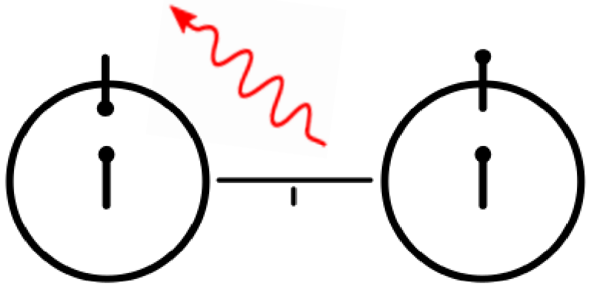
SCUOLA
NORMALE
SUPERIORE

Machine Learning of the Cosmic 21-cm Signal

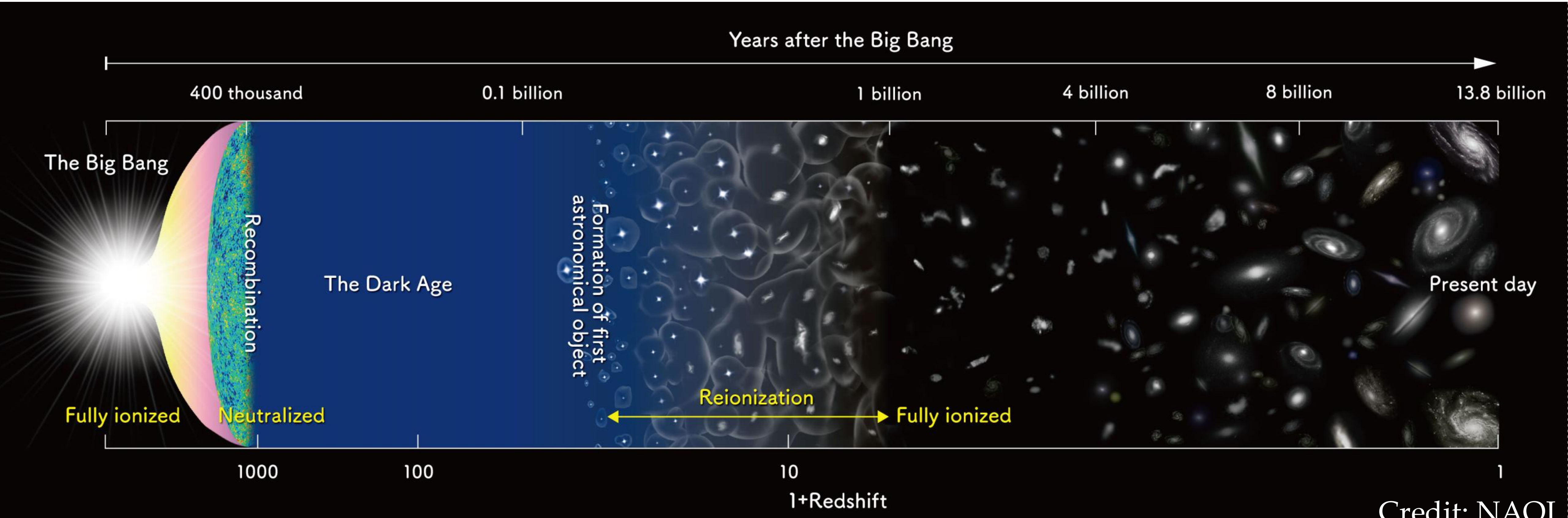
David Prelogović

PostDoc @ SISSA, Trieste, IT

1. Cosmic 21-cm Signal



- Hydrogen atoms abundant throughout the Universe’s evolution
- Encoding the first billion years

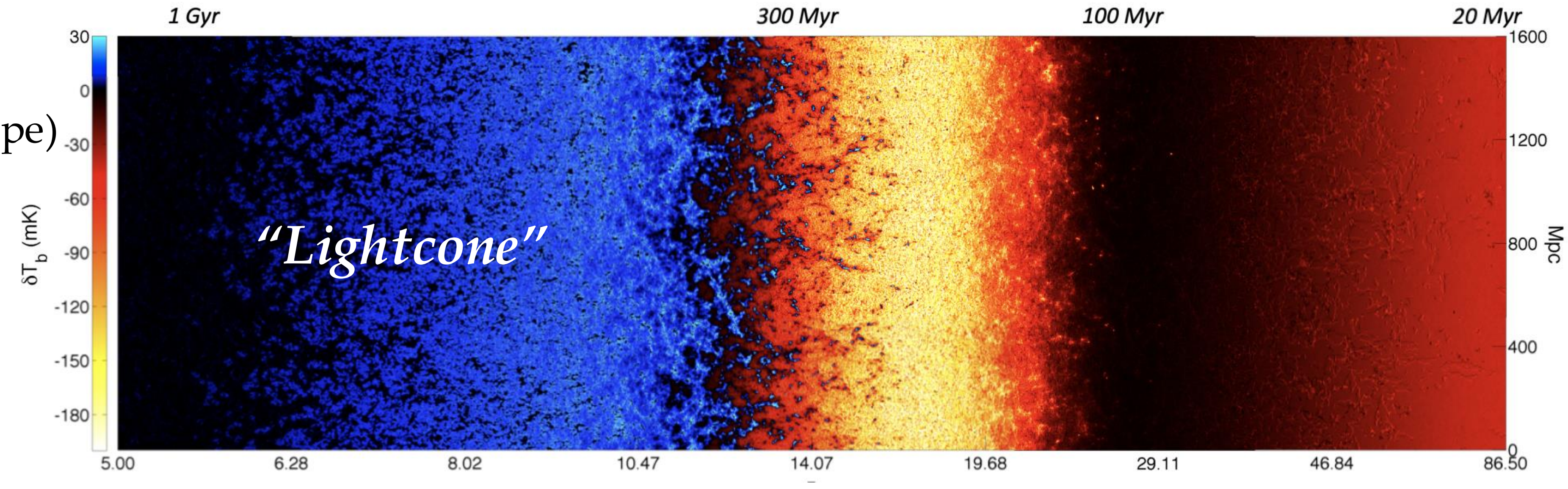


Credit: NAOJ

2. Cosmic 21-cm Signal

View to the past →

observer
(radio-telescope)



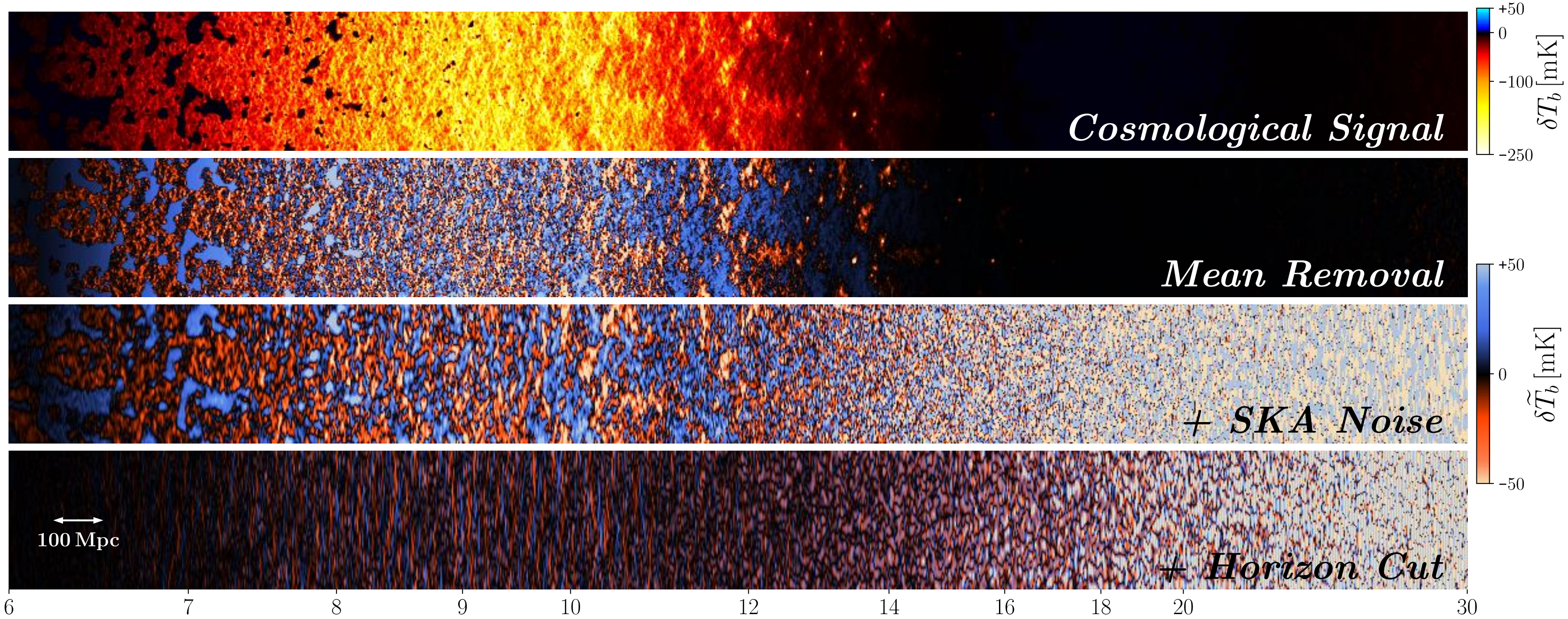
- cosmo. + astro.

$$\delta T_b \approx 30\, x_{\rm HI}\, \Delta \left(\frac{H}{dv_r/dr + H}\right)\left(1-\frac{T_\gamma}{T_{\rm S}}\right)\left(\frac{1+z}{10}\,\frac{0.15}{\Omega_{\rm M}h^2}\right)^{1/2}\left(\frac{\Omega_b h^2}{0.023}\right)\,{\rm mK}$$

