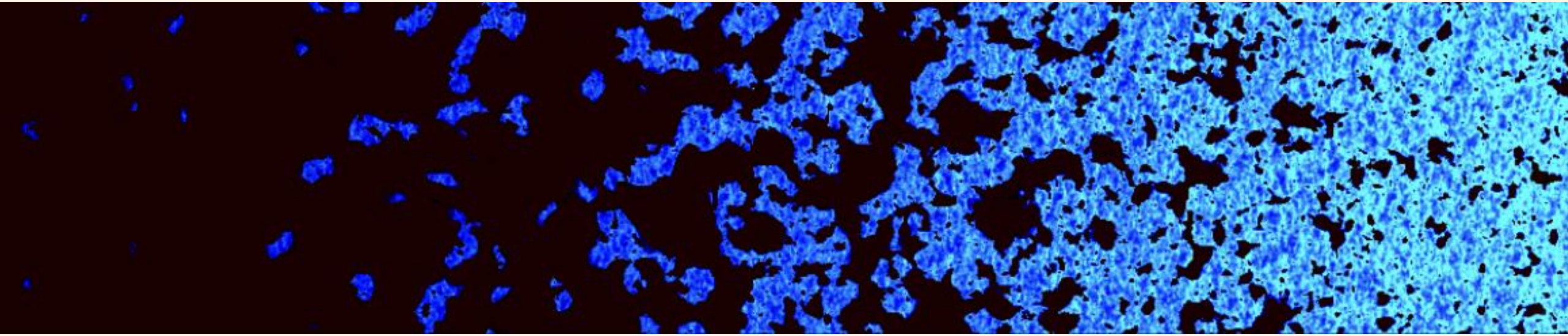
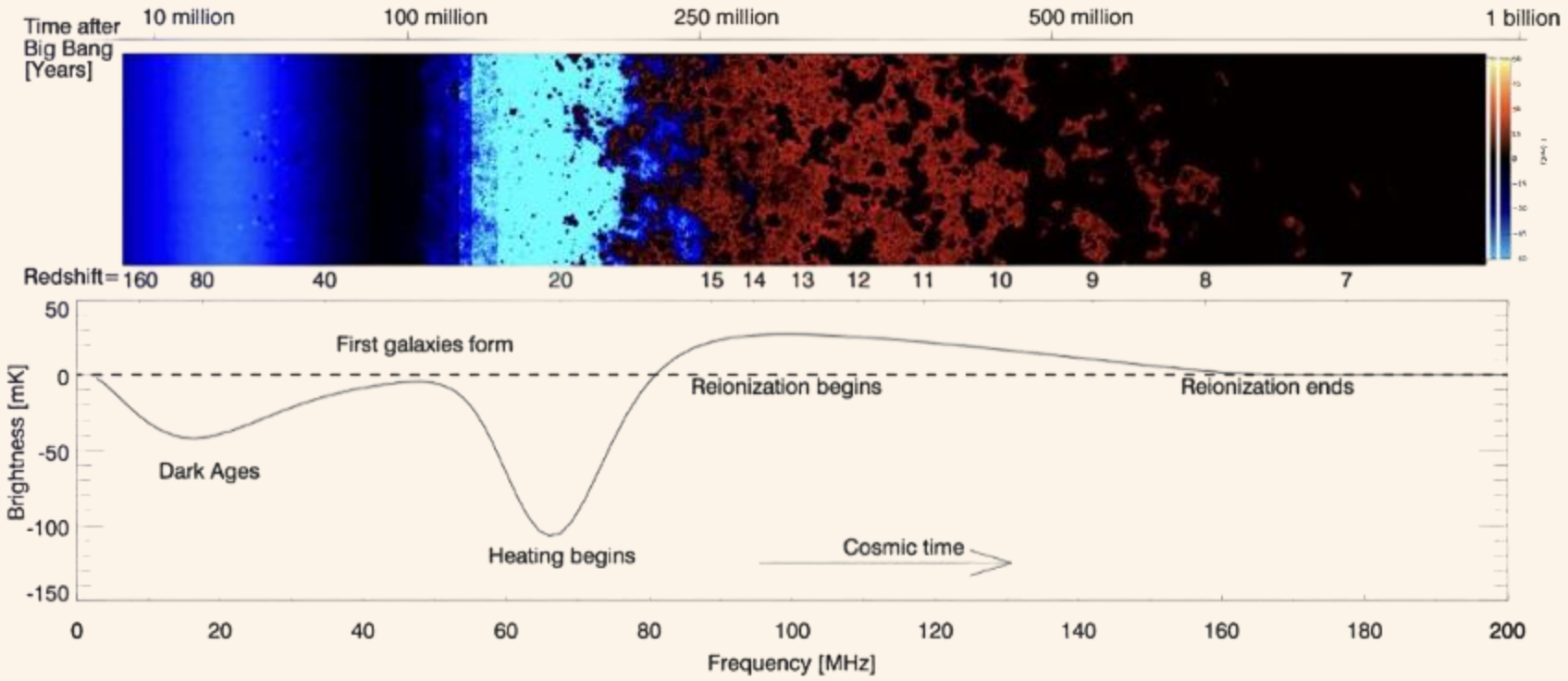
Hybrid Summary Statistics



Nadia Cooper: nhc19@ic.ac.uk
Supervisor: Jonathan Pritchard
Cosmology in the Alps 17th March 2026

