Deciphering the Epoch of Reionization using Neural Networks

Madhurima Choudhury

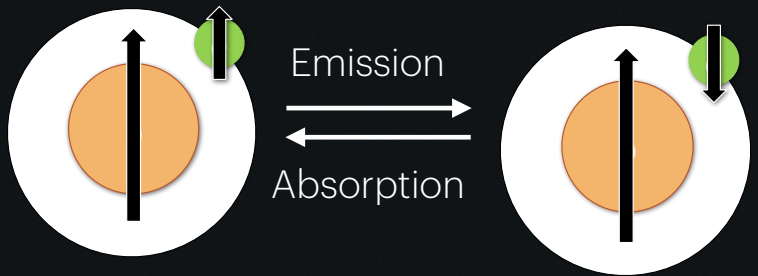
Center for Fundamental Physics of the Universe,
Brown University

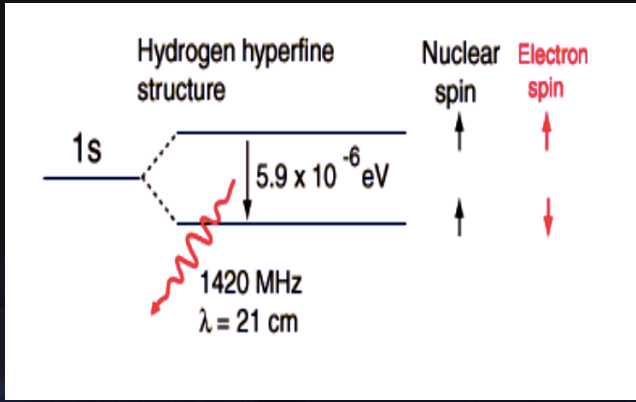
BROWN

AI/ML in Astrophysics & Astronomy, IUCAA
January 2025

SKAIC
Indian Consortium for the SKAO

The 21 cm Hydrogen Line





THIS SPECIAL TRANSITION LINE OF HYDROGEN SERVES AS AN EXCELLENT PROBE INTO THE VARIOUS EPOCHS OF THE EVOLUTION OF THE UNIVERSE

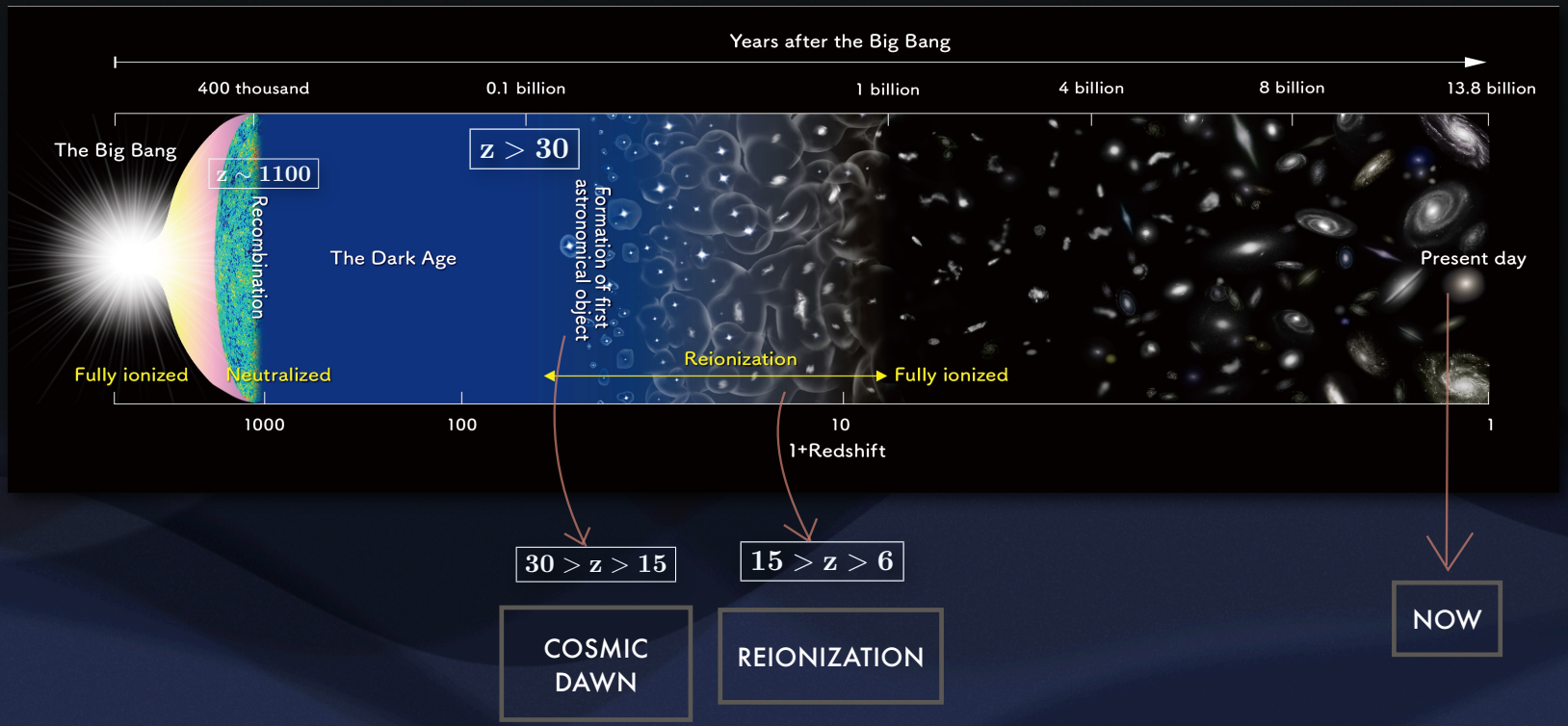
- This hyperfine transition line of atomic hydrogen (in the ground state) arises due to the interaction between the electron and proton spins.
  - The excited triplet state — spins are parallel
  - lower energy (singlet) state — spins are antiparallel.
- A spin temperature is defined

$$\frac{n_2}{n_1} = 3 \exp \frac{-T_*}{T_s}$$

$T_* = 0.068$ K

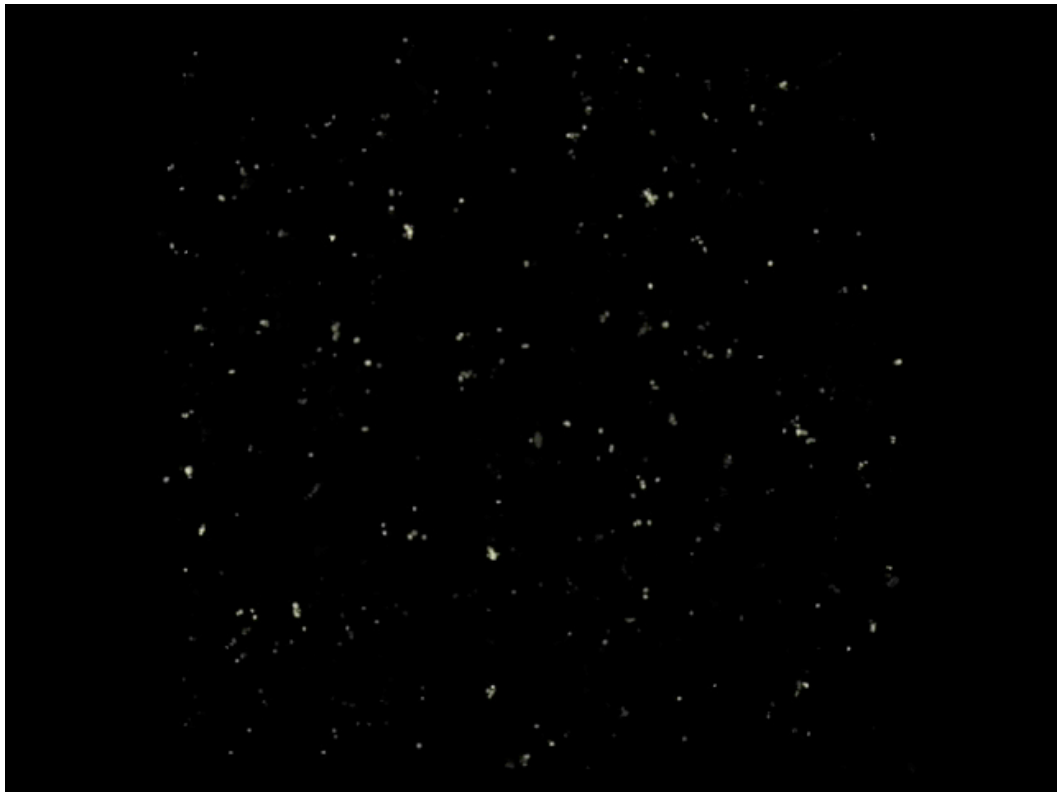
Population ratios of the two states

The Cosmic Timeline & Evolution of neutral Hydrogen

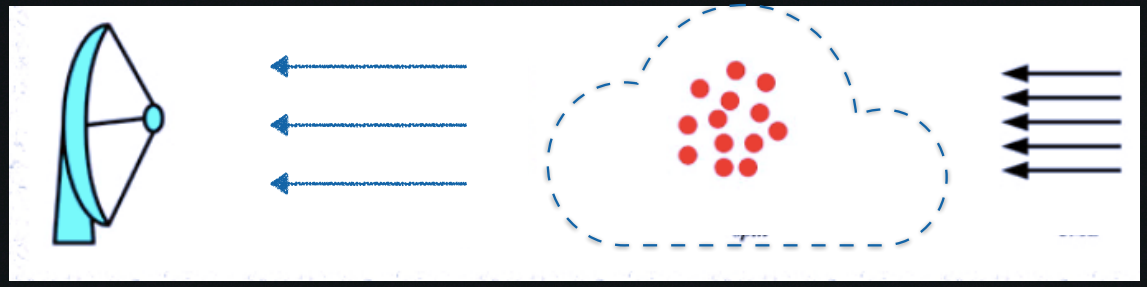


The evolution of neutral Hydrogen

The Epoch of Reionization (EoR) and the the redshifted 21-cm line of neutral Hydrogen as a probe



Observations of the redshifted 21cm signal



δT_b T_s, T_k T_{CMB}

Differential brightness temp.

CMB temp.

$$\delta T_b \propto (1+\delta_b) \cdot x_{HII} \cdot \frac{T_\gamma}{T_S}$$

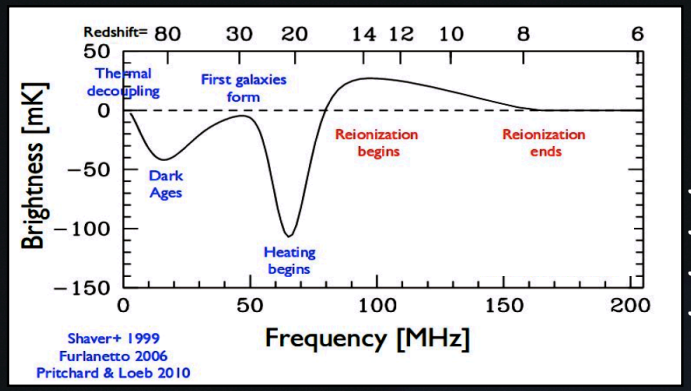
Ionised fraction

Spin temp.

Observations of the redshifted 21cm signal

Observations using single dish

Sky averaged 21-cm Global Signal

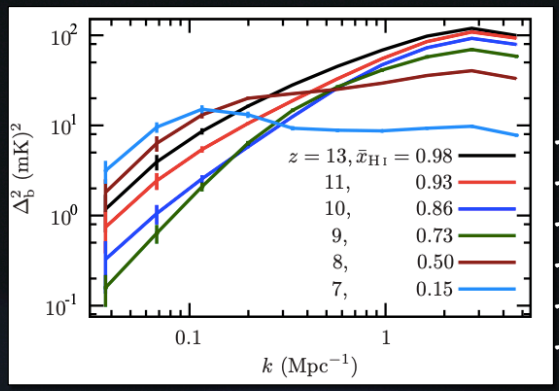


Experiments

- EDGES
- SARAS
- REACH
- ...

Observations using interferometers

21-cm Power Spectrum

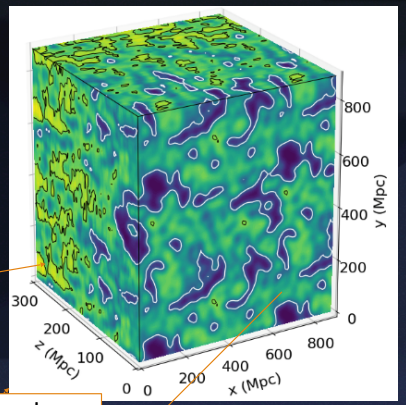


Experiments

- MWA
- LOFAR
- GMRT
- HERA
- SKA
- ...