Mesinger+2016

SCUOLA
NORMALE
SUPERIORE

3. Forward modeling pipeline

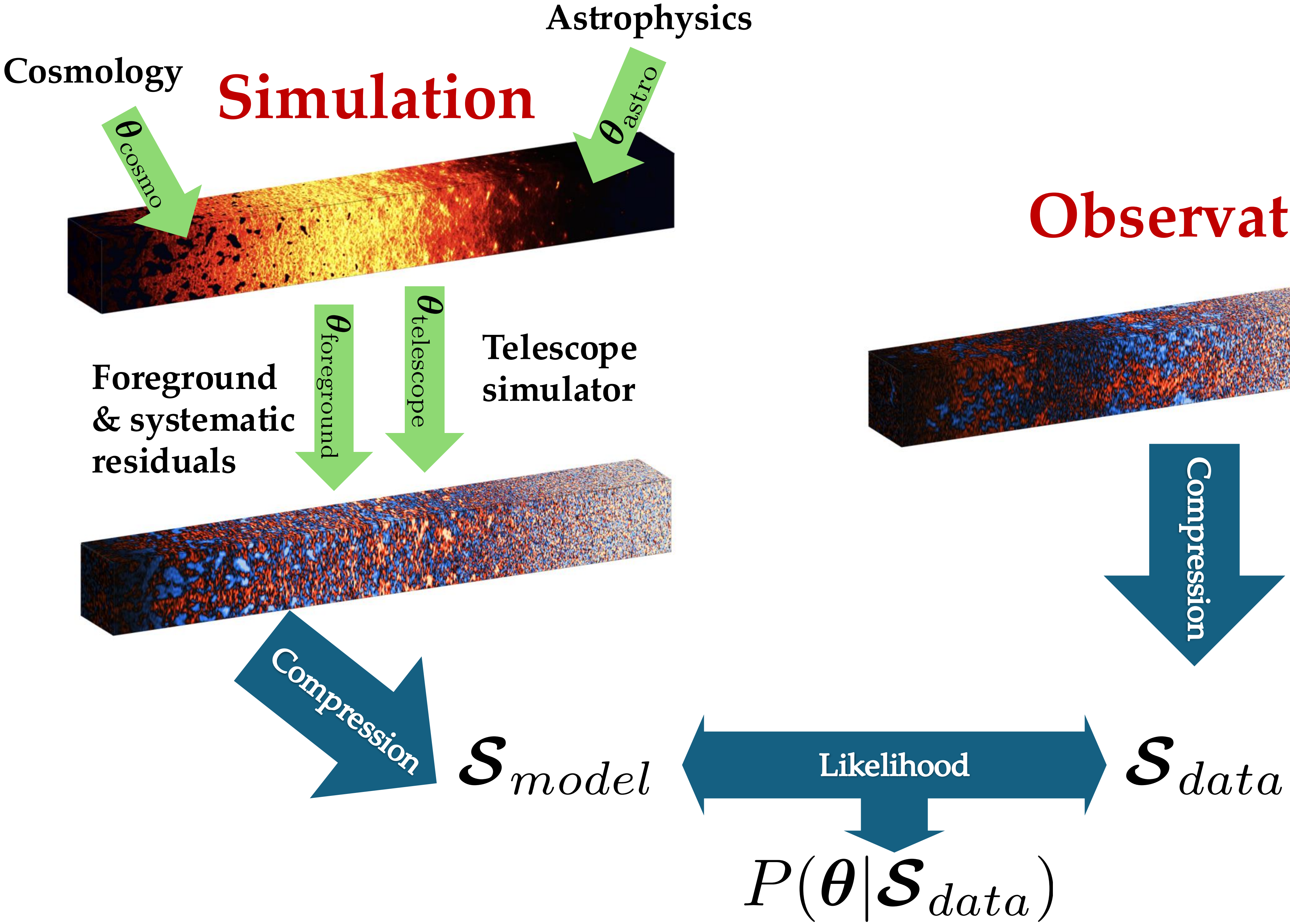


21cmFAST

SKA simulator

Foreground avoidance

Prelogović+2022



Cosmology

Simulation

Astrophysics

$\boldsymbol{\theta}_{\rm cosmo}$

$\boldsymbol{\theta}_{\rm astro}$

Foreground
& systematic
residuals

$\boldsymbol{\theta}_{\rm foreground}$

$\boldsymbol{\theta}_{\rm telescope}$

Telescope
simulator

Observation

Compression

Compression

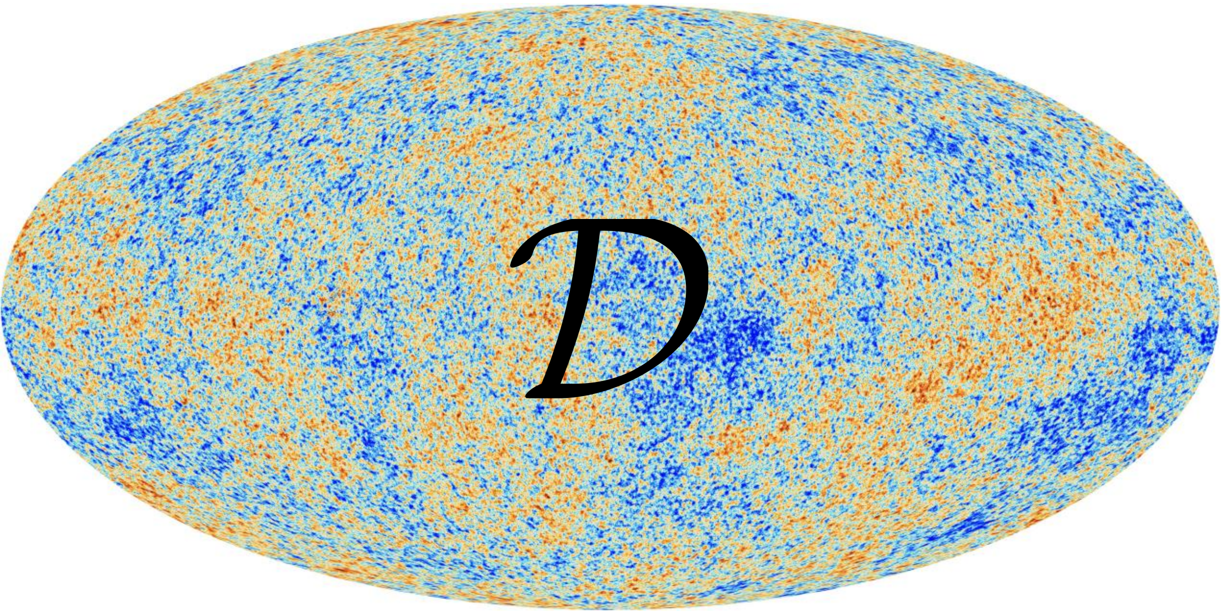
$\boldsymbol{s}_{model}$

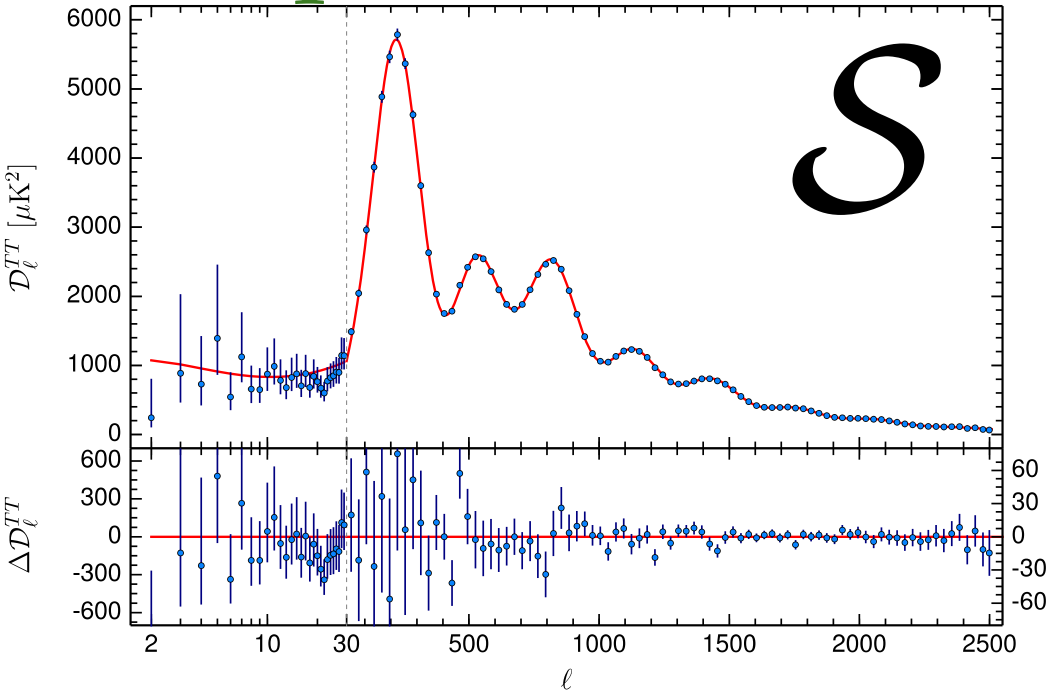
Likelihood

$\boldsymbol{s}_{data}$

$P(\boldsymbol{\theta}|\boldsymbol{s}_{data})$

4.1 Classical Inference Example: CMB

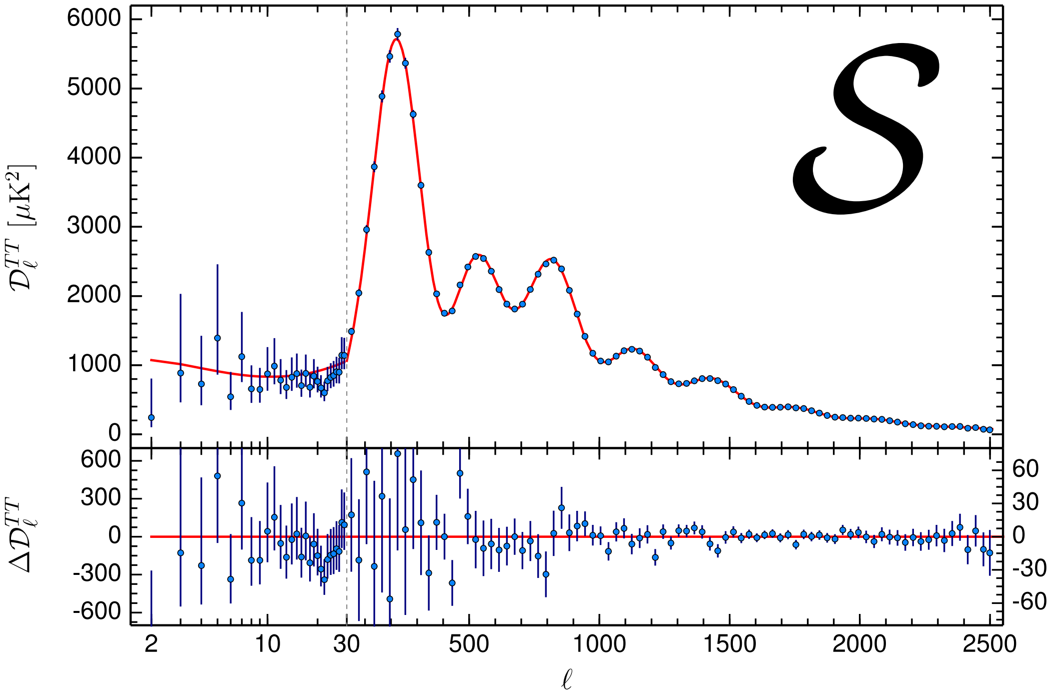




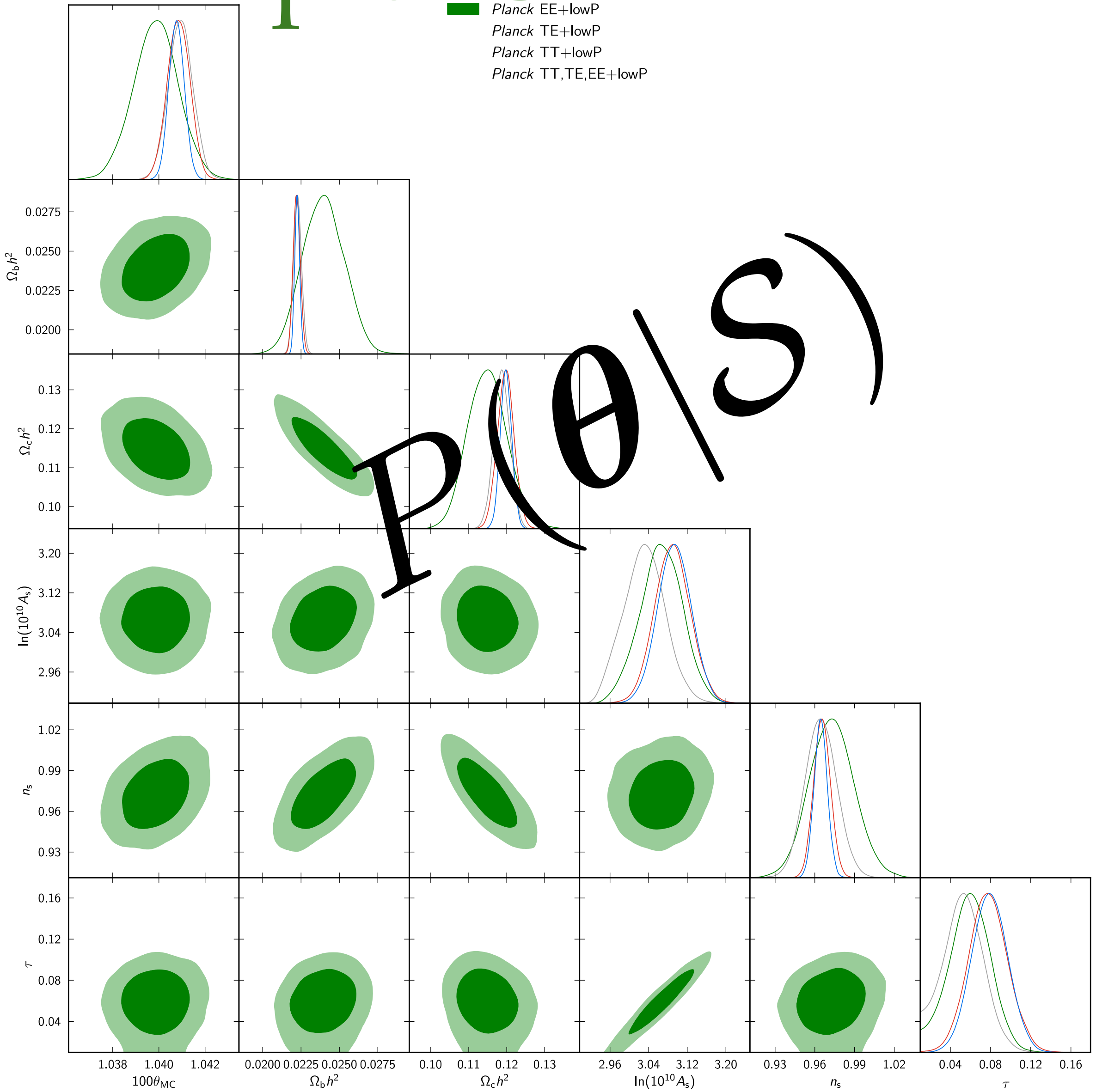
- Full sky map compressed to 1DPS
  - Known, optimal compression

Planck 2015

4.1 Classical Inference Example: CMB



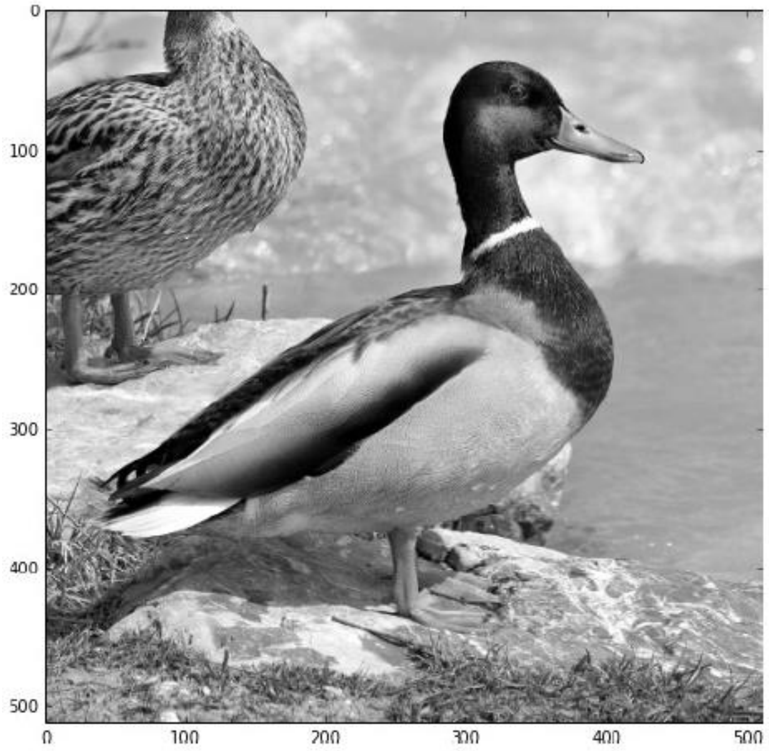
→



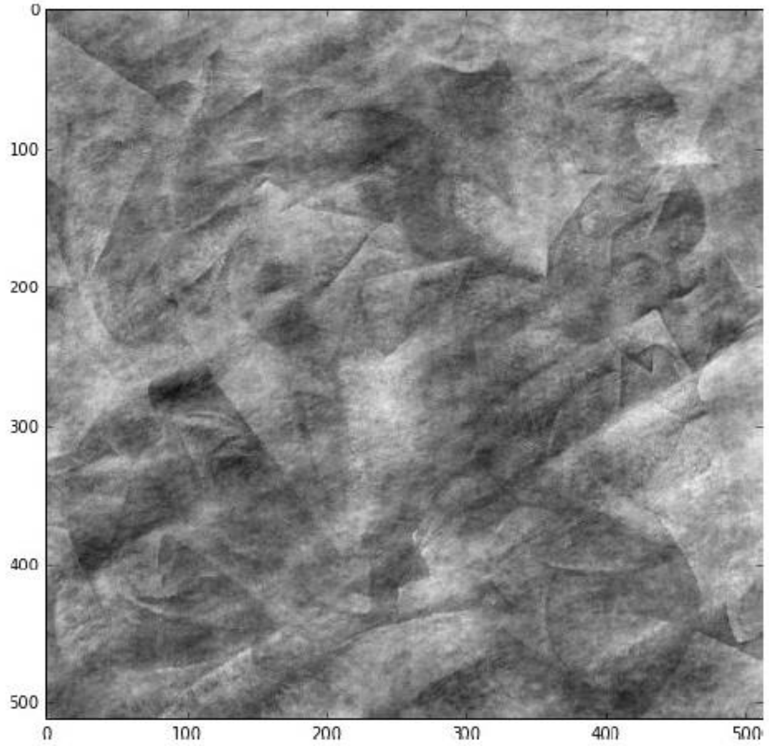
- Full sky map compressed to 1DPS
  - Known, optimal compression
- From it we infer the cosmology
  - Known likelihood

Planck 2015

4.2 Compression for a Duck



change phases

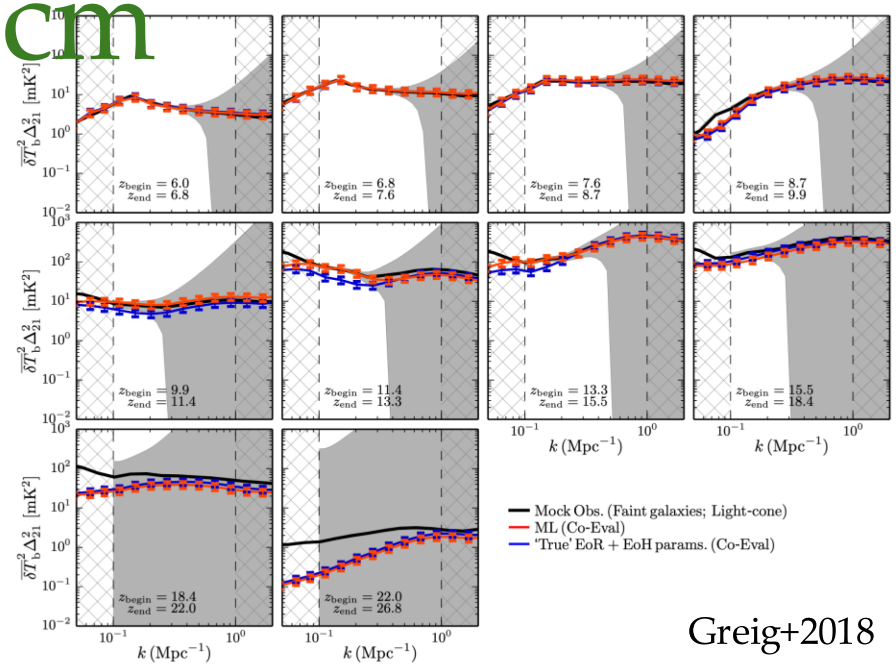


- Same 2D PS
- Highly non-Gaussian

Credit: G. Bernardi

4.3 Compression for the 21-cm

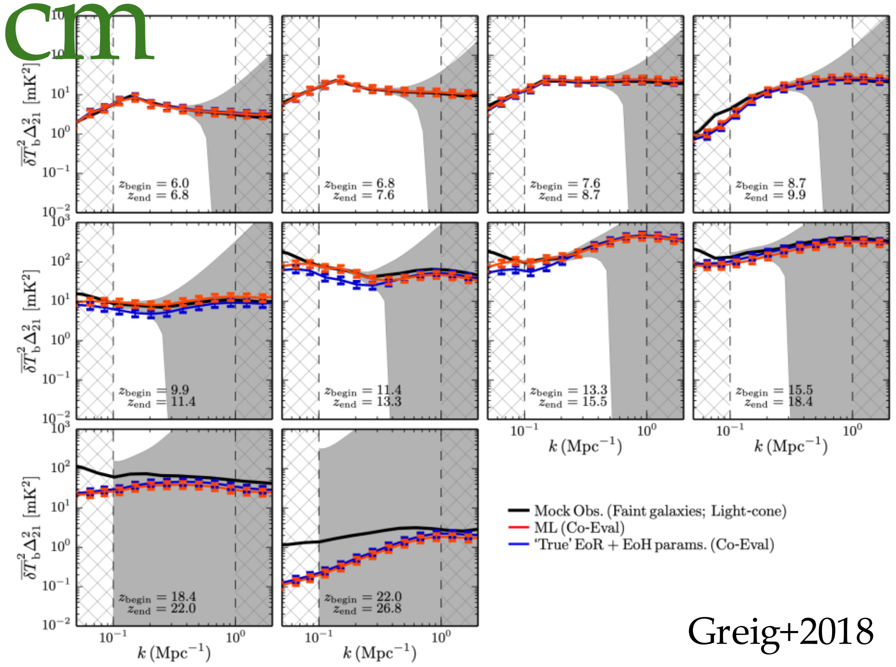
- Simpler than a duck
  - Power spectrum