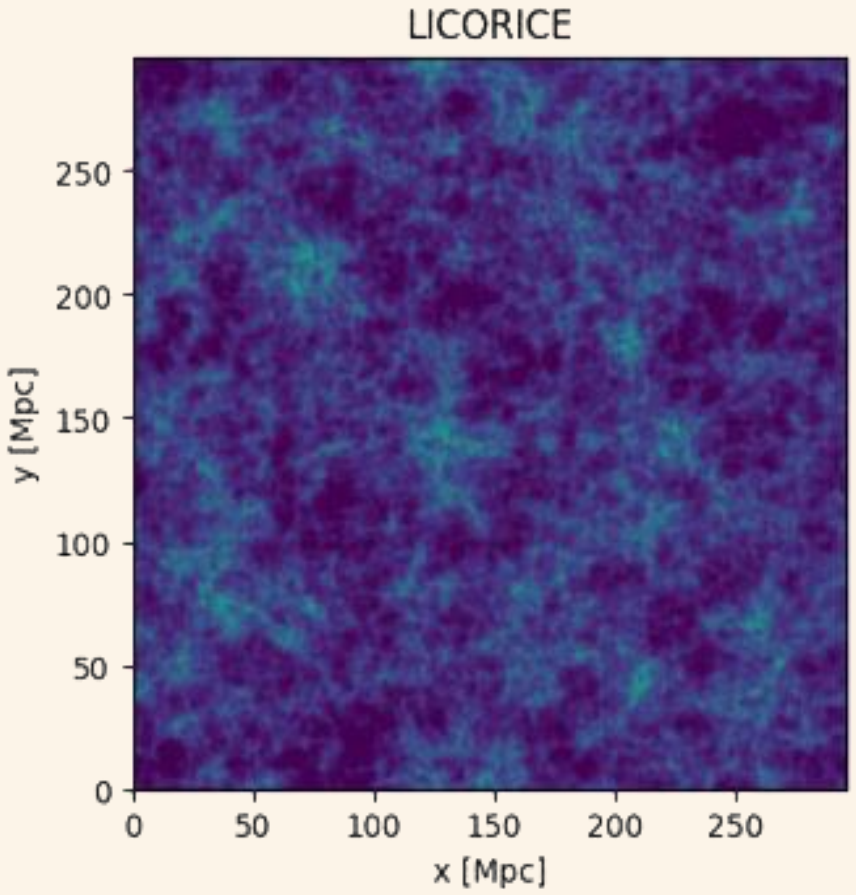
IMPERIAL

Motivation

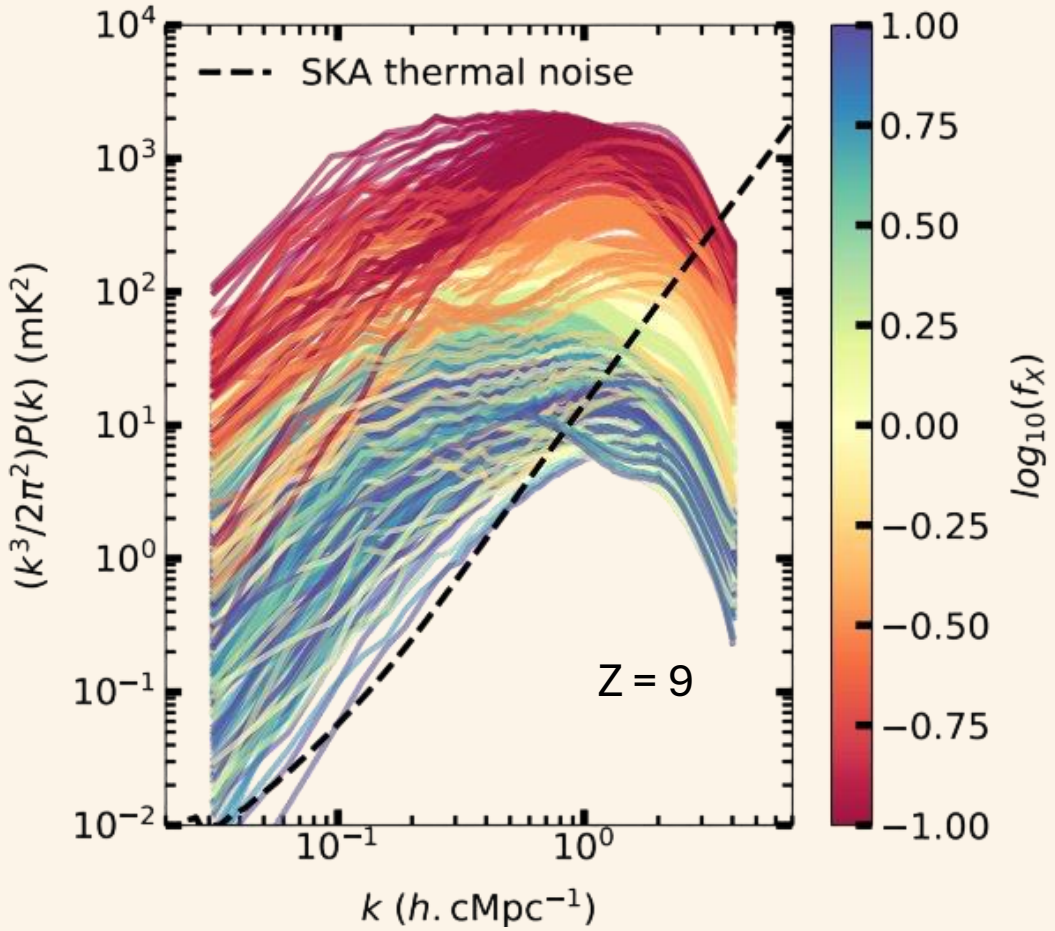


Pritchard et al, 2012

Summary Statistics of the 21 cm Field



LICORICE: Semelin et al, 2017



Meriot and Semelin et al, 2024

Inference from Power Spectra

STEP ONE – Generate Training set

20 K
Simulations from 21cmFAST (Mesinger et al, 2010)

8.5 K
Samples from Licorice (Semelin et al, 2017)

STEP TWO – Summary statistic

Calculate the power spectrum of the field

STEP THREE - SBI

Train neural density estimator to learn astrophysical parameters from power spectrum

Sample from your posterior density estimator

STEP FOUR - SBC

Validate Pipeline Using Posterior Calibration Tests (Talts et al, 2018)

Hybrid Summary Statistic Inference

STEP ONE – Generate Training set

20 K Simulations from 21cmFAST (Mesinger et al, 2010)

8.5 K Samples from Licorice (Semelin et al, 2017)

STEP TWO – Learn summaries

Train CNN to learn optimal additional summaries

Concatenate with PS

STEP THREE - SBI

Train neural density estimator to learn astrophysical parameters from hybrid summaries

Contrast with only power spectra constraints

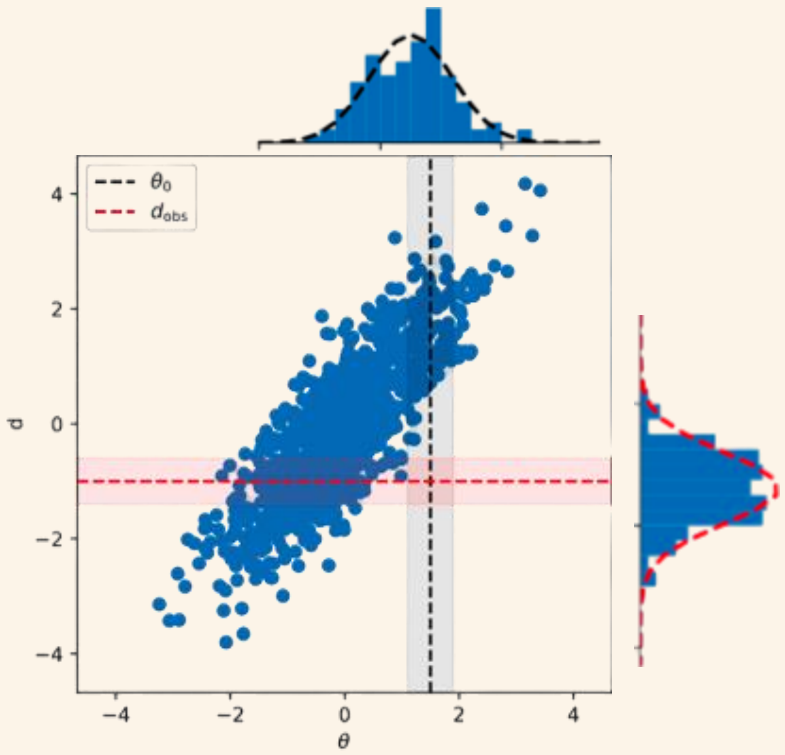
STEP FOUR - SBC

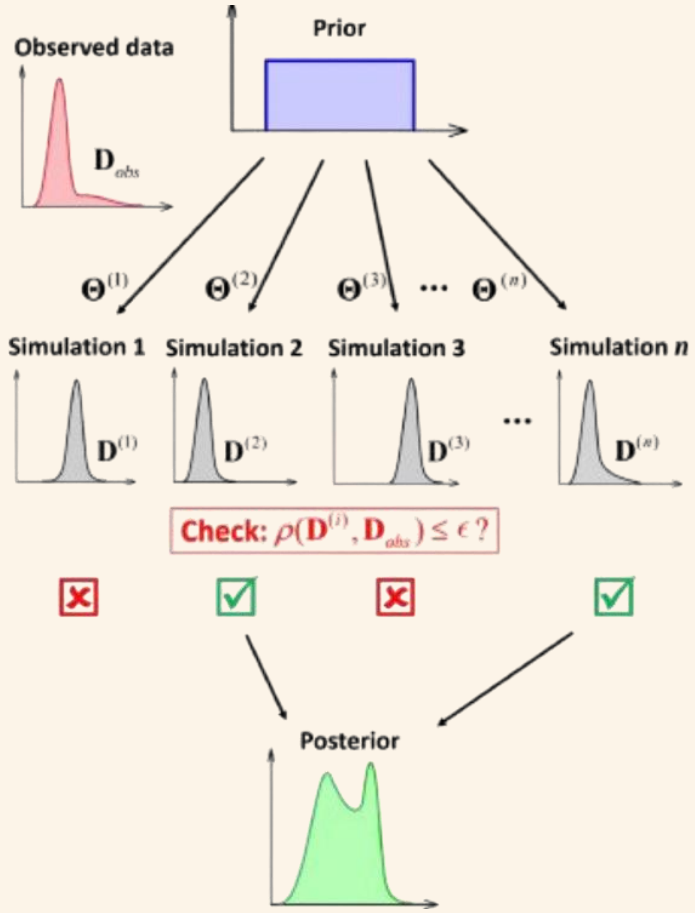
Validate Pipeline Using Posterior Calibration Tests (Talts et al, 2018)

