LERMA l'Observatoire de Paris | PSL

SORBONNE UNIVERSITÉ

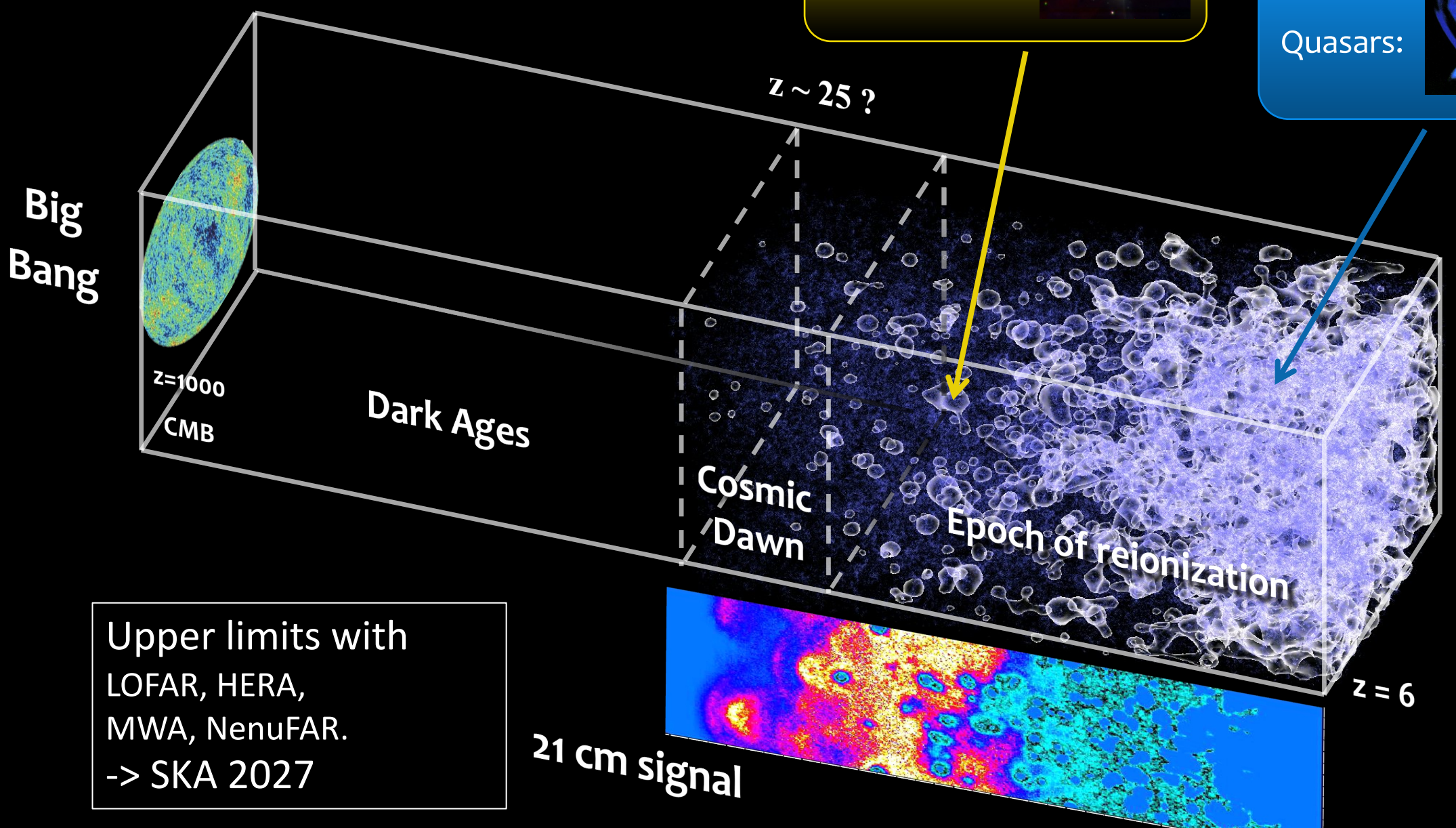
cnrs

Simulation-based inference with radiative hydrodynamics simulations for SKA

B. Semelin, R. Mériot and D. Cornu

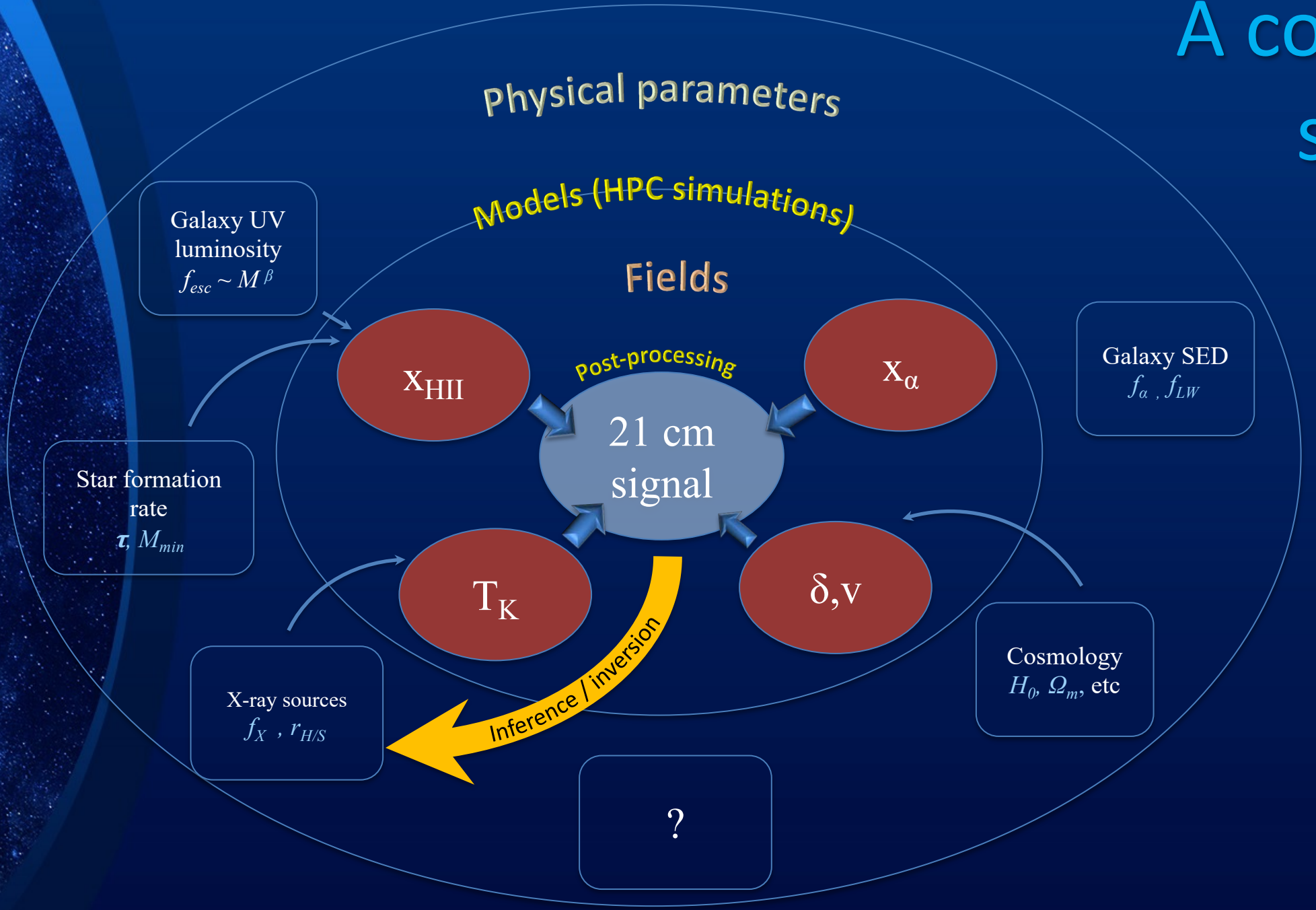
AI for HPC@Exascale, Oct 2024

The first billion years



Upper limits with
LOFAR, HERA,
MWA, NenuFAR.
-> SKA 2027

A complex system



Forward model: The LICORICE code

- Dynamics (Treecode + SPH)
- Monte Carlo Ray-Tracing RT: UV and X
- Lyman-alpha 3D RT
- MPI+OpenMP parallelization

HIRRAH-21 simulation (2018):

- 300 Mpc box
- 10^{10} particles, > 10^9 $M_\odot$, resolution ~ 3 kpc
- 4 x 10^{12} photons
- 5 Mh CPU. 4096 MPI domains, 16384 core.

