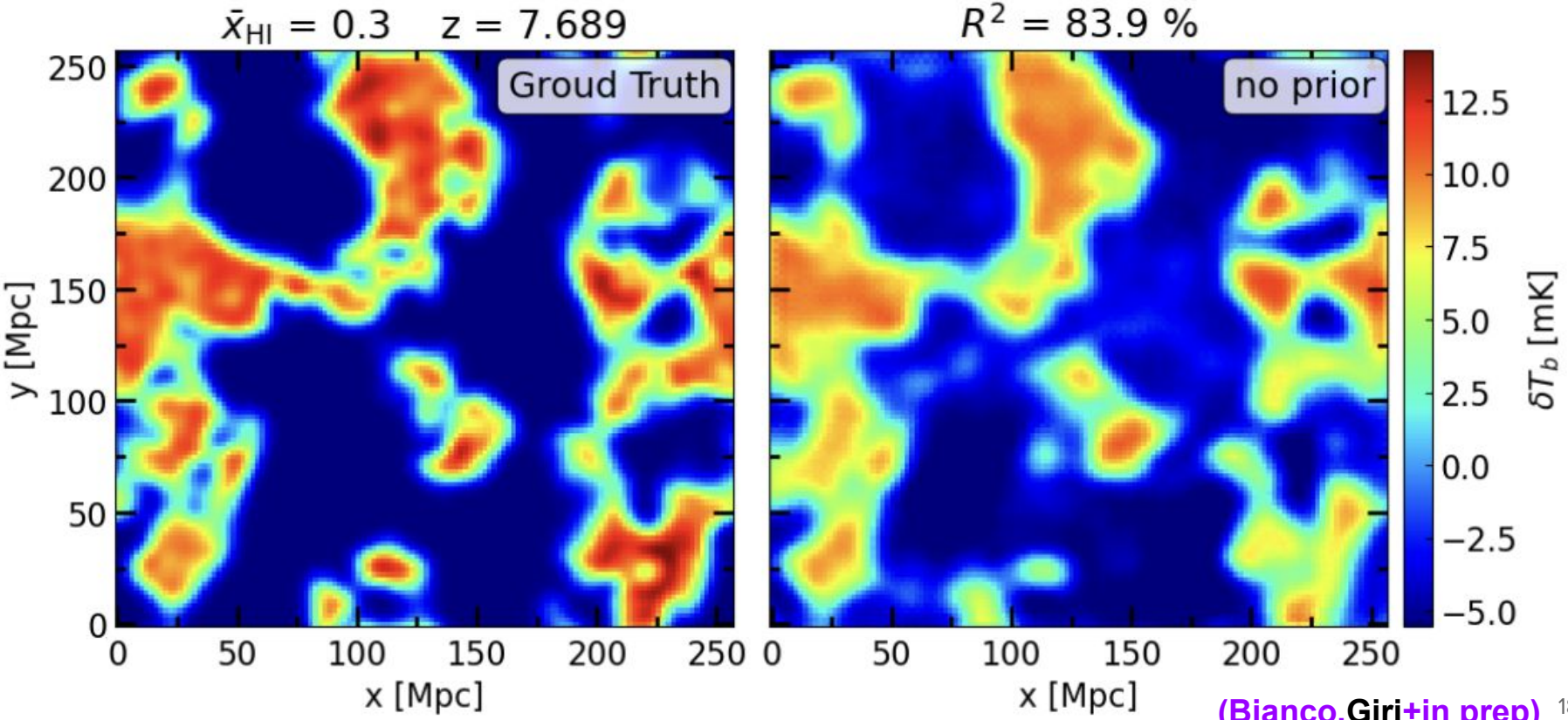
(Bianco,Giri+in prep)