Simulation Based Inference

Advantages

- Doesn't require an analytic form of the likelihood
- Can be done on an amortized bank of simulations
- Typically requires fewer simulations to converge

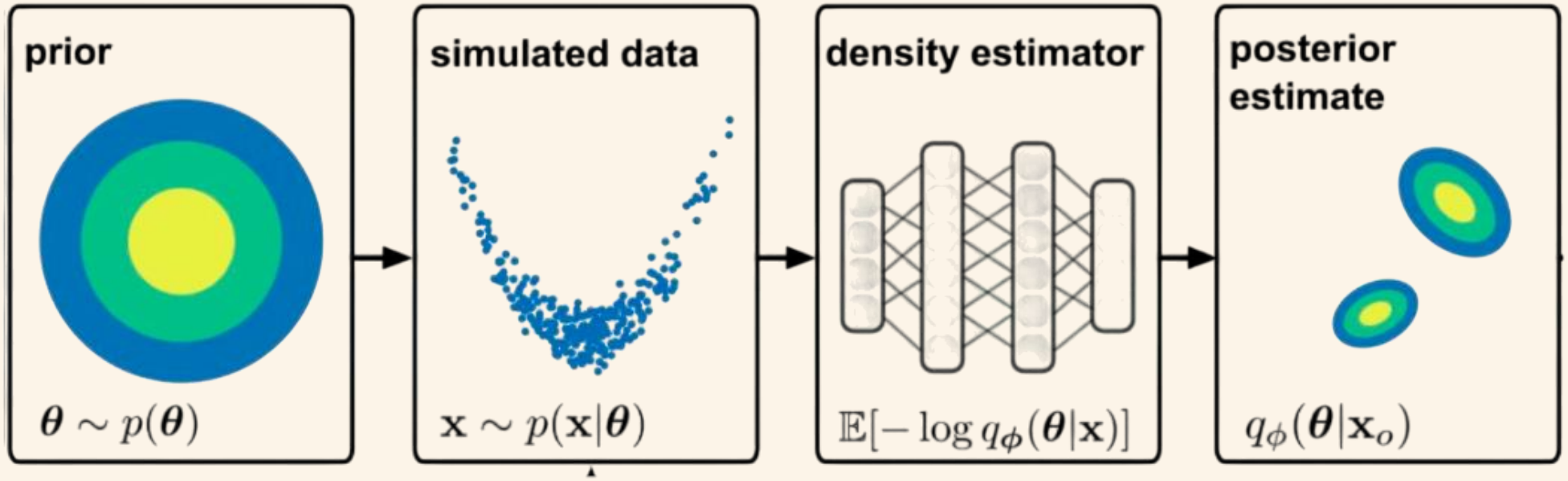




M. Sunnåker, 2013

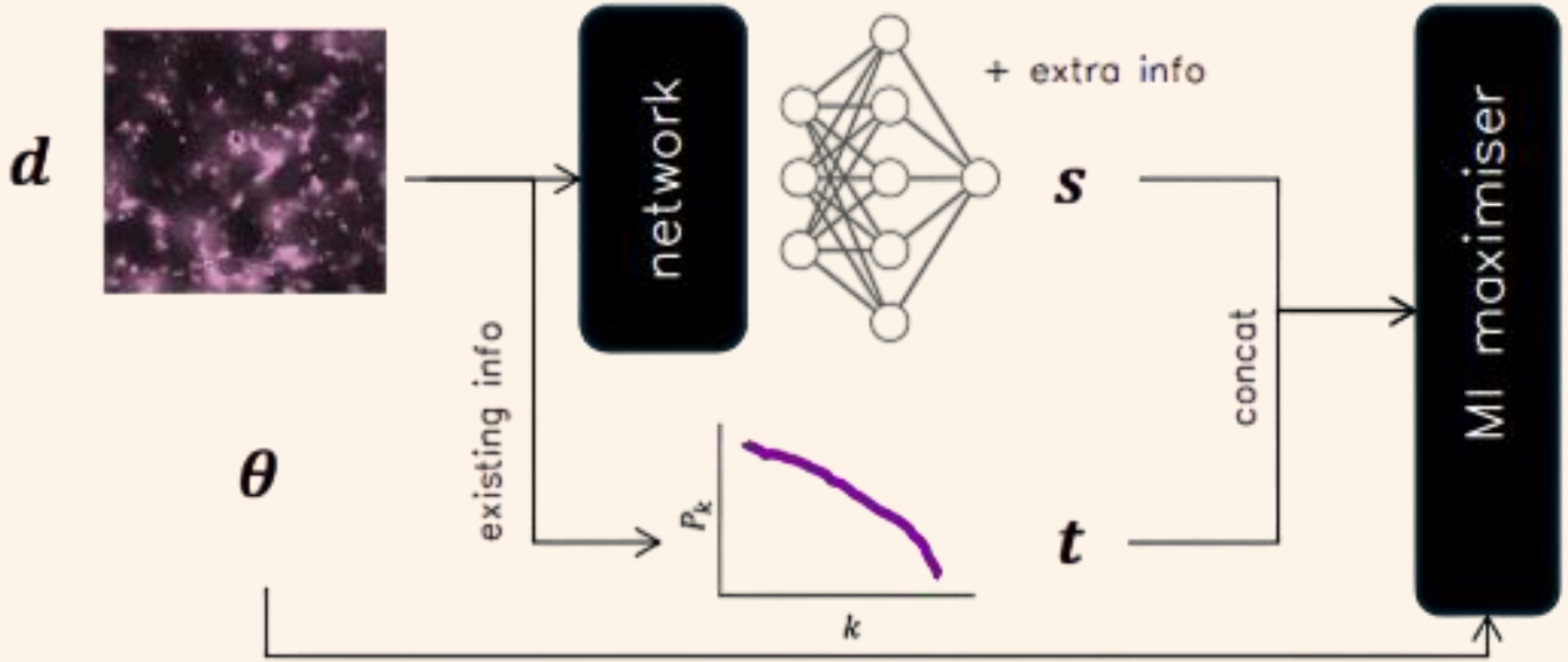
See Zhao et al 2021, 2022, Prelogovic and Mesinger 2023, Saxena et al 2024, Grieg et al 2024, Meriot et al 2025, Cooper, Norregaard et al 2025

Neural Density Estimation



Deistler, Michael and Goncalves, 2022

Hybrid Summary Statistics



Makinen et al, 2024

Learning the Summaries

Mutual Information

$$I(\mathbf{s};\theta|\mathbf{t}) = I([\mathbf{s},\mathbf{t}];\theta) - I(\mathbf{t};\theta)$$

