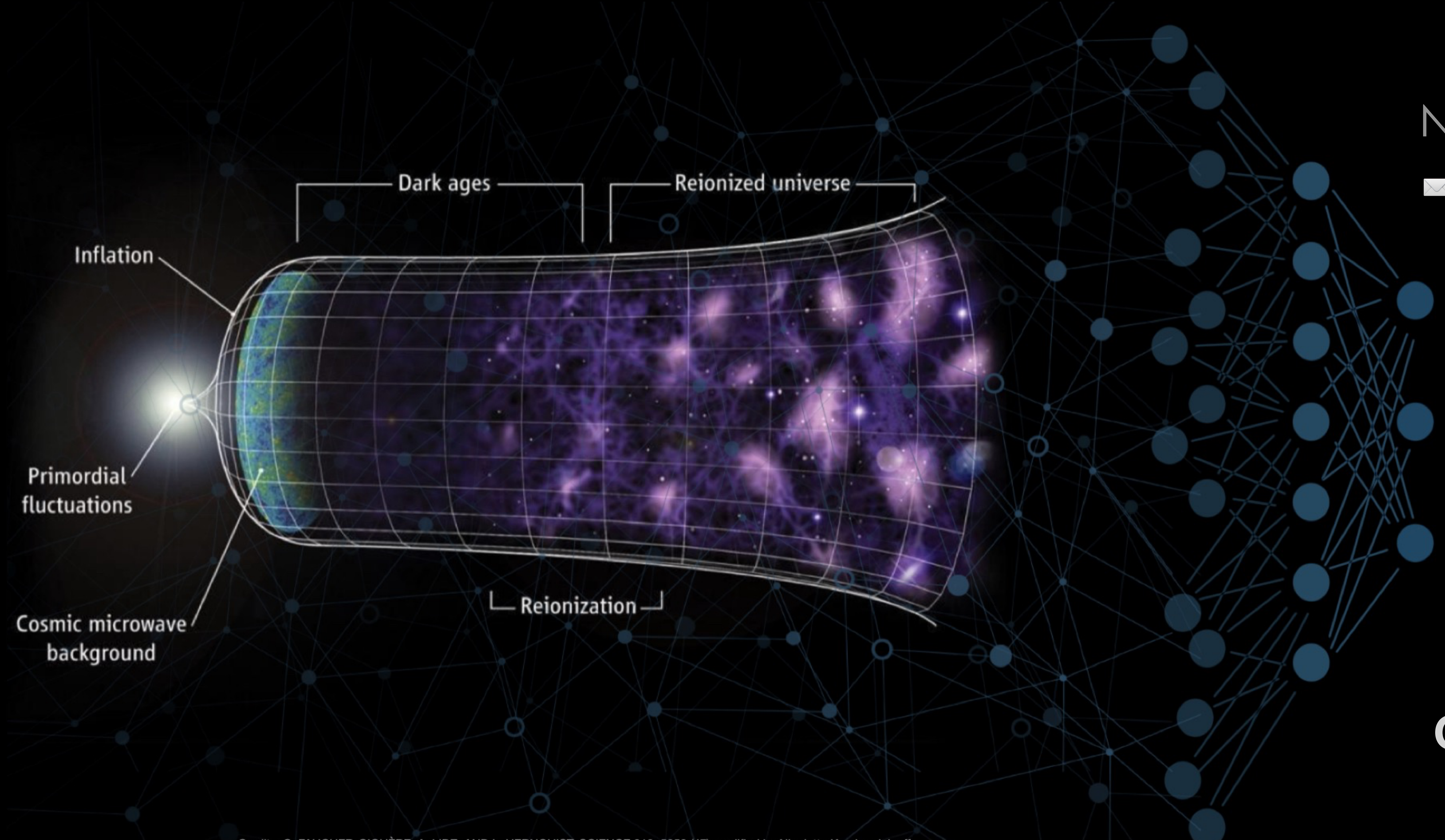
Mining the Universe

Machine Learning in Cosmology



Credits: C. FAUCHER-GIGUÈRE, A. LIDZ, AND L. HERNQUIST, SCIENCE 319, 5859 (47) modified by Nicoletta Krachmalnicoff