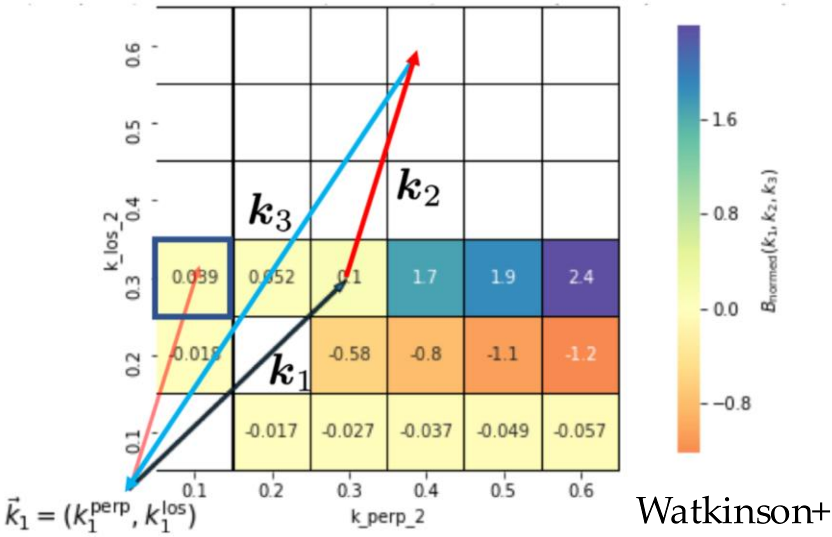
Greig+2018

4.3 Compression for the 21-cm

- Simpler than a duck
  - Power spectrum
  - Bispectrum



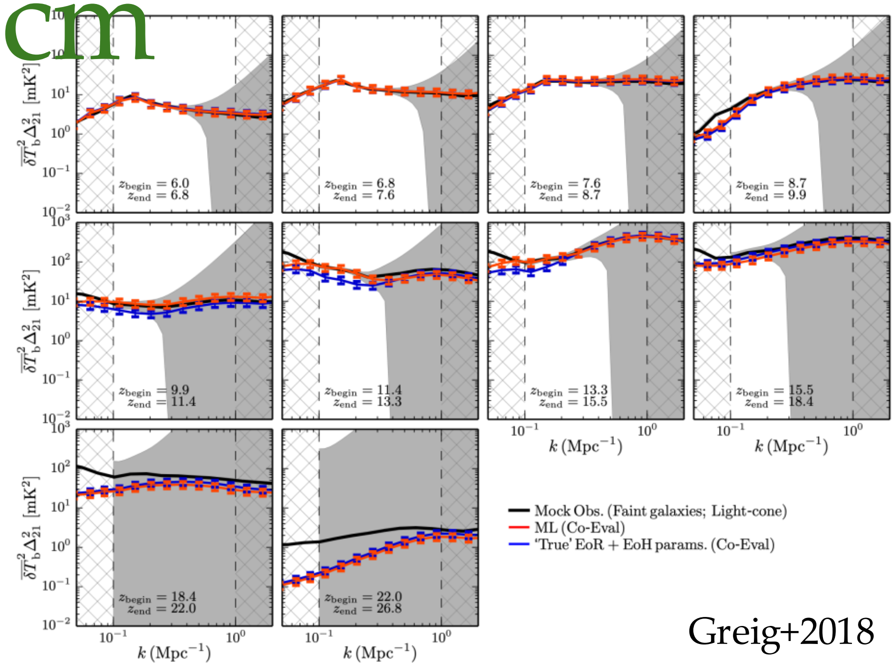
Greig+2018



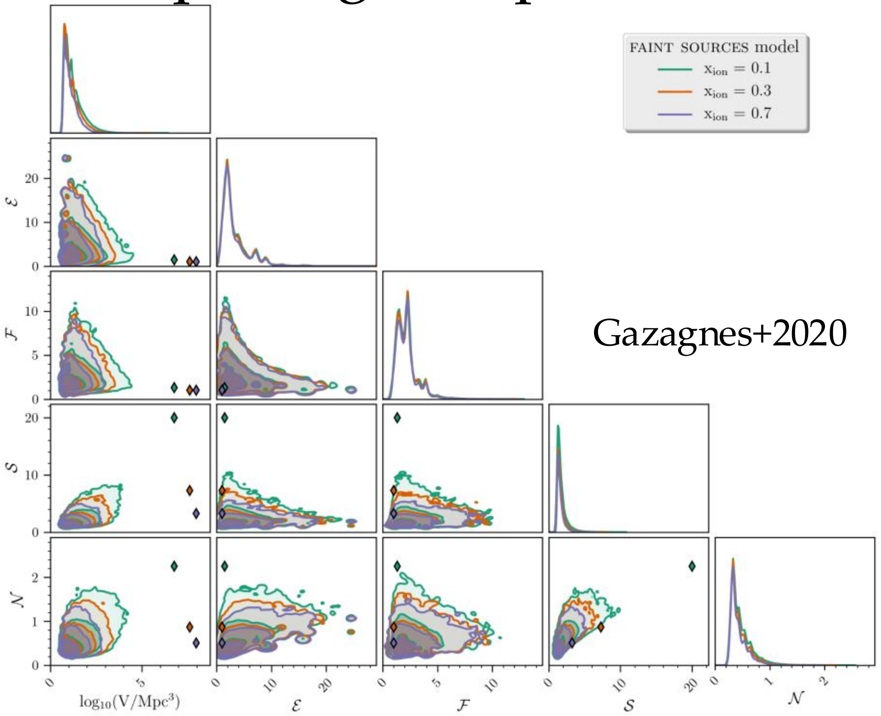
Watkinson+2020

4.3 Compression for the 21-cm

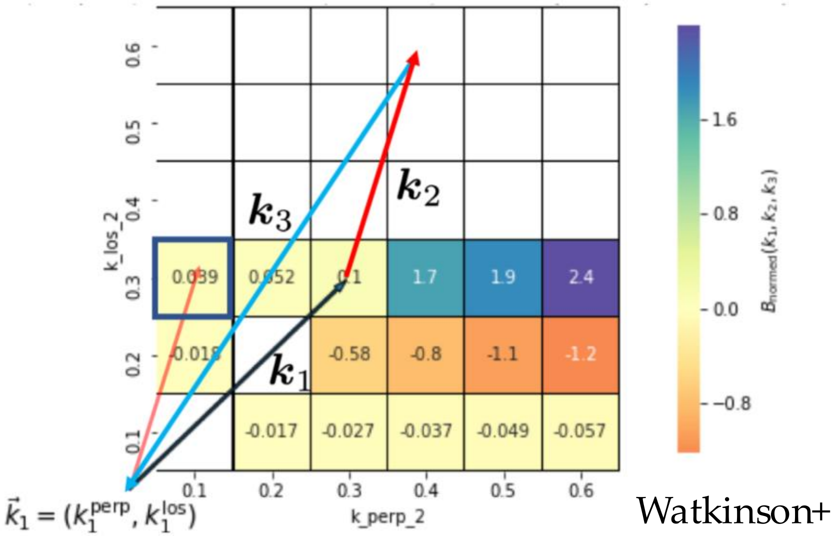
- Simpler than a duck
  - Power spectrum
  - Bispectrum
  - Morphological spectra



Greig+2018



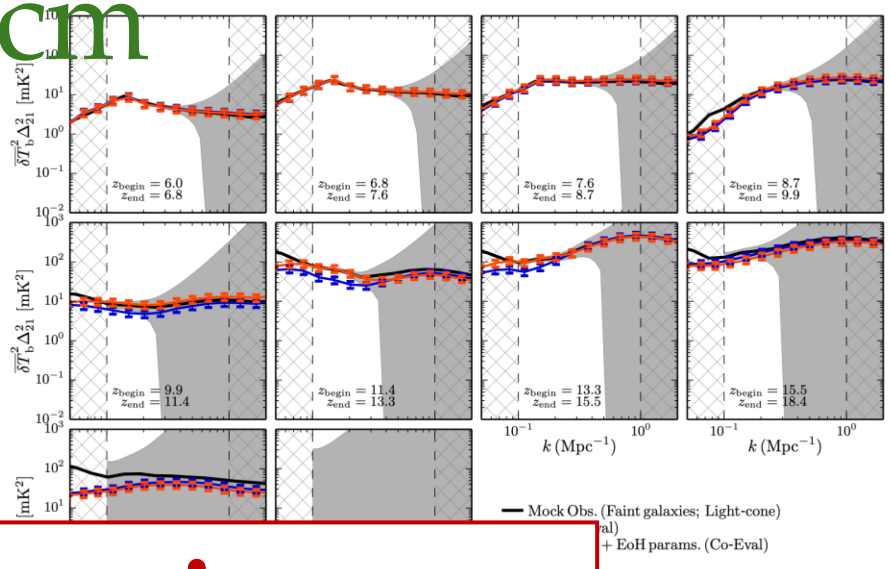
Gazagnes+2020



Watkinson+2020

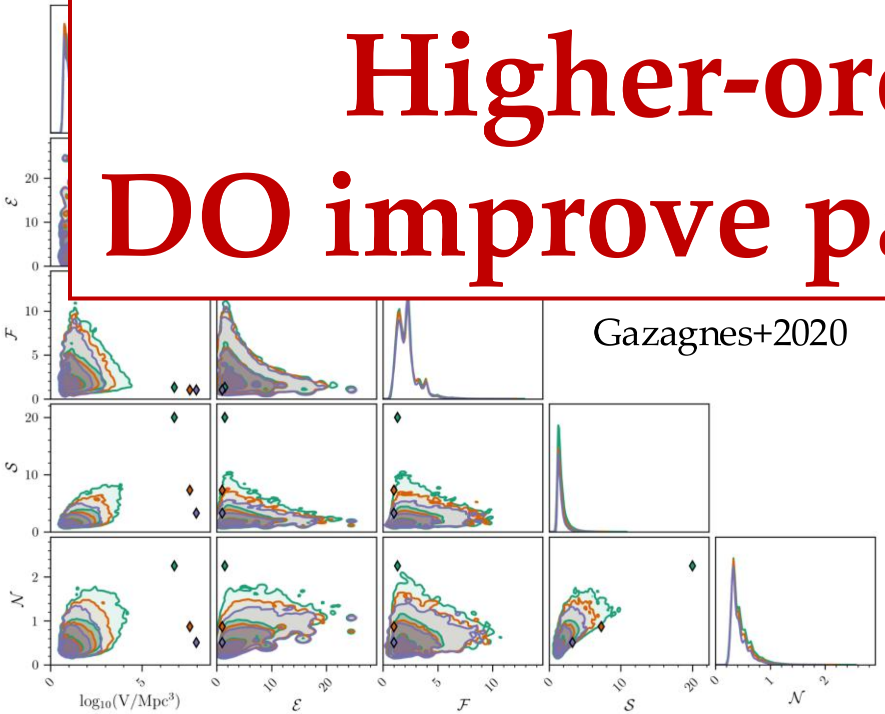
4.3 Compression for the 21-cm

- Simpler than a duck
  - Power spectrum
  - Bispectrum
  - Morphological spectra

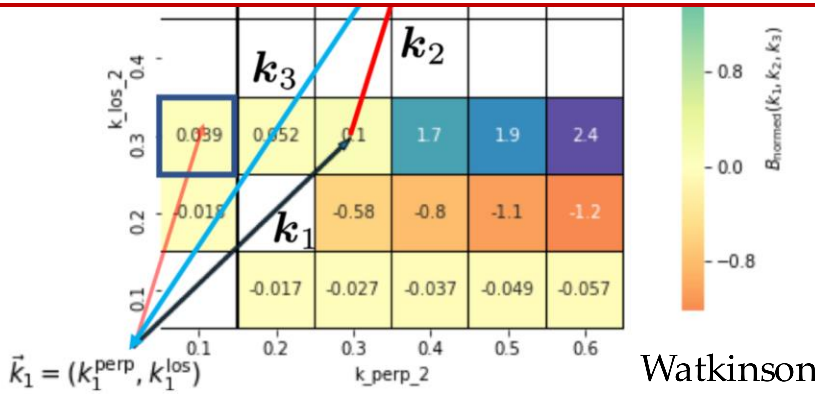


Greig+2018

Higher-order summaries
DO improve parameter inference!



Gazagnes+2020



Watkinson+2020

5.1 ML role #1 - Compression

- 21-cm – no good a-priori physical motivation for a compression
- We cannot know THE optimal compression/summary

5.1 ML role #1 - Compression

- 21-cm – no good a-priori physical motivation for a compression
- We cannot know THE optimal compression/summary

Solution:

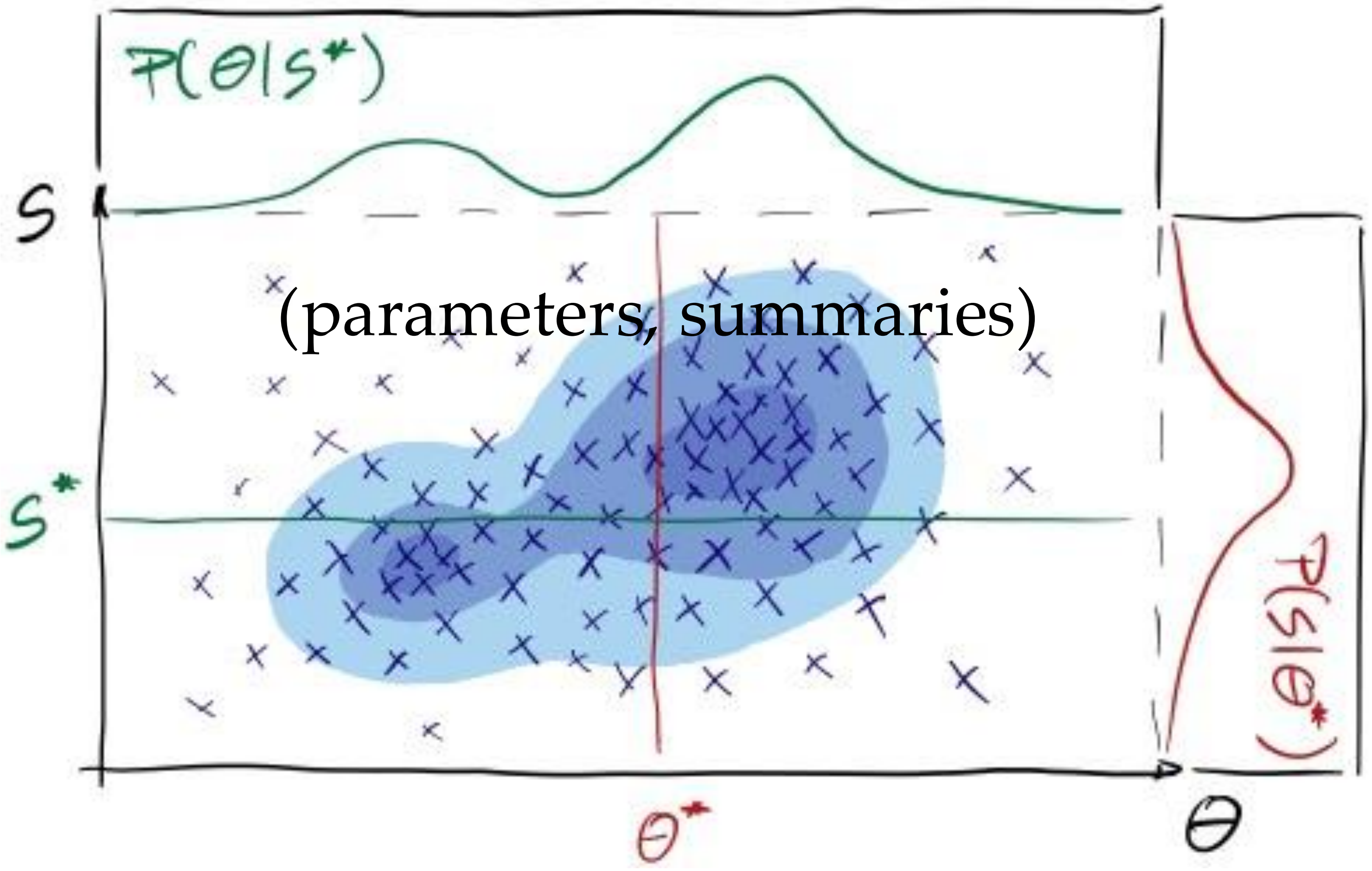
Let the machines figure it out for us!