How to perform inversion?

Bayesian Inference with 3D RT simulations?

Possible inference methods:

1) Bayesian MCMC: > 10^5 model evaluations => Not with 3D RT!
2) Model inversion with ML (bayesian or not): a few 10^3 models
3) Trained ML emulator (a few 10^3 models) + MCMC
4) SBI: ML Density Estimator (a few 10^3 models AND implicit likelihood)

=> We need to build a training set with 3D RT simulations!

LoReLi II database

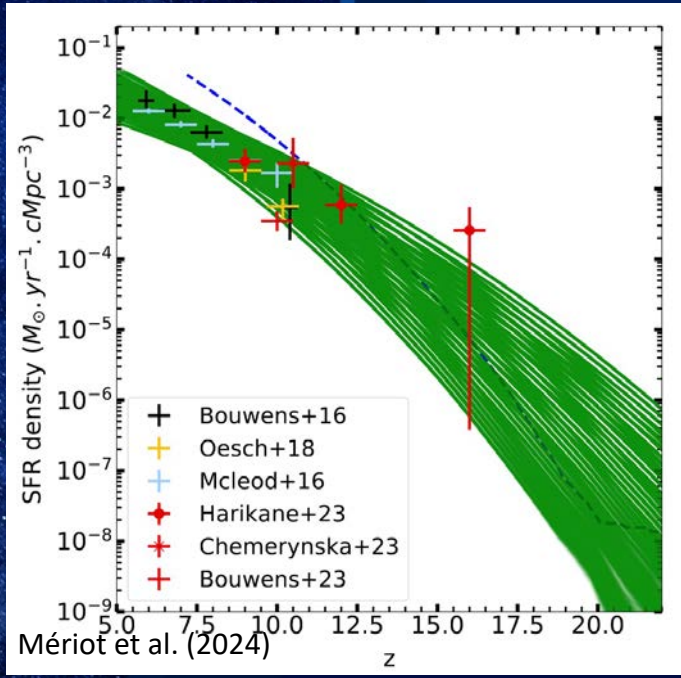
(Mériot, Semelin and Cornu, in prep, 2024)

10 000 simulations (1.5 Po, 5 Mh CPU)

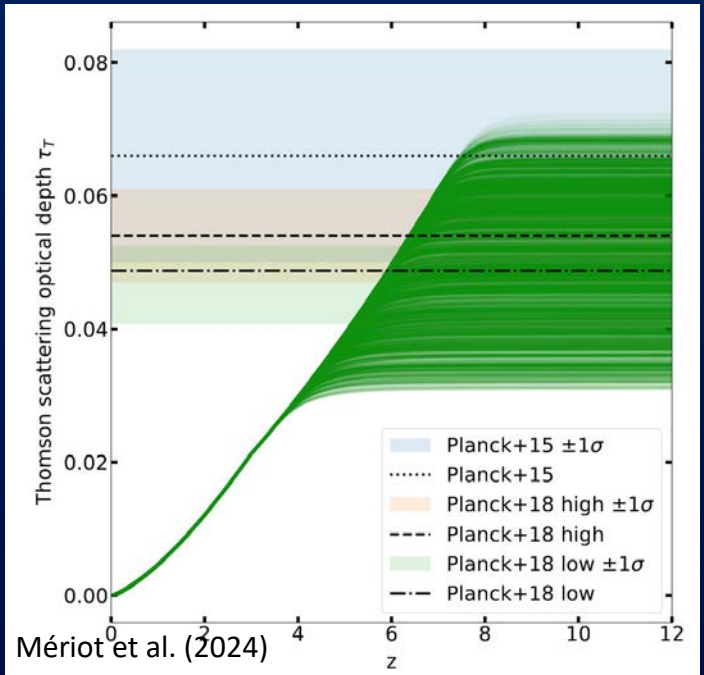
500 000 21-cm cubes (80 To)