Nicoletta Krachmalnicoff
nkrach@sissa.it

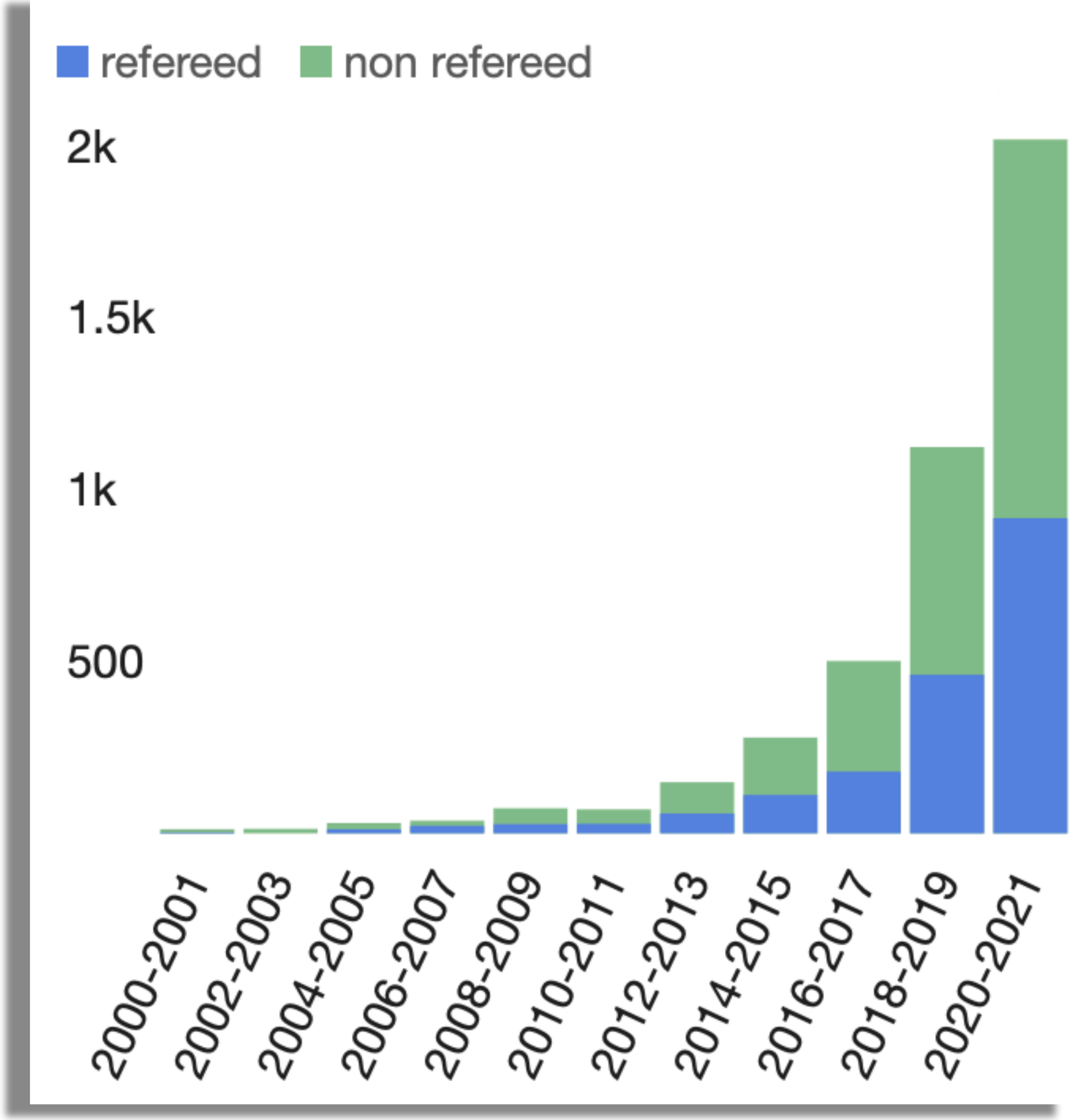
Cosmology in Miramare

August 28th 2023

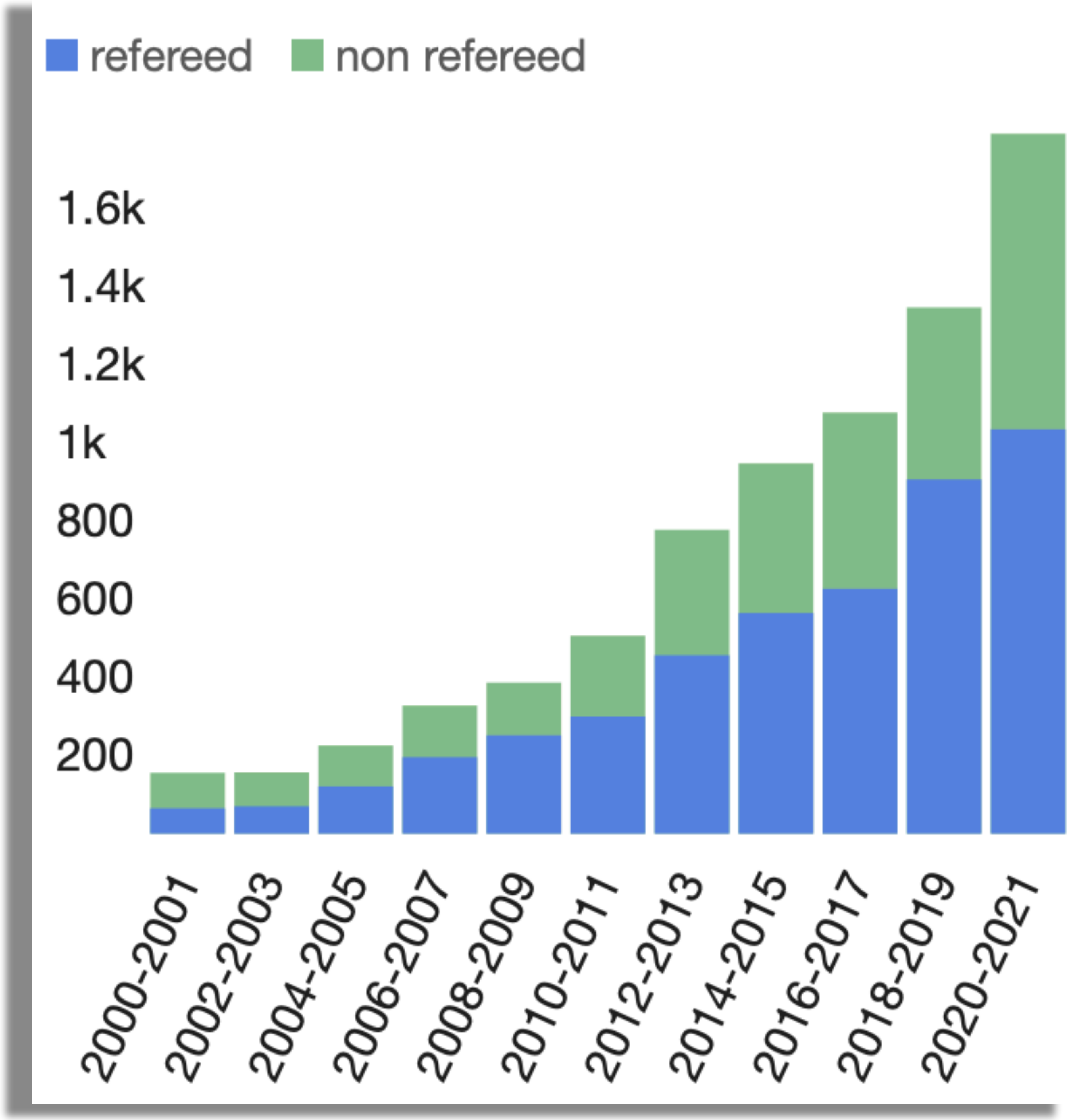
Some numbers

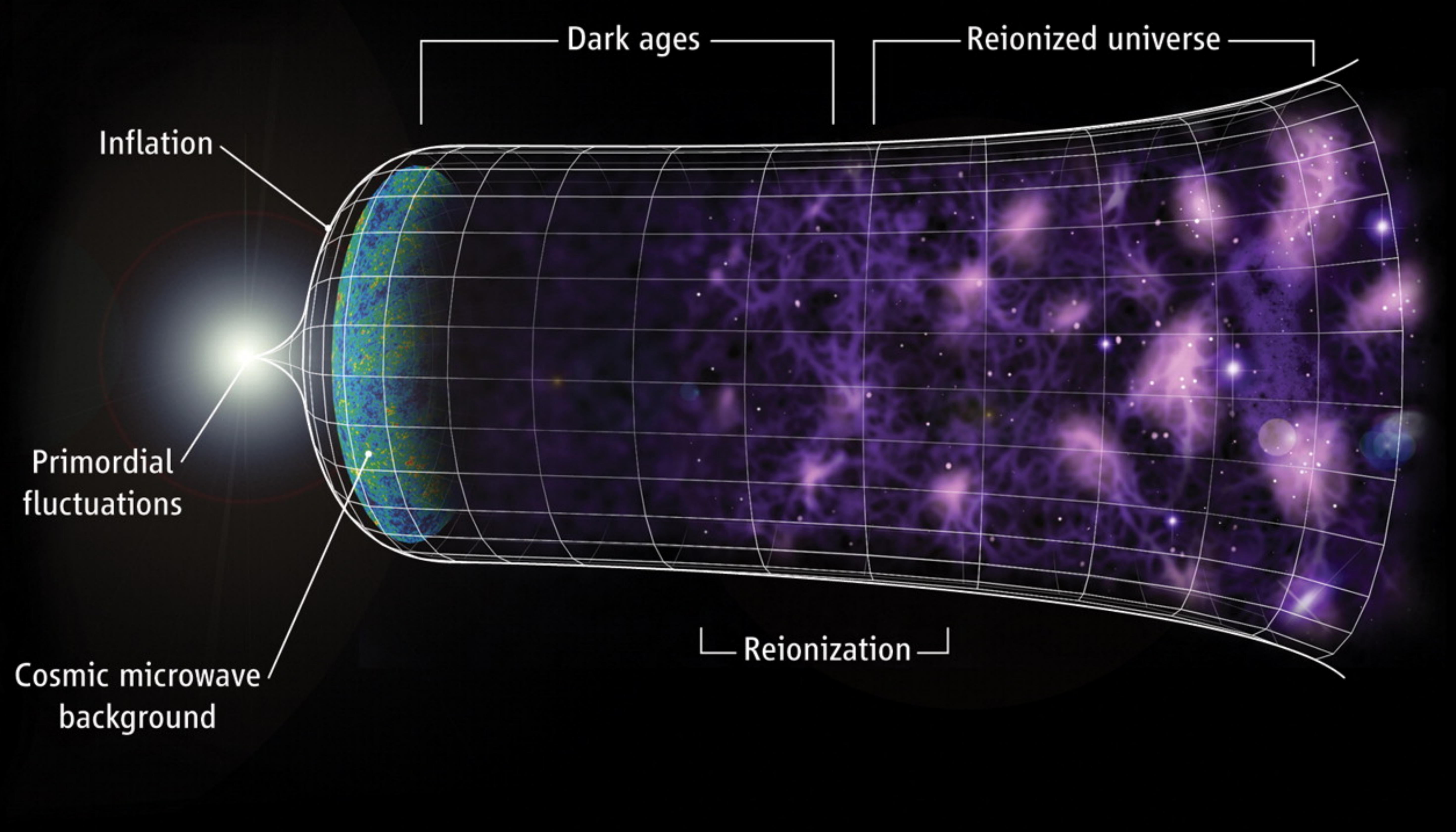
source: NASA/ADS

“Neural Networks” in abstract



“Bayesian” in abstract





Many open questions:

- What is Dark Matter?
- What is the nature of Dark Energy?
- What is the correct theory of Inflation?
- Which are the neutrino masses?
- Tensions
- ….

CMB:

“simple”, almost perfectly Gaussians…but faint and highly contaminated (foregrounds and instrumental systematics)

Large Scale Structure:

Complex signal, involving highly non linear physical process

CMB experiments

Early Universe - **faint signal**

SIMONS OBSERVATORY

CMB-S4
Next Generation CMB Experiment

LiteBIRD

Galaxy surveys

Large Scale Structure - **complex signal**

DARK ENERGY SPECTROSCOPIC INSTRUMENT

VERA C. RUBIN OBSERVATORY

euclid

How to fully exploit data?

Are current methodologies sufficient, given the amount of data, the signal complexity and the precision we want to achieve?