Information we had before

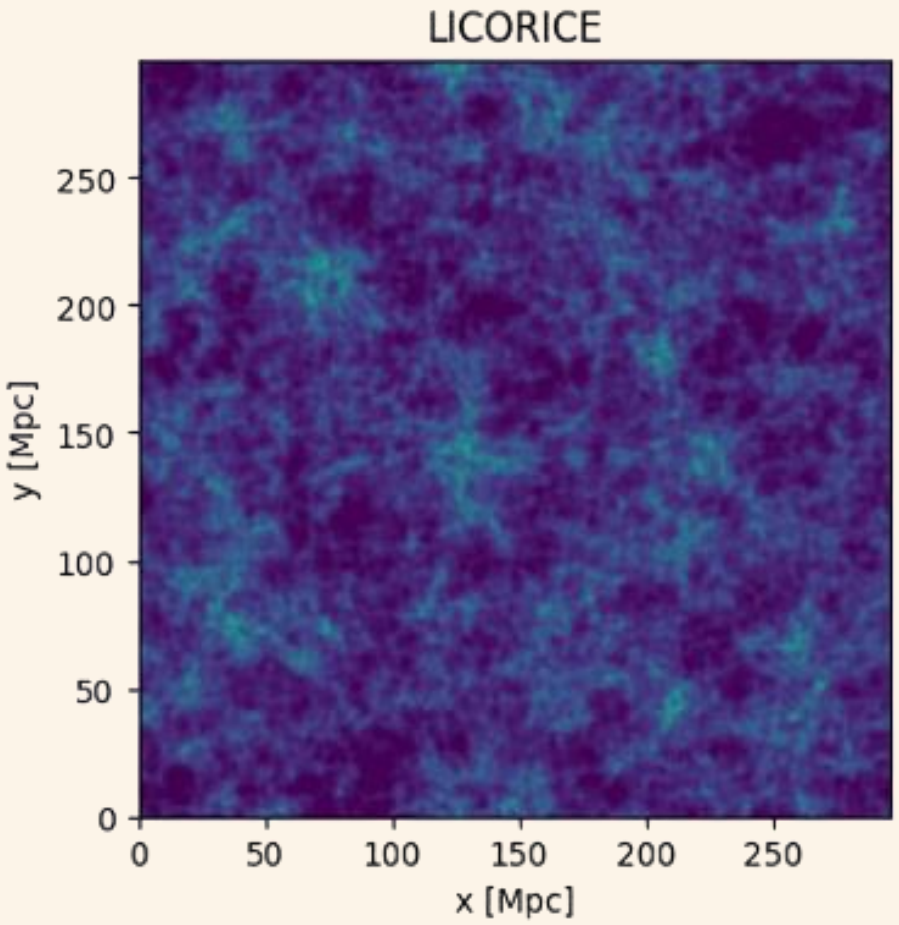
Total information

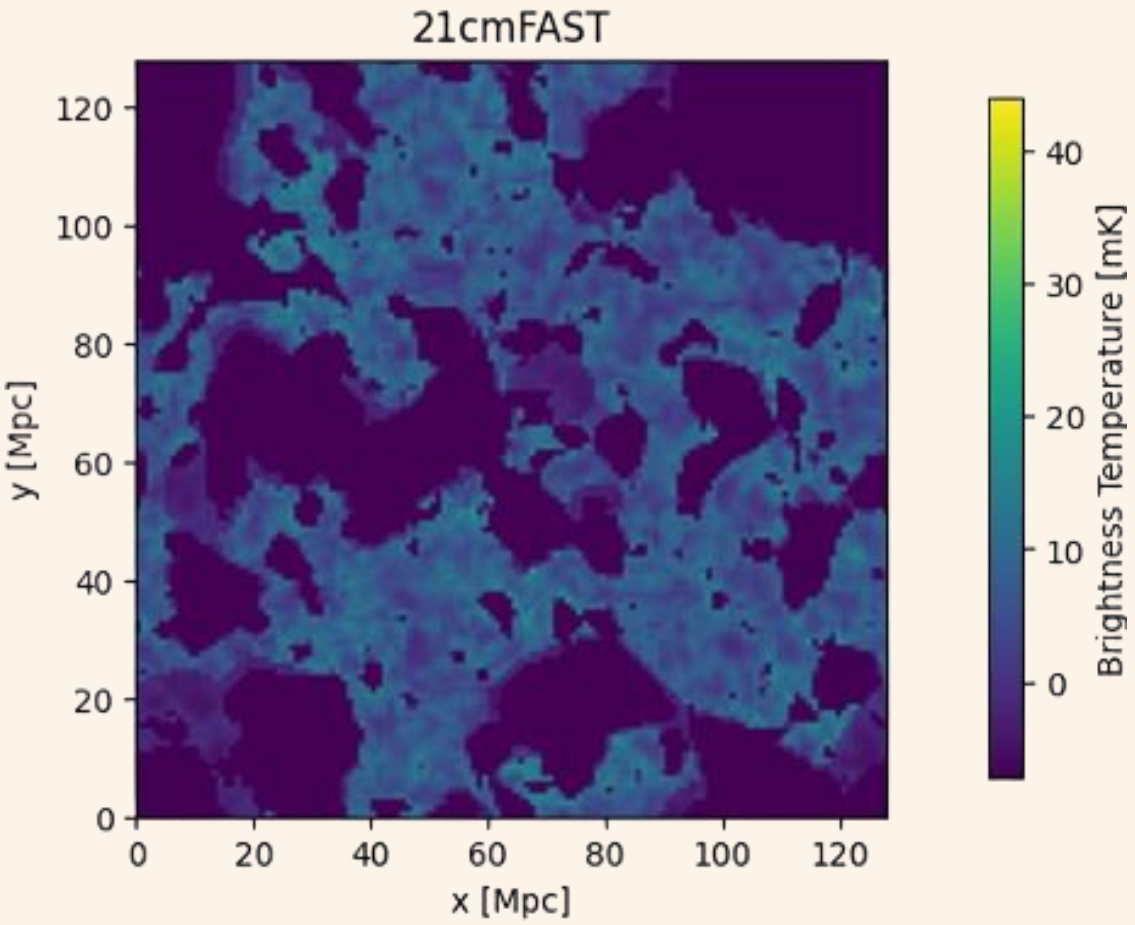
t – existing summary
s – learnt summary

Minimise the expected posterior entropy:

$$\min_{\mathbf{s},q} \mathcal{L} = -\mathbb{E}_{p(\theta,\mathbf{d})}\left[\log q\left(\theta|[\mathbf{s}(\mathbf{d}),\mathbf{t}(\mathbf{d})]\right)\right]$$