Explore a 5-param space: f_X, M_{min}, τ_{SF}, f_{esc} , and $R_{H/S}$

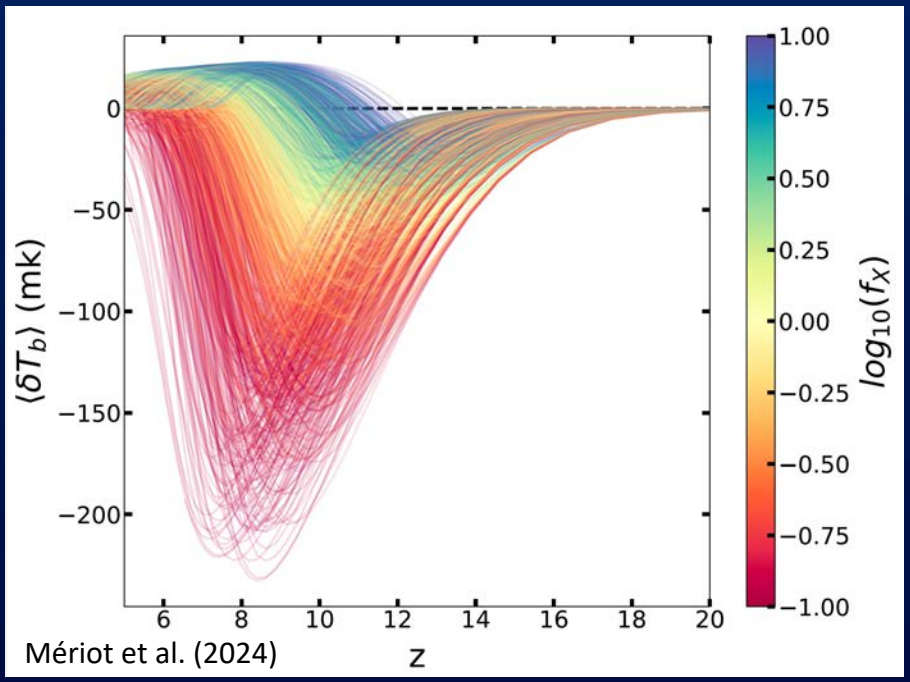
Non-hypercubic domain (prior) to account for observational constraints



Mériot et al. (2024)



Mériot et al. (2024)



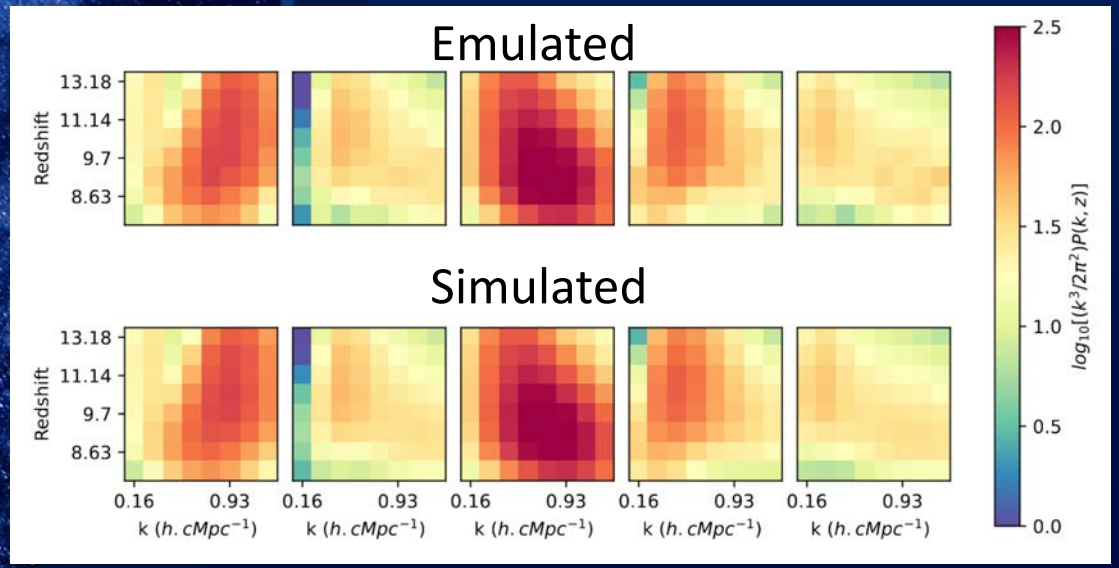
Mériot et al. (2024)