Standard way of analyzing data

Theory and simulations

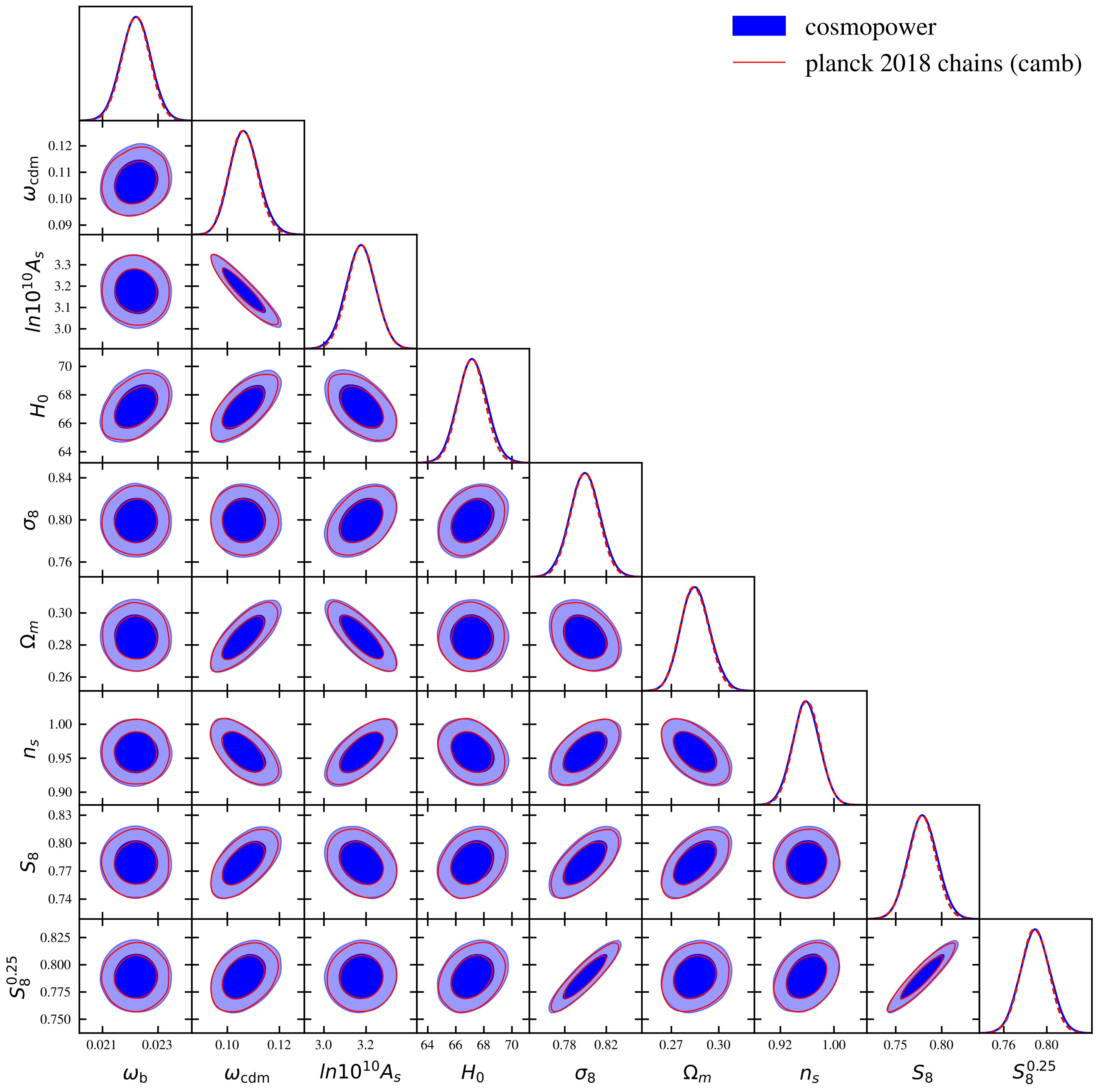
Definition and computation of summary statistics

Definition of the likelihood model and inference

Machine Learning have the **potential** to help in all these steps (except theory) by being more efficient or faster than traditional methods

Definition and **computation of summary statistics**

Emulators



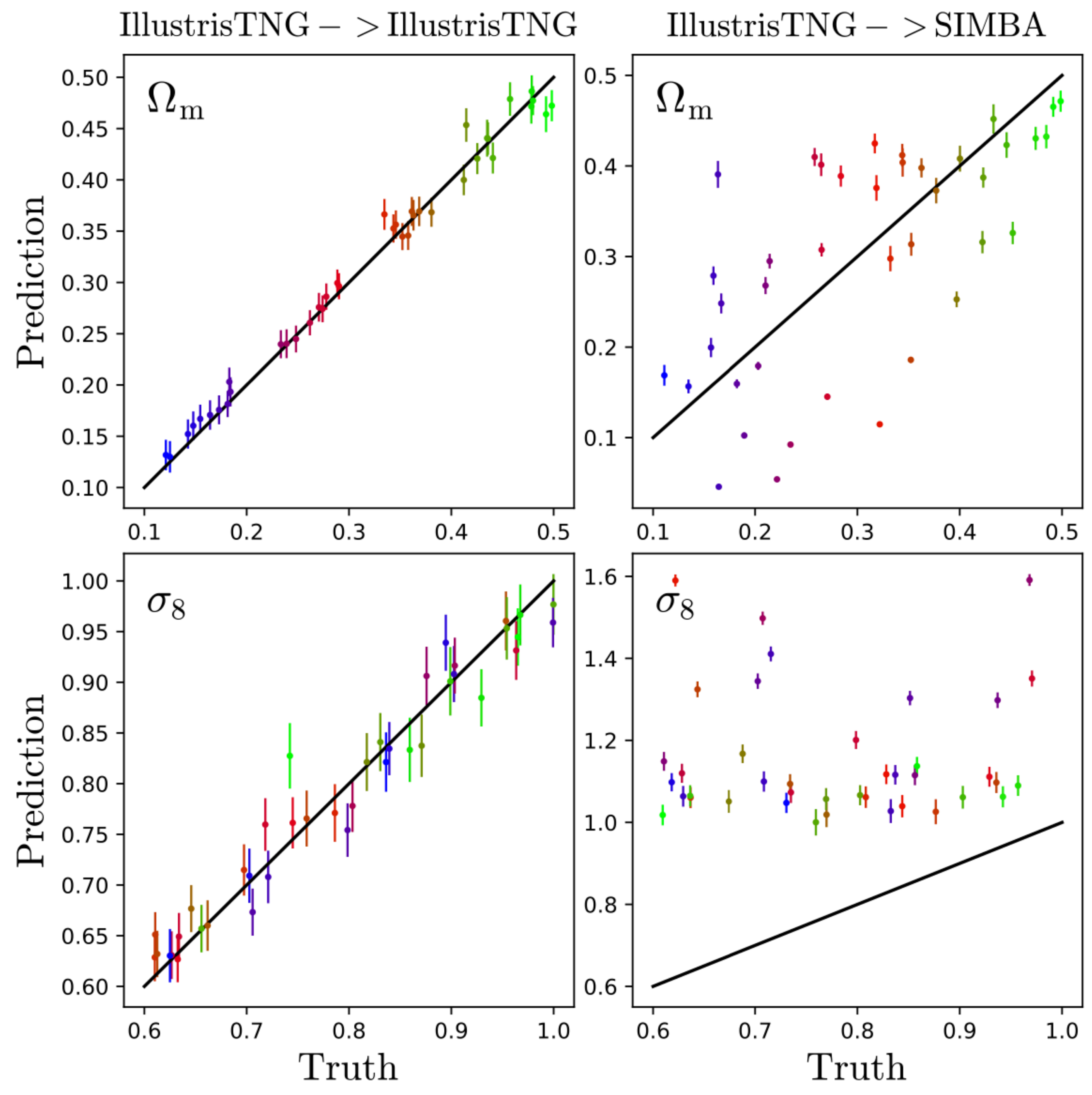
Bolliet, B et al. 2023, https://arxiv.org/abs/2303.01591

- Algorithms that approximate the outputs of computationally expensive models (Einstein-Boltzmann codes, e.g `CAMB`/`CLASS`) at significantly lower computational costs.
- Often based on simple Neural Networks architectures (fully-connected, few layers)
- Efficient way to considerably speed up the sampling of parameter's space in standard MCMC inference
- Already proved to be a valuable tool also on analysis of real data.

Definition of the likelihood model and **inference**

ML-based/likelihood free inference

- No need to define and compute summary statistics from the data, in principle no loss of information
- No need of an analytical likelihood model, trained only on simulations
- Potentially powerful for both LSS (complex non-Gaussian signal) and CMB (Gaussian signal, but highly contaminated by non-Gaussian foregrounds and instrumental systematic effects)
- Many different implementations, mostly applied, tested and validated on simulations
- Application to real data still lacking!



Villaescusa-Navarro, F et al. https://arxiv.org/abs/2109.09747

Inference of the optical depth to reionization τ from *Planck* CMB maps with convolutional neural networks