SERENet framework



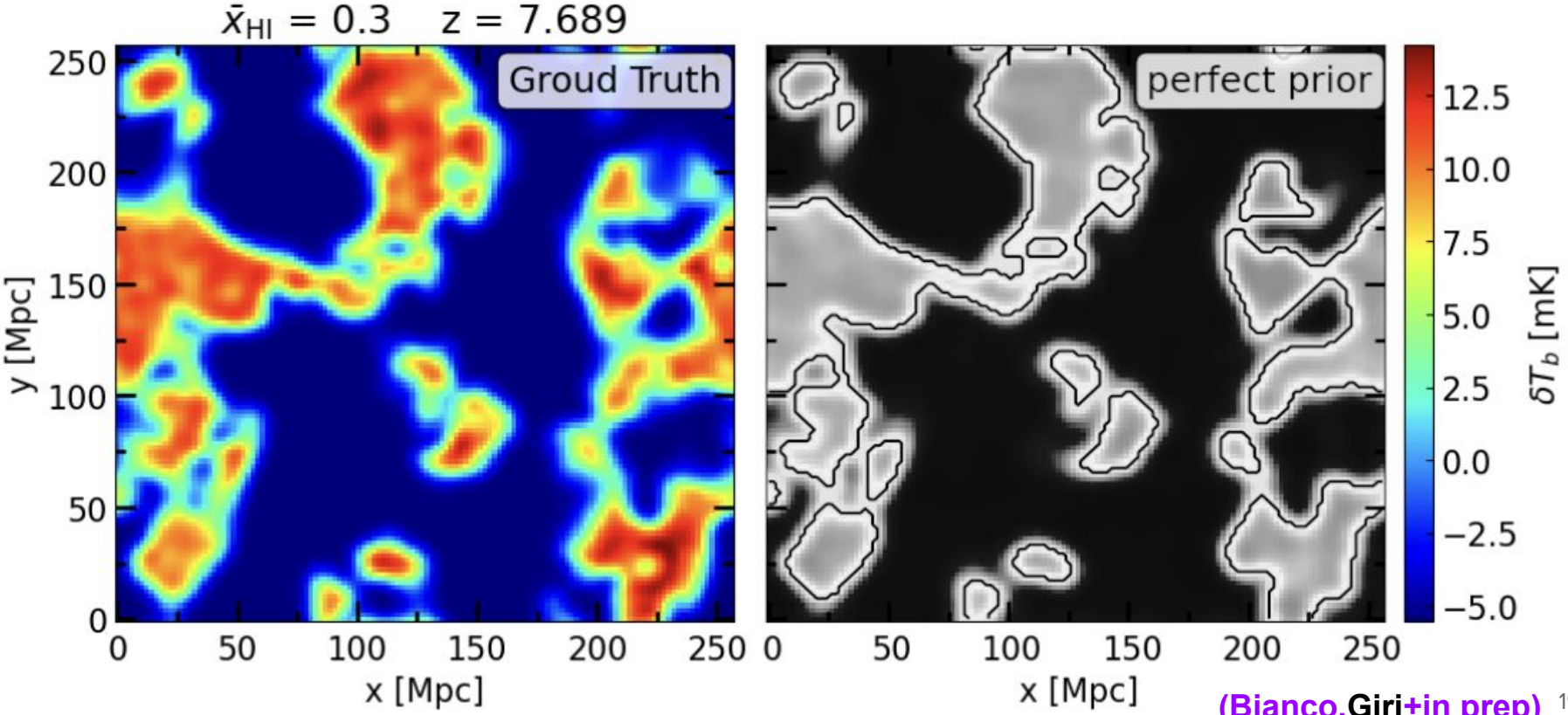
(Bianco,Giri+in prep)