(Neural Network)

- Gillet+2018
- La Plante & Ntampaka 2019
- Makinen+2020
- Mangena+2020
- Hortúa+2020
- Prelogović+2021
- +++

5.2 ML role #2 – Simulation Based Inference

- Joint space of



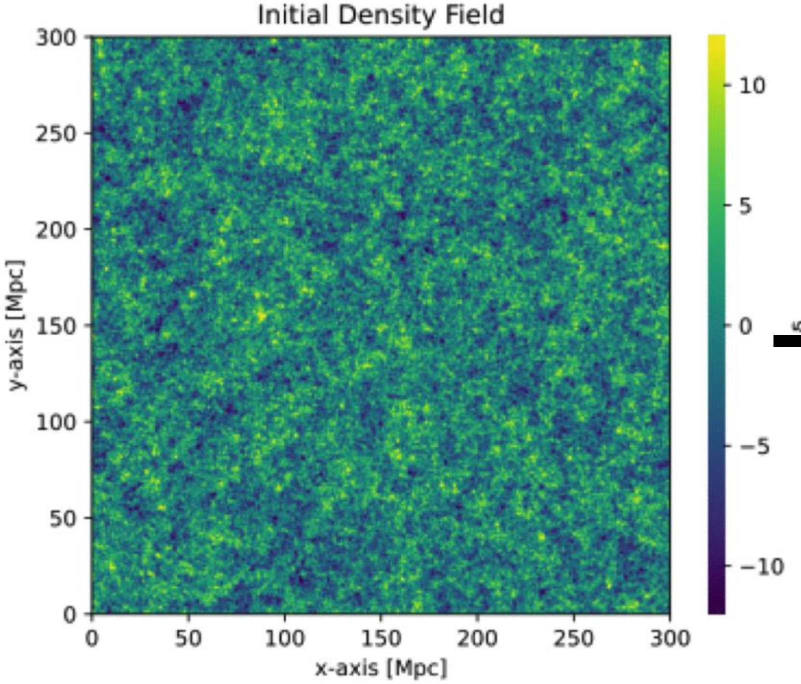
(parameters, summaries)

$$P(\mathcal{S},\boldsymbol{\theta}) = P(\mathcal{S}|\boldsymbol{\theta})\,P(\boldsymbol{\theta})$$

Credit:
D. Breitman
N. Triantafyllou

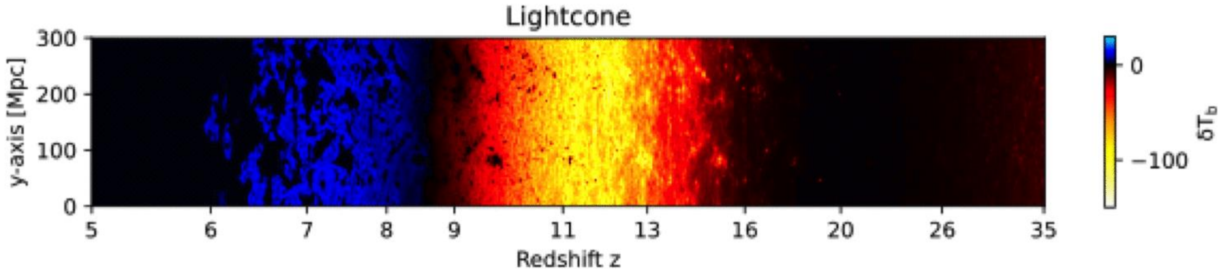
SCUOLA
NORMALE
SUPERIORE

$\theta^* \rightarrow$

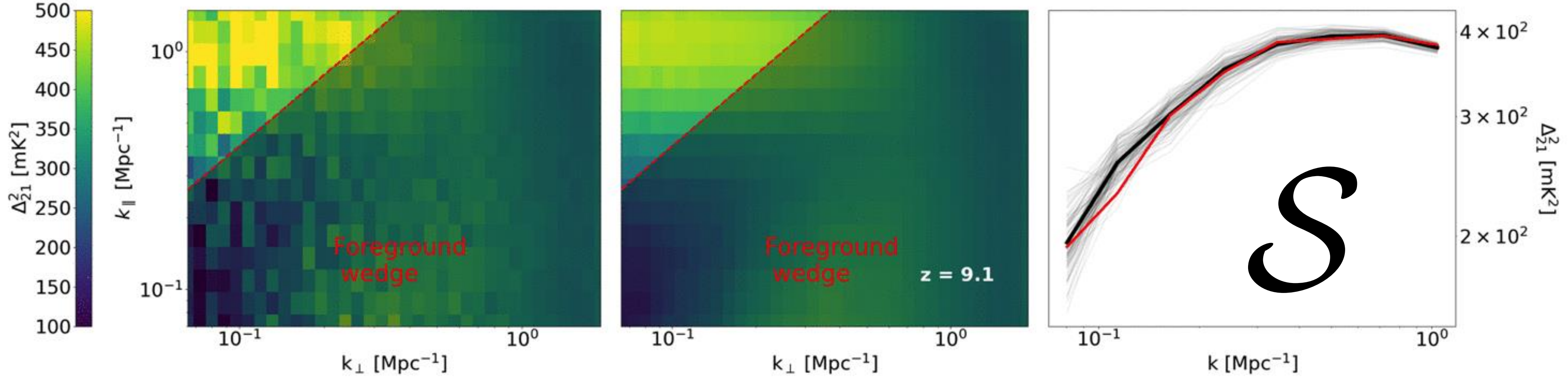


$\rightarrow$

$\mathcal{D}$



$\downarrow$

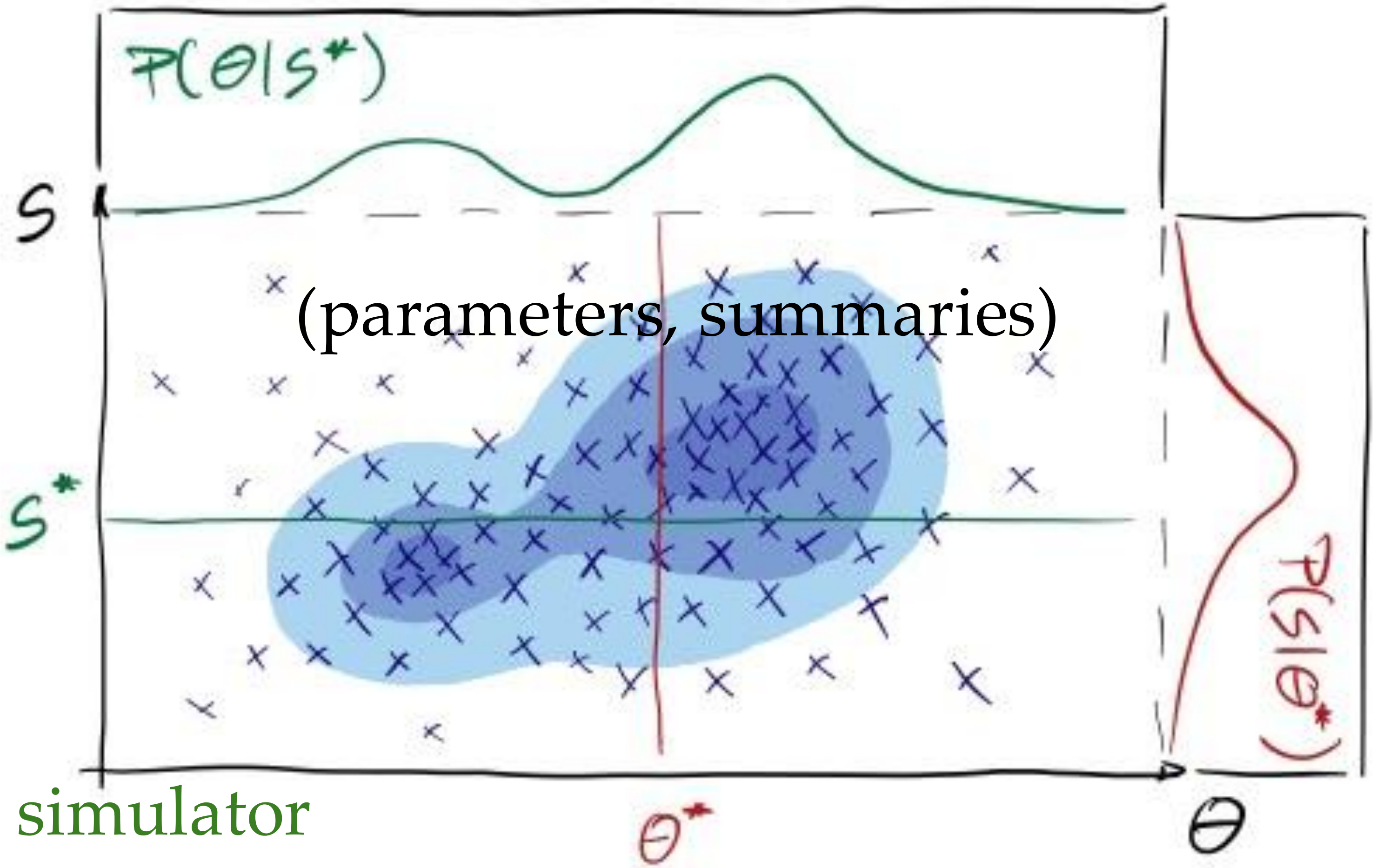


$\mathcal{S}$

6. ML role #2 – Simulation Based Inference

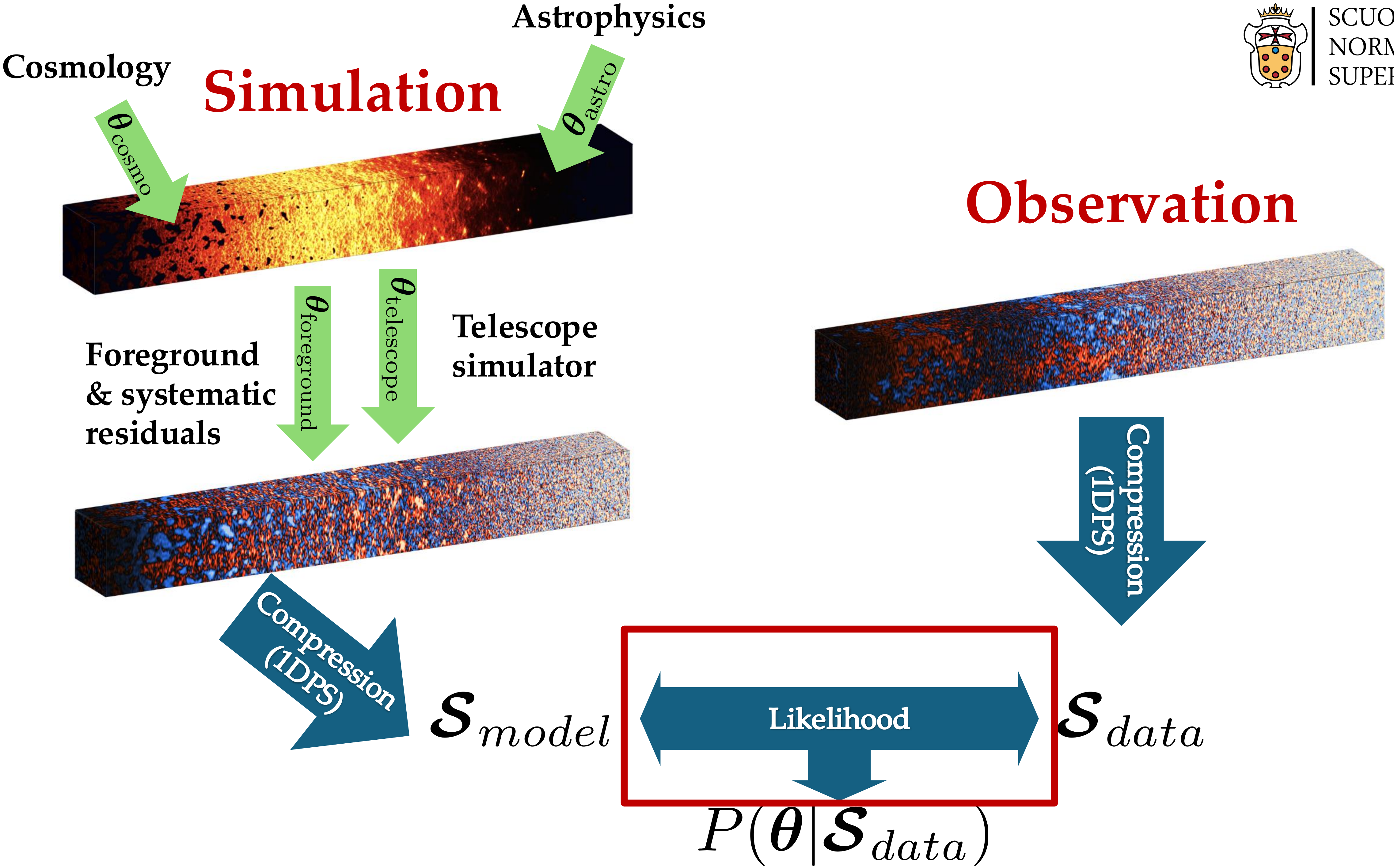
- Joint space of (parameters, summaries)

$$P(\mathcal{S}, \boldsymbol{\theta}) = P(\mathcal{S}|\boldsymbol{\theta})\, P(\boldsymbol{\theta})$$



- Assumption – perfect data simulator
- Fitting the distribution with Neural Density Estimators (NDE)

What is the likelihood of the 21-cm 1D power spectrum?



Cosmology

Simulation

Astrophysics

$\boldsymbol{\theta}_{\rm cosmo}$

$\boldsymbol{\theta}_{\rm astro}$

Foreground
& systematic
residuals

$\boldsymbol{\theta}_{\rm foreground}$

$\boldsymbol{\theta}_{\rm telescope}$

Telescope
simulator

Observation

Compression (1DPS)

Compression (1DPS)

$\boldsymbol{S}_{model}$

Likelihood

$\boldsymbol{S}_{data}$

$P(\boldsymbol{\theta}|\boldsymbol{S}_{data})$

1. 1DPS has a non-Gaussian likelihood

SCUOLA
NORMALE
SUPERIORE

Gaussian data

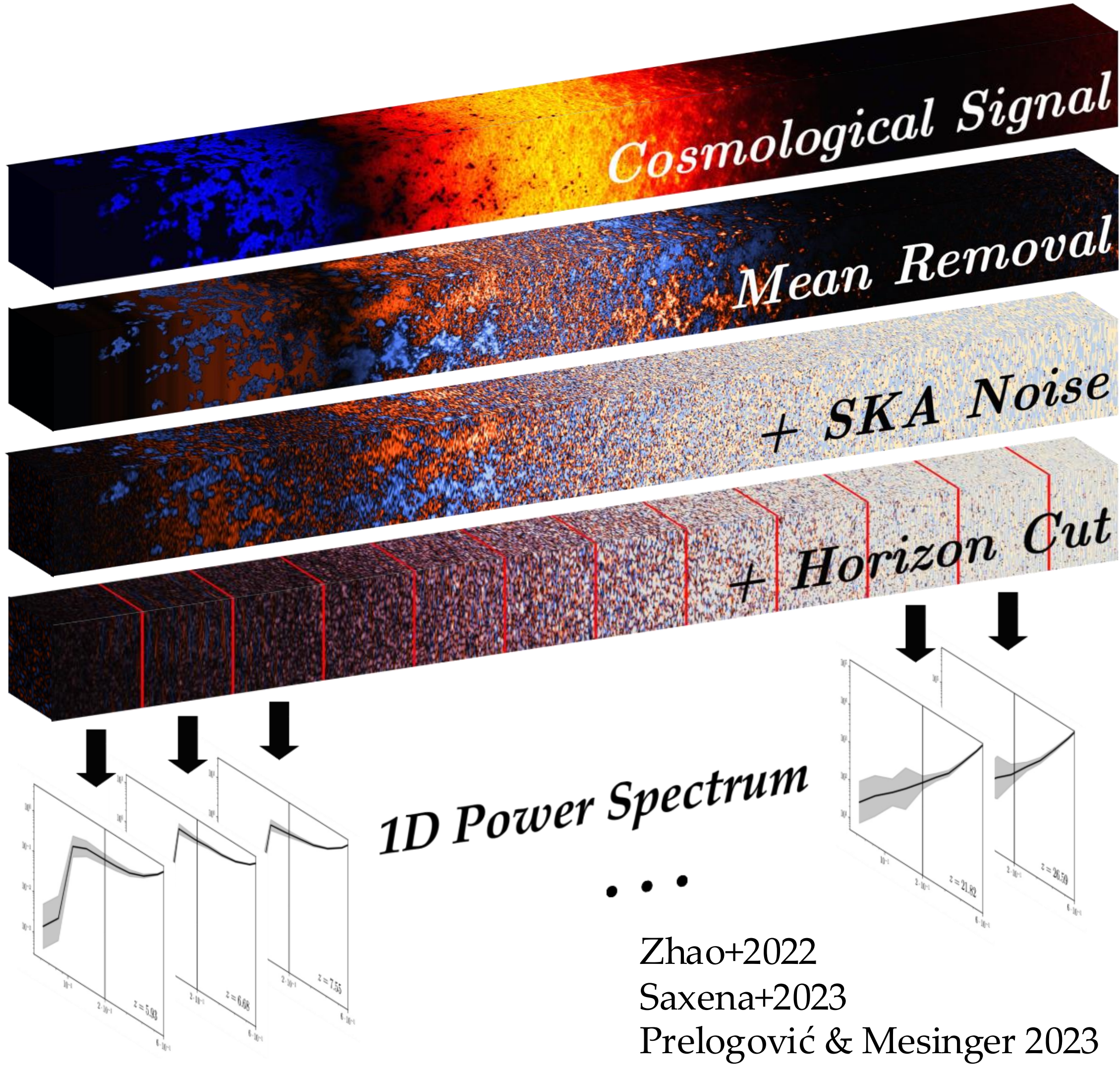
=

Gaussian likelihood in the PS

Non-Gaussian data

=

Non-Gaussian likelihood, even in the PS



Zhao+2022
Saxena+2023
Prelogović & Mesinger 2023

2. Classical inference (MCMC)

- Possible by approximating the PS likelihood with a Gaussian
  - Usually wrongly justified through the central limit theorem

$$
\begin{aligned}
P(\mathcal{S}|\boldsymbol{\theta}) &= \mathcal{N}\left(\boldsymbol{\Sigma}_{\mathcal{S}}(\boldsymbol{\theta}), \boldsymbol{\mu}_{\mathcal{S}}(\boldsymbol{\theta})\right) \\
&= \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{\Sigma}_{\mathcal{S}}(\boldsymbol{\theta})|}} e^{-\frac{1}{2}(\mathcal{S}-\boldsymbol{\mu}_{\mathcal{S}}(\boldsymbol{\theta}))^T\,\boldsymbol{\Sigma}_{\mathcal{S}}^{-1}(\boldsymbol{\theta})\,(\mathcal{S}-\boldsymbol{\mu}_{\mathcal{S}}(\boldsymbol{\theta}))}
\end{aligned}
$$