Kevin Wolz, Nicoletta Krachmalnicoff, Luca Pagano

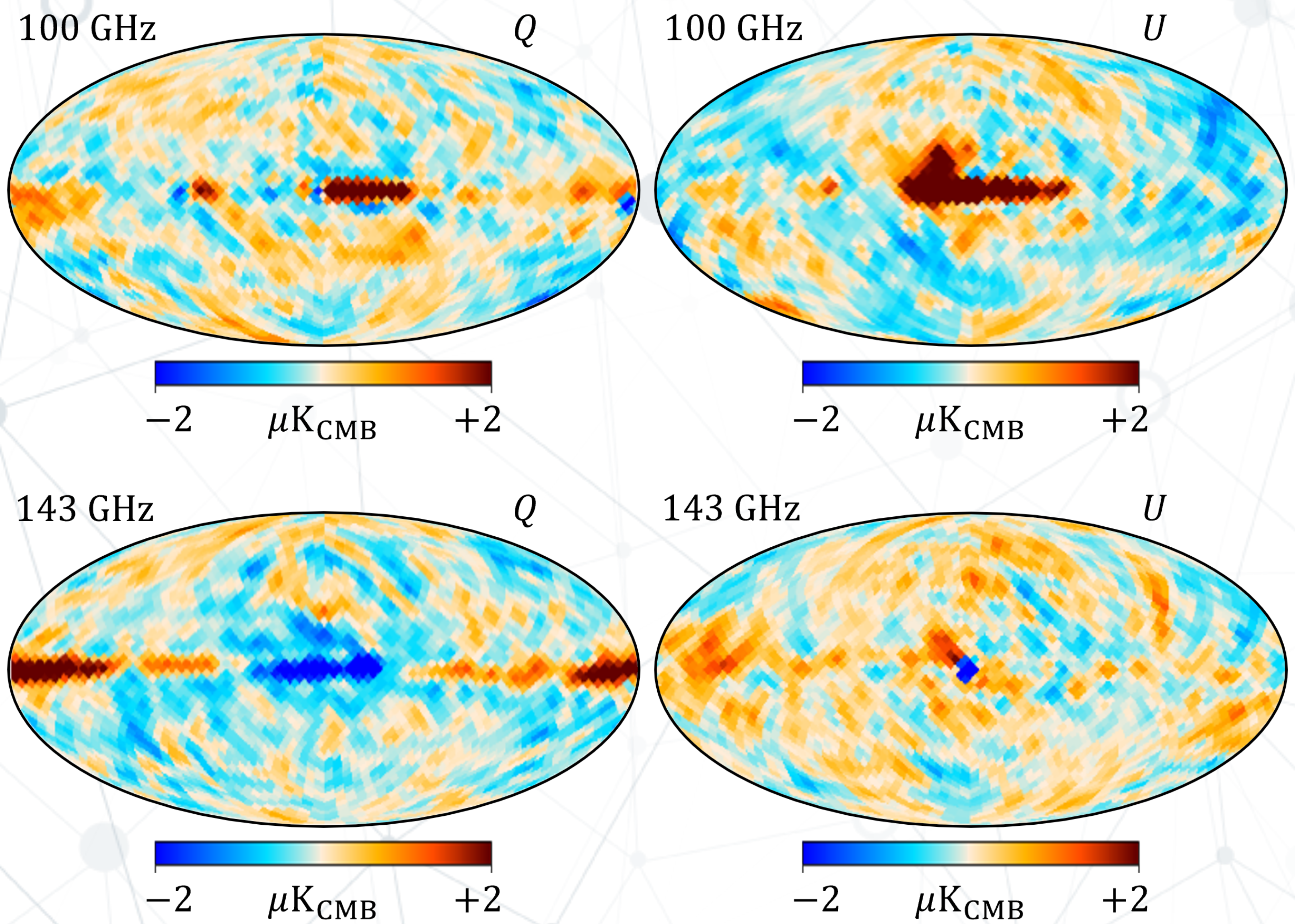
https://arxiv.org/abs/2301.09634

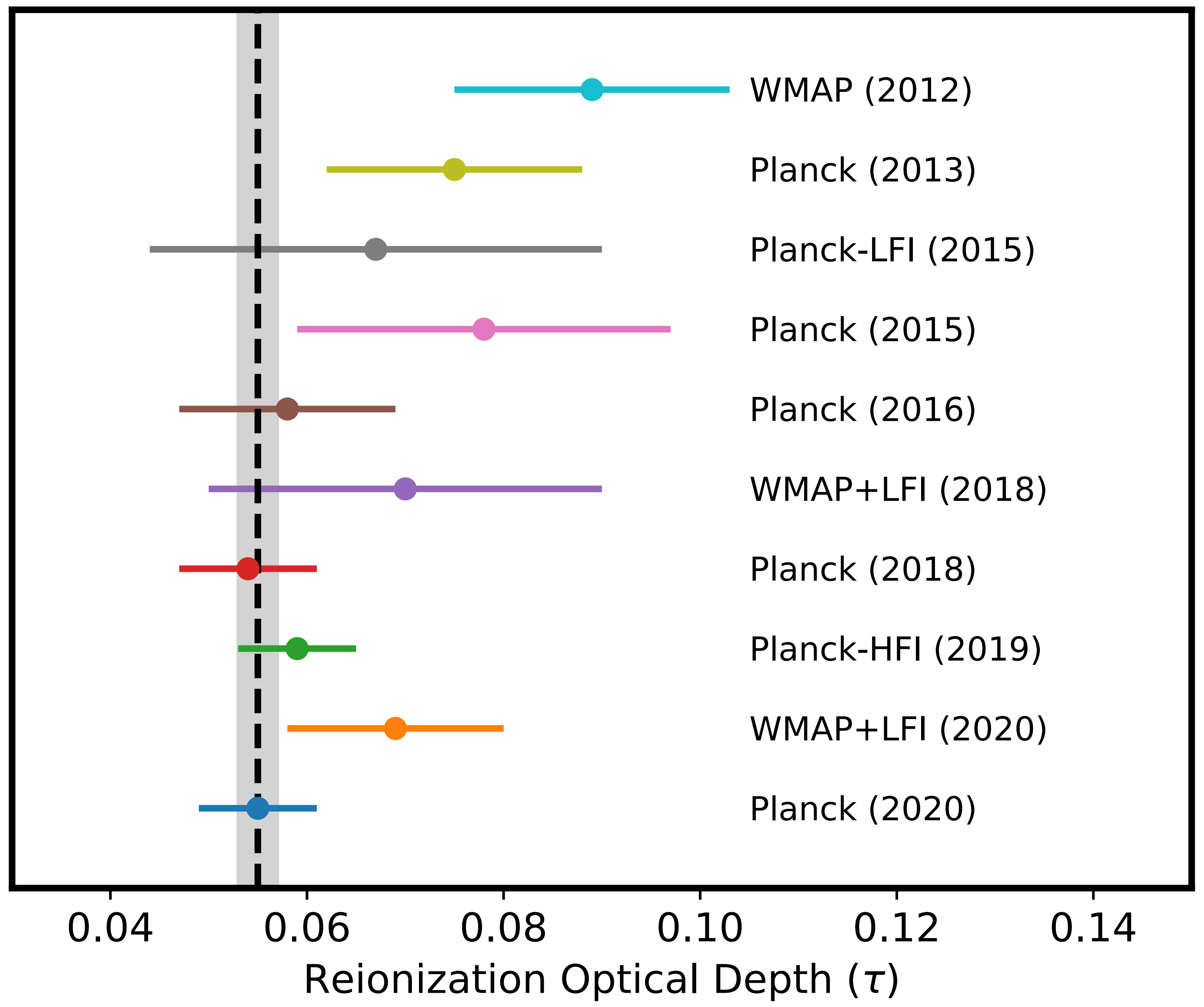
Motivations:

- One of the first work in CMB field that robustly applies likelihood-free inference fully based on Convolutional NNs to real, non ideal, data!
- Tested the applicability of this approach on the estimation of the optical depth to reionization, τ
- τ impacts the very large angular scales of CMB E-modes, largely affected by instrumental systematics and Foreground residuals
- First instructive test before applying the method to primordial B-modes for future experiments

Planck maps and τ estimates

- Planck maps at 100 and 143 GHz are known to contain significant level of residual systematic effects (mainly due to T-to-P leakage) at large angular scales
- Although mitigated by the optimization of the map making procedure (e.g. Sroll2 maps), they cannot be considered negligible