SERENet recovering the 21-cm image



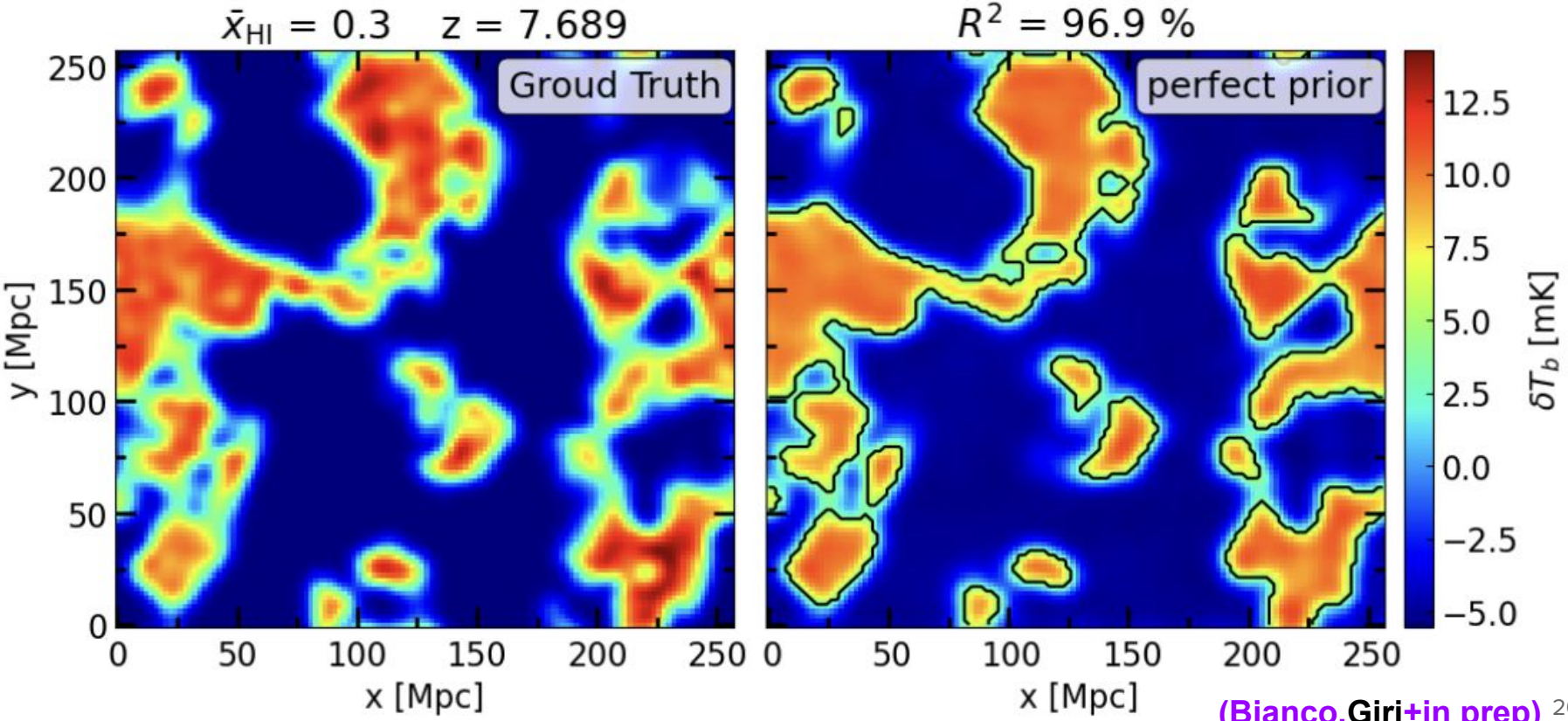
(Bianco,Giri+in prep)