2. Classical inference (MCMC)

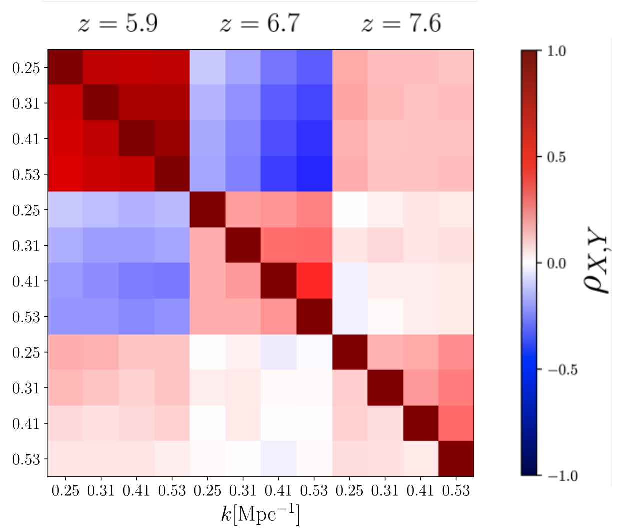
- Possible by approximating the PS likelihood with a Gaussian
  - Usually wrongly justified through the central limit theorem

$$P(\mathcal{S}|\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\Sigma}_{\mathcal{S}}(\boldsymbol{\theta}), \boldsymbol{\mu}_{\mathcal{S}}(\boldsymbol{\theta})\right)$$

$$= \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{\Sigma}_{\mathcal{S}}(\boldsymbol{\theta})|}} e^{-\frac{1}{2}(\mathcal{S}-\boldsymbol{\mu}_{\mathcal{S}}(\boldsymbol{\theta}))^T \boldsymbol{\Sigma}_{\mathcal{S}}^{-1}(\boldsymbol{\theta})\,(\mathcal{S}-\boldsymbol{\mu}_{\mathcal{S}}(\boldsymbol{\theta}))}$$

- Common additional simplifications
  1) ignoring correlations by using diagonal Σ

Prelogović&Mesinger 2023



2. Classical inference (MCMC)

- Possible by approximating the PS likelihood with a Gaussian
  - Usually wrongly justified through the central limit theorem

$$P(\mathcal{S}|\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\Sigma}_{\mathcal{S}}(\boldsymbol{\theta}), \boldsymbol{\mu}_{\mathcal{S}}(\boldsymbol{\theta})\right)$$

$$= \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{\Sigma}_{\mathcal{S}}\cancel{(\boldsymbol{\theta})}|}} e^{-\frac{1}{2}(\mathcal{S}-\boldsymbol{\mu}_{\mathcal{S}}(\boldsymbol{\theta}))^T \boldsymbol{\Sigma}_{\mathcal{S}}^{-1}\cancel{(\boldsymbol{\theta})}\,(\mathcal{S}-\boldsymbol{\mu}_{\mathcal{S}}(\boldsymbol{\theta}))}$$

- Common additional simplifications
  1) ignoring correlations by using diagonal Σ
  2) Fixing the covariance at fiducial parameters $\Sigma = \Sigma_{\theta\mathrm{fid}}$

Greig&Mesinger 2018
Trott+2020
Mertens+2020
HERA+2023

2. Classical inference (MCMC)

- Possible by approximating the PS likelihood with a Gaussian
  - Usually wrongly justified through the central limit theorem

$$P(\mathcal{S}|\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\Sigma}_\mathcal{S}(\boldsymbol{\theta}), \boldsymbol{\mu}_\mathcal{S}(\boldsymbol{\theta})\right)$$

$$= \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{\Sigma}_\mathcal{S}(\boldsymbol{\theta})|}} e^{-\frac{1}{2}(\mathcal{S}-\boldsymbol{\mu}_\mathcal{S}(\boldsymbol{\theta}))^T \boldsymbol{\Sigma}_\mathcal{S}^{-1}(\boldsymbol{\theta})\,(\mathcal{S}-\boldsymbol{\mu}_\mathcal{S}(\boldsymbol{\theta}))}$$

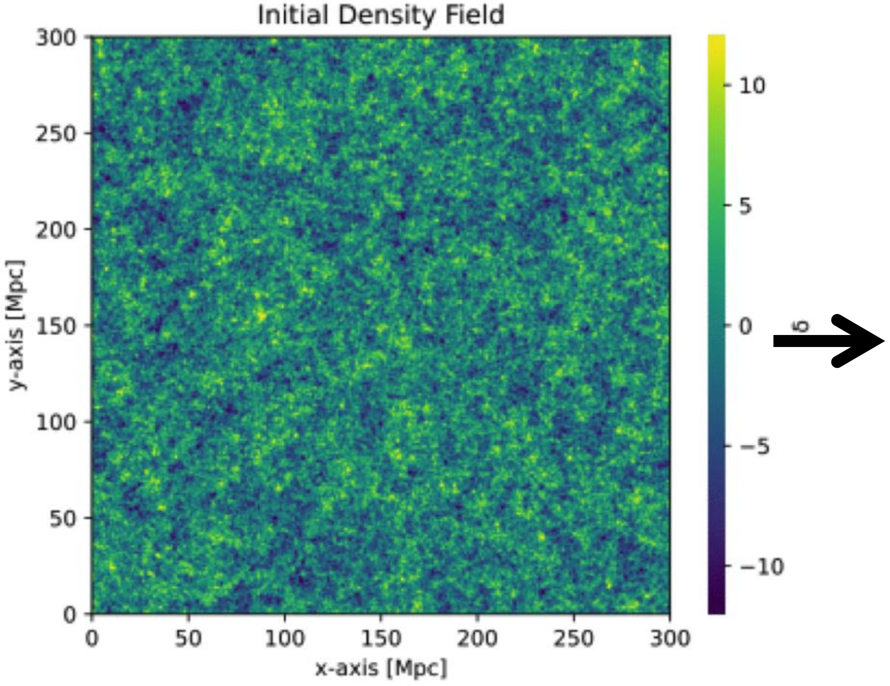
- Common additional simplifications
  1) ignoring correlations by using diagonal Σ
  2) Fixing the covariance at fiducial parameters $\Sigma = \Sigma_{\theta\mathrm{fid}}$
  3) μ estimated from one simulation