Simulation Codes





LICORICE: Semelin et al, 2017

21cmFAST: Mesinger et al, 2010

Simulation Code Comparison

21cmFAST

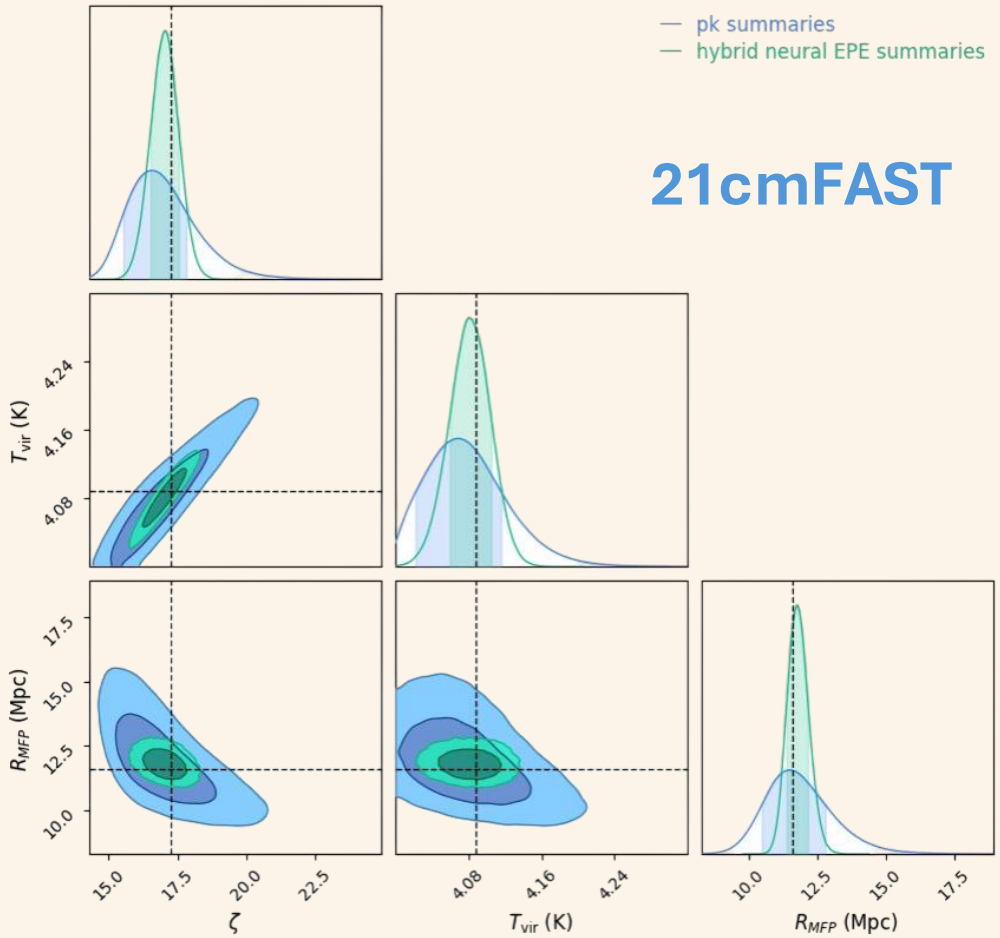
- Semi-numerical
- 20K simulation data base
- 3 parameters of interest
- Sampled on a Latin hypercube
- Densely sampled

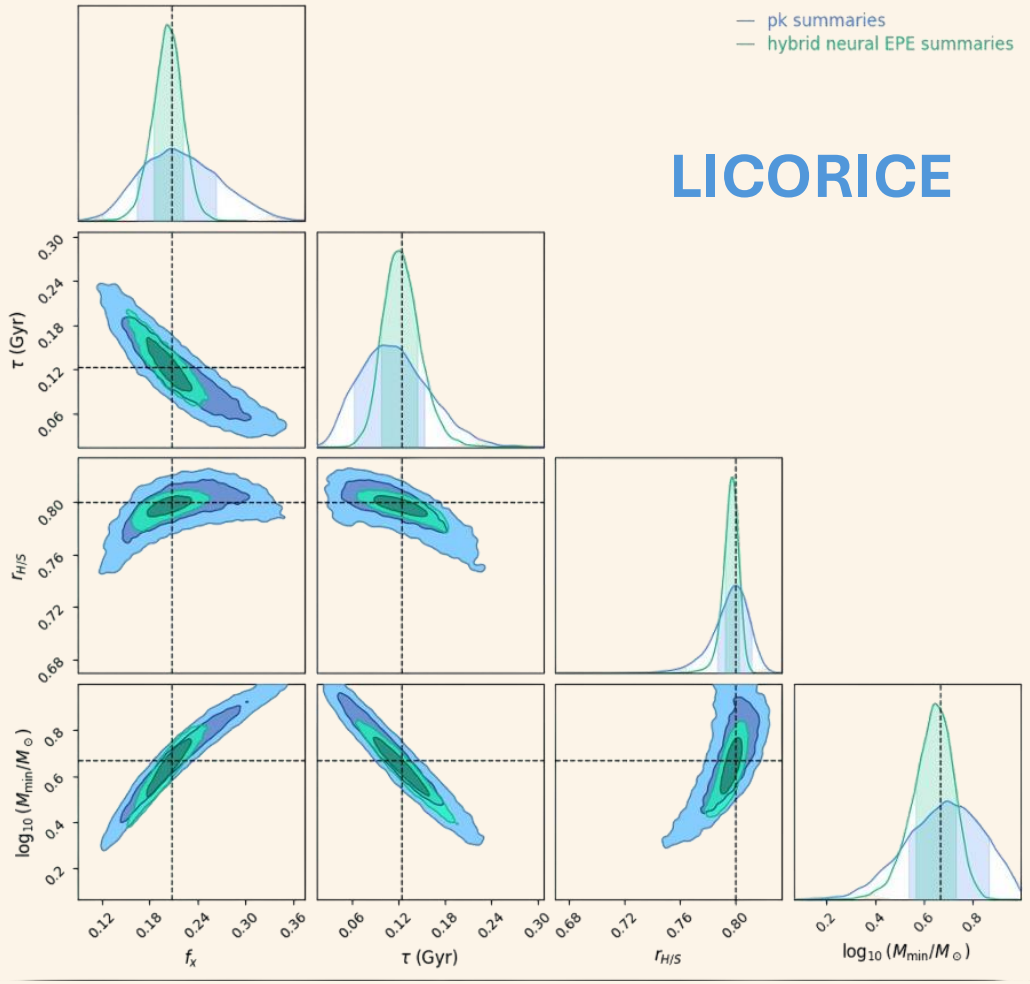
LICORICE

- Radiative transfer
- ~8.5K simulation data base
- Four parameters of interest
- Sampled on a grid
- Sparsely sampled

.. will it still work?