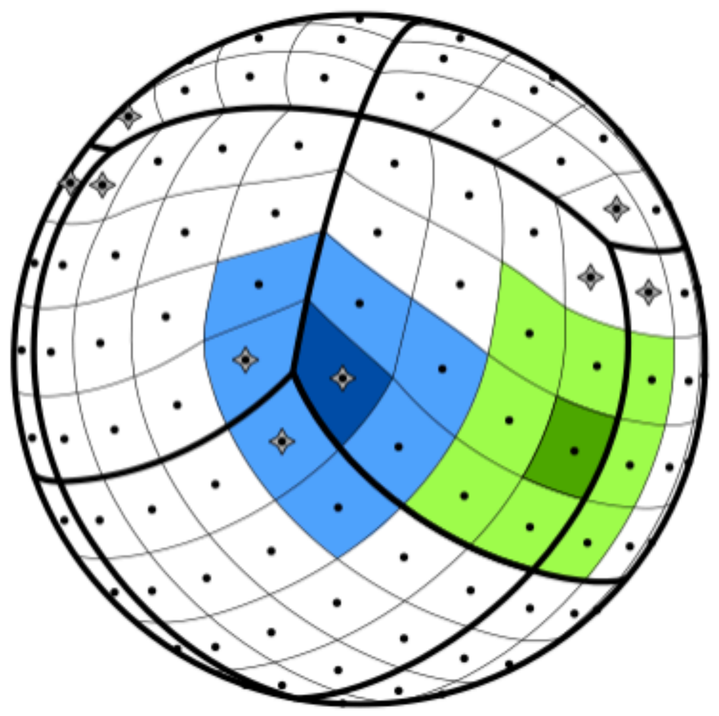


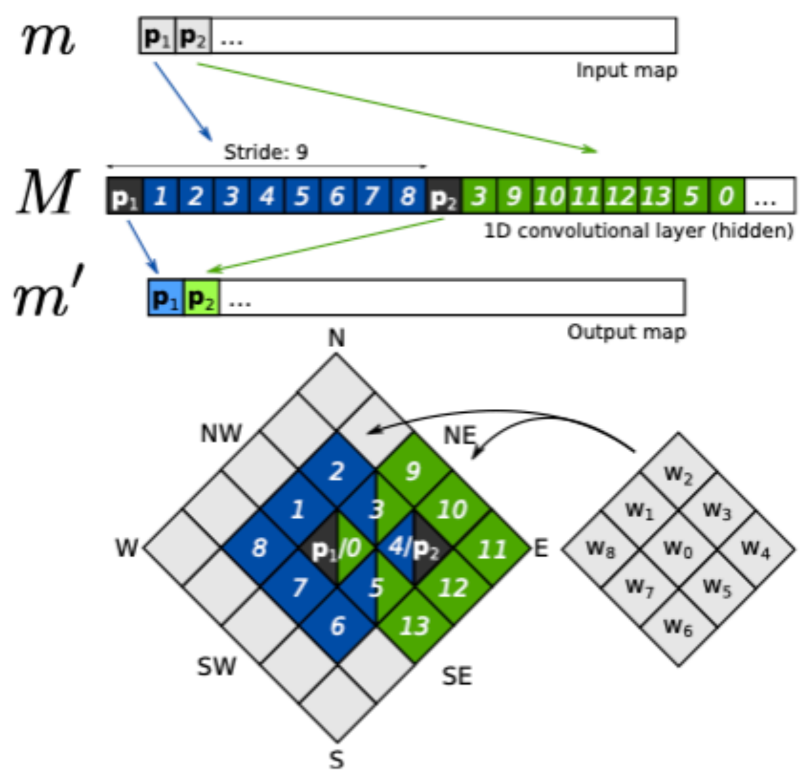


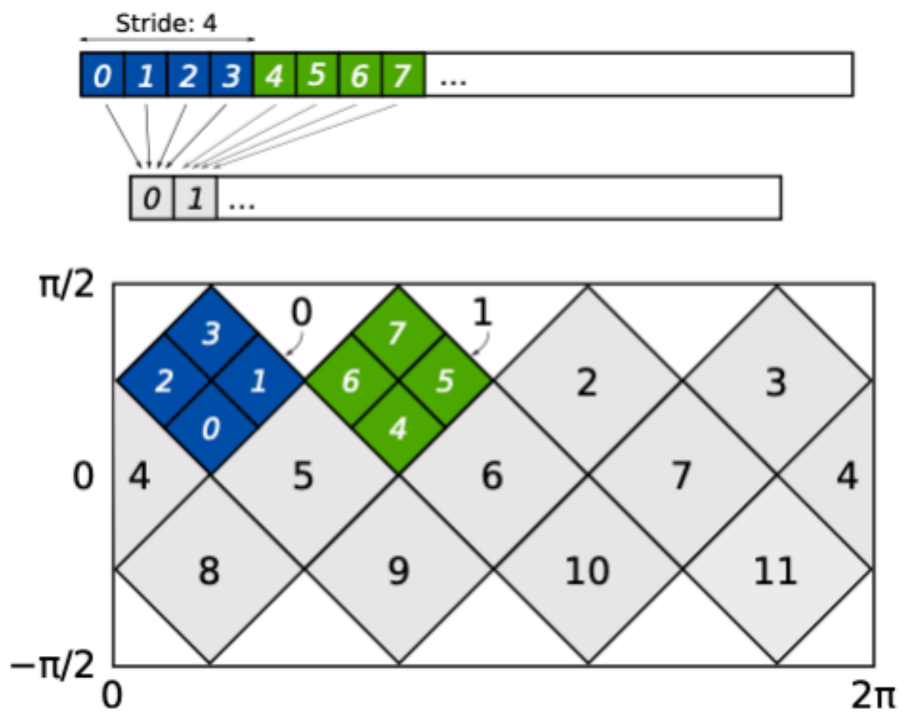
- These **residual non-Gaussian signals are hard to be analytically modeled**
- Current constraints on τ are obtained from an empirical likelihood based on cross-spectra

NN approach

- Estimate τ with convolutional NN directly from maps (without computation of power spectra), combining information from multiple channels
- No need of a likelihood model, but only large set of simulations to train the NN
- **Two types of simulations:**
  - **CMB + Gaussian correlated noise (from Planck covariances)**
  - **CMB + Gaussian + Systematics (limited to 500 realizations!)**
- Convolution of the sphere, using the NNhealpix algorithm, since we are dealing with super-degree angular scales