Greig&Mesinger 2018
Trott+2020
Mertens+2020
HERA+2023

Credit:
D. Breitman
N. Triantafyllou

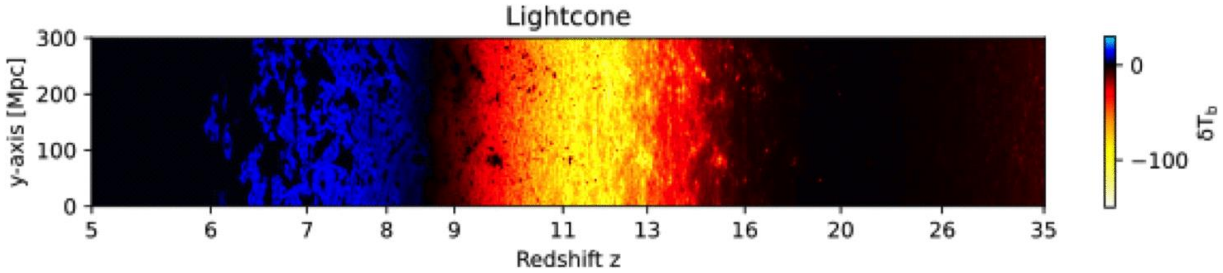
SCUOLA
NORMALE
SUPERIORE

$\theta^* \rightarrow$

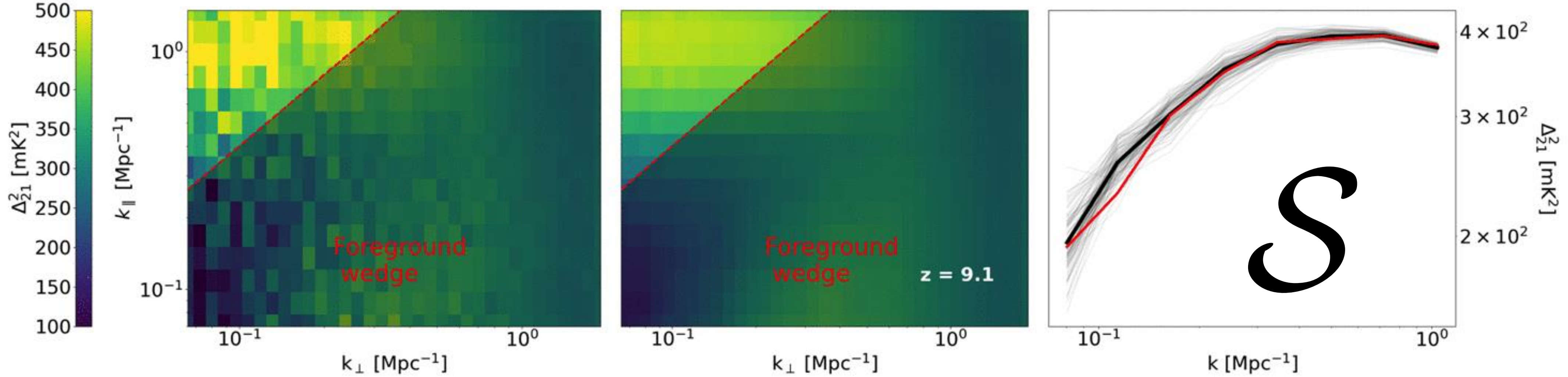


$\rightarrow$

$\mathcal{D}$



$\downarrow$

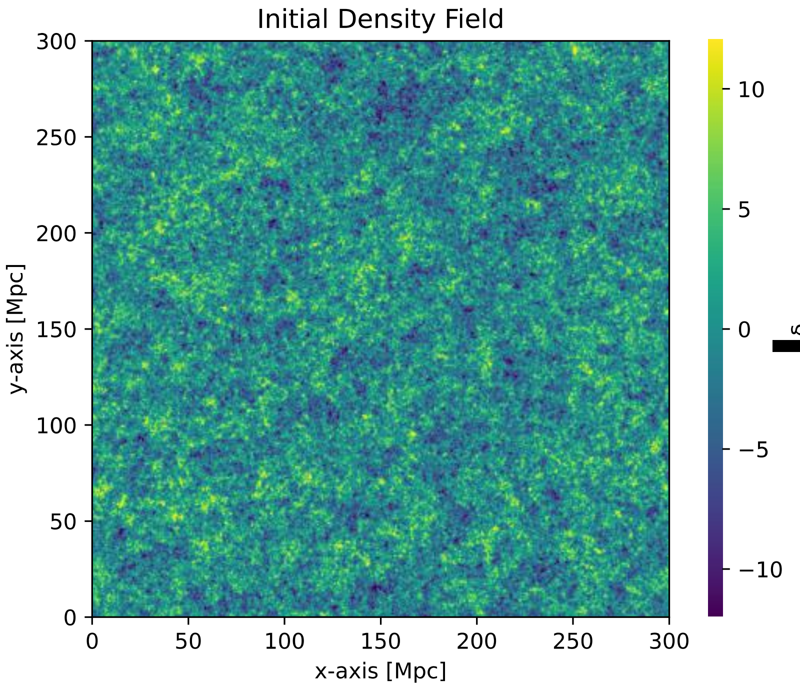


$\mathcal{S}$

Credit:
D. Breitman
N. Triantafyllou

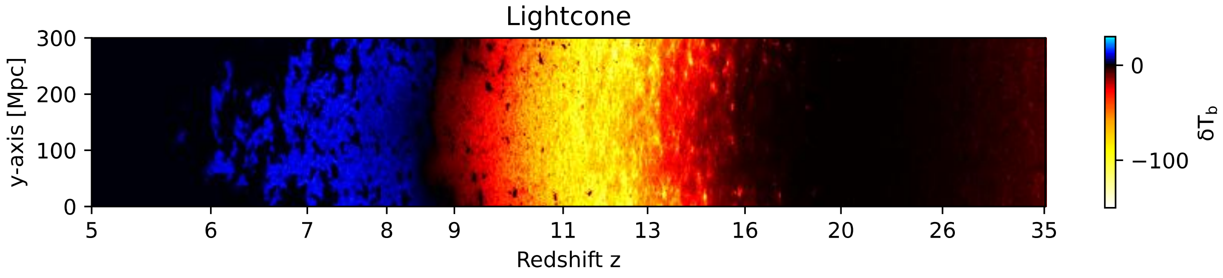
SCUOLA NORMALE SUPERIORE

$\theta^* \rightarrow$

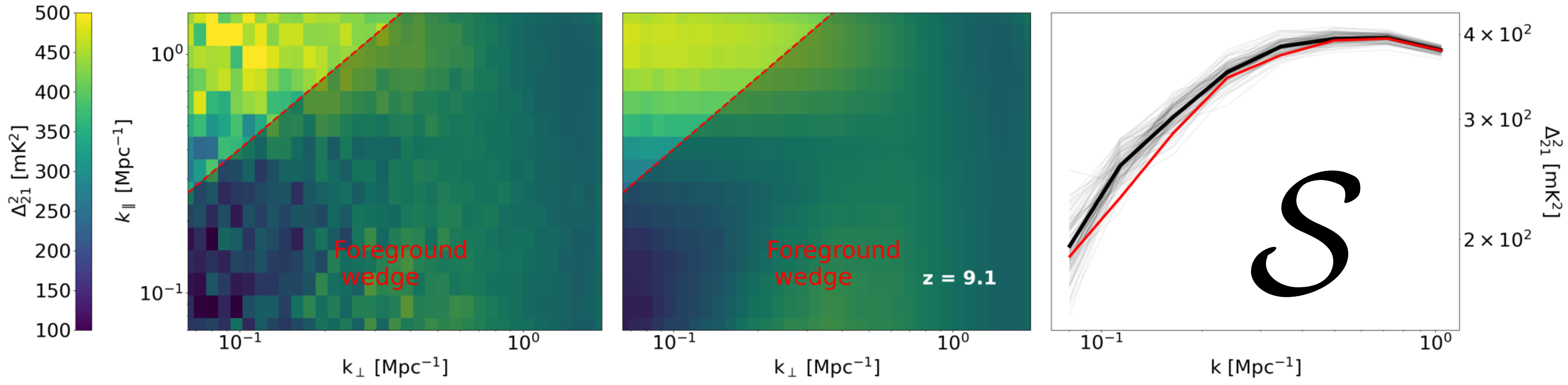


$\rightarrow$

$\mathcal{D}$

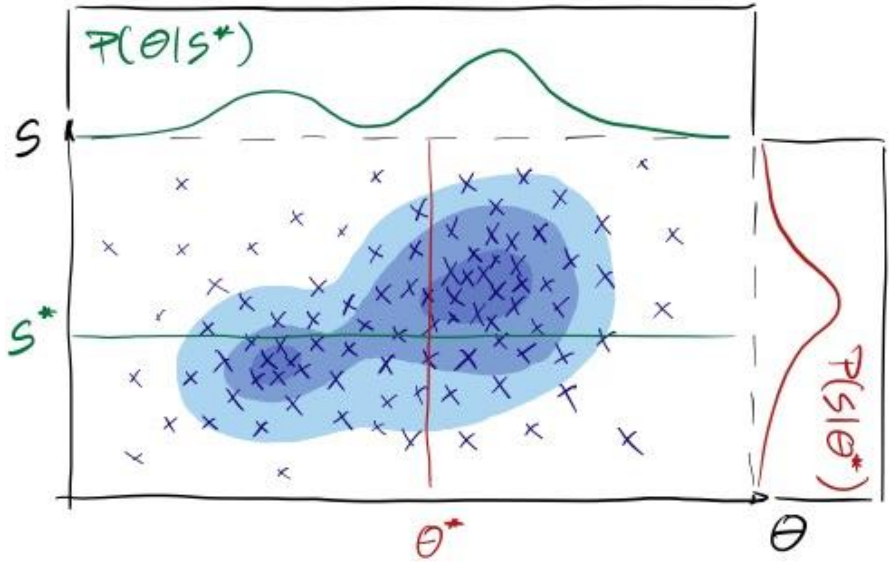


$\downarrow$



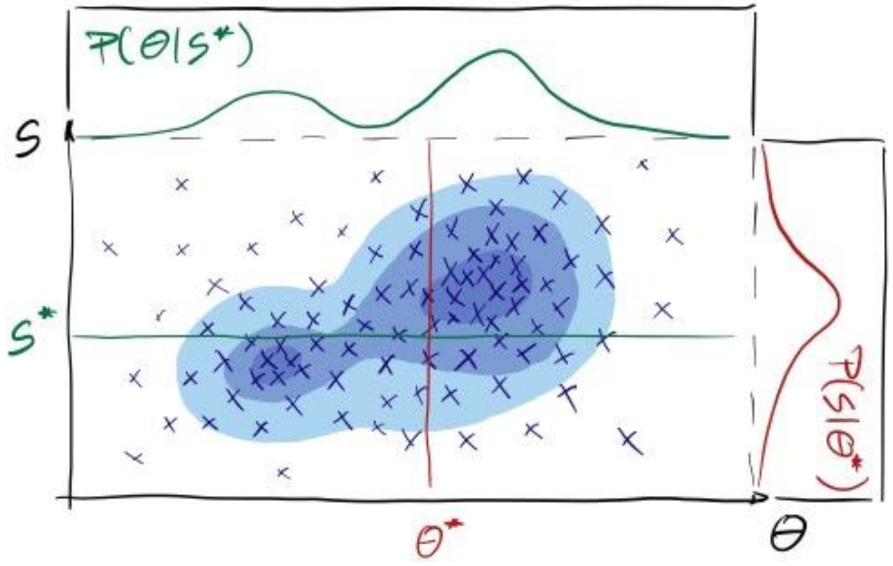
SCUOLA
NORMALE
SUPERIORE

3. Simulation Based Inference

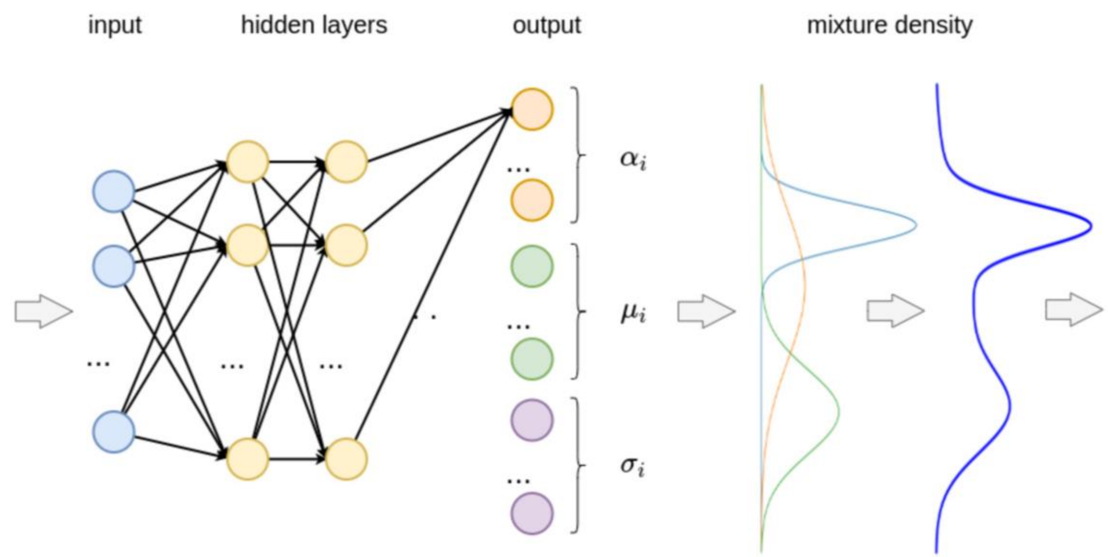


3. Simulation Based Inference

- Train a neural density estimator (NDE)
  - Gaussian mixture



$\boldsymbol{\theta}$ ⇒

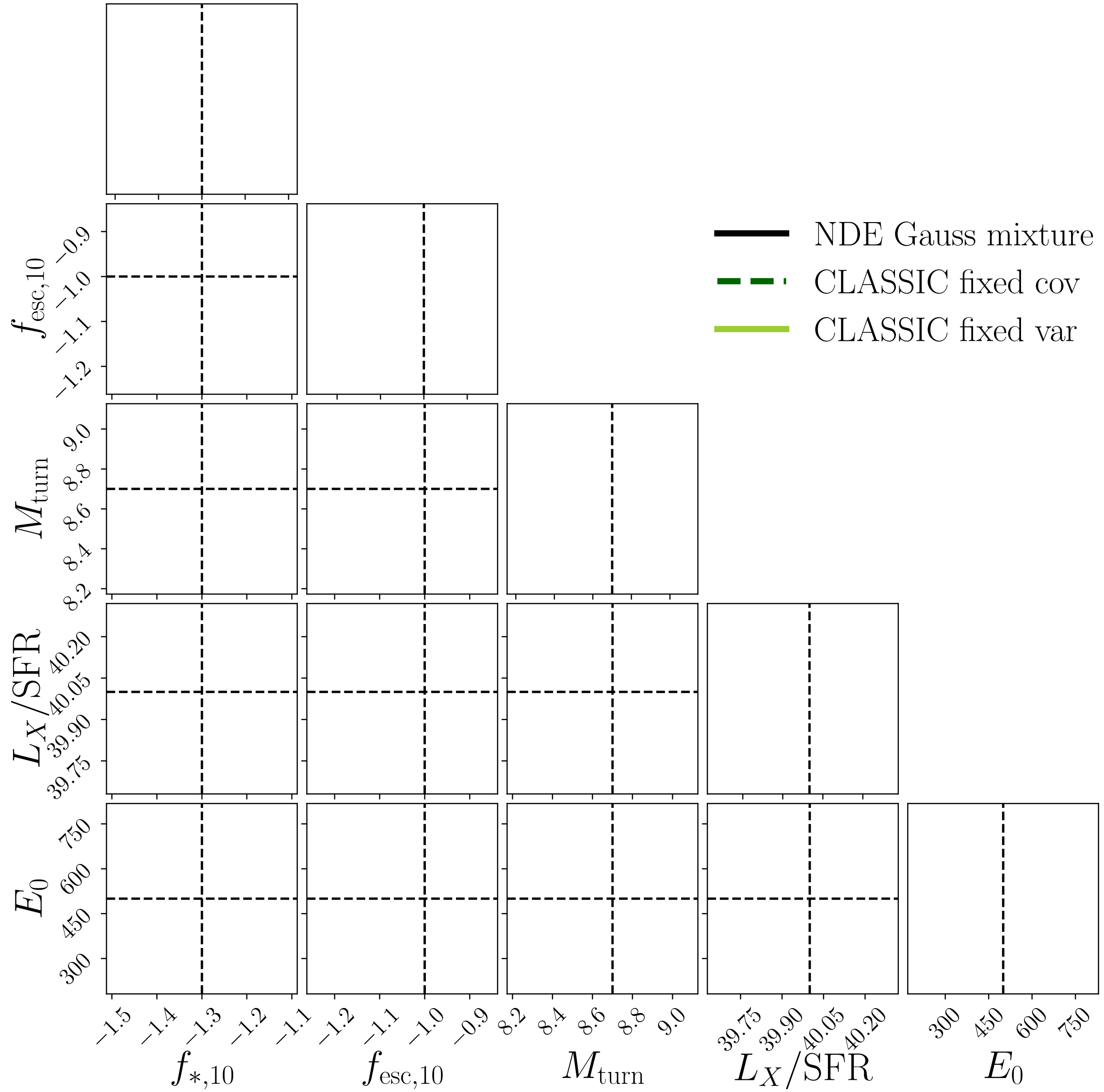


⇒ $P(\mathcal{S}|\boldsymbol{\theta})$

Repin+2021

Prelogović & Mesinger 2023

4. Results

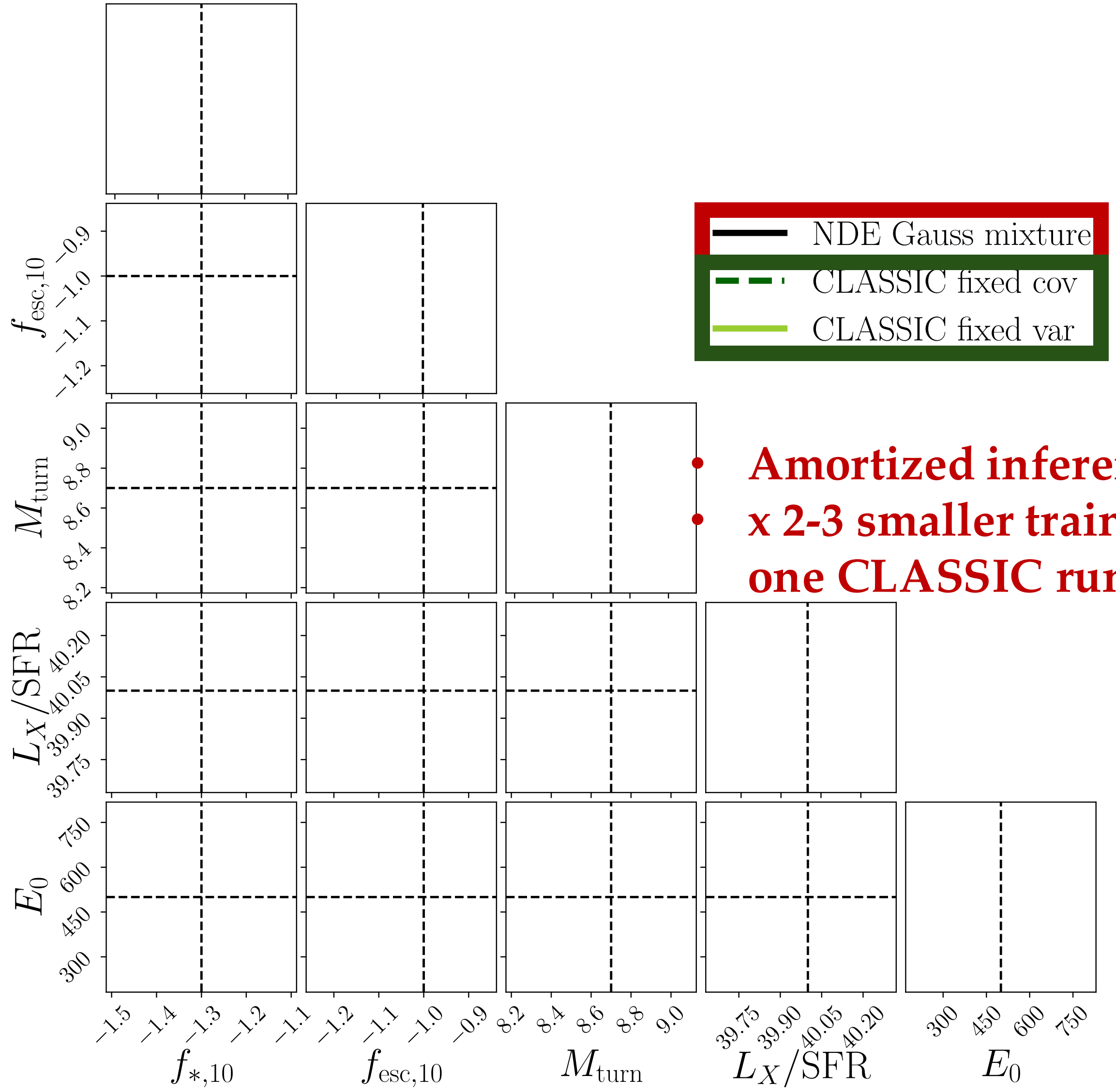


SCUOLA NORMALE SUPERIORE

Including more realistic likelihood ≠ more constraining posterior

Prelogović & Mesinger 2023

4. Results



0.1-1 CPU h

100 000 CPU h

- **Amortized inference**
- **x 2-3 smaller training DB vs. one CLASSIC run**

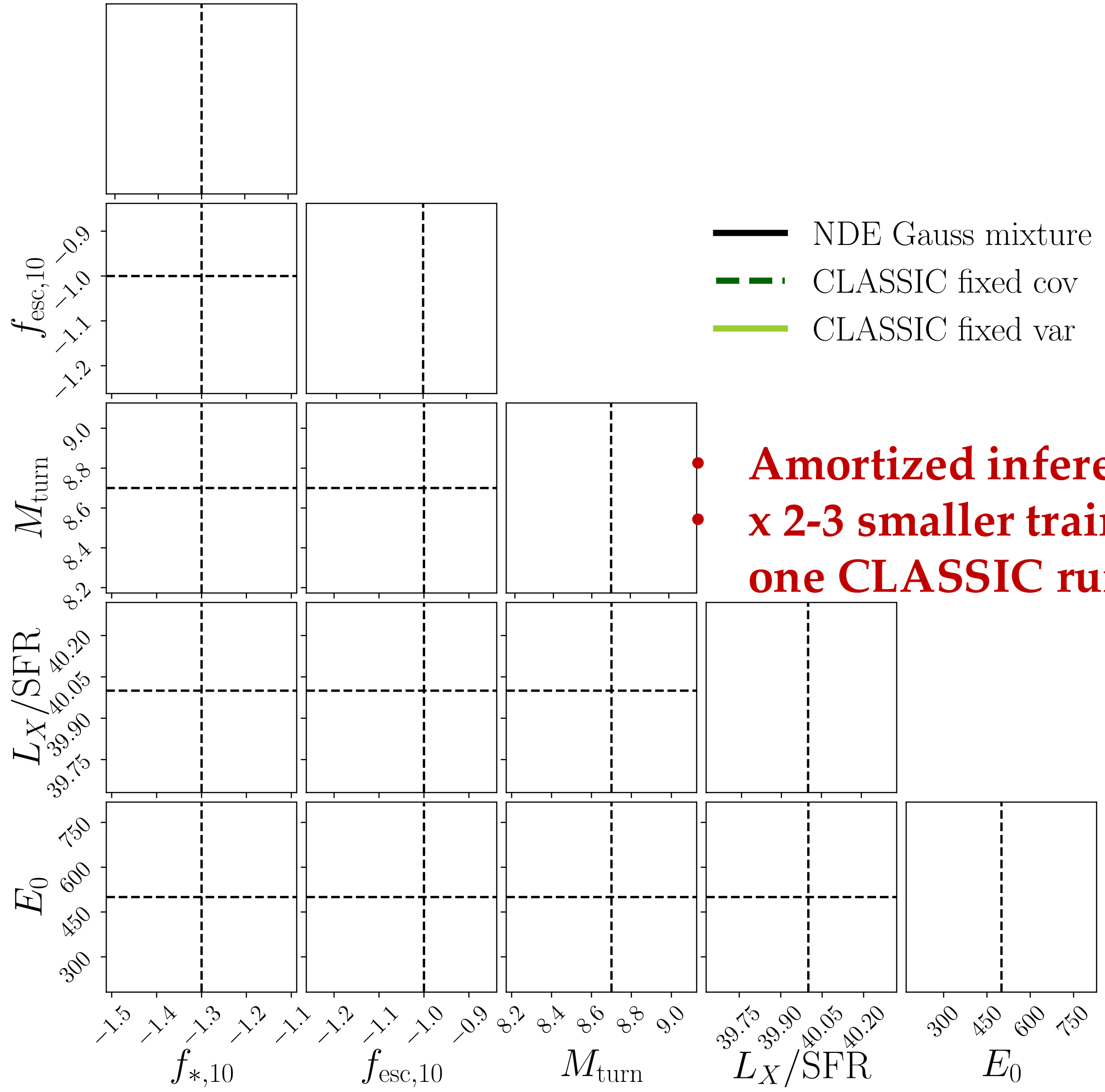
Including more realistic likelihood
≠
more constraining posterior