SERENet: binary maps of ionized regions as prior



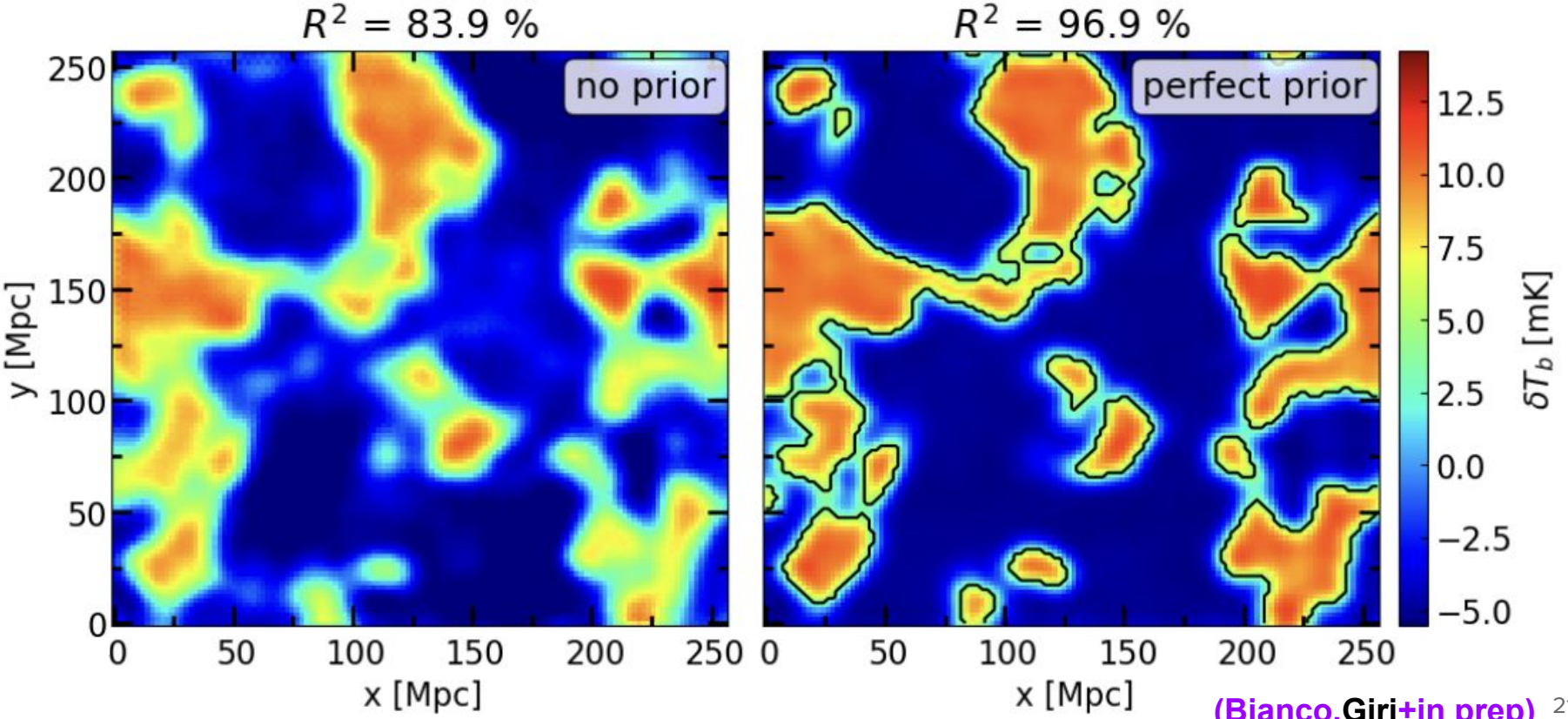
(Bianco,Giri+in prep)

SERENet recovering the 21-cm image with prior



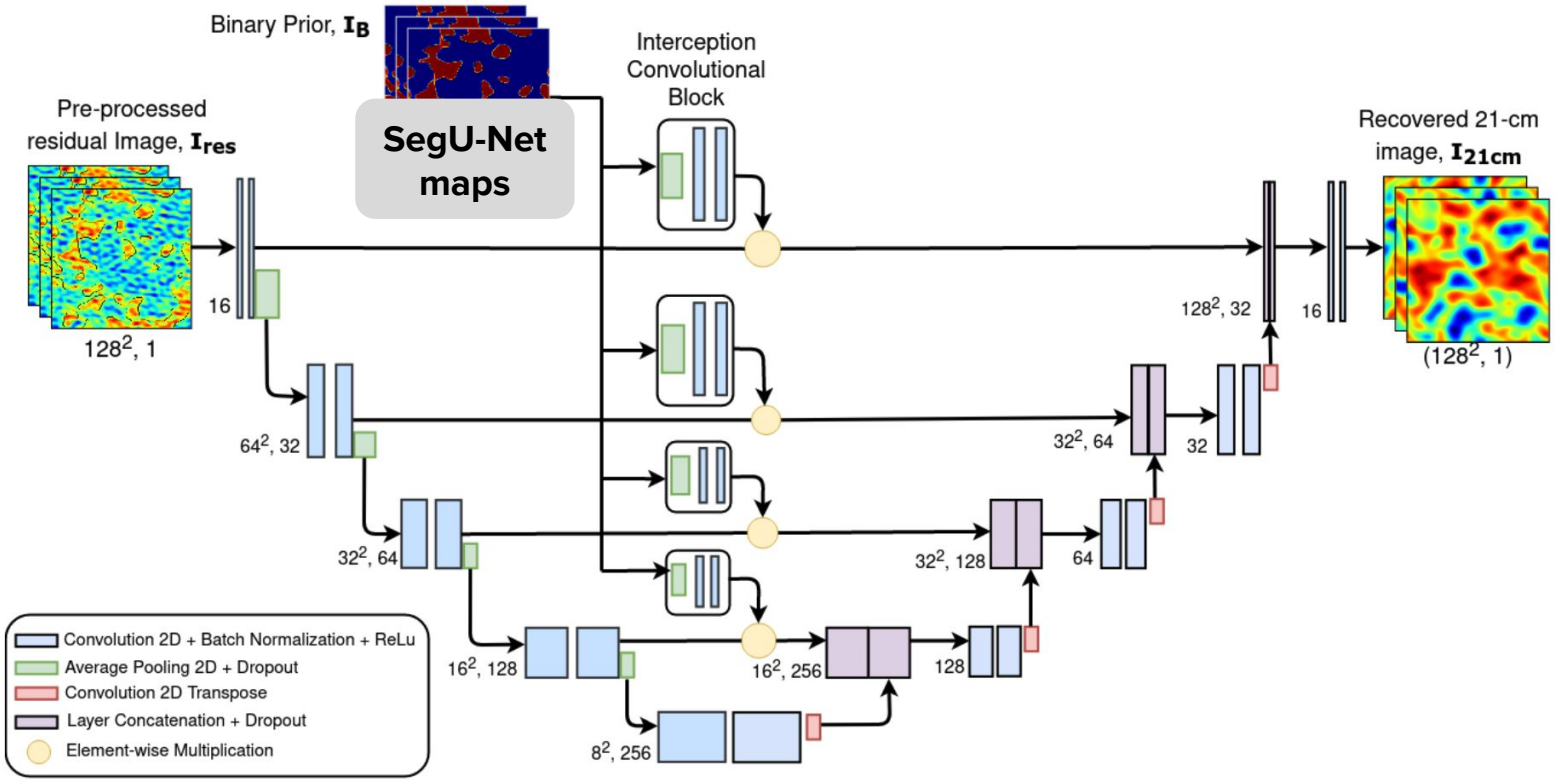
(Bianco,Giri+in prep)