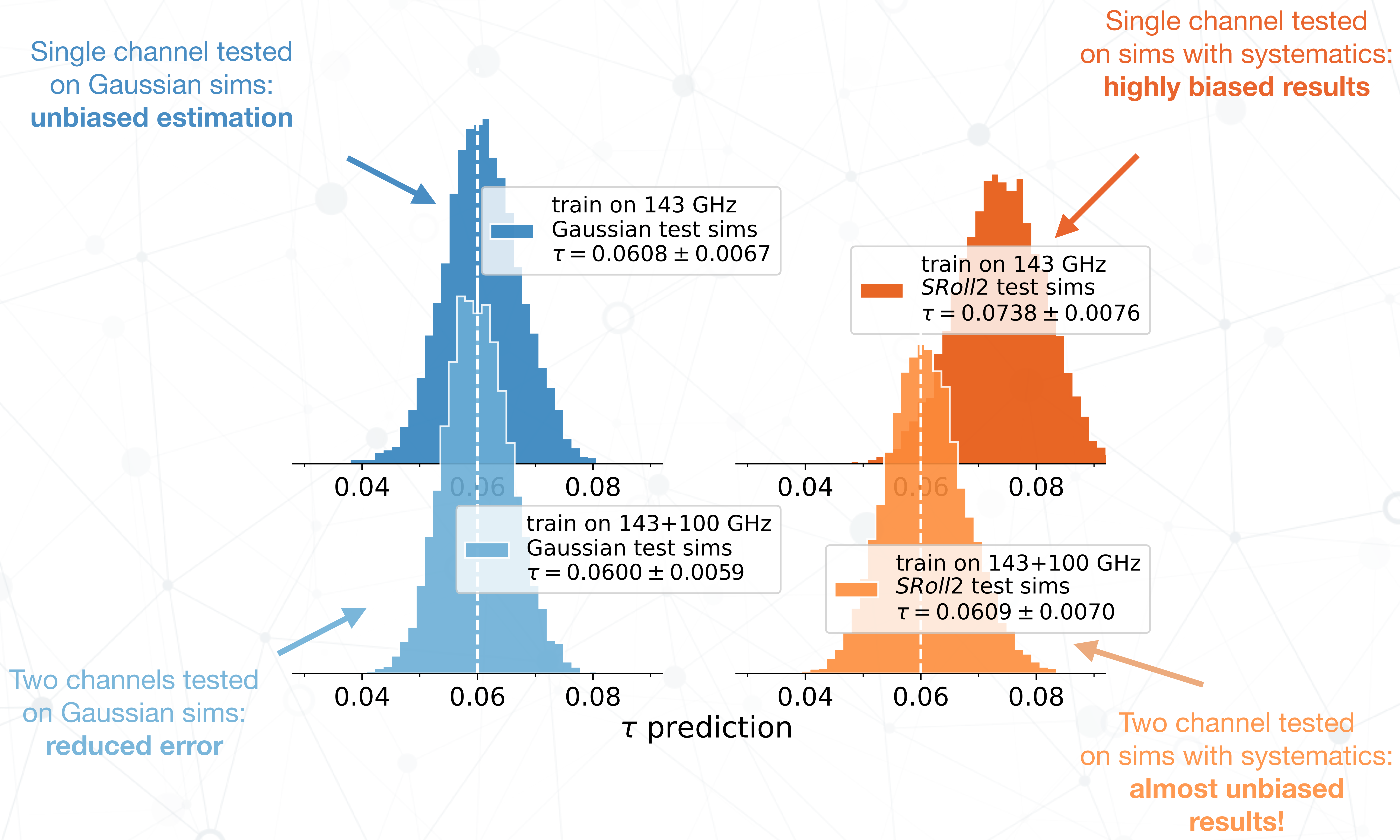




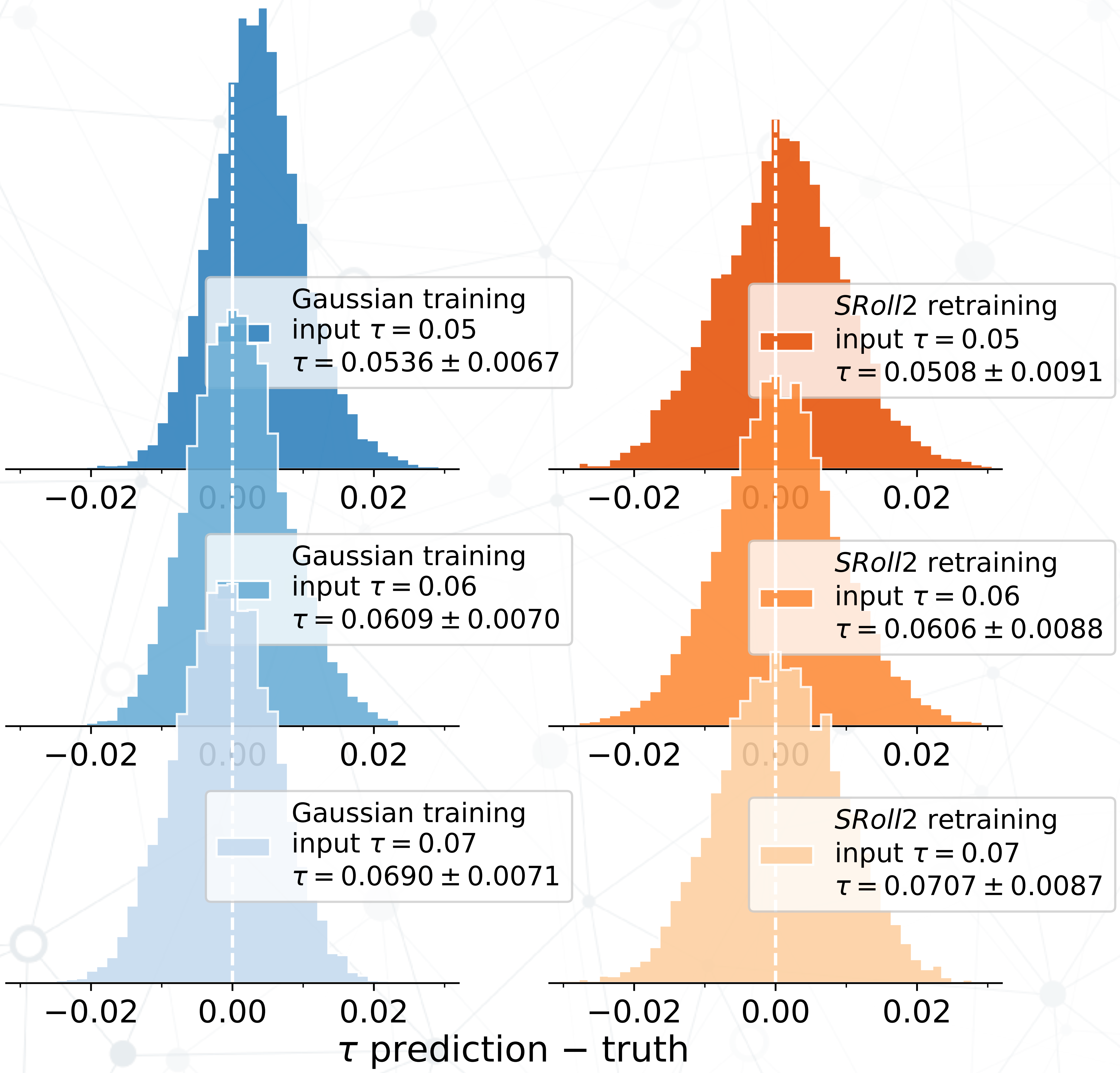


Results on simulations

- NN trained on Gaussian simulations (CMB + Gaussian correlated noised)
- Having one channel (100 GHz) or two channels input (100 and 143 GHz)

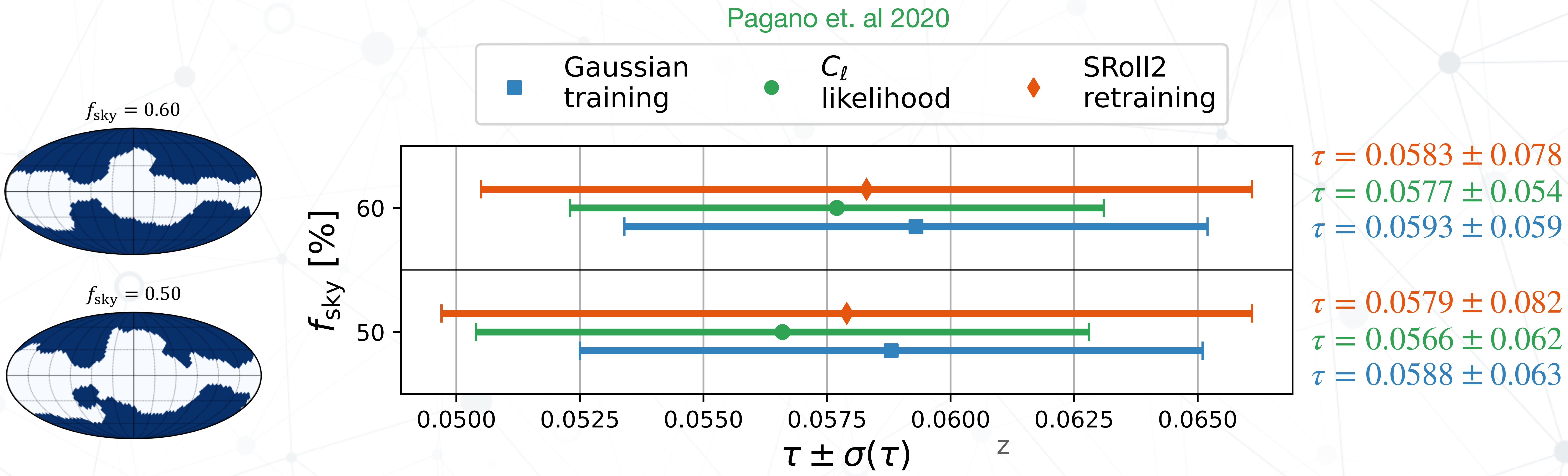


Results on simulations



- Arriving to fully unbiased results on maps that include systematic effects requires to include those systematics in the training procedure
- Limited by the number of realizations (only 500)
- *Minimal retraining* procedure:
  - ➡ Starting with Gaussian NN we use 400 realization of systematics to update the NN weight
  - ➡ Update must be large enough to arrive to unbiased results but not too big to destroy what already learnt

Results on Planck data

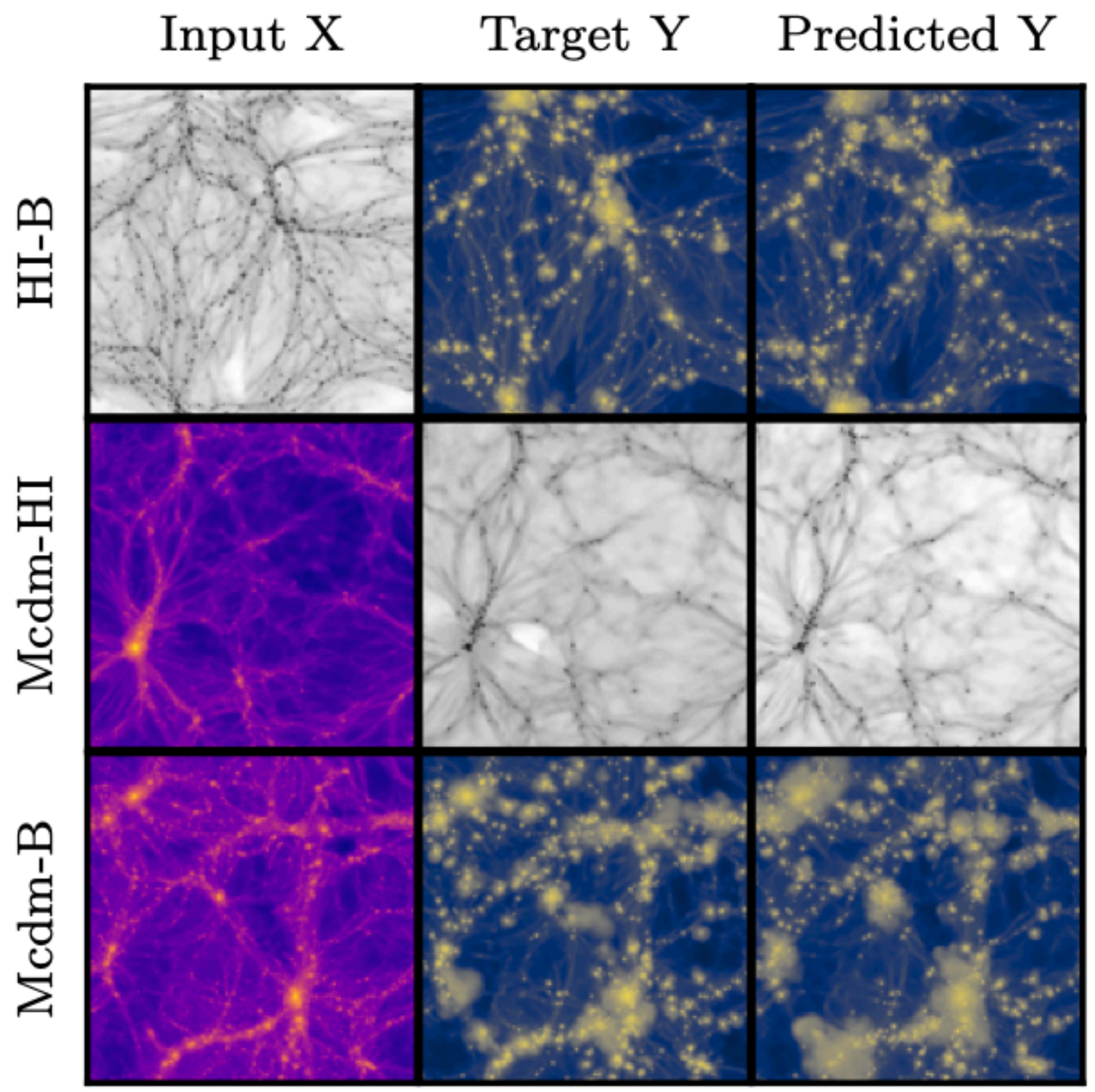


- High level of agreement with cross-spectrum likelihood value, but with ~30% larger errorbars
- Optimization of NN architecture and procedure needed to improve
- Combination with other dataset is possible (no need of a common data model, just simulations)
- **Application of NN to real data is challenging!!!!!!!**