Emulator based inference

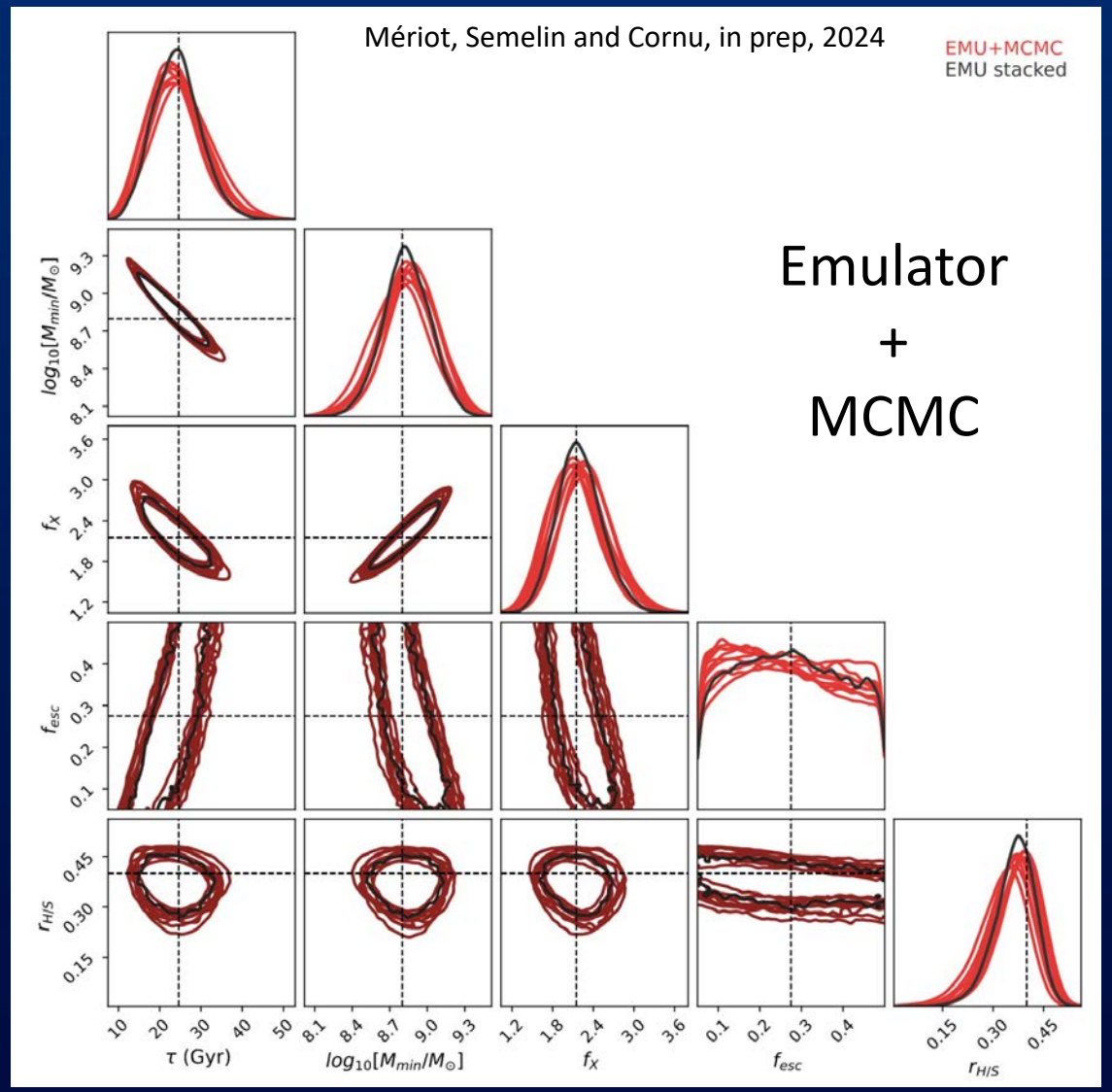
Train NN with Loreli2:

Astro parameters → Neural Network → Powerspectrum P(k,z)

Percent level accuracy:



Mériot, Semelin and Cornu, in prep, 2024



Principle of Simulation-based inference (SBI)

Emulator based inference:

- Assume uncorrelated noise with known variance and mean
=> $P(\ data \mid parameters\ )$ = analytical gaussian

But correlations exist and 21-cm signal is non gaussian

If we can simulate the signal and the noise (yes we can!)