21-cm tomography

3D imaging of all the Hydrogen in the Universe



Experiments

- MWA
- SKA
- ...

Depth direction comes from the spectrum (i.e. the dependence of the signal versus frequency/ wavelength)

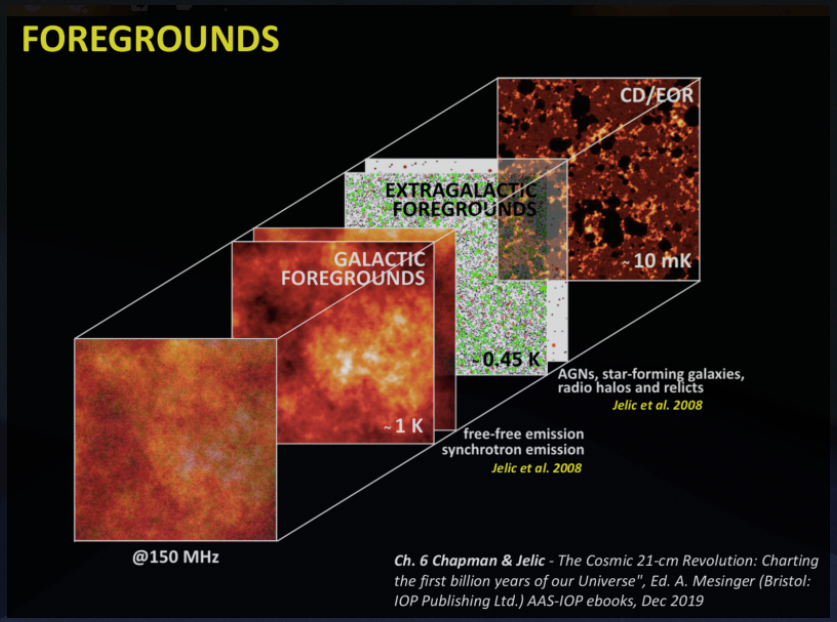
Two directions on plane of the sky (i.e. make an image)

What the radio telescopes aim to measure!

Challenges

THE 21CM SIGNAL IS VERY FAINT! DETECTION IS NOT EASY!

- The foregrounds: Galactic/Extra-galactic foregrounds (dominated by synchrotron radiation)
- Earth's ionosphere (introduces direction dependent effects)
- Radio frequency interference (RFI)
- The instrument response and systematics

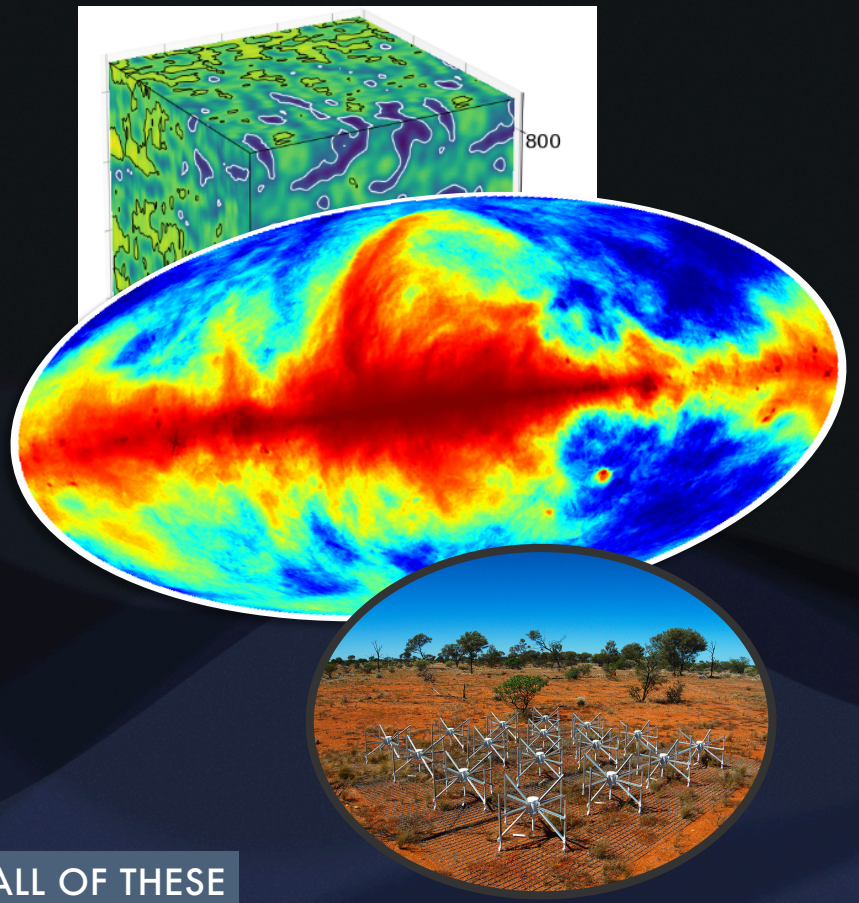


How to deal with these challenges?

1. We are looking at a very faint signal, efficiently modeling and removing the foregrounds is crucial.

2. RFI indentification and mitigation, understanding systematics are also major challenges

3. Theoretical modelling and simulations :
   - ✓ very large dynamical scales involved
   - ✓ Uncertain and unconstrained astrophysics
   - ✓ Very large and uncertain parameter space.



ML CAN BE USEFUL IN ALL OF THESE CASES!

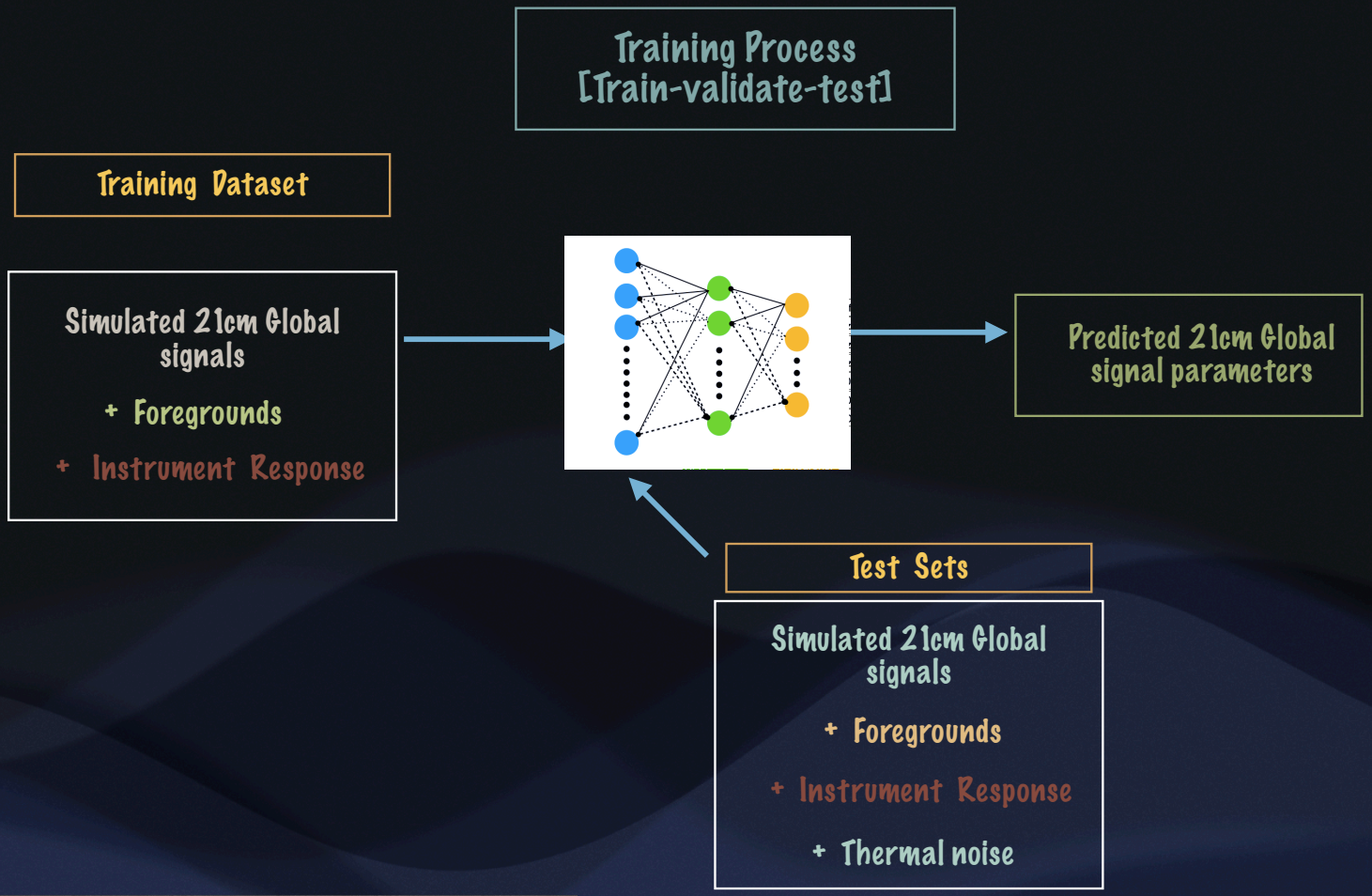
DOES A SUPERVISED ML BASED FRAMEWORK WORK FOR FOREGROUND REMOVAL AND PARAMETER ESTIMATION SIMULATANEOUSLY?

Extracting the 21-cm **Global signal** from mock datasets using Machine Learning

Choudhury, M+ 2019, MNRAS

Choudhury M+ MNRAS, 2021

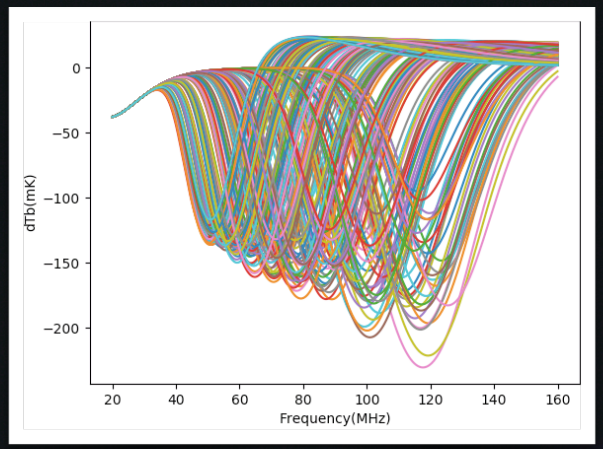
Extracting the 21cm Global Signal using ANN



A PROOF OF CONCEPT DEMONSTRATION OF AN ML FRAMEWORK FOR PARAMETER ESTIMATION

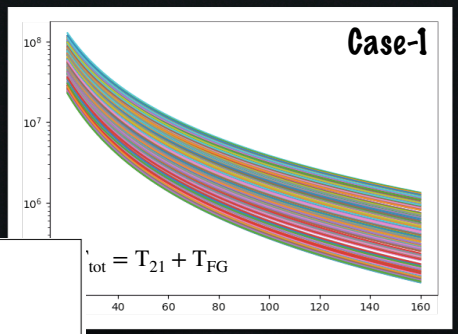
Choudhury, M+ 2019, MNRAS

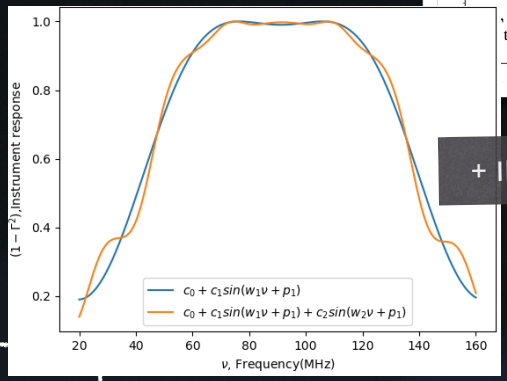
Preparing the training sets



$$ln\ T_{FG} = \sum_{i}^{n} a_i\ [ln(\nu/\nu_0)]^i$$

+ Foregrounds





+ INSTRUMENT EFFECT

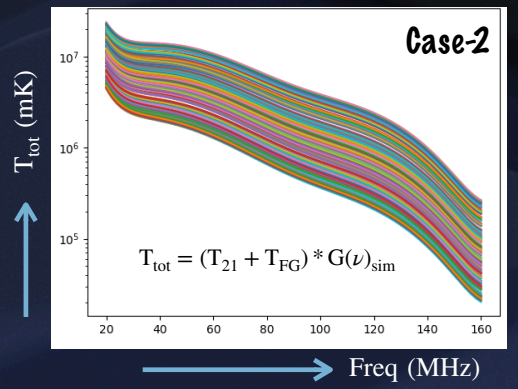
$$G(\nu) = |1 - \Gamma^2|$$

Parameters evolve according to a tanh model

- J(z)—Lyman-alpha background (which determines the strength of W-F coupling)
- Xi(z)— Ionized fraction of hydrogen
- T(z)—temperature of the IGM