Prelogović & Mesinger 2023

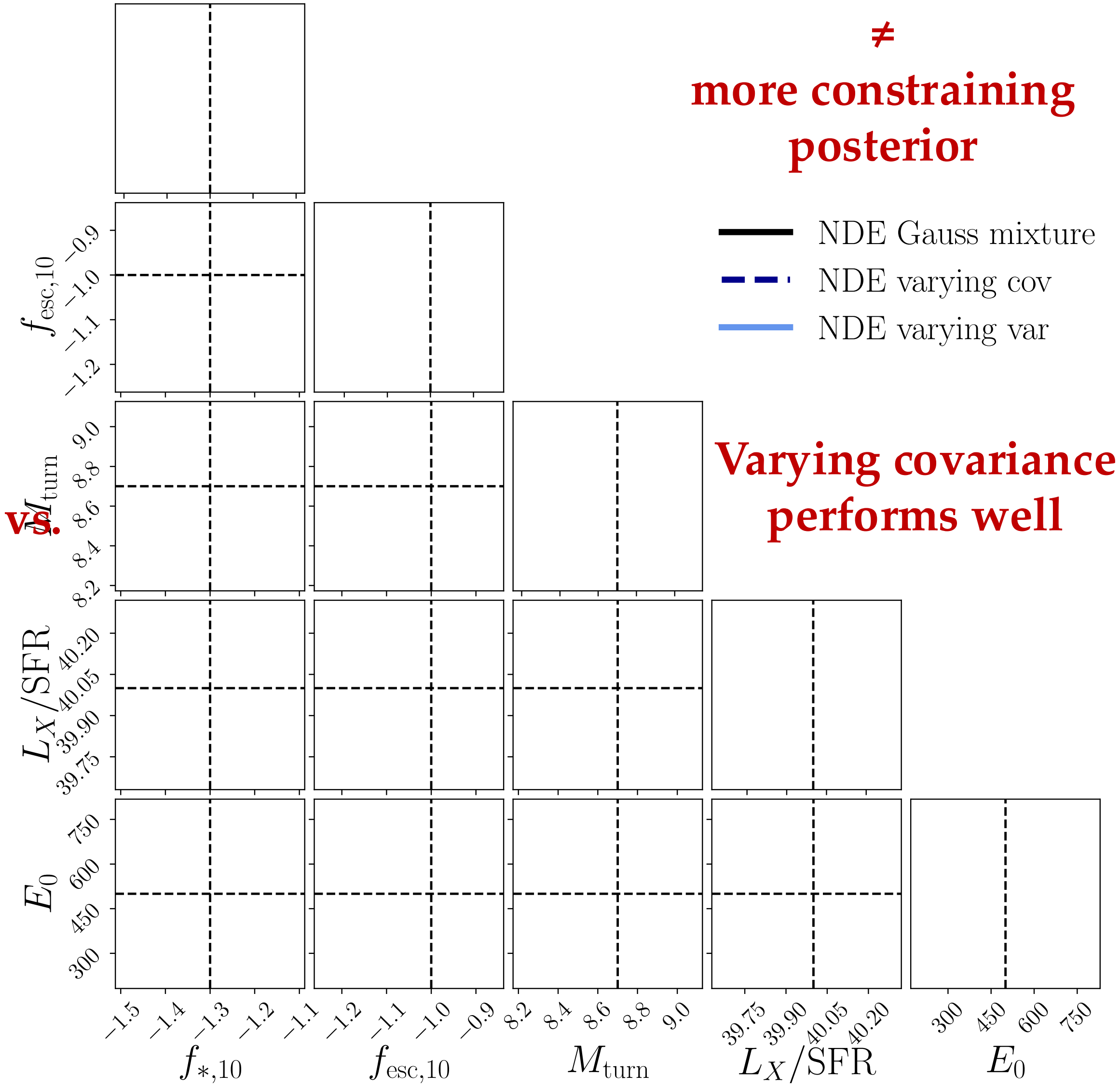
SCUOLA NORMALE SUPERIORE

4. Results

Including more realistic likelihood ≠ more constraining posterior



- **Amortized inference**
- **x 2-3 smaller training DB vs. one CLASSIC run**



Varying covariance performs well

4. Results

BUT:

This is only qualitative description, and only for the mock observation

- How does it perform for other points in the parameter space?
- Did the training converge?
- Can we quantify the best model?

–> Simulation Based Calibration

SCUOLA
NORMALE
SUPERIORE

5. Simulation Based Calibration (SBC)

- ***"prior" = "data averaged posterior"*** $P(\boldsymbol{\theta}) = \int P(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})\, P(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{\theta}}) P(\tilde{\boldsymbol{\theta}})\, \mathrm{d}\tilde{\boldsymbol{y}}\, \mathrm{d}\tilde{\boldsymbol{\theta}}$

5. Simulation Based Calibration (SBC)

- ***"prior" = "data averaged posterior"*** $P(\boldsymbol{\theta}) = \int P(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})\, P(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{\theta}}) P(\tilde{\boldsymbol{\theta}})\, \mathrm{d}\tilde{\boldsymbol{y}}\, \mathrm{d}\tilde{\boldsymbol{\theta}}$

1. Pull from prior $\tilde{\boldsymbol{\theta}} \sim P(\boldsymbol{\theta})$

5. Simulation Based Calibration (SBC)

- ***"prior" = "data averaged posterior"*** $P(\boldsymbol{\theta}) = \int P(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})\, P(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{\theta}})P(\tilde{\boldsymbol{\theta}})\,\mathrm{d}\tilde{\boldsymbol{y}}\,\mathrm{d}\tilde{\boldsymbol{\theta}}$

1. Pull from prior $\tilde{\boldsymbol{\theta}} \sim P(\boldsymbol{\theta})$
2. Pull the data from the likelihood $\tilde{\boldsymbol{y}} \sim P(\boldsymbol{y}|\tilde{\boldsymbol{\theta}}) \quad \Leftrightarrow \quad \tilde{\boldsymbol{y}} = \mathrm{simulator}(\tilde{\boldsymbol{\theta}})$

5. Simulation Based Calibration (SBC)

- ***"prior" = "data averaged posterior"*** $P(\boldsymbol{\theta}) = \int P(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})\, P(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{\theta}}) P(\tilde{\boldsymbol{\theta}})\, \mathrm{d}\tilde{\boldsymbol{y}}\, \mathrm{d}\tilde{\boldsymbol{\theta}}$

1. Pull from prior $\tilde{\boldsymbol{\theta}} \sim P(\boldsymbol{\theta})$
2. Pull the data from the likelihood $\tilde{\boldsymbol{y}} \sim P(\boldsymbol{y}|\tilde{\boldsymbol{\theta}}) \quad \Leftrightarrow \quad \tilde{\boldsymbol{y}} = \mathrm{simulator}(\tilde{\boldsymbol{\theta}})$
3. Calculate the posterior the sample $P(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})$

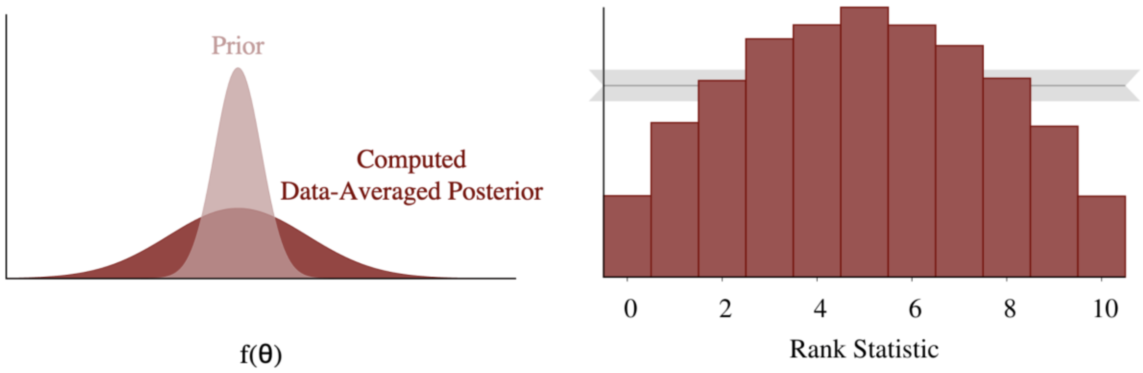
5. Simulation Based Calibration (SBC)

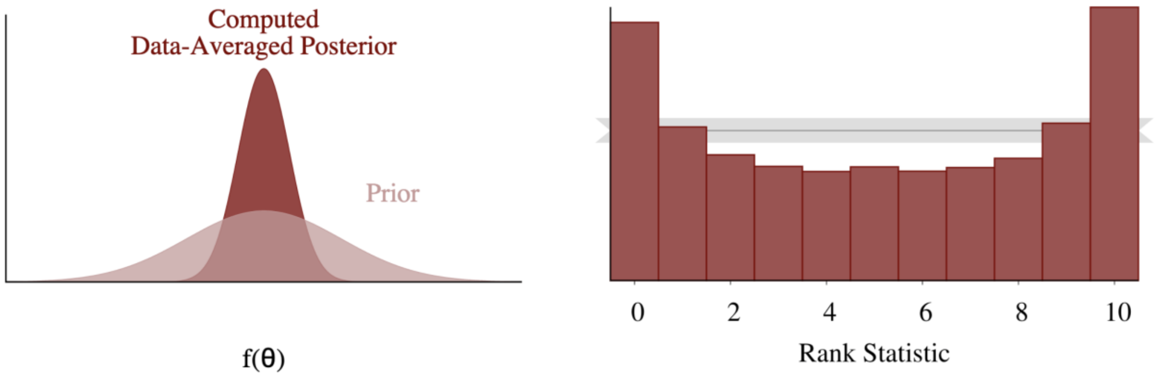
- ***"prior" = "data averaged posterior"*** $P(\boldsymbol{\theta}) = \int P(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})\, P(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{\theta}}) P(\tilde{\boldsymbol{\theta}})\, \mathrm{d}\tilde{\boldsymbol{y}}\, \mathrm{d}\tilde{\boldsymbol{\theta}}$

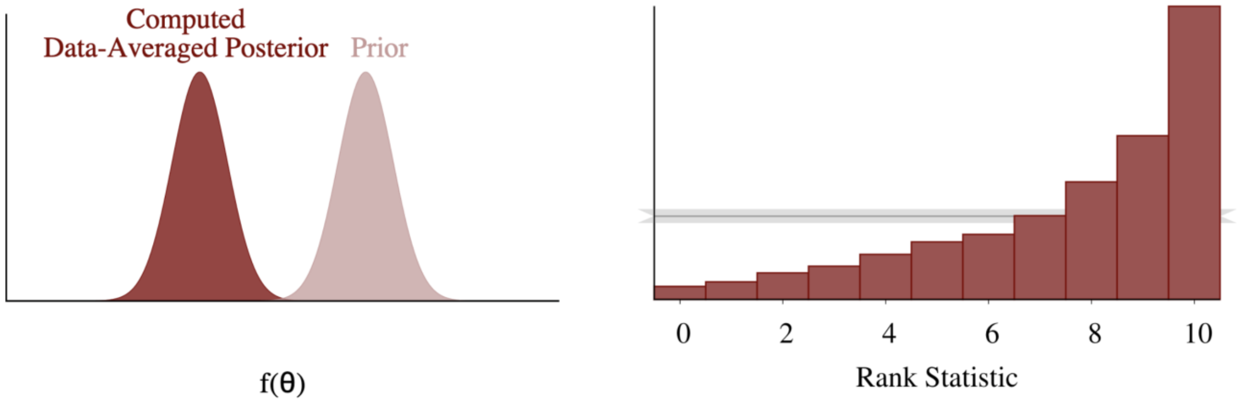
1. Pull from prior $\tilde{\boldsymbol{\theta}} \sim P(\boldsymbol{\theta})$
2. Pull the data from the likelihood $\tilde{\boldsymbol{y}} \sim P(\boldsymbol{y}|\tilde{\boldsymbol{\theta}}) \quad \Leftrightarrow \quad \tilde{\boldsymbol{y}} = \mathrm{simulator}(\tilde{\boldsymbol{\theta}})$
3. Calculate the posterior the sample $P(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})$
4. Repeat and average posteriors $P(\boldsymbol{\theta}) \approx \frac{1}{N} \sum_{i=1}^{N} P_i(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}_i)$

5. Simulation Based Calibration (SBC)

- ***"prior" = "data averaged posterior"*** $P(\boldsymbol{\theta}) = \int P(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})\, P(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{\theta}})P(\tilde{\boldsymbol{\theta}})\, \mathrm{d}\tilde{\boldsymbol{y}}\, \mathrm{d}\tilde{\boldsymbol{\theta}}$
- SBC – casting integral into 1D rank statistics distribution







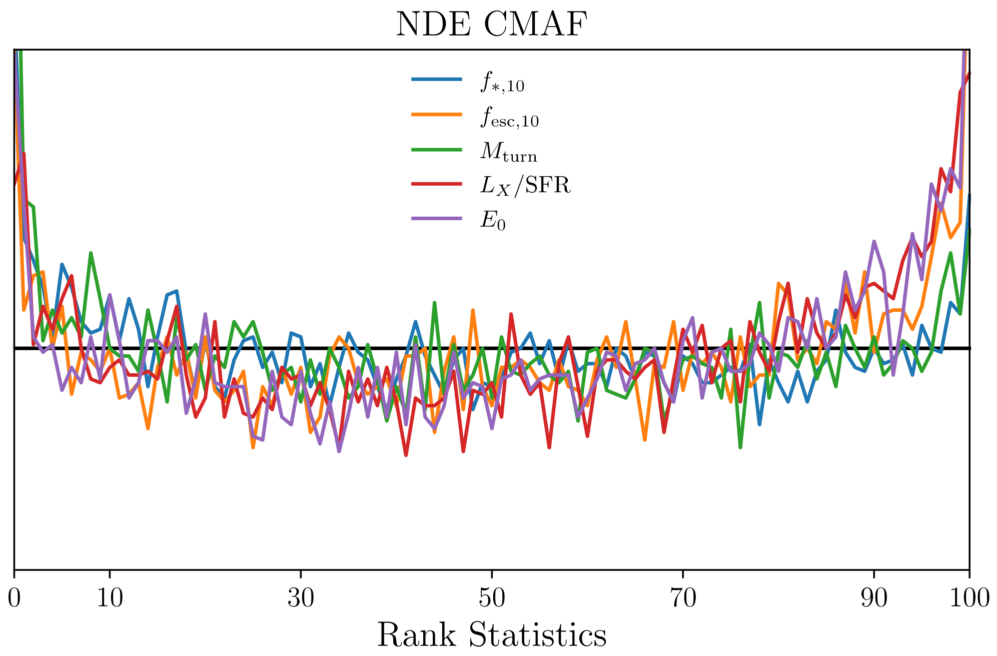
Talts+2018

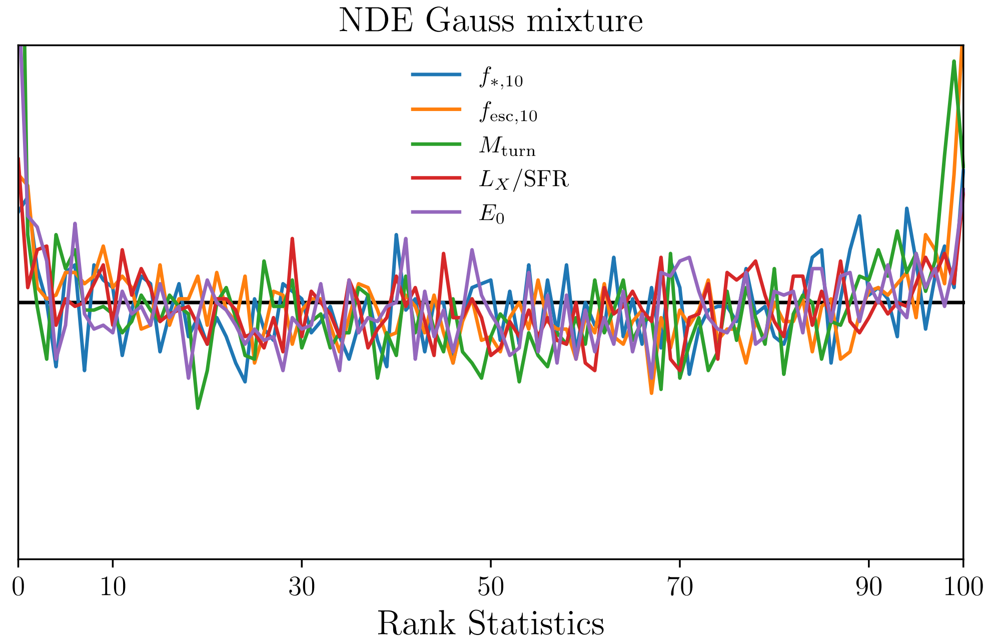
6. SBC for 21-cm PS

- 10 000 posteriors

- Would be useful for classic inference, but is too expensive to compute

- NDE Gauss mixture – the best





Prelogović & Mesinger 2023

Conclusions

- SBI – current and future frontier in the 21-cm inference
  - Cheaper and more precise, by recovering a data-driven likelihood
  - Convergence / performance tests crucial!