Theory and **simulations**

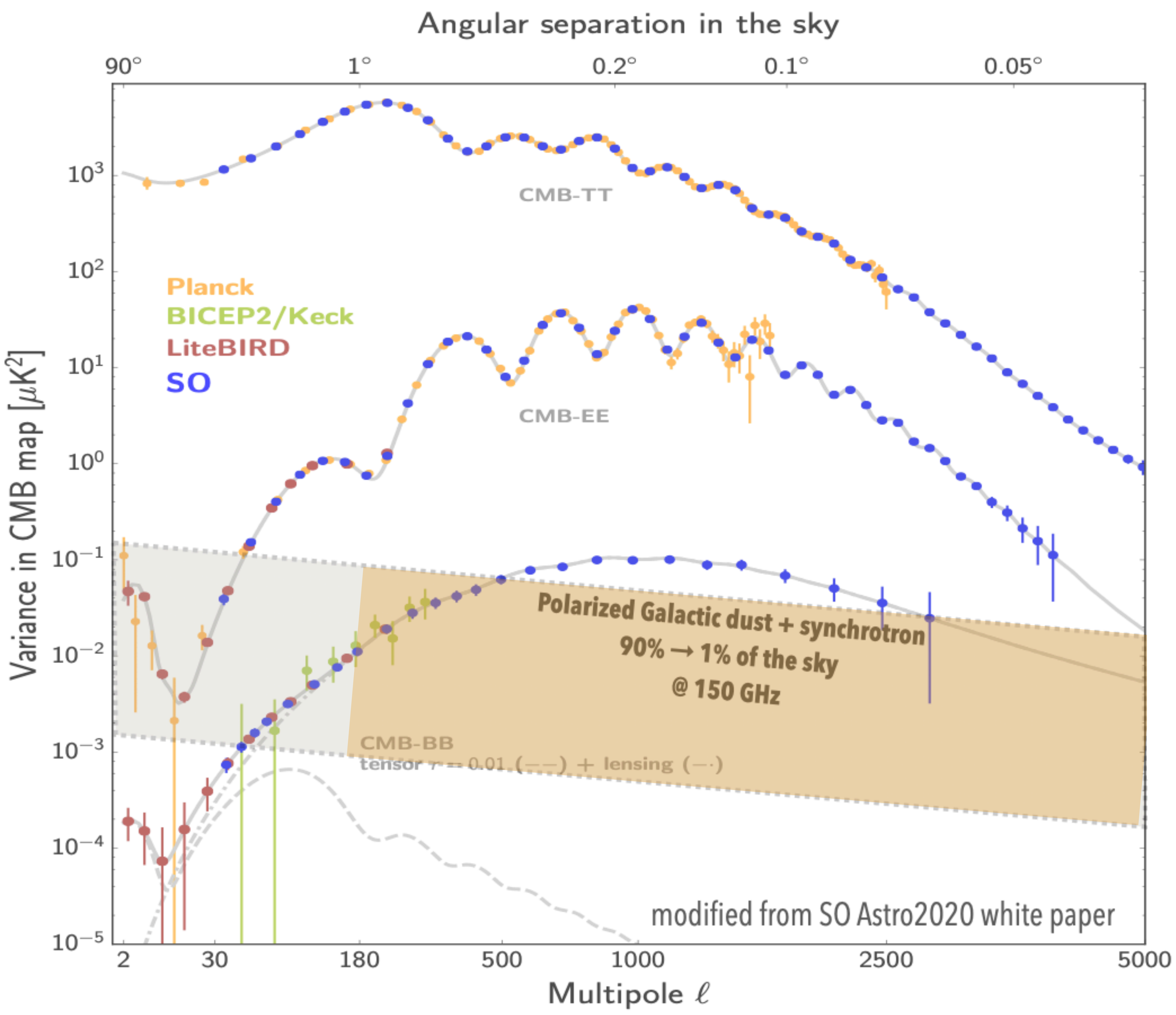
Simulations with NNs

- Inference in Cosmology relies on the existence of large number of simulations
- On going effort in trying to take advantage of ML to generate simulations more efficiently, enhance exiting ones while learning physical properties/correlations.



Andrianomena, S. et al. https://arxiv.org/abs/2303.07473

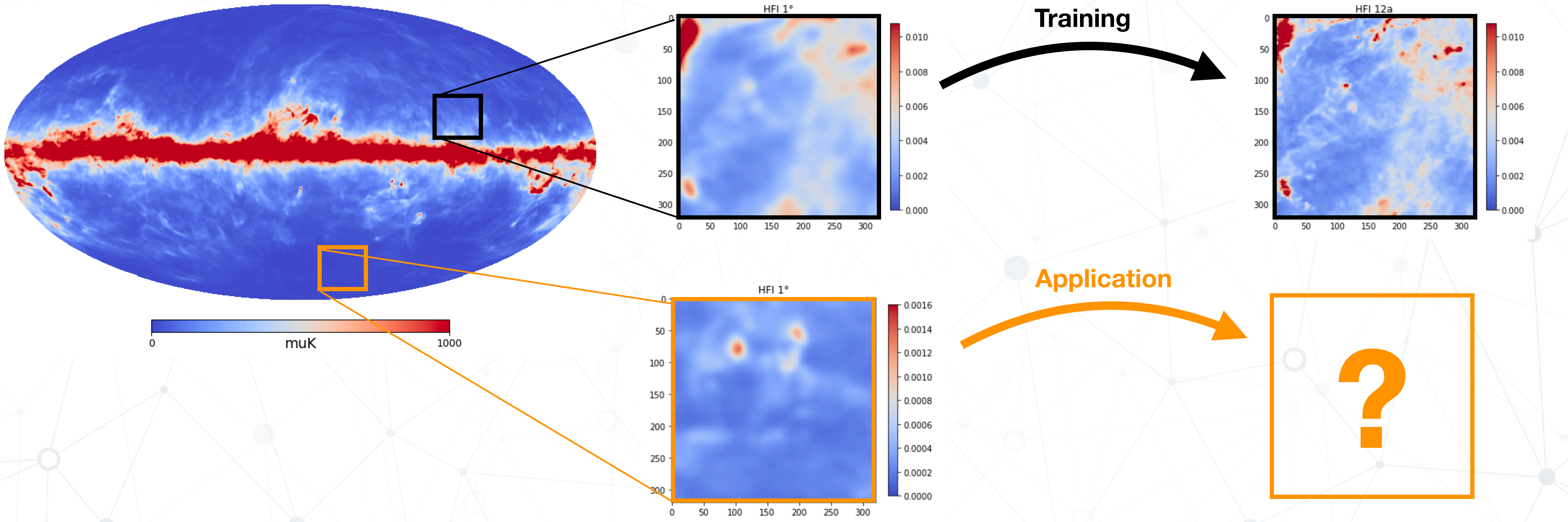
CMB observations and foregrounds



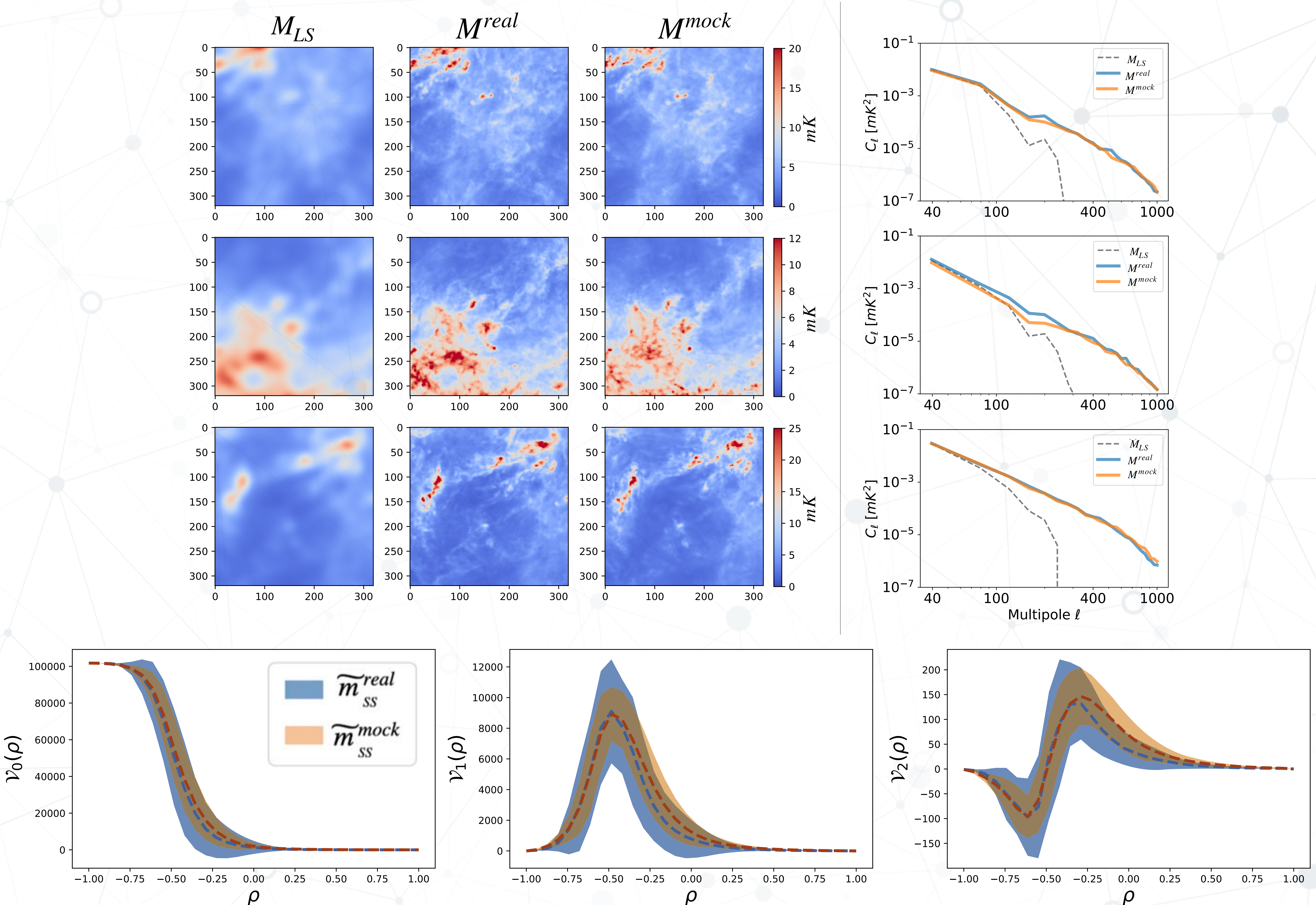
- Galactic foregrounds are the main contaminant to CMB observations in polarization
- we have **FG data only at angular scale > 1°**
- Important to understand the **impact** of Non-Gaussian sub-degree foreground emission **on lensing reconstruction, de-lensing.**