No prior vs Perfect prior



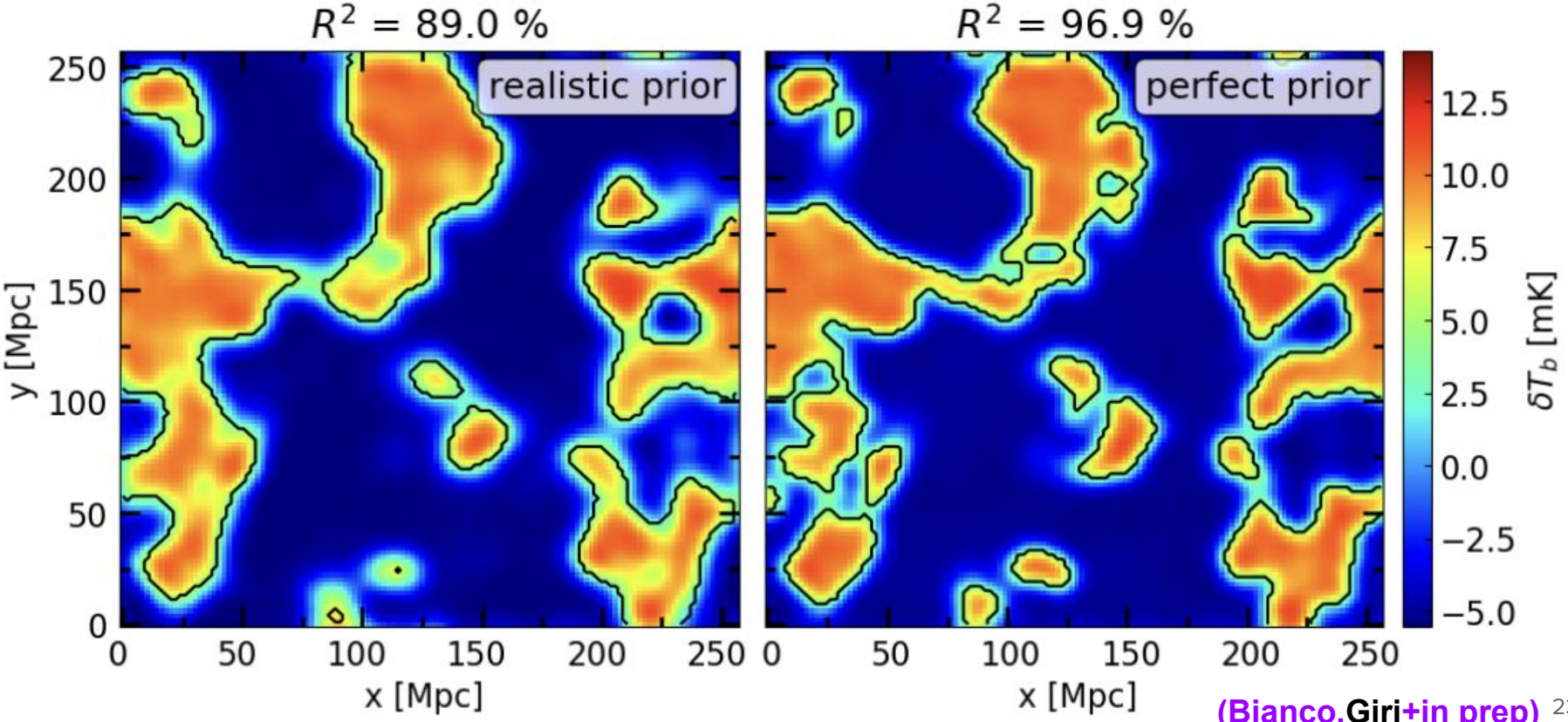
(Bianco,Giri+in prep)