=> One simulation = one draw from $P(\ data \mid parameters\ )$

=> Train NN to fit P (or the posterior) from a collection of draws (LoReLi II)

=> Use trained NN for very fast MCMC inference

SBI with Loreli II

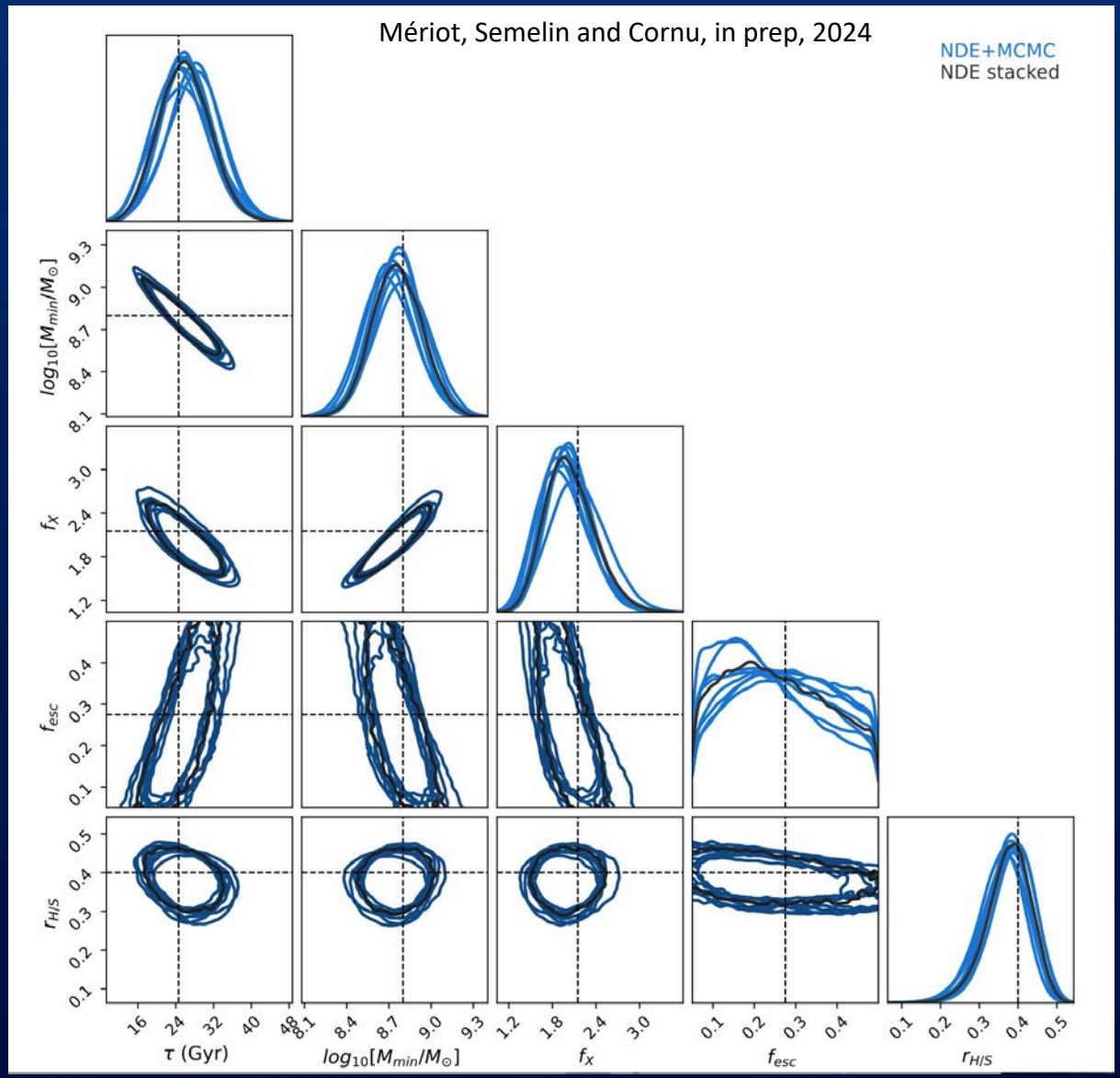
- Assume 100h SKA thermal noise
- Noiseless mock target signal
  => true params +/- centered

- Train 10 NN to evaluate stability
- Also infer with stacked NNs

Control validity with SBC (~1000 inferences!):

=> Bias < 0.2 σ
=> underconfident by ~20%

… a promising approach!