GANs to simulate small scale foregrounds

i. Train **Neural Networks to learn the statistics of foregrounds at the sub-degree scale** in total intensity (in the regions where we have enough sensitivity)



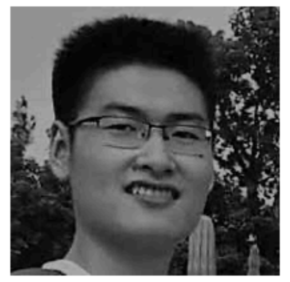
ii. Reproduce the same statistics starting from large scales in other regions of the sky and in polarization



M_{LS}
M^{real}
M^{mock}
mK
$C_\ell\ [mK^2]$
Multipole ℓ
M_{LS}
M^{real}
M^{mock}
$\widetilde{m}_{ss}^{real}$
$\widetilde{m}_{ss}^{mock}$
$\mathcal{V}_0(\rho)$
$\mathcal{V}_1(\rho)$
$\mathcal{V}_2(\rho)$
ρ

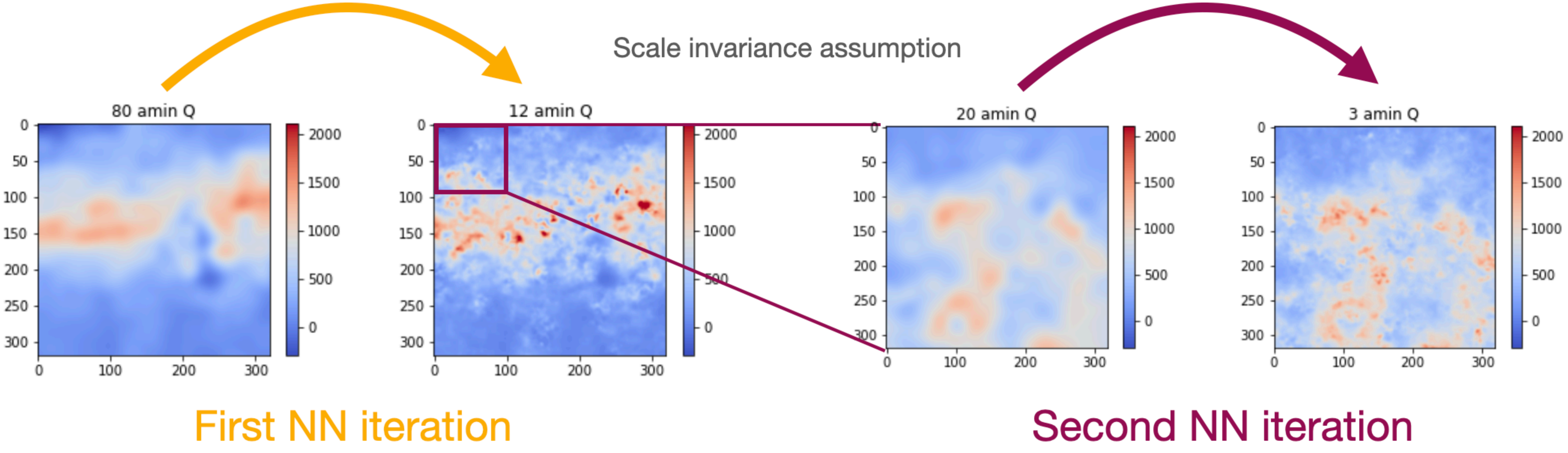
GANs to simulate small scale foregrounds

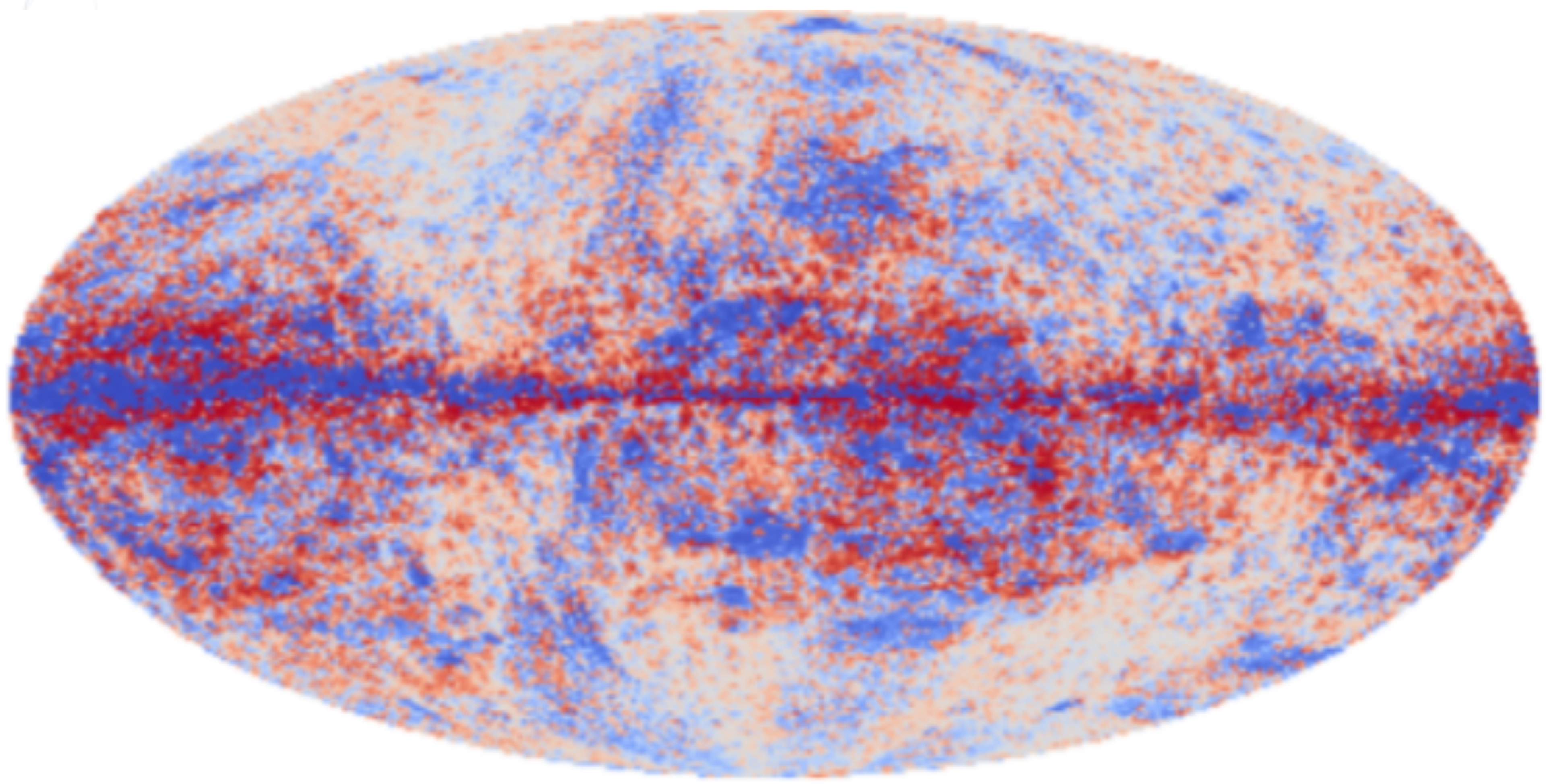
Jian Yao



Giuseppe Puglisi

Marianna Foschi





- Polarization full sky map with stochastic non-Gaussian small scales up to 3 arcmin

Concluding remarks

- ML offers diverse applications in cosmology, with the potential of enhancing data analysis efficiency for upcoming experiments .
- The field is currently in an exploratory phase. Feasibility tests are ongoing, but real-world application on data is limited.
- Progress from simulation success to reliable data outcomes is challenging.
- Complementary tool, not yet revolutionary