Each of these parameters, is expressed as:

$$A(z) = \frac{A_{ref}}{2}\{1 + \tanh[(z_0 - z)/\Delta z]\}$$



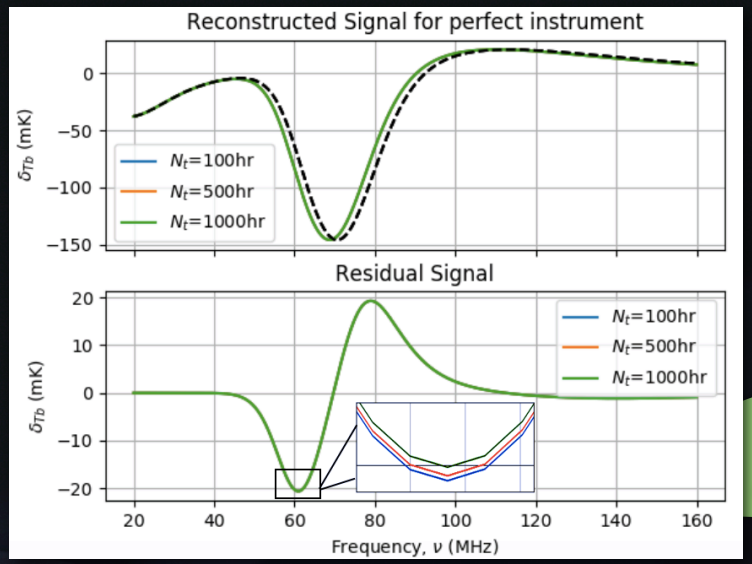
SUCH TRAINING SETS WERE THEN USED TO PREDICT THE TANH MODEL PARAMETERS OF THE 21CM GLOBAL SIGNAL

Testing the ANN with noisy test sets

We build a test set, unseen to the network, corrupted with thermal noise corresponding to 1000 hrs of observation time.

(21cmGS + FG)*G + thermal noise

$$\frac{T_{FG}}{\sqrt{\Delta\nu \cdot 10^6 \cdot 3600 \cdot N_t}}$$



Reconstructed Global signal with the predicted parameters

93-98% ACCURACY

Parameters	Perfect Instrument Case 1	Fixed(simple) Instrument Case 2A
J_{ref}	0.0245	0.0705
dz_J	0.0209	0.0575
dz_T	0.0230	0.0668
dz_X	0.0207	0.0709
$z0_J$	0.0216	0.0550
$z0_T$	0.0194	0.0650
$z0_X$	0.0278	0.0739

THIS IS A VERY SIMPLE APPROACH TO IMPLEMENT ANN AS A 21CM GLOBAL SIGNAL EXTRACTION TECHNIQUE

The normalised RMSE,

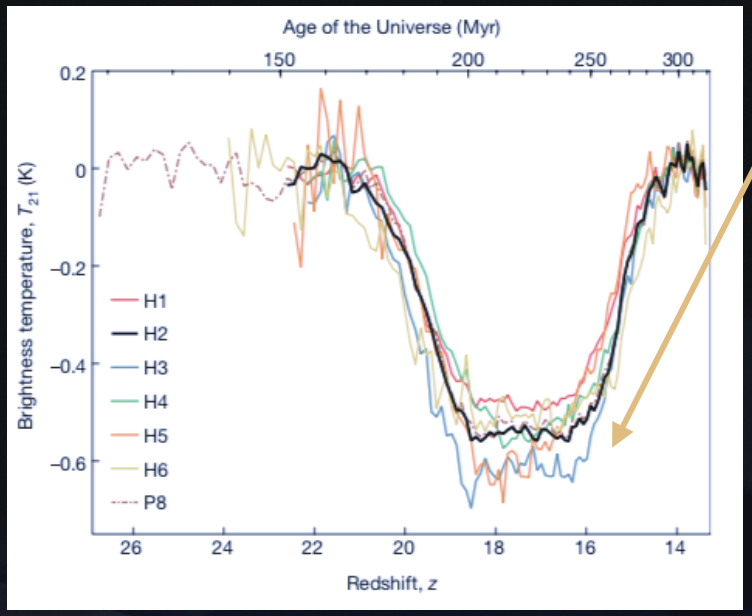
$$\sqrt{\frac{1}{N_{test}}\sum^{N}\left(\frac{y_{pred}-y_{ori}}{y_{ori}}\right)^2}$$

Choudhury M + MNRAS, 2019

WOULD THIS KIND OF A SUPERVISED ML FRAMEWORK WORK ON REAL DATA FROM OBSERVATIONS?

Using ANN to extract the 21cm Global Signal from EDGES data

Recent detection from EDGES experiment



Bowman et al (2018)

Unexpectedly deep and flat absorption trough

Explanation?

Usually, $T_{bg} \sim T_{CMB}$.
T_{bg} could be much larger,
$T_{bg} \sim (T_{CMB} + T_{excess})$

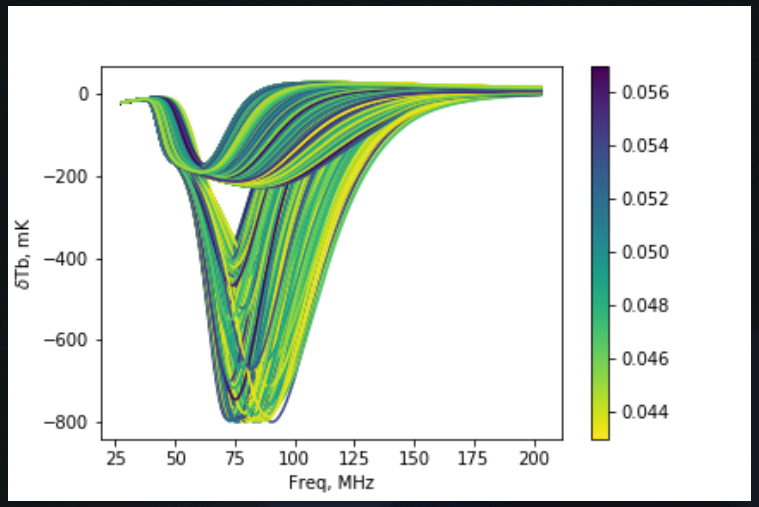
$$T_b \propto \left(\frac{T_s - T_{bg}}{T_s}\right)$$

T_s could be lower than expected.
→ IGM could be cooler,
→ possible DM, Baryon interactions

So we now build training sets, which includes such EDGES-like sample signals

Choudhury M+ MNRAS, 2021

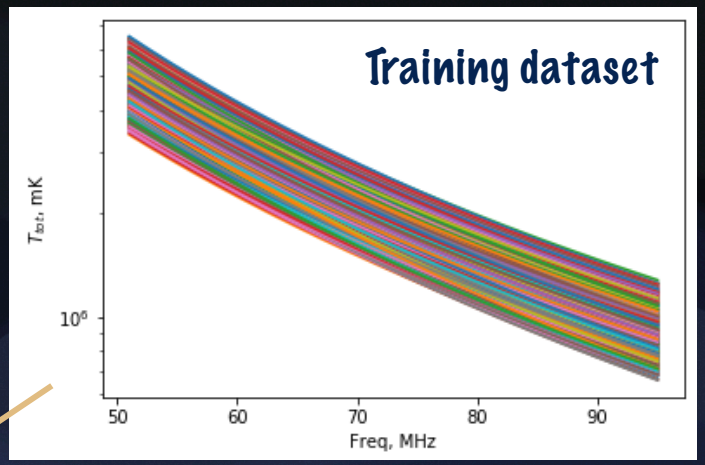
Generating training datasets by adding foregrounds



Bowman-Rogers foreground model (Bowman et al 2018)

$$T_{FG} = b_0 \cdot (\frac{\nu}{\nu_c})^{-2.5} + b_1 \cdot (\frac{\nu}{\nu_c})^{-2.5} \ln(\frac{\nu}{\nu_c})$$
$$+ b_2 \cdot (\frac{\nu}{\nu_c})^{-2.5} \cdot [\ln(\frac{\nu}{\nu_c})]^2 + b_3 \cdot (\frac{\nu}{\nu_c})^{-4.5} + b_4 \cdot (\frac{\nu}{\nu_c})^{-2}$$

+ Foregrounds



Training data: $T_{train} = T_{21} + T_{FG}$

OUTPUT ASTROPHYSICAL PARAMETERS

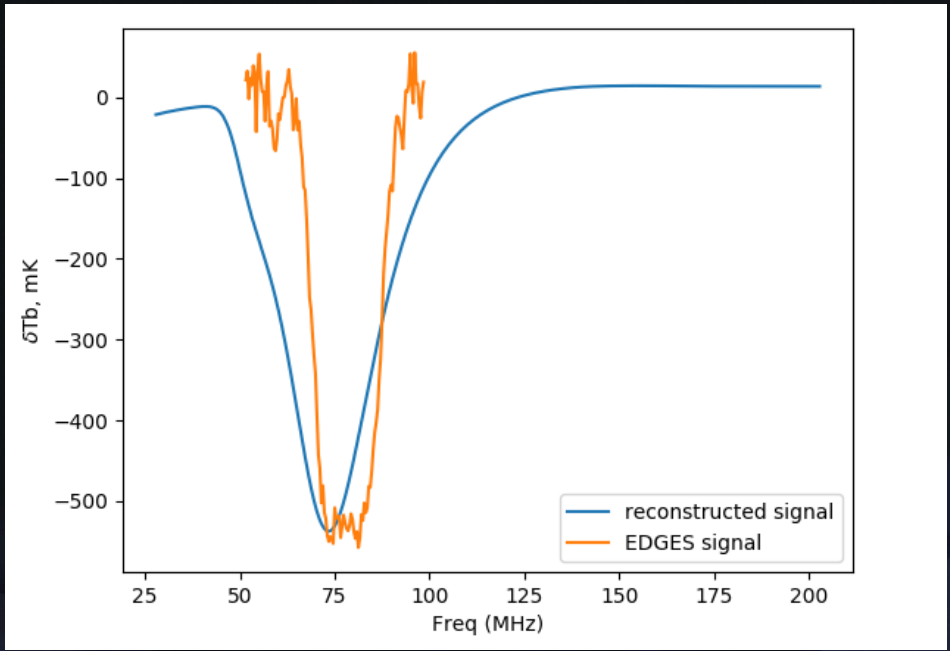
f_* f_{esc} f_{xh} f_X f_R N_α

Very good acuuracies

Choudhury M+ MNRAS, 2021

WE INPUT THE EDGES 2018 DATA (THE OBSERVED TOTAL SKY TEMPERATURE FOR THE OBSERVED BAND) AS THE TEST DATA TO THE ANN

Reconstructed 21cm signal from the predicted parameters



Parameter	Predicted values from ANN
f_*	0.006
f_{esc}	0.139
f_{Xh}	0.3
f_X	16.24
N_α	9728
f_R	1E+04

A strong radio background could explain the large dip of the EDGES signal

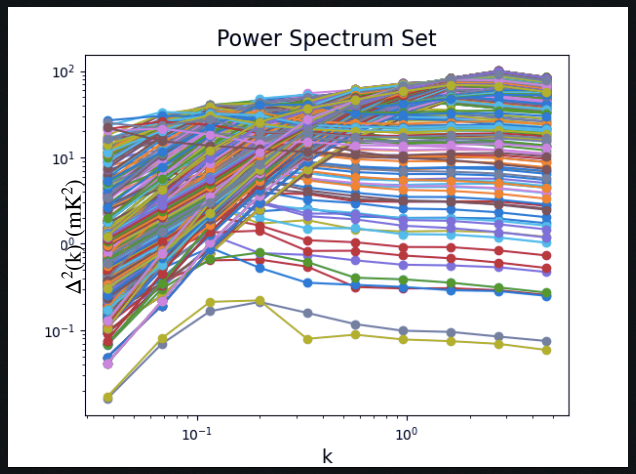
THIS IS AN ALTERNATE CONFIRMATION THAT THE OBSERVED SKY TEMPERATURE MIGHT CONTAIN SIGNATURES OF A DEEP ABSORPTION PROFILE.