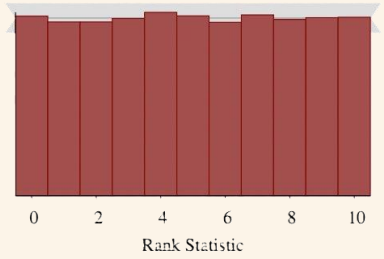
Improvements in Constraints



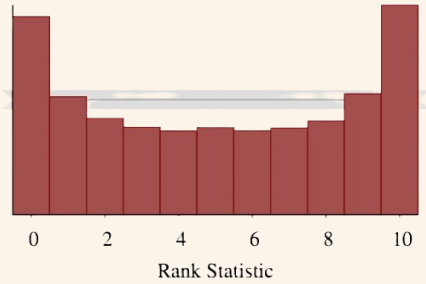


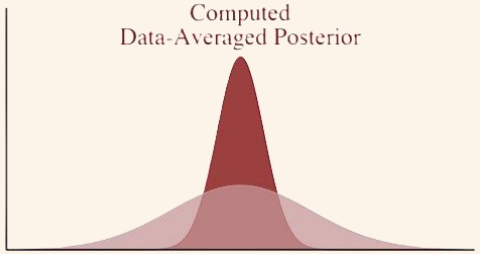
Simulation Based Calibration

Well Calibrated

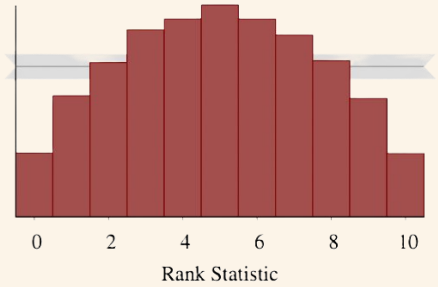


Overconfident

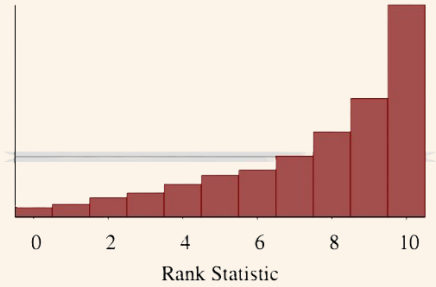


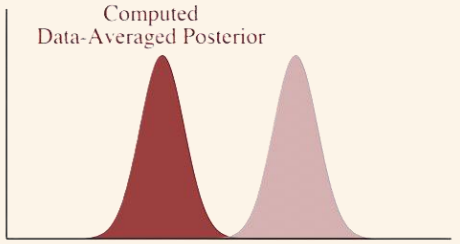


Underconfident



Biased



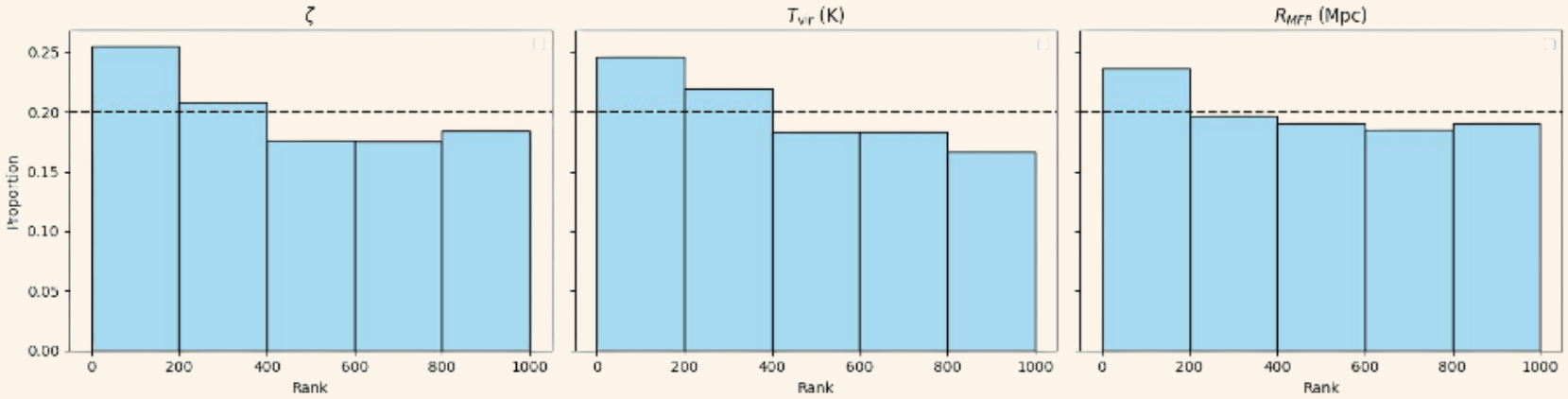


Talts et al 2020

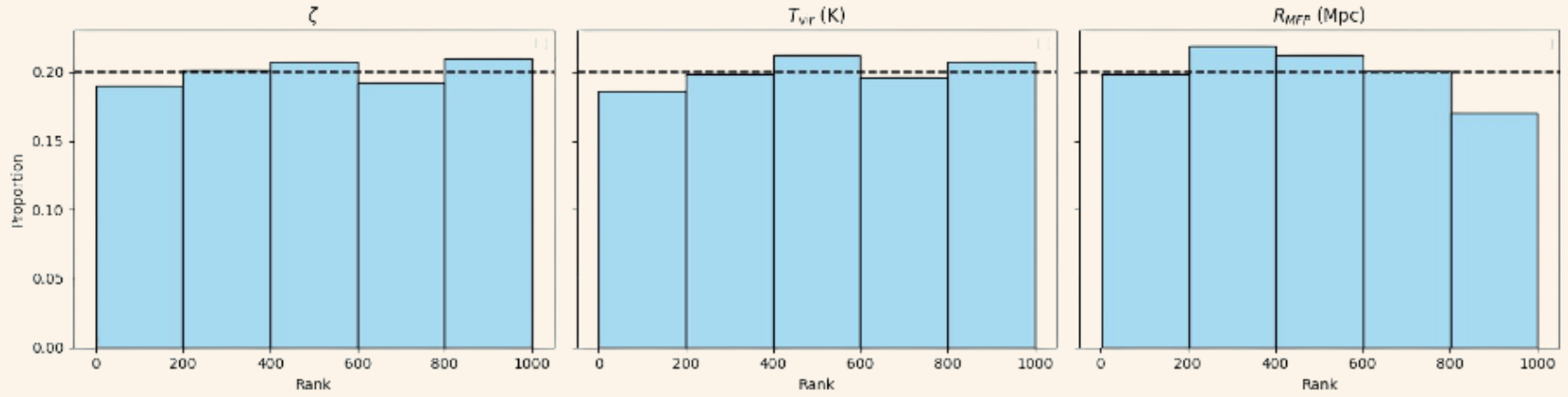
Simulation Based Calibration

21cmFAST

Hybrid Summaries



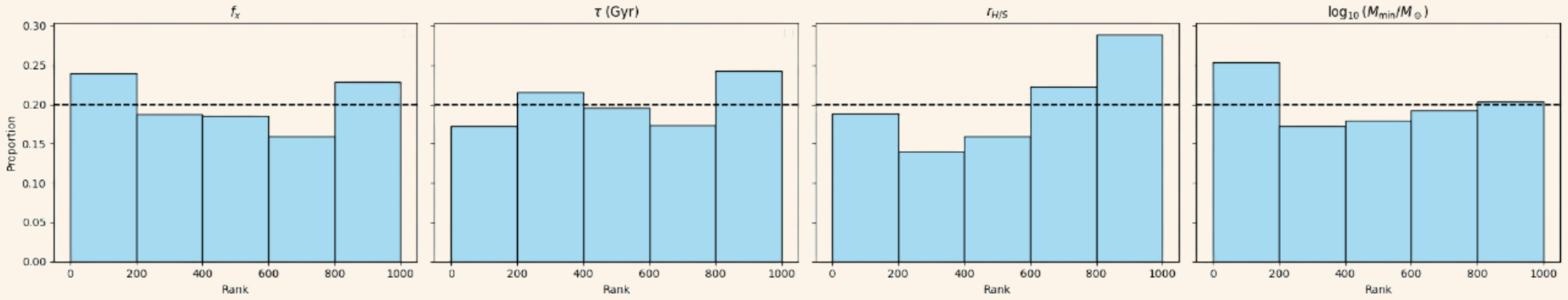
Power Spectra only Summaries