How informative are summaries of the 21-cm signal?

1. Fisher information matrix

- If we label data space as $\boldsymbol{d}$ and its likelihood as $P(\boldsymbol{d}|\boldsymbol{\theta})$

$$\boldsymbol{F}(\boldsymbol{\theta}^*)_{mn} = \mathrm{E}_{P(\boldsymbol{d}|\boldsymbol{\theta}^*)}\left[\frac{\partial}{\partial \boldsymbol{\theta}_m} \ln P(\boldsymbol{d}|\boldsymbol{\theta}^*) \cdot \frac{\partial}{\partial \boldsymbol{\theta}_n} \ln P(\boldsymbol{d}|\boldsymbol{\theta}^*)\right]$$

- The usefulness comes from

$$\text{1D: } \mathrm{Var}\left(\hat{\boldsymbol{\theta}}_m\right) \geq (\boldsymbol{F}^{-1})_{mm} \qquad \text{ND: } \det \mathrm{Cov}(\hat{\boldsymbol{\theta}}) \geq \det \boldsymbol{F}^{-1}$$

How well we can estimate a parameter is fundamentally limited by its Fisher information.

(i.e. one cannot go below it)

Fisher 1935

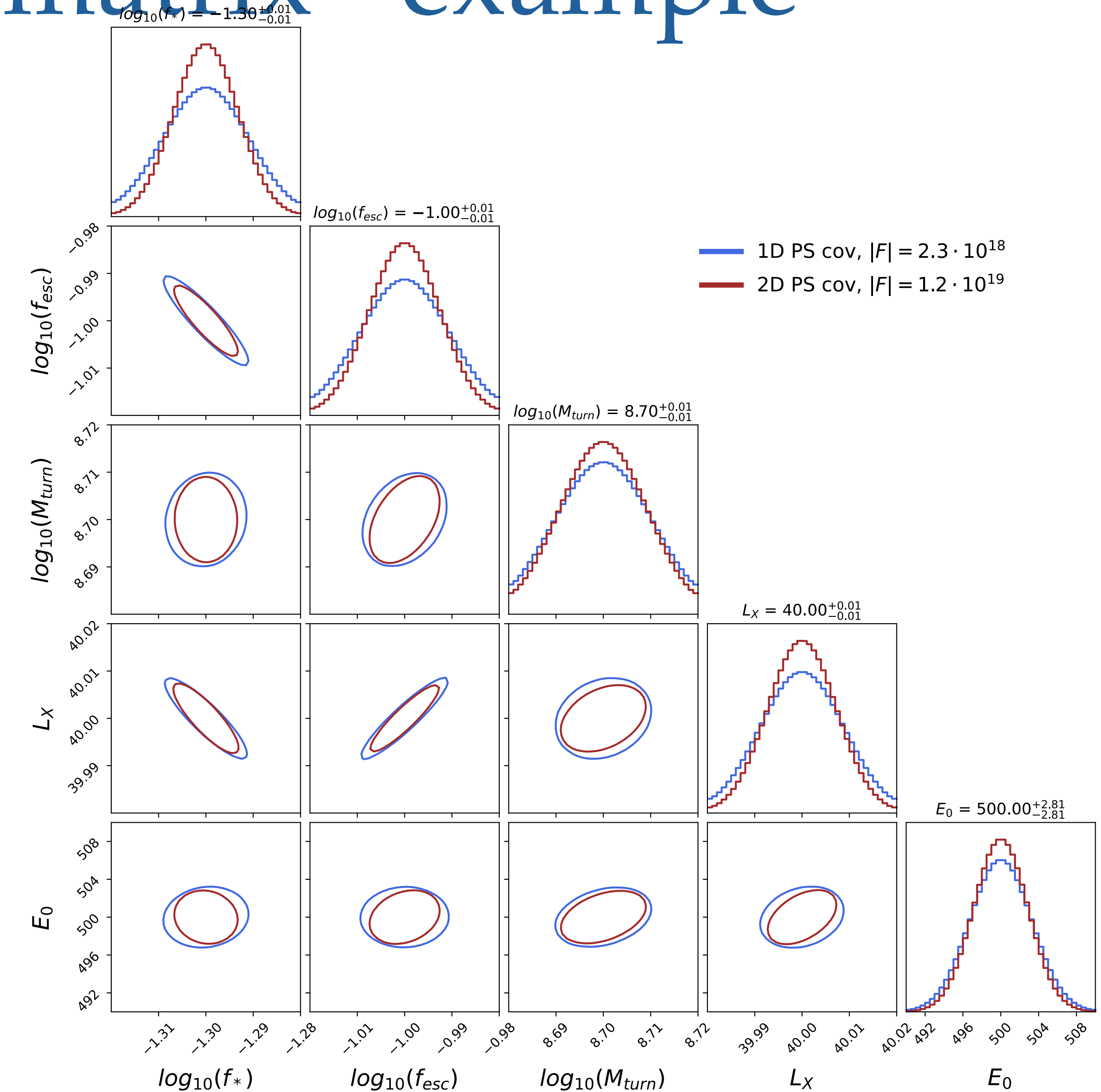
SCUOLA NORMALE SUPERIORE

1. Fisher information matrix - example

- We cannot perform better than the shown ellipse

- Different summary, different Fisher matrix

- det $\mathbf{F}^{-1}$ = volume of the ellipse
  - det $\mathbf{F}^{-1}$ smaller the better
  - det $\mathbf{F}$ bigger the better



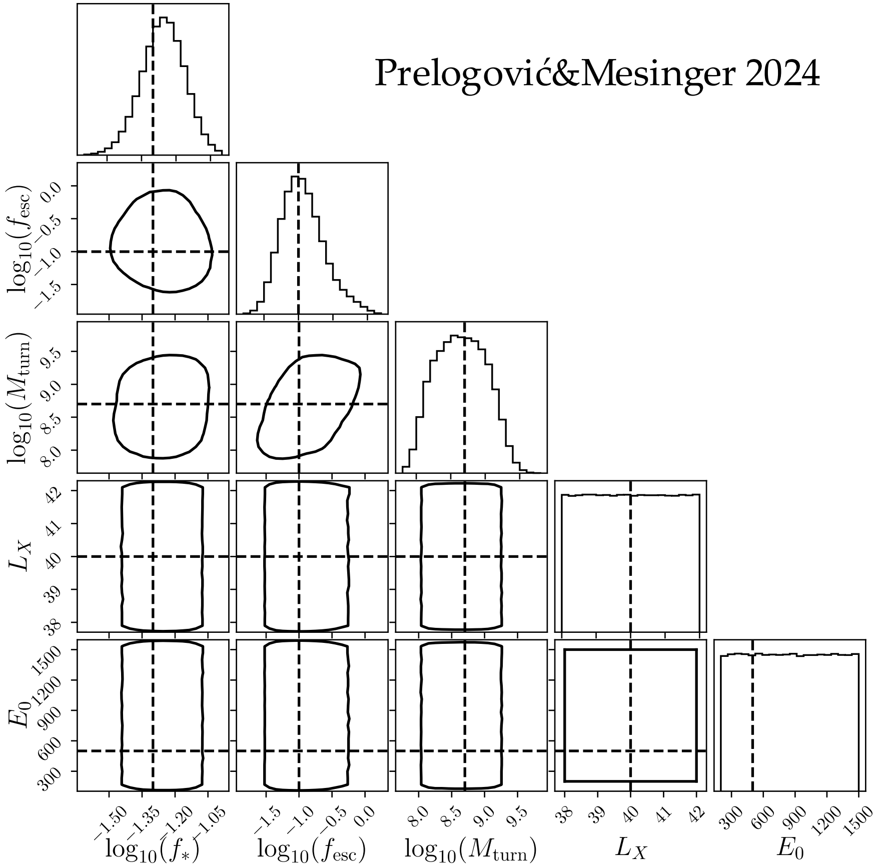
3. Distribution of the Fisher information

- $\det \boldsymbol{F}(\boldsymbol{\theta}^*)$ is information measure just around one point
- Calculating around many different points is better

3. Distribution of the Fisher information

- $\det \boldsymbol{F}(\boldsymbol{\theta}^{*})$ is information measure just around one point
- Calculating around many different points is better

- Sample ~150 points from the prior
- Around each point construct simulations. needed to compute the Fisher matrix

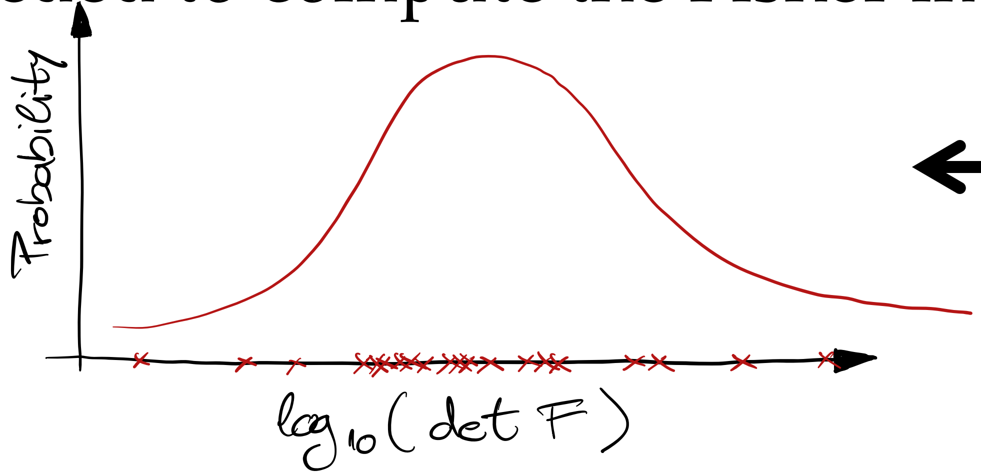

Prelogović&Mesinger 2024

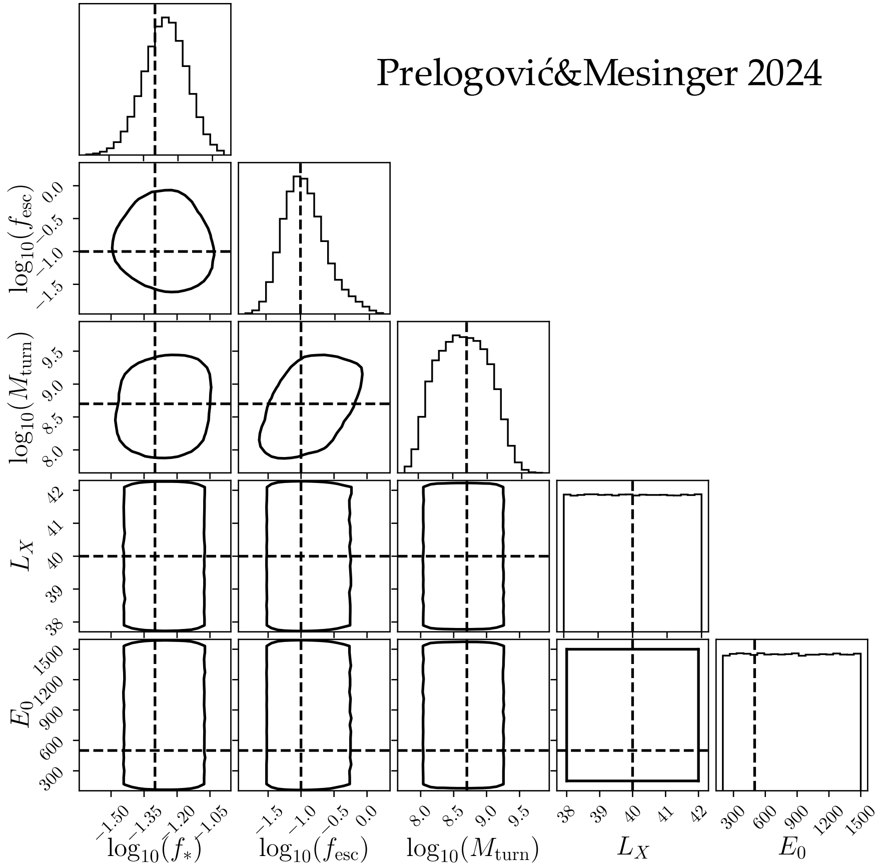
SCUOLA
NORMALE
SUPERIORE

3. Distribution of the Fisher information

- $\det \boldsymbol{F}(\boldsymbol{\theta}^*)$ is information measure just around one point
- Calculating around many different points is better

- Sample ~150 points from the prior
- Around each point construct simulations. needed to compute the Fisher matrix



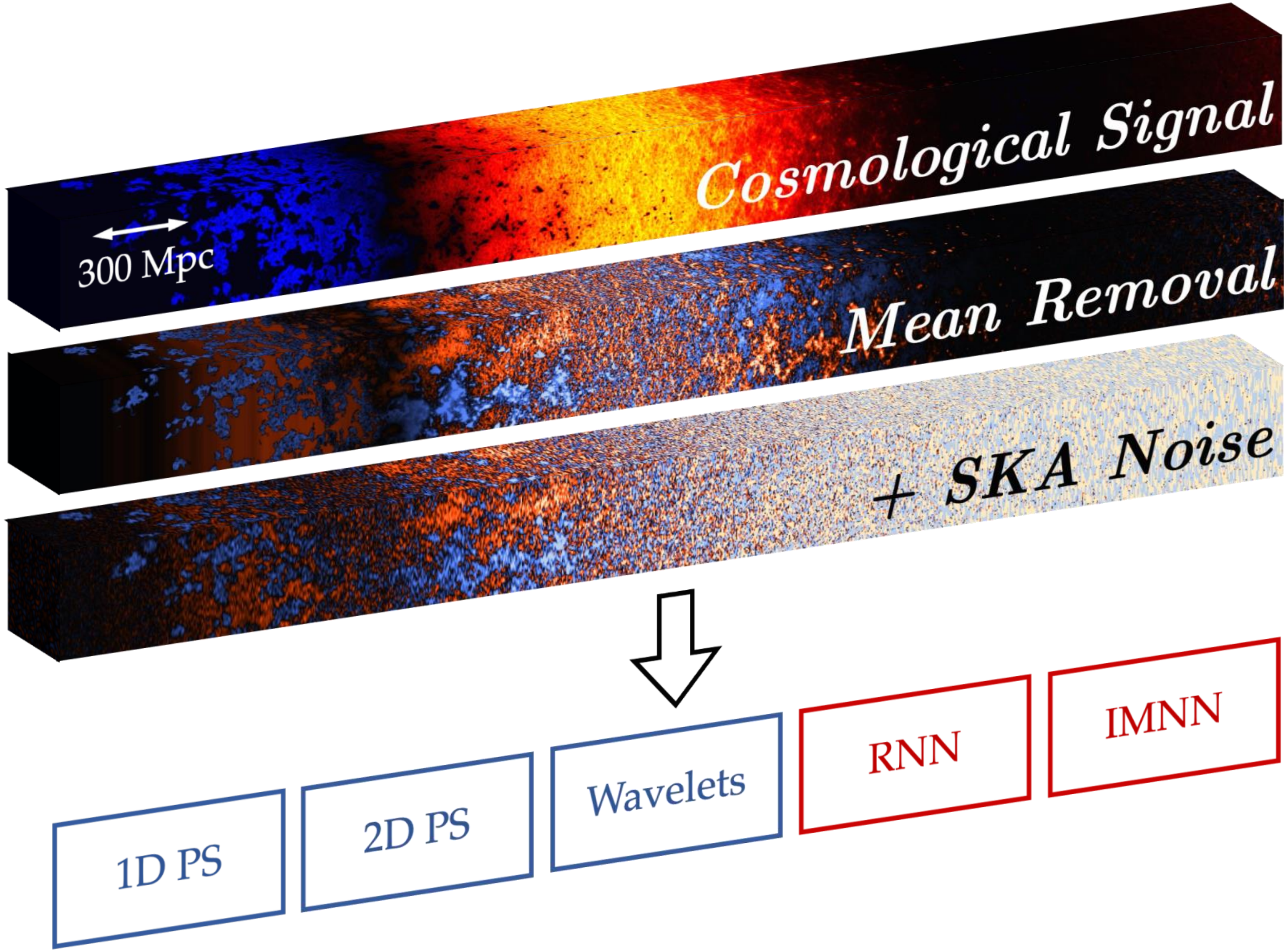


Prelogović&Mesinger 2024

Ce Sui+2023

Later talk!

4. Considered summaries



Prelogović&Mesinger 2024