Choudhury M+ MNRAS, 2021

ANOTHER SUPERVISED ML BASED FRAMEWORK FOR FOREGROUND REMOVAL AND PARAMETER ESTIMATION FOR 21-CM POWER SPECTRUM

Choudhury M+ MNRAS, 2022

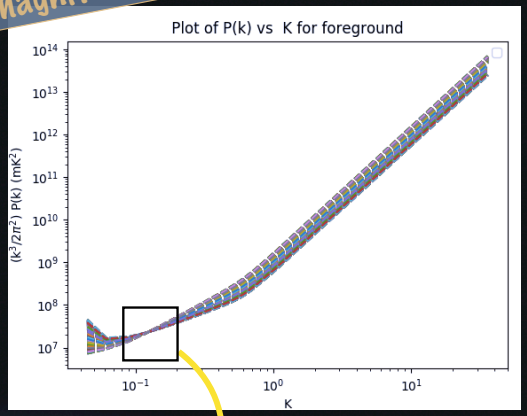
CONSTRUCTING THE TRAINING DATASETS CORRUPTED WITH FOREGROUNDS



Foregrounds orders of magnitude brighter

$$C_l(\nu) = A(\ell/\ell_0)^{-\beta}(\nu/\nu_0)^{-2\alpha}$$

+



For a flat-sky approximation,

$C_\ell(\Delta\nu) \xrightarrow{\text{F.T.}} P(k)$

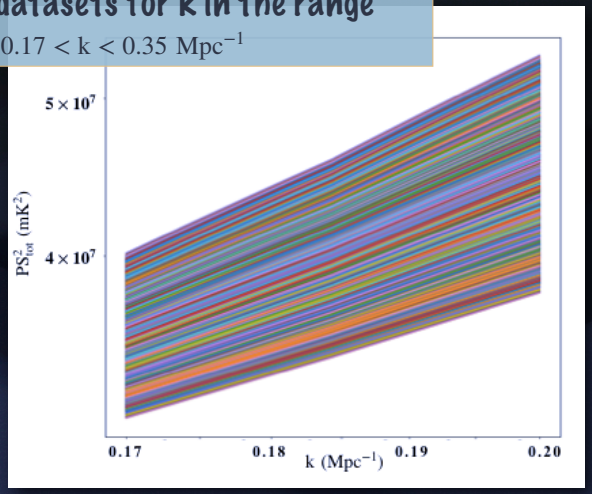
Following the formalism in Datta, K et al,2006

Simultaneous prediction of signal and foreground parameters give us best accuracies of only ~35-40%

While the FG parameters are obtained with good accuracy, the associated signal parameters are not predicted accurately enough

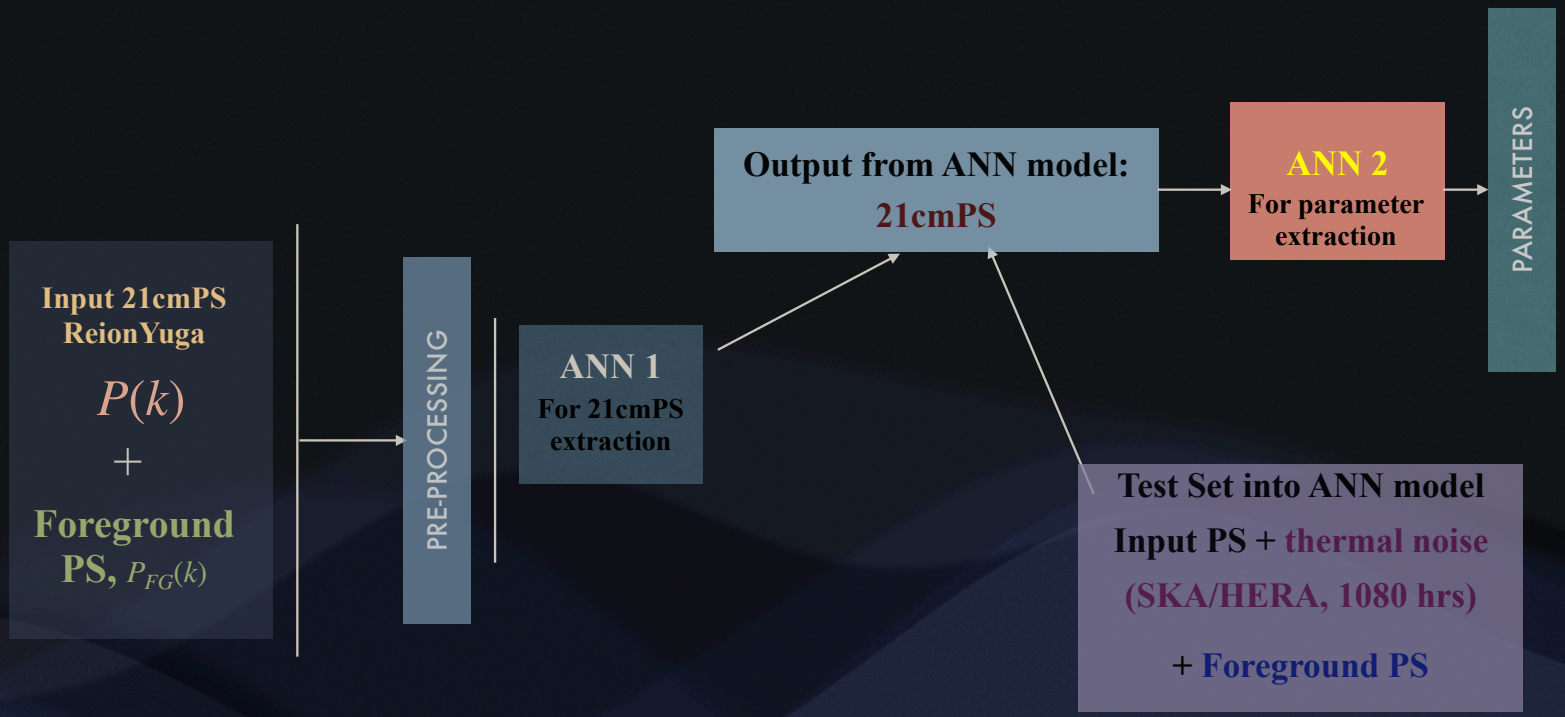
Training datasets for k in the range

$0.17 < k < 0.35$ Mpc^{-1}



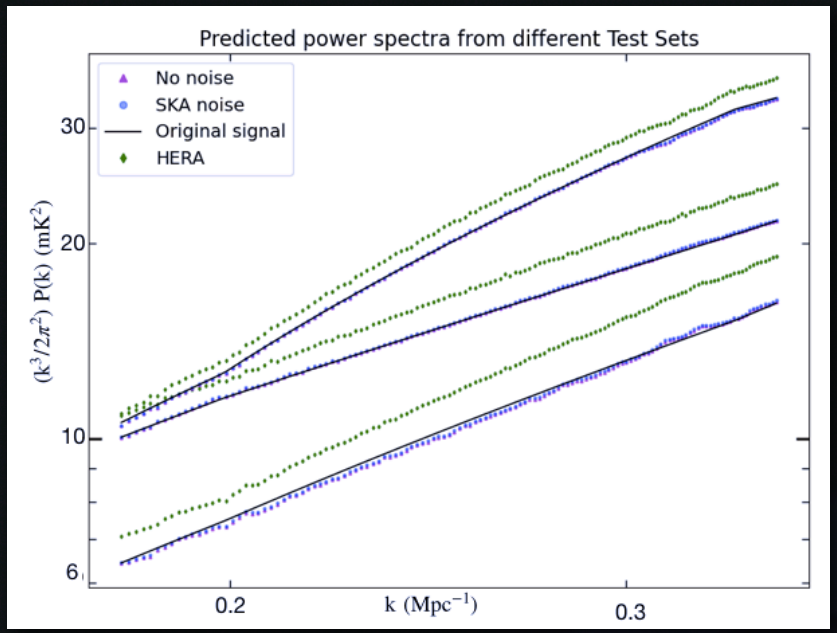
WE HAVE PROPOSED A TWO-STEP ANN FRAMEWORK TO EXTRACT THE HI 21CM POWER SPECTRUM AS WELL AS THE ASTROPHYSICAL PARAMETERS

21cm Power Spectrum Extraction Framework



Choudhury M+ MNRAS, 2022

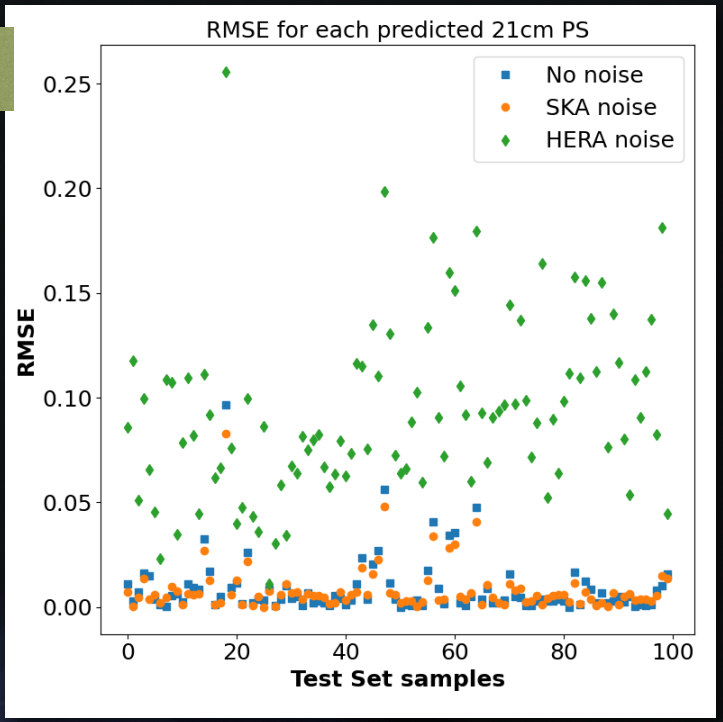
PREDICTIONS OF ANN1: 21CM POWER SPECTRUM



VERY GOOD PERFORMANCE AND NETWORK ACCURACY!

RMSE computed for each sample in the test set implies:

~95-99% accuracy (for SKA-noise added test sets)



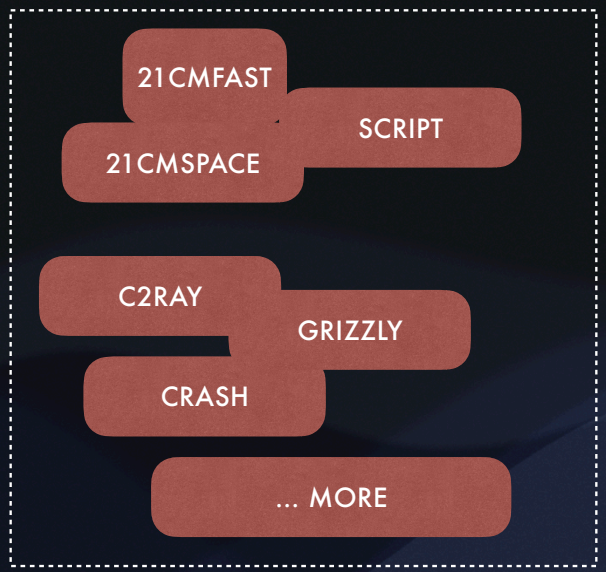
WE DIRECTLY PREDICT THE PS AND ITS ASSOCIATED PARAMETERS FROM FOREGROUND DOMINATED MOCK-DATASETS

From the predicted 21cmPS, we can predict the reionization parameters, using ANN2.

R2 scores for parameters for different test sets:	ζ	$\log M_{h,min}$
No-noise	0.9	0.8
SKA-noise	0.9	0.8
HERA-noise	0.6	0.5

THEORETICAL SIMULATIONS AND INFERENCE FRAMEWORKS WITH MACHINE LEARNING

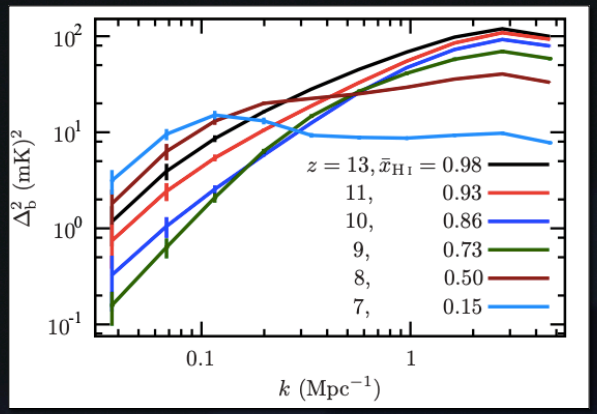
Observations



Parameters

MOST PARAMETER ESTIMATION FRAMESWORKS AIM TO CONSTRAIN THE ASTROPHYSICAL PARAMETERS ASSOCIATED WITH THE SOURCE MODEL

The 21-cm signal directly probes the state of the intergalactic medium (IGM)



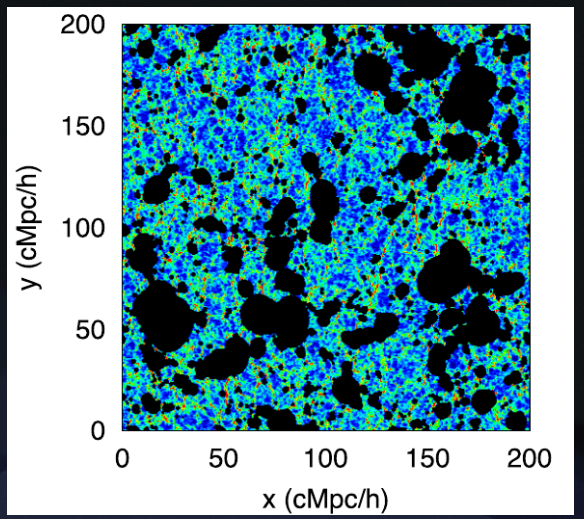
Ionised fraction, xHII

Distribution of the ionised regions

Evolution of Lyman alpha flux

Evolution of kinetic temperature, T_k

More quantities…



A slice from **GRIZZLY**.