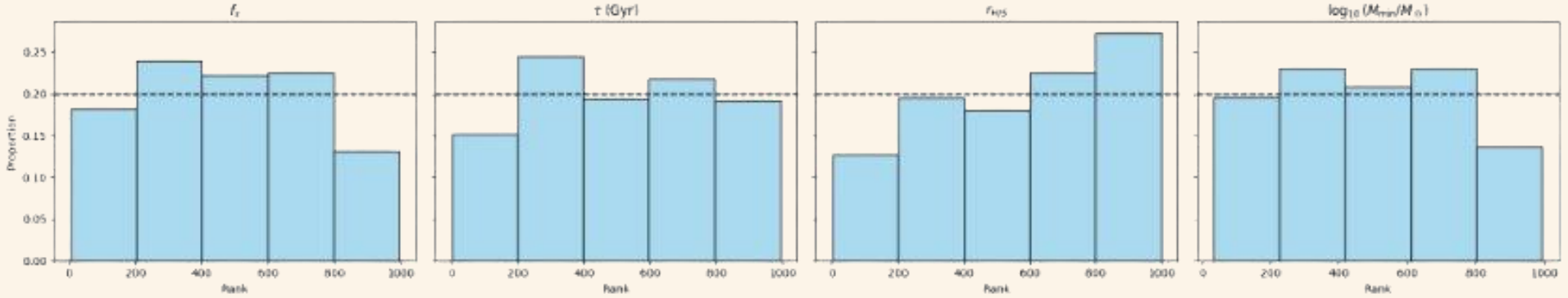
Simulation Based Calibration

LICORICE

Hybrid Summaries

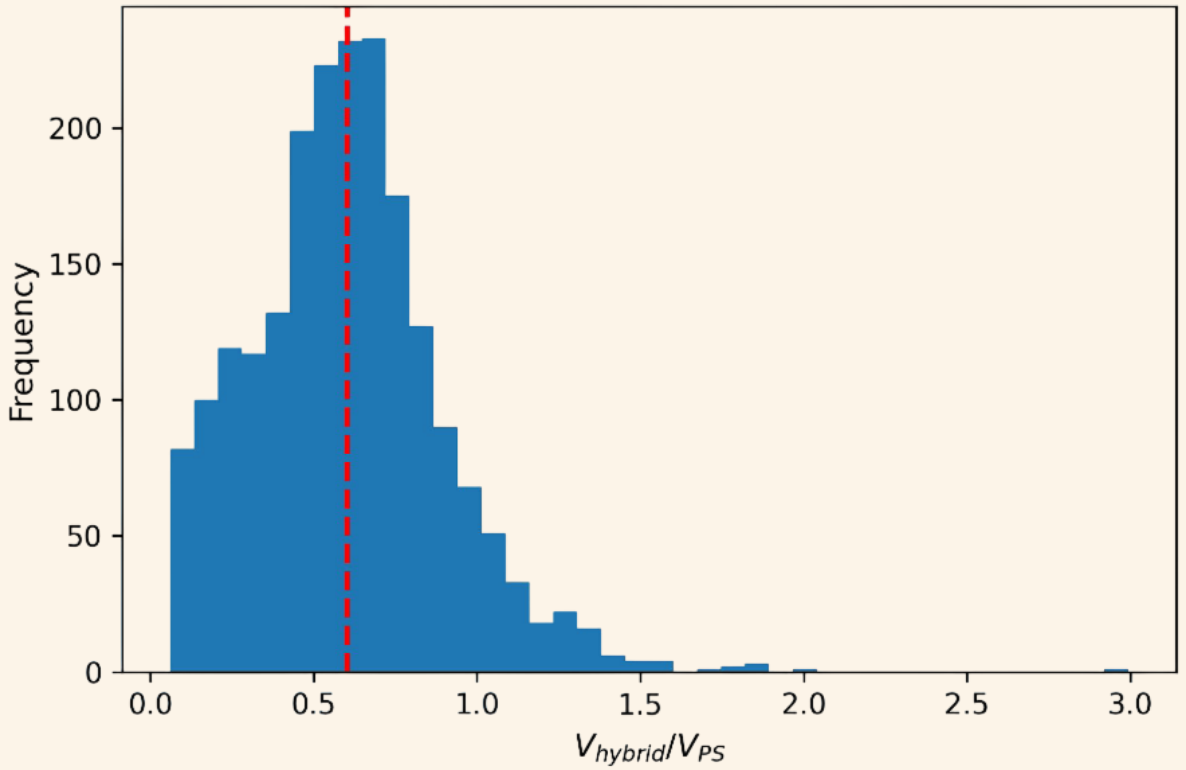


Power Spectra only Summaries

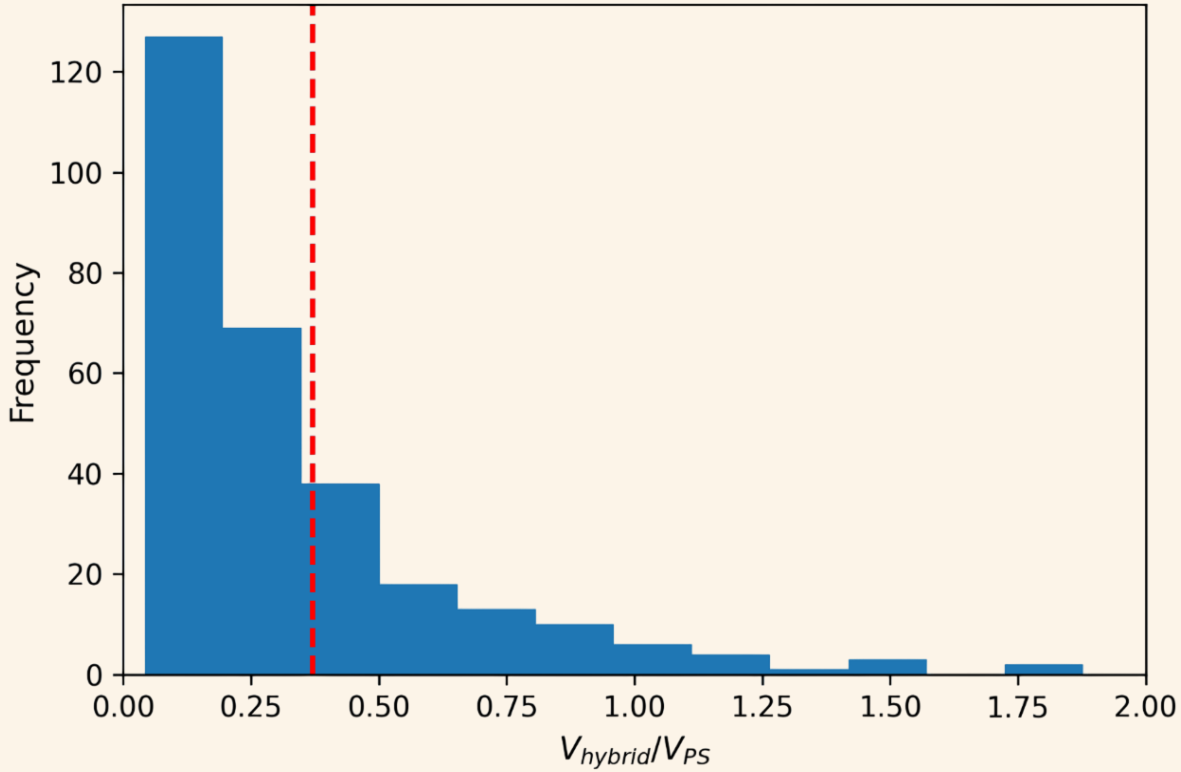


Reduction in posterior volume

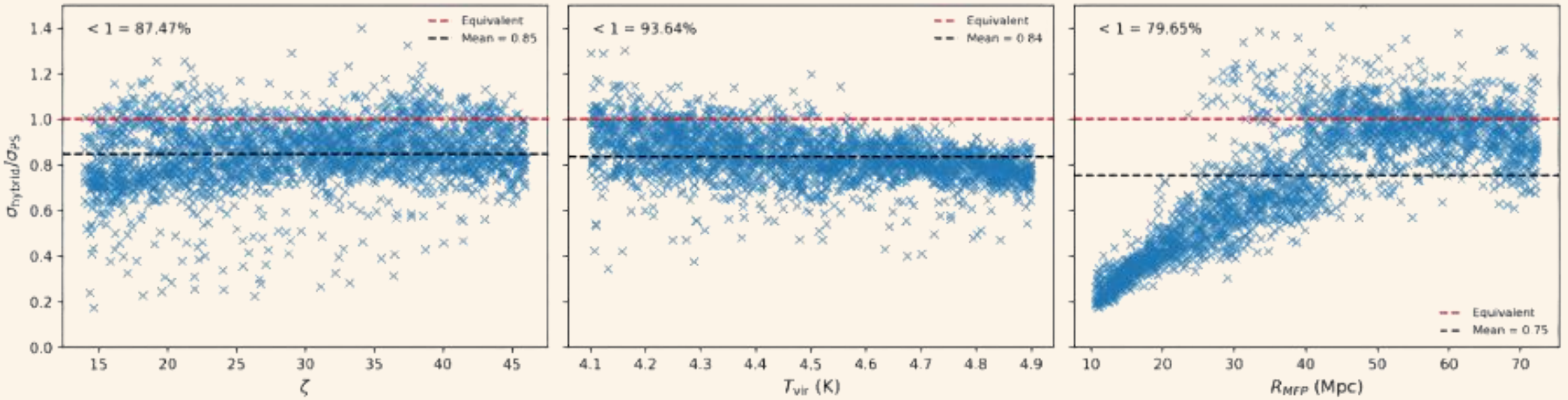
21cmFAST



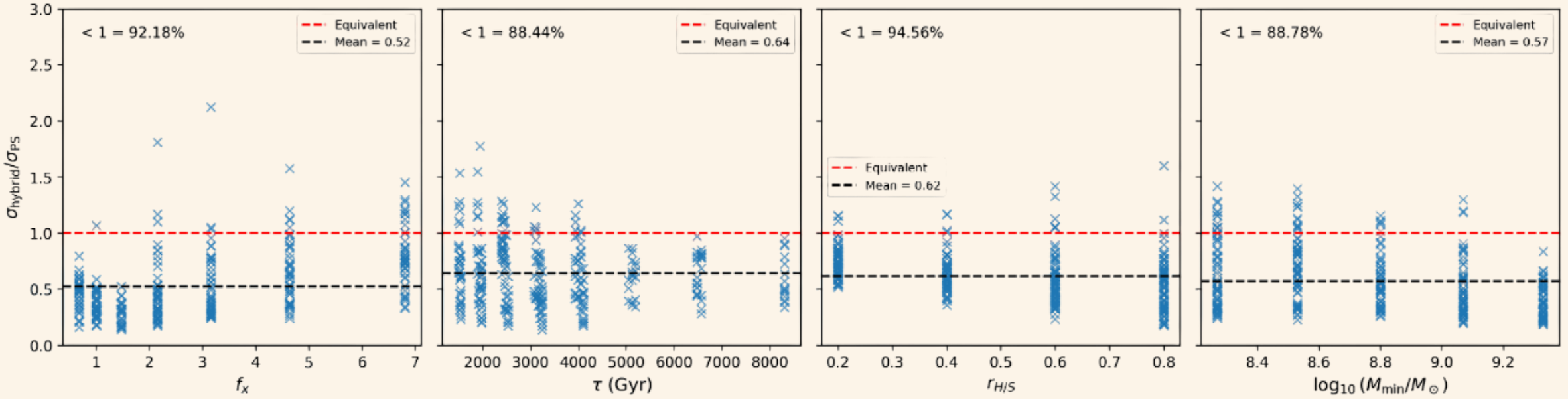
LICORICE



Marginal Improvement with 21cmFAST

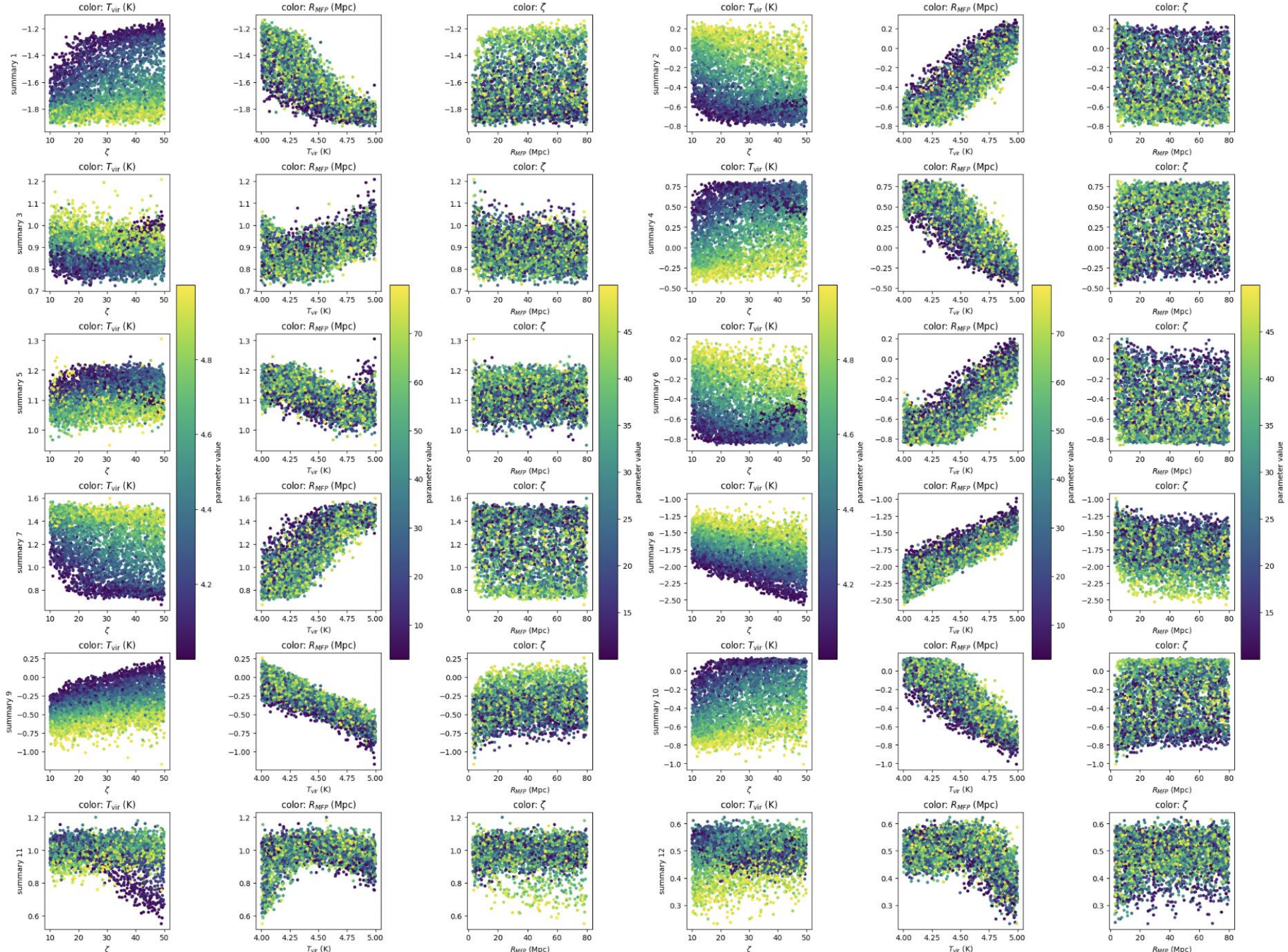


Marginal Improvement with LICORICE



Conclusions

- 21 cm signal contains a wealth of information about the early universe
- Commonly, astrophysical parameters are constrained by power spectra of the field
- This loses all non-gaussian information
- CNN learnt summary statistics are learnt with the objective of finding information not already contained in the power spectra
- Inference is then performed on the concatenated hybrid summary statistics
- Hybrid summary statistics can be used to tighten constraints
- We find hybrid summary statistics on average can reduce posterior volume by ~50%


color: T_{vir} (K)
color: R_{MFP} (Mpc)
color: ζ
summary 1
summary 2
summary 3
summary 4
summary 5
summary 6
summary 7
summary 8
summary 9
summary 10
summary 11
summary 12
ζ
T_{vir} (K)
R_{MFP} (Mpc)
parameter value