Maximizing information with SBI

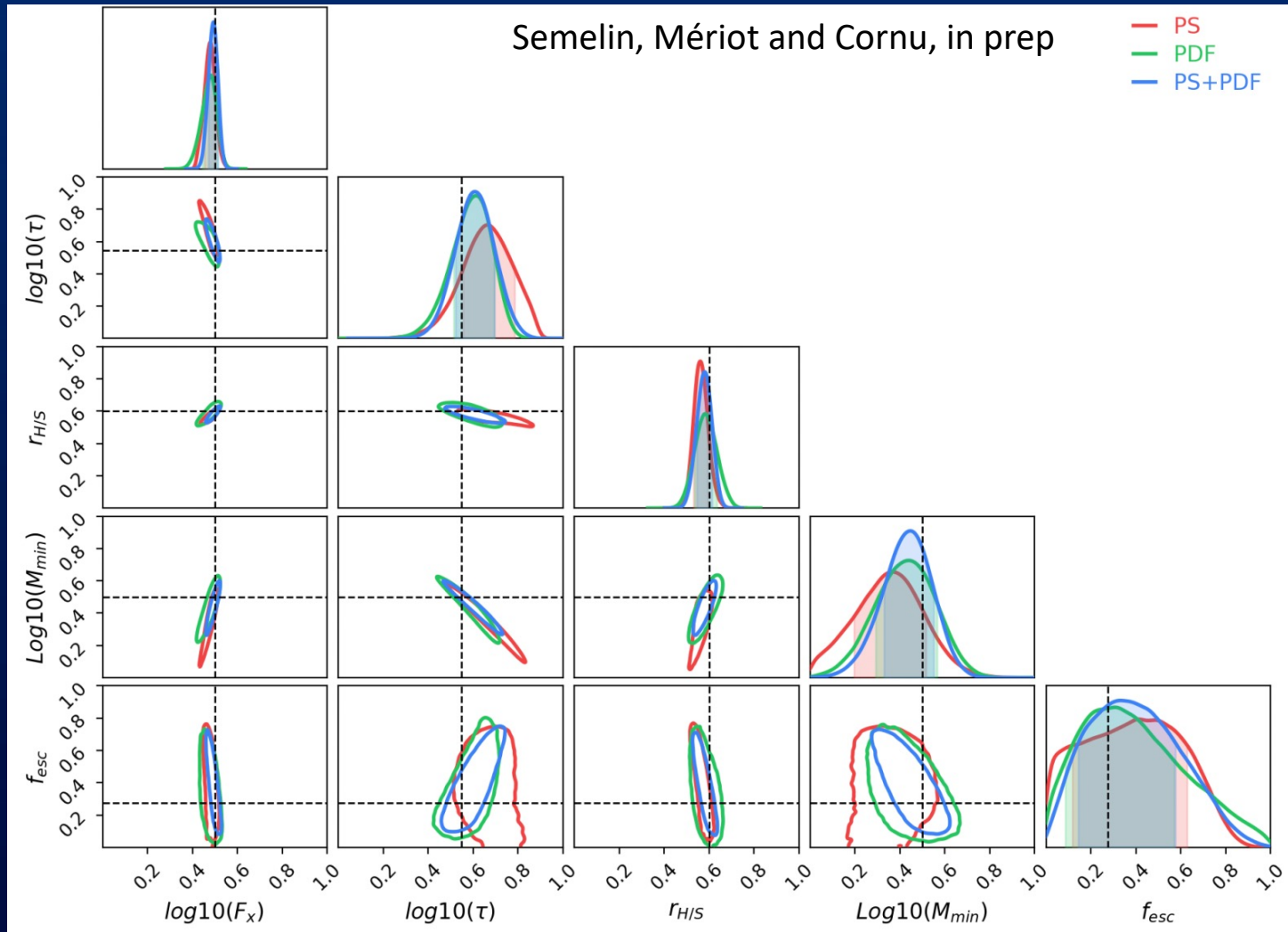
Combining **several** summary statistics (or observables!):

- no analytic form for the likelihood
- not independent => correlations

Example with 21-cm signal:

- Power spectrum + Pixel Distrib Func)
- Fit joint likelihood with NDE
- Train NDE on LoReLi II
- MCMC inference

=> A net gain of information



Conclusions and perspectives

Current accuracy: ~ 0.2 x variance, ~0.2 x training grid step

Reduced variance (longer obs time, less noise in data)

Expand training Database (Loreli III)

- Further improve physical modelling
- Streamline data production pipeline
- Reduce stored data volume
- Refine prior, perform adaptive sampling

Thank you!