Why do we need an IGM-based framework?

Source parameters

f_*, f_{esc}, f_x, $M_{h,min}$, etc

What is the connection?

IGM parameters

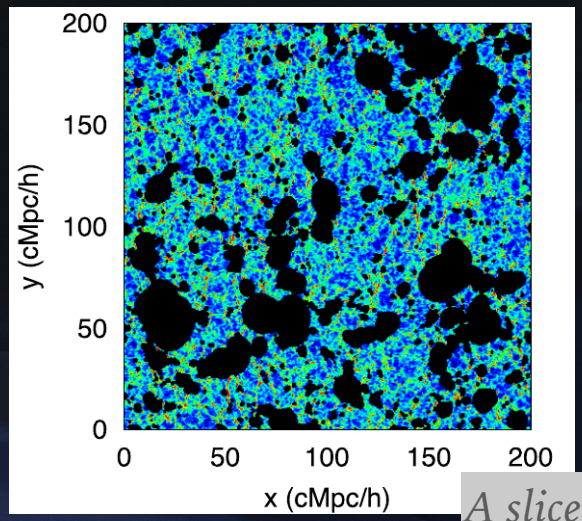
x_{HII}, BSD, Lyα, T_k, etc

SIMULATIONS

↓

21-cm Power Spectrum

How will be able to extract the most information from the 21cm PS?



A slice from **GRIZZLY**.

As the analytical connection between the IGM and the PS is not very straightforward, we can use ML methods to establish this connection.

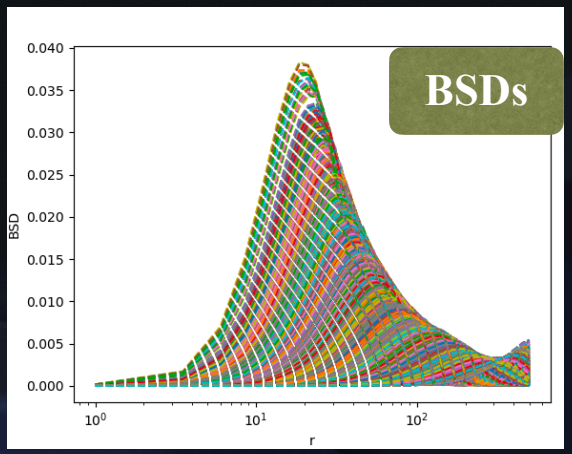
Emulator connecting IGM params and the 21-cm Power Spectrum

Building the training sets

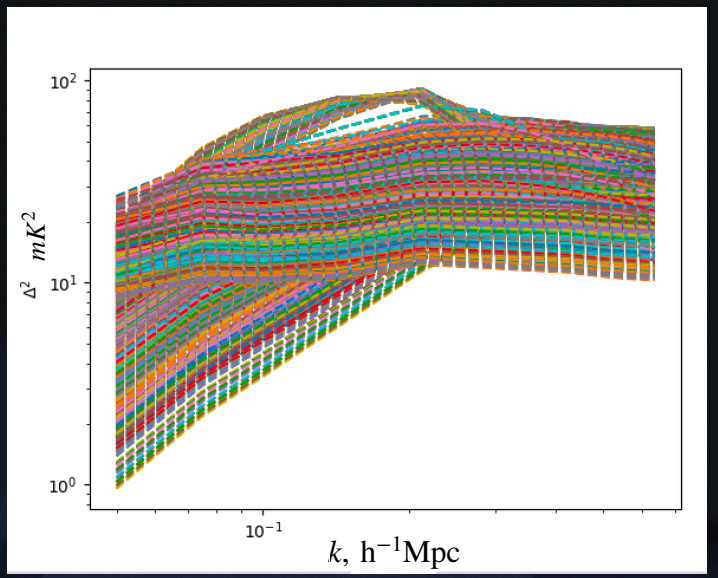
Simulation details: GRIZZLY (T_S fluctuations not considered)

$\zeta \in (1.5, 300)$, $M_{min,halo} \in 10^9, 10^{11} M_\odot$ *at* $z=9.1$

$x_{HII} \in (0.20, 0.90)$, *bubble size distributions (BSD)*



ANN



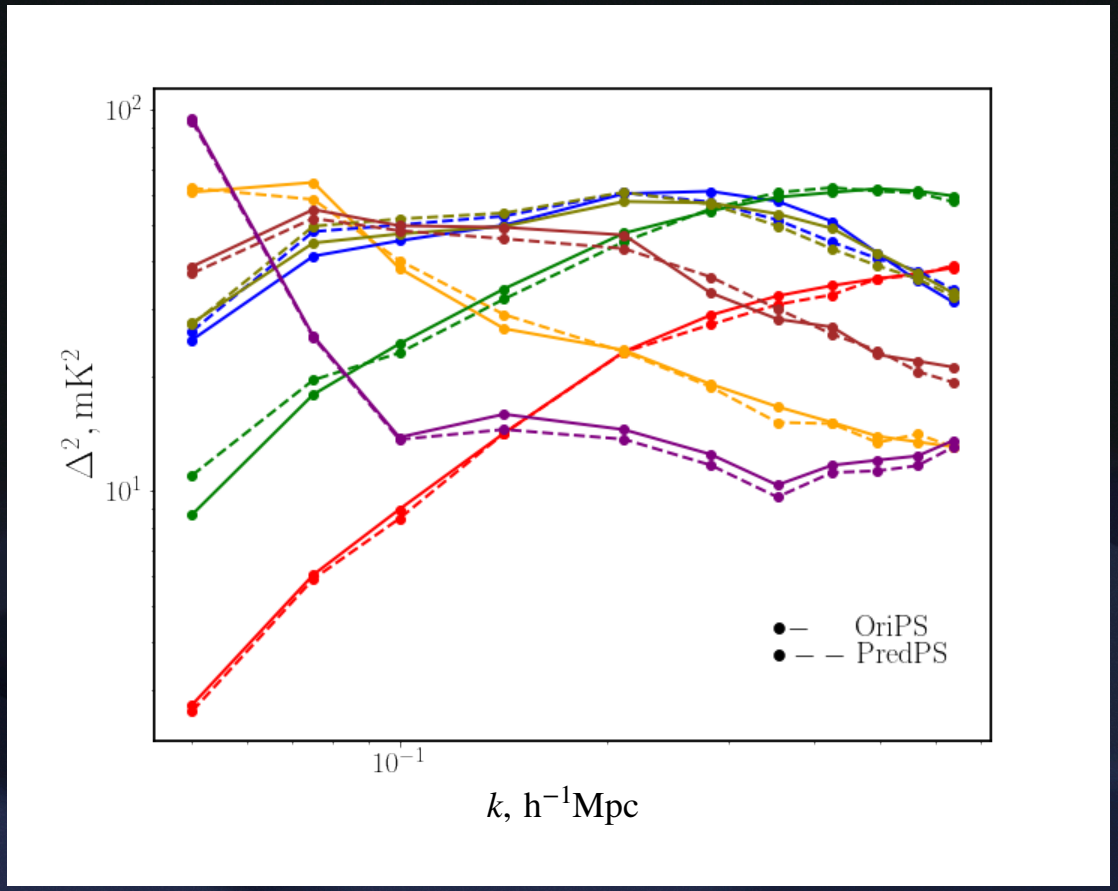
OUTPUT 21-CM POWER SPECTRUM

& x_{HII}

TRAINING SET

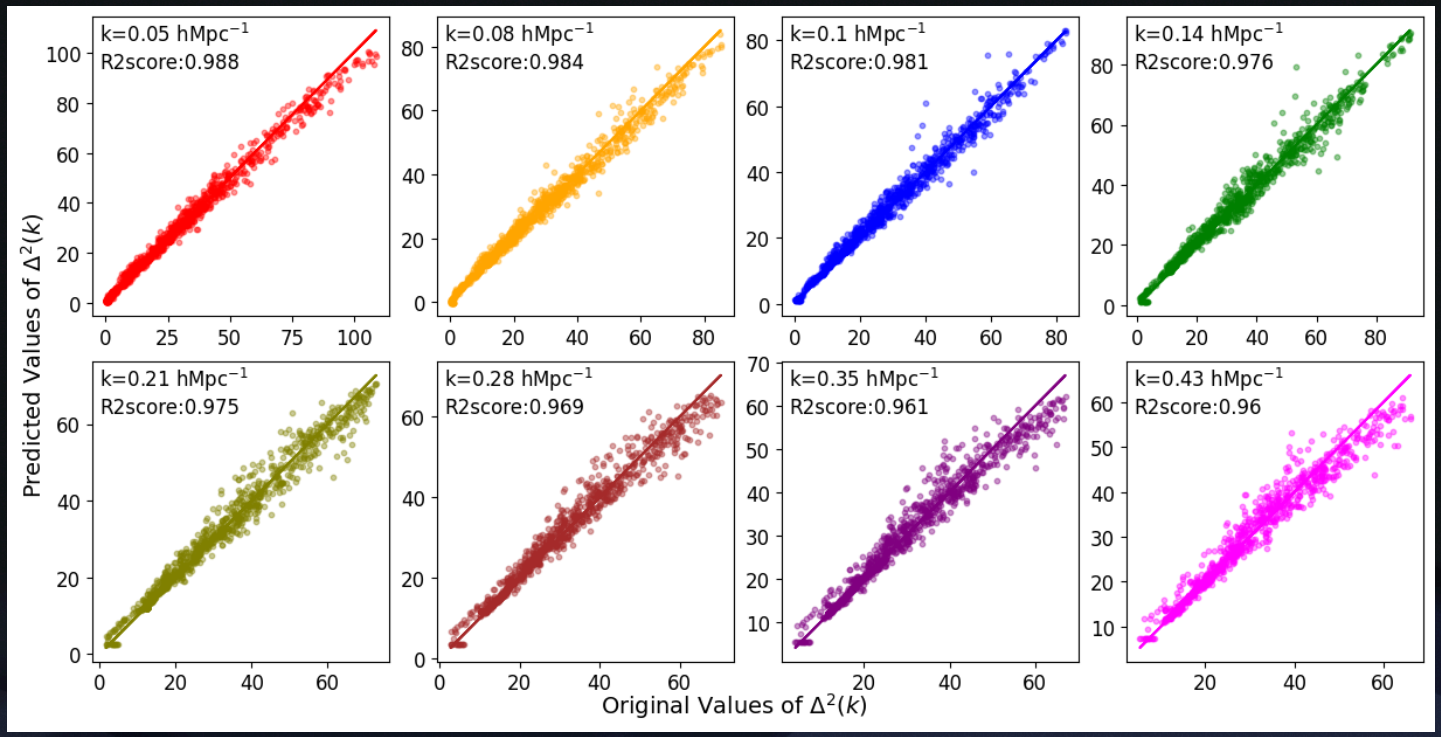
~17000 training samples generated using GRIZZLY