SERENet: realistic prior with SegU-Net



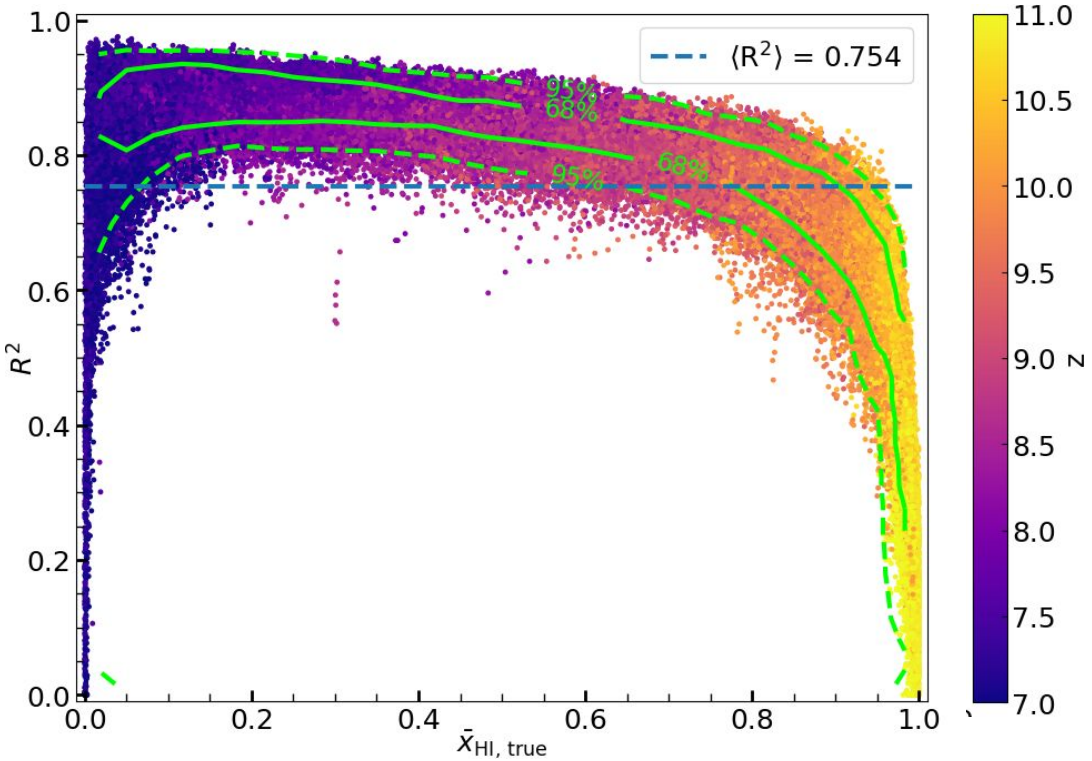
(Bianco,Giri+in prep)