Results from the ANN-Emulator: Predicted power spectra from GRIZZLY testset



Choudhury M+ **LOFAR**, 2024

R2 scores from the ANN predictions for each kmode

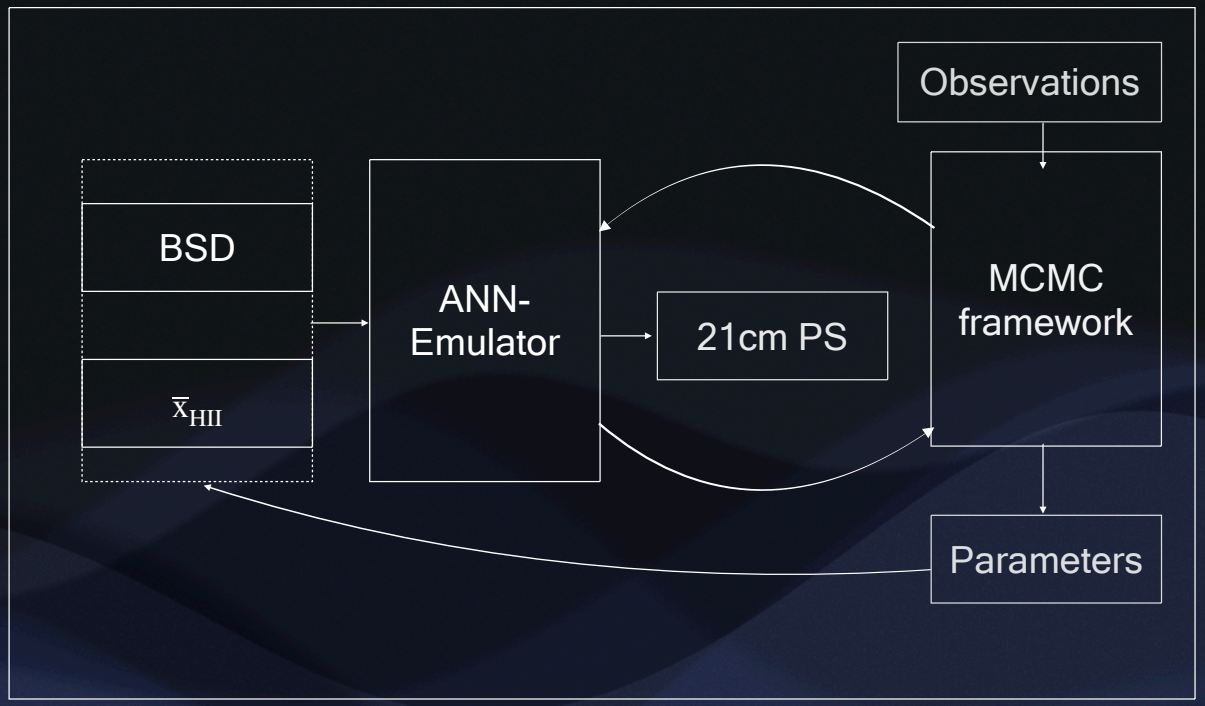


$$R^2 = \frac{\Sigma(y_{pred} - \overline{y}_{ori})^2}{\Sigma(y_{ori} - \overline{y}_{ori})^2}$$

R2 score closer to 1 implies a good prediction!

R2 scores calculated for a test set comprising of ~1000 GRIZZLY samples

Implementing the ANN Emulator within an MCMC framework

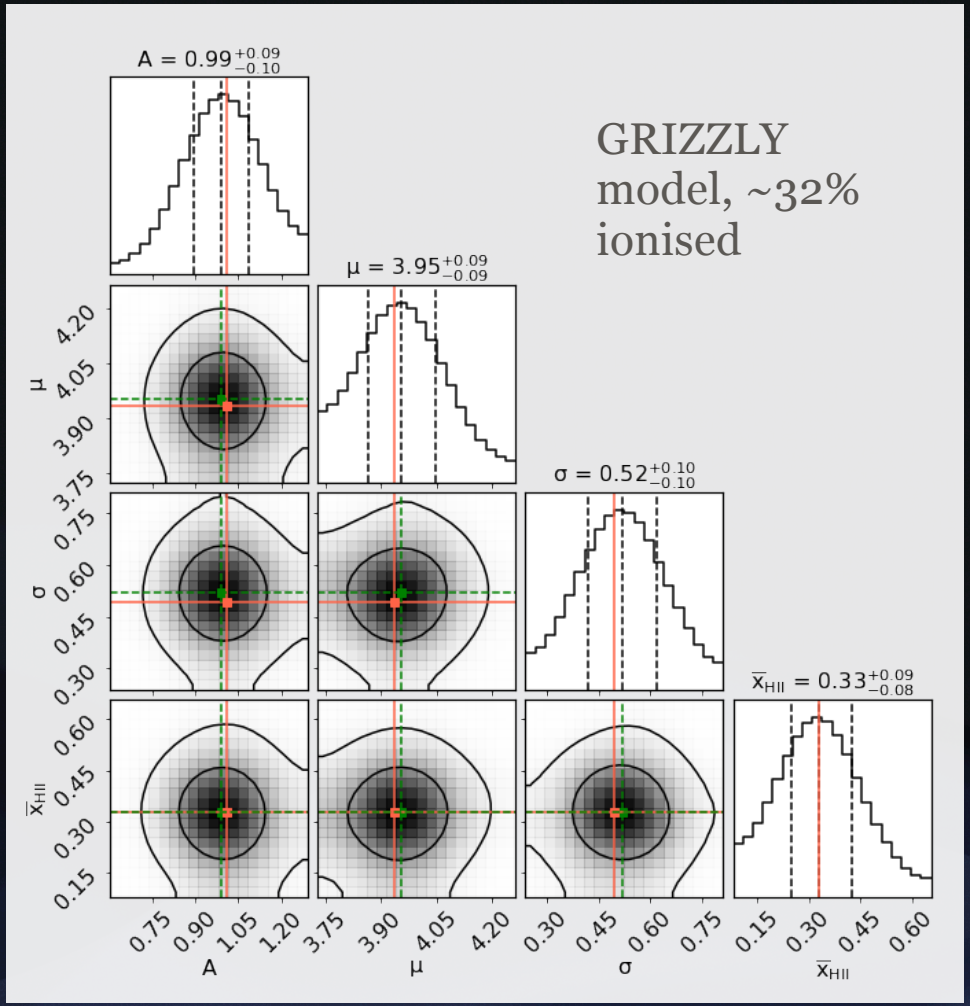


Input 21cm power spectrum

Volume averaged ionizaed fraction: $\overline{x}_{HII}$

(A, μ, σ): BSD parametrized as a lognormal distribution

PARAMETERS	TRUE VALUES
A	1.008
μ	3.935
σ	0.493
x_{HII}	0.329



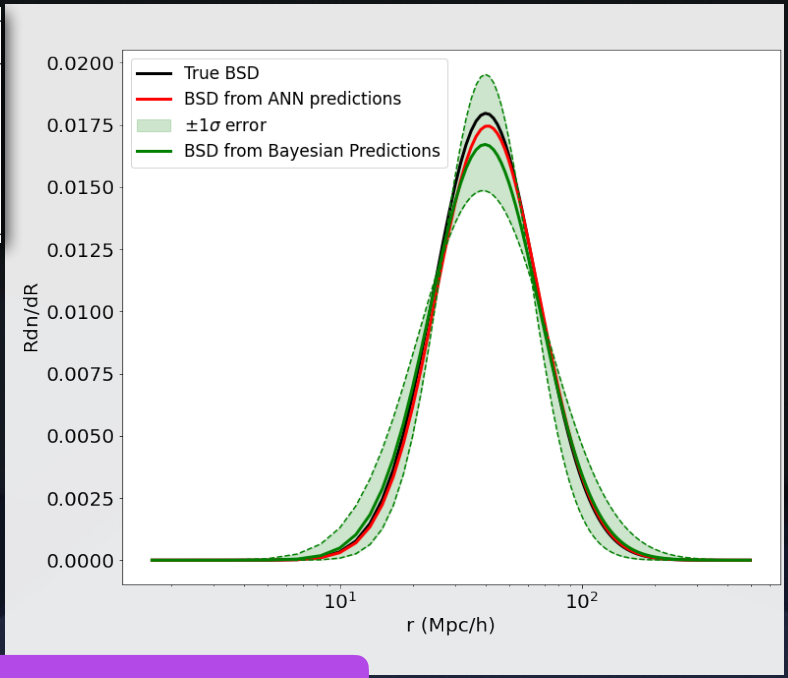
Comparison between MCMC values and the predictions from ANN

Parameter	True values	Bayesian method	ANN method
A	1.008	$0.99^{+0.14}_{-0.14}$	0.999
μ	3.935	$3.95^{+0.13}_{-0.13}$	3.954
σ	0.493	$0.52^{+0.13}_{-0.14}$	0.495
$\overline{x}_{HII}$	0.329	$0.33^{+0.10}_{-0.13}$	0.312

Quite close!

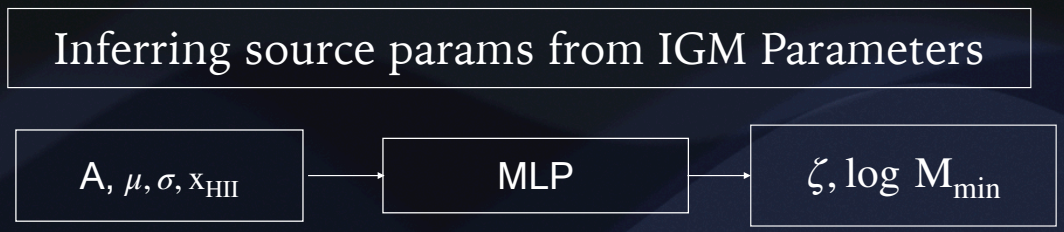
Takes only a few seconds to predict.

Bypasses the computation of the likelihood several times.

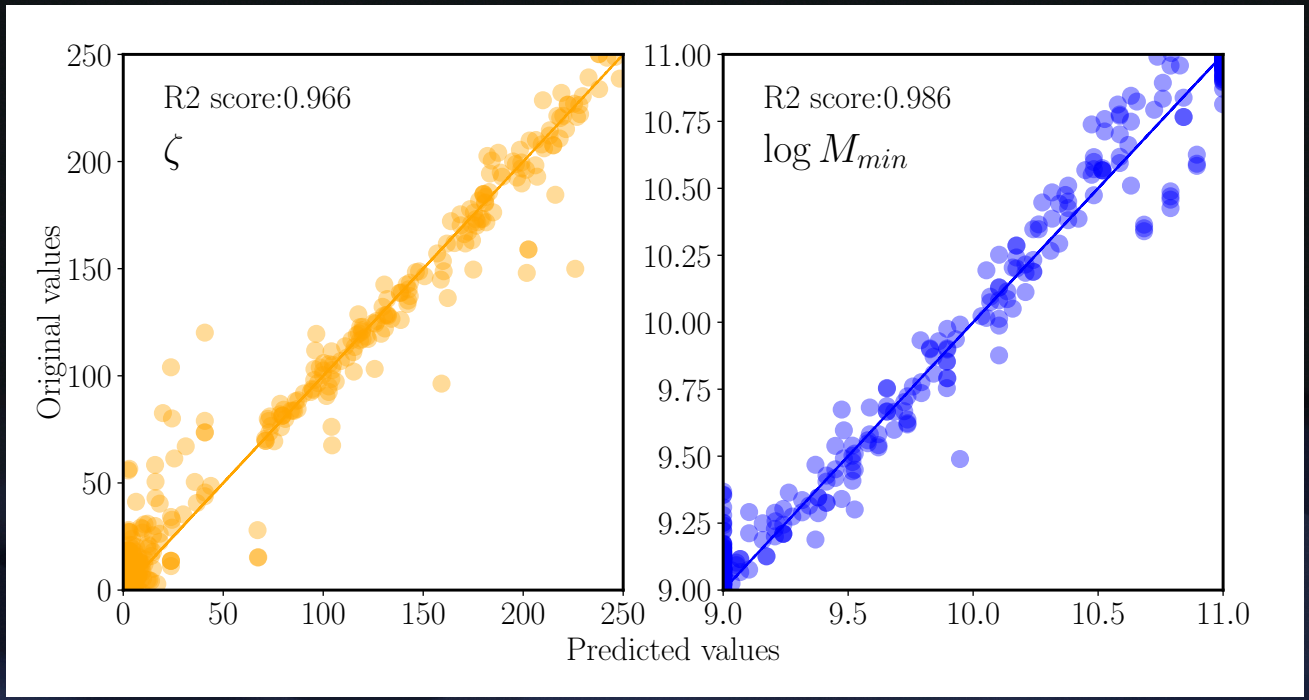


Important to note!
The IGM parameter space is a derived parameter space and is difficult to sample well.

CAN WE INFER THE SOURCE PARAMETERS ONCE THE IGM PARAMETERS ARE KNOWN?



Source parameters from the IGM parameters?



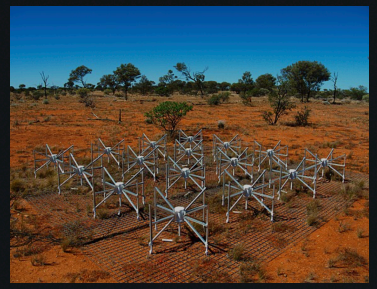
We find that the source parameters can be predicted with good accuracy, when inferred from the IGM parameters directly.

Regression from 21-cm images using ML?

Current and upcoming 21-cm experiments

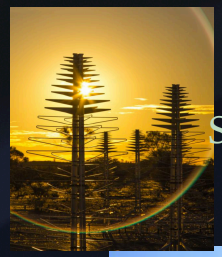
Ground-based experiments

HERA



MWA

GMRT

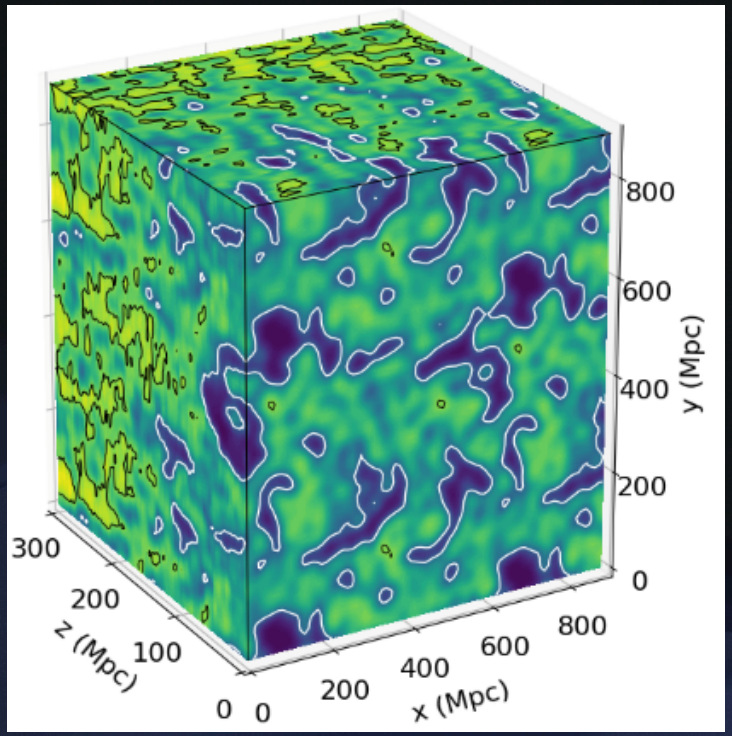


SKA-Low

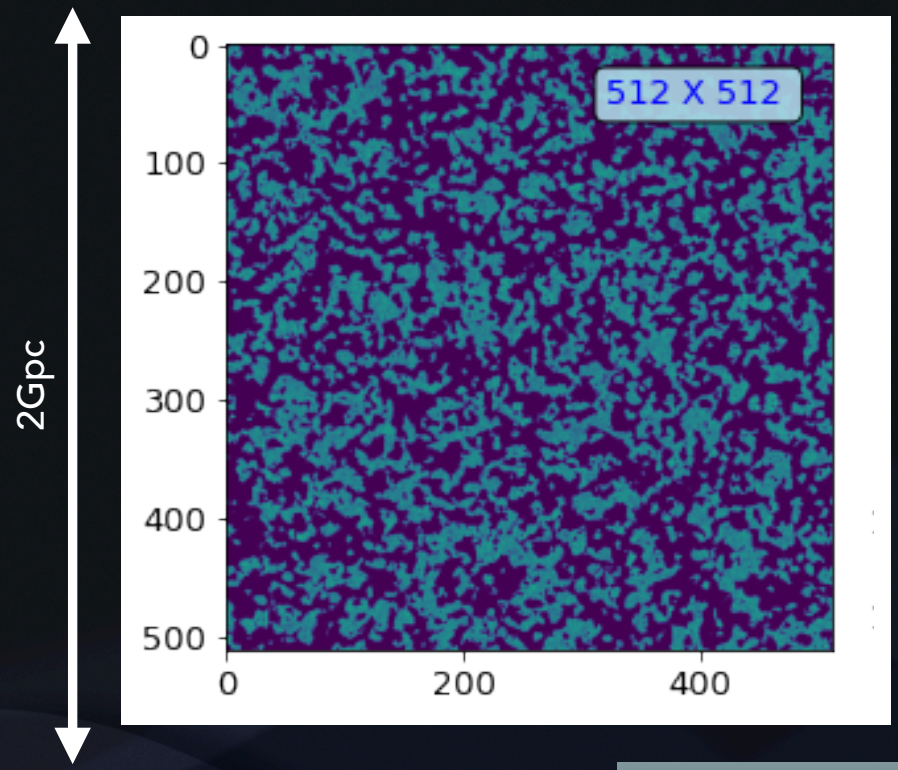
SKA-Mid

https://www.skao.int/en/explore/telescopes/

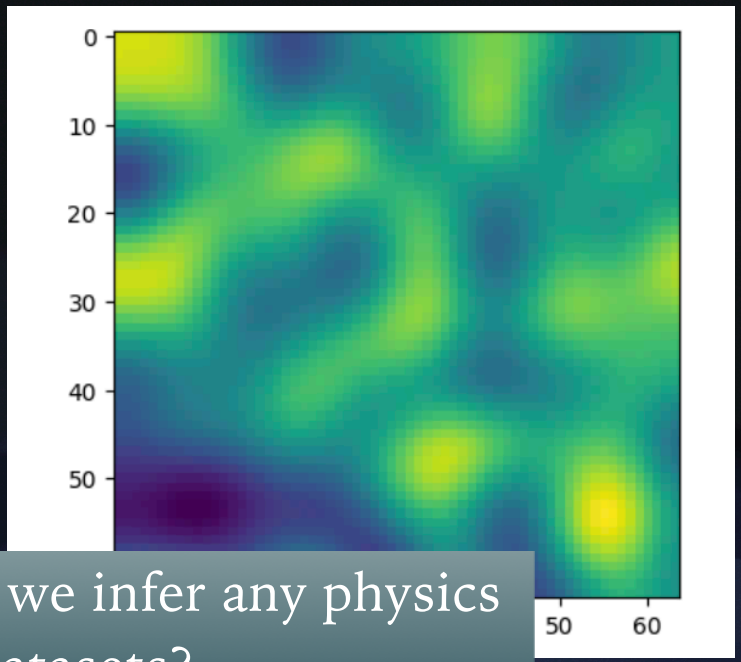
21cm tomography is the science goal!



A sample 2D δT_b slice from a light cone

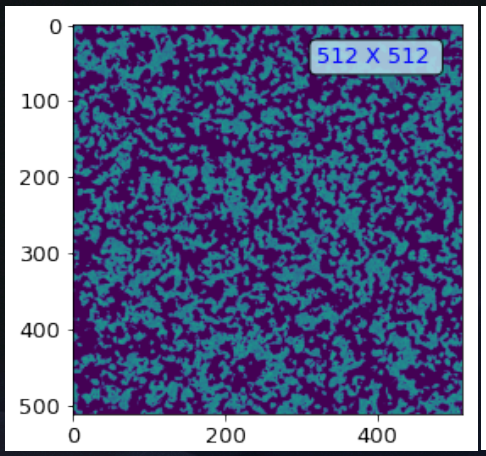


We'll be looking at something like this!? (Exaggerated effect)

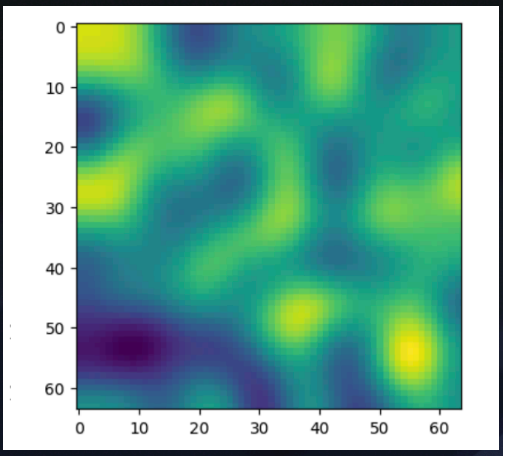


How would we infer any physics from such datasets?

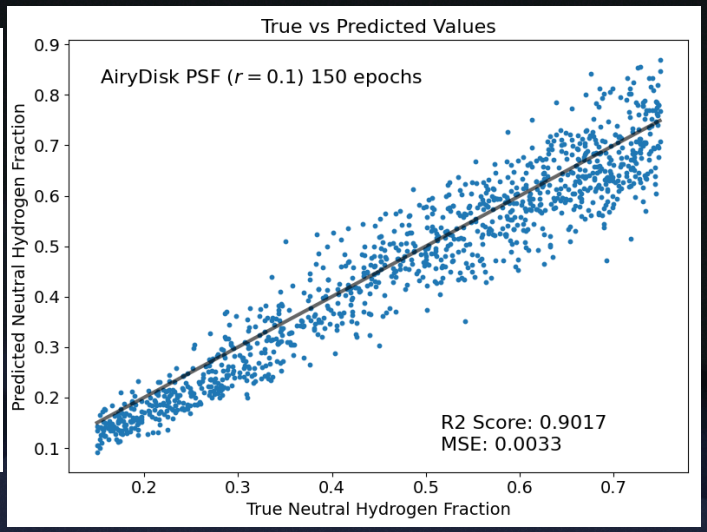
Preliminary test to check how well ML can be used to predict the x_{HII} from these synthesised images



Original sample dTb map

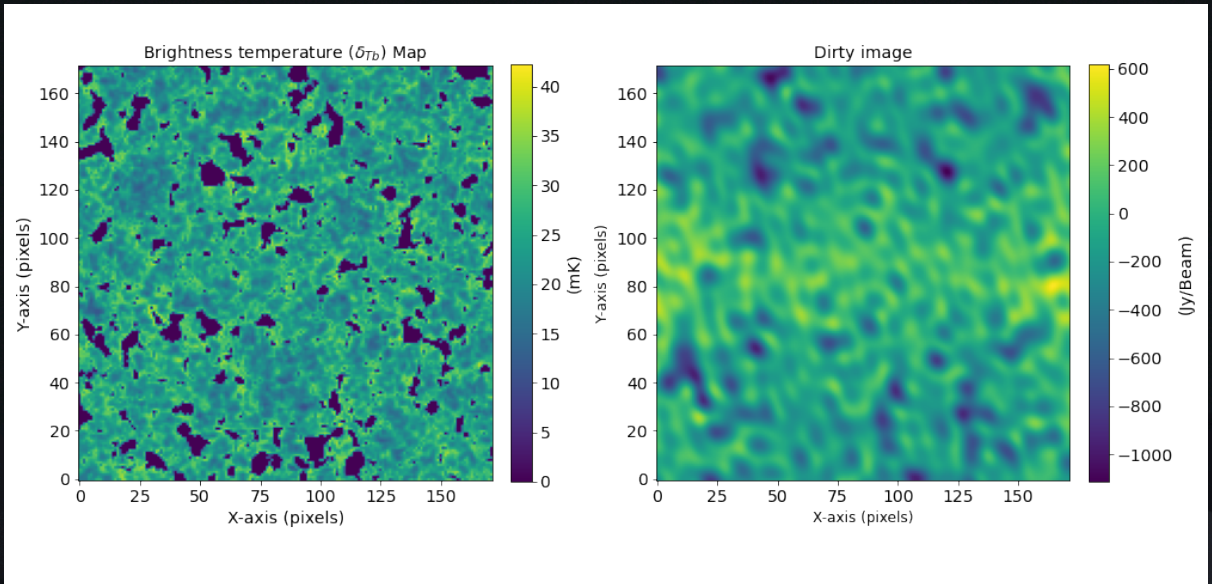


Dirty images/ convolved with the sythesized beam



Predictions from separately trained CNN with an Airy disk PSF

Simulated observations of the brightness temperature slices δT_b with the MWA Phase II

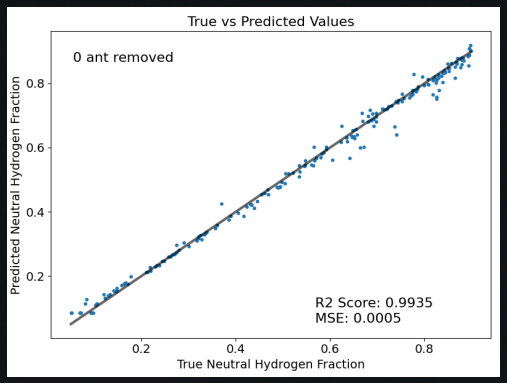


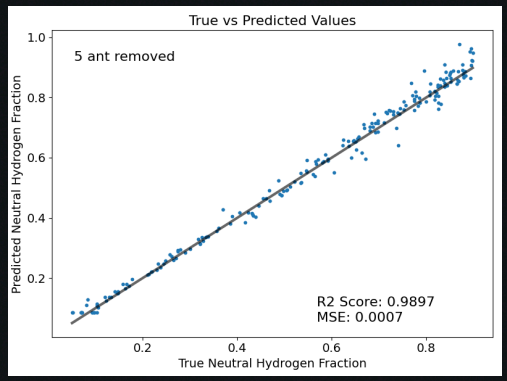
A 2D CNN is then trained on these simulated observations to predict the neutral Hydrogen fraction (x_{HI})

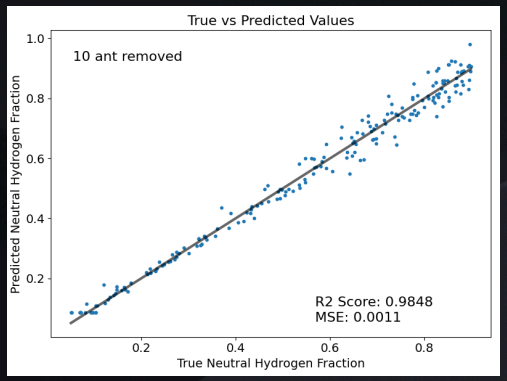
R2 scores for various test sets

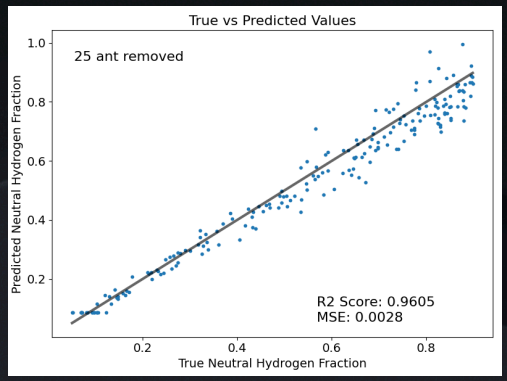
Test sets are constructed by randomly switching off a few antennae form the MWA Phase II configuration.