SERENet recovering the 21-cm image with prior



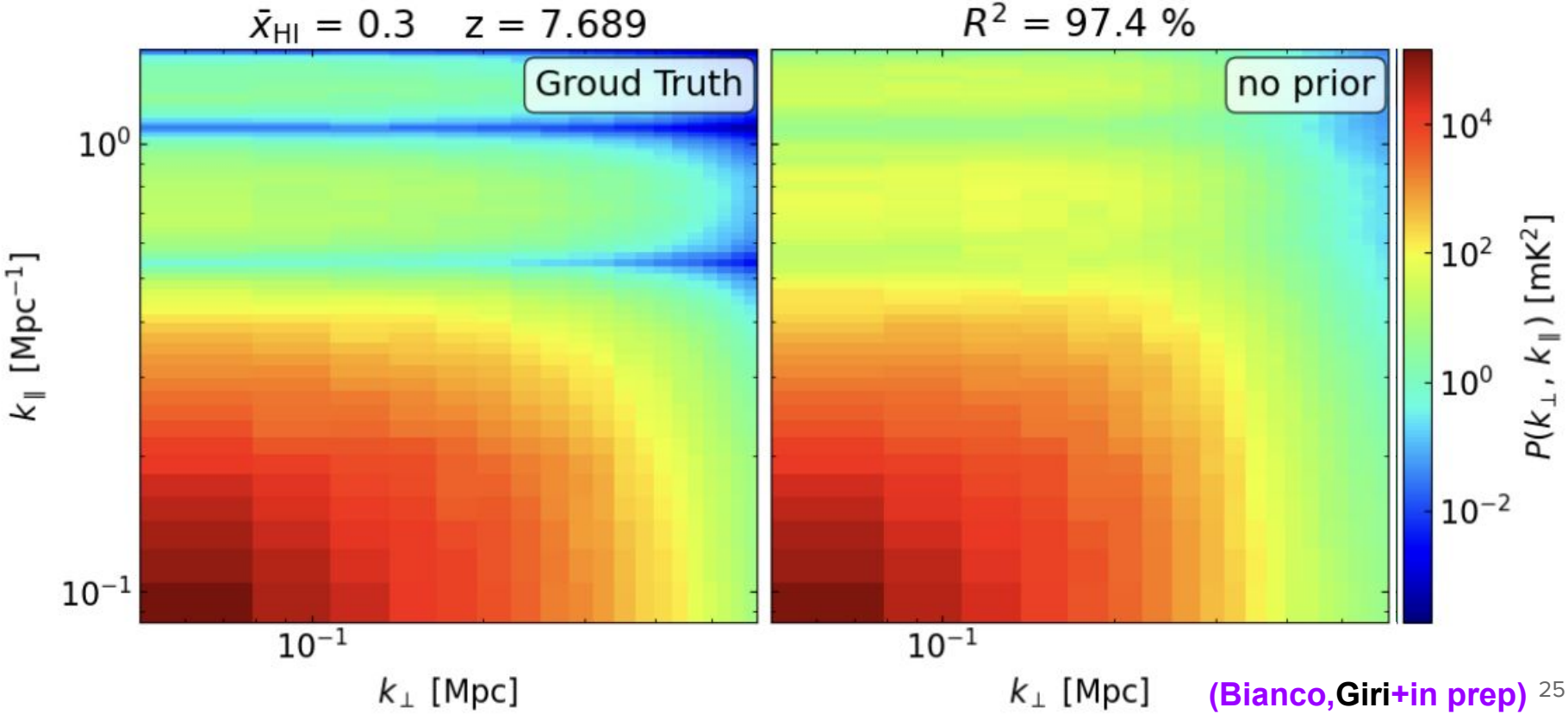
(Bianco,Giri+in prep)

R^2-score of the entire testing set



(Bianco,Giri+in prep)

SERENet recovering the 21-cm power spectrum



(Bianco,Giri+in prep)

Summary

- Data science methods will maximise the scientific outcomes of the huge amount of data that SKA will produce
- SegU-Net: deep learning-based framework to probe ionization state of the intergalactic medium
- SERENet: deep learning-based to mitigate foreground contamination in the 21-cm signal during the epoch of reionization