R2 scores are very good!



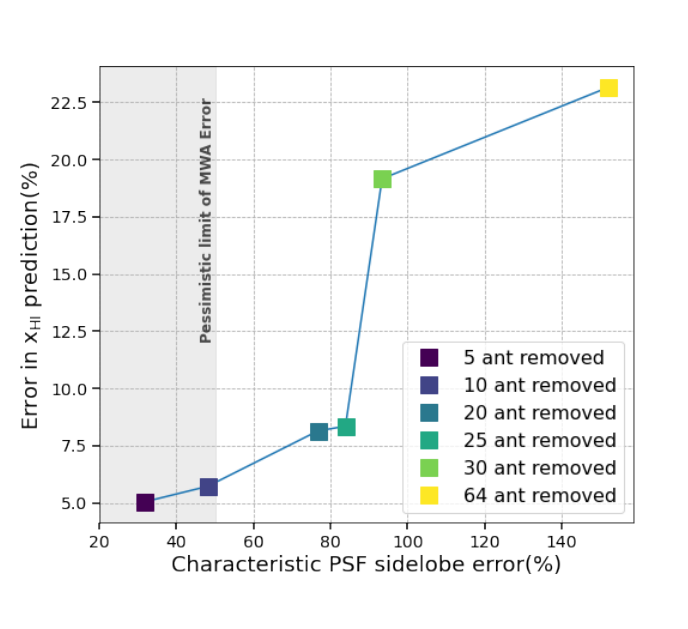






Choudhury & Pober (In prep)

Effects of mismodeling the PSF in predicting x_{HI}



Choudhury & Pober (In prep)

We are extending this work to include the effects of the primary beam model and systemtics while incorporating the frequency axis as well.

Once we have enough measurements of the 21-cm power spectra..

Regressing on evolving physical quantities from 21-cm power spectrum using ML?

REGRESSING ON REDSHIFT DEPENDENT HISTORIES

EXTRACTING THERMAL AND IONISING HISTORIES FROM 21-CM POWER SPECTRUM DATA USING NEURAL NETWORKS

$P_n(k, z)$

Input 21cmPS information

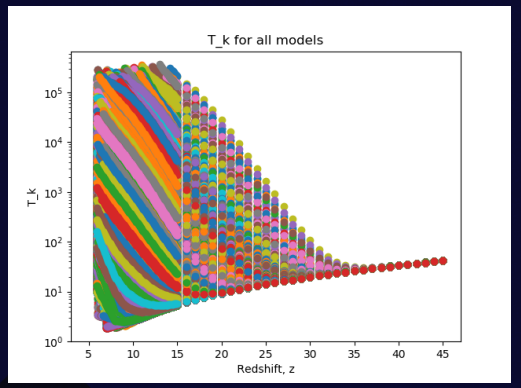
Pre-processing

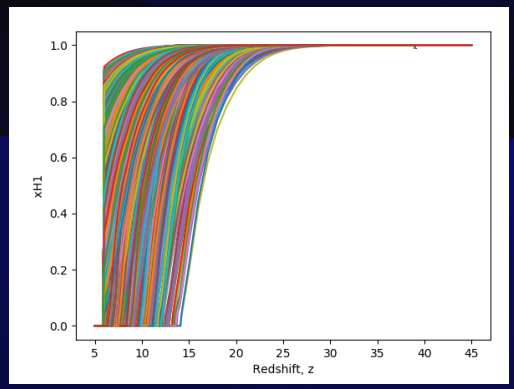
Neural networks

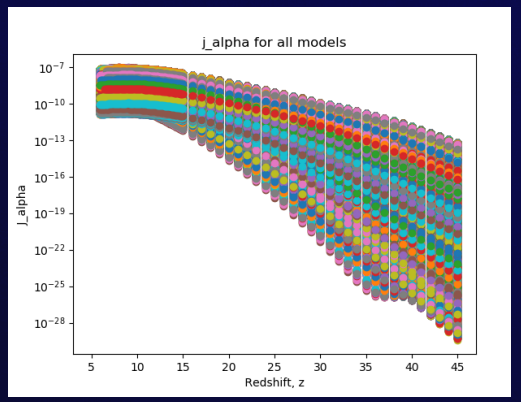
$T_k(z)$

$x_{HI}(z)$

$J_\alpha(z)$







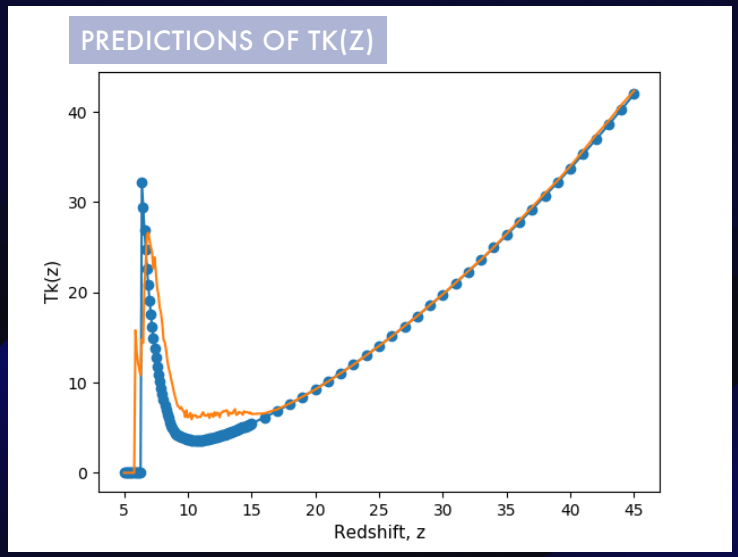
Associated Tk(z),Ja(z), xH1(z) for each model: these are the targets/outputs for the entire set

Choudhury M+ (In prep)

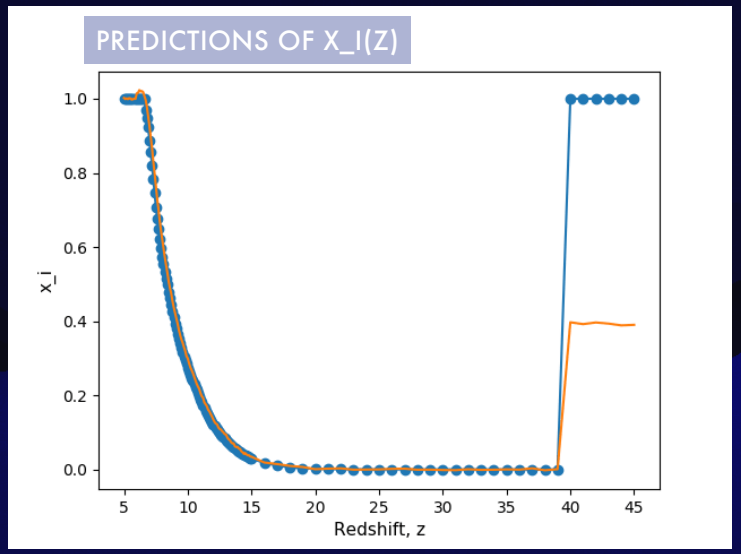
Flowchart

Predictions of extracted histories from 21cm PS information

PREDICTIONS OF TK(Z)



PREDICTIONS OF X_I(Z)

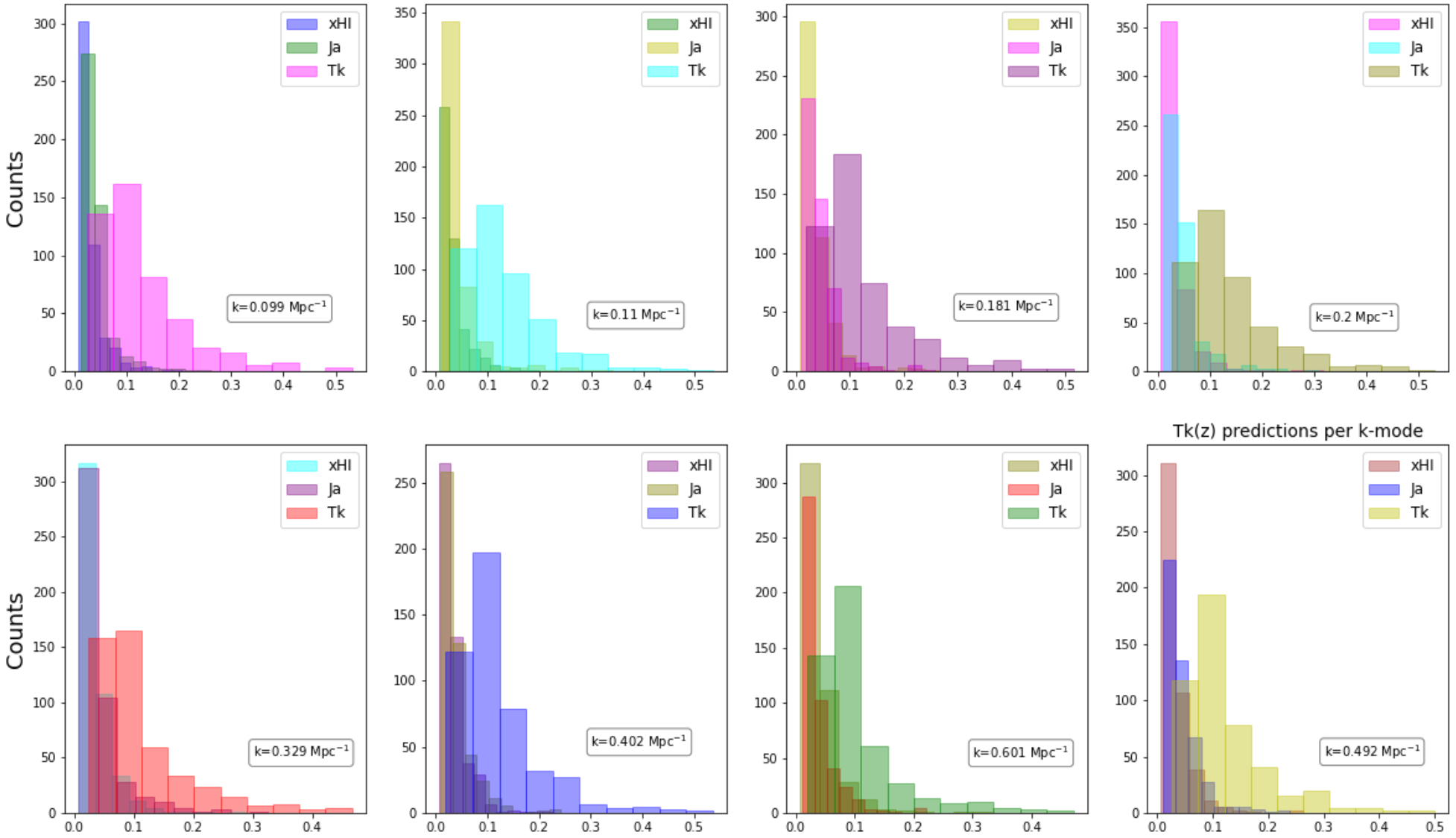


Original data:

Predictions:

THIS FRAMEWORK WOULD BE EXTREMELY USEFUL TO PREDICT IONISATION HISTORIES AND THERMAL HISTORIES FROM POWER SPECTRUM INFORMATION

Train several NNs separtately for different k's and bin the results!



Conclusions & road ahead!

- Infering the IGM parameters directly from the 21-cm power spectrum, would lead to developing a ‘source-parameter-free’ framework.
- We are developing a more generalised inference framework, which would work for other simulations not being dependent on source models or the initial density fields.
- The MCMC framework can be replaced by an ANN for estimating the IGM parameters. Computationally fast, bypasses explicit likelihood calculation.
- We can use image based regresssion methods to infer the parameters describing the state of the IGM directly from future images!

Thanks!

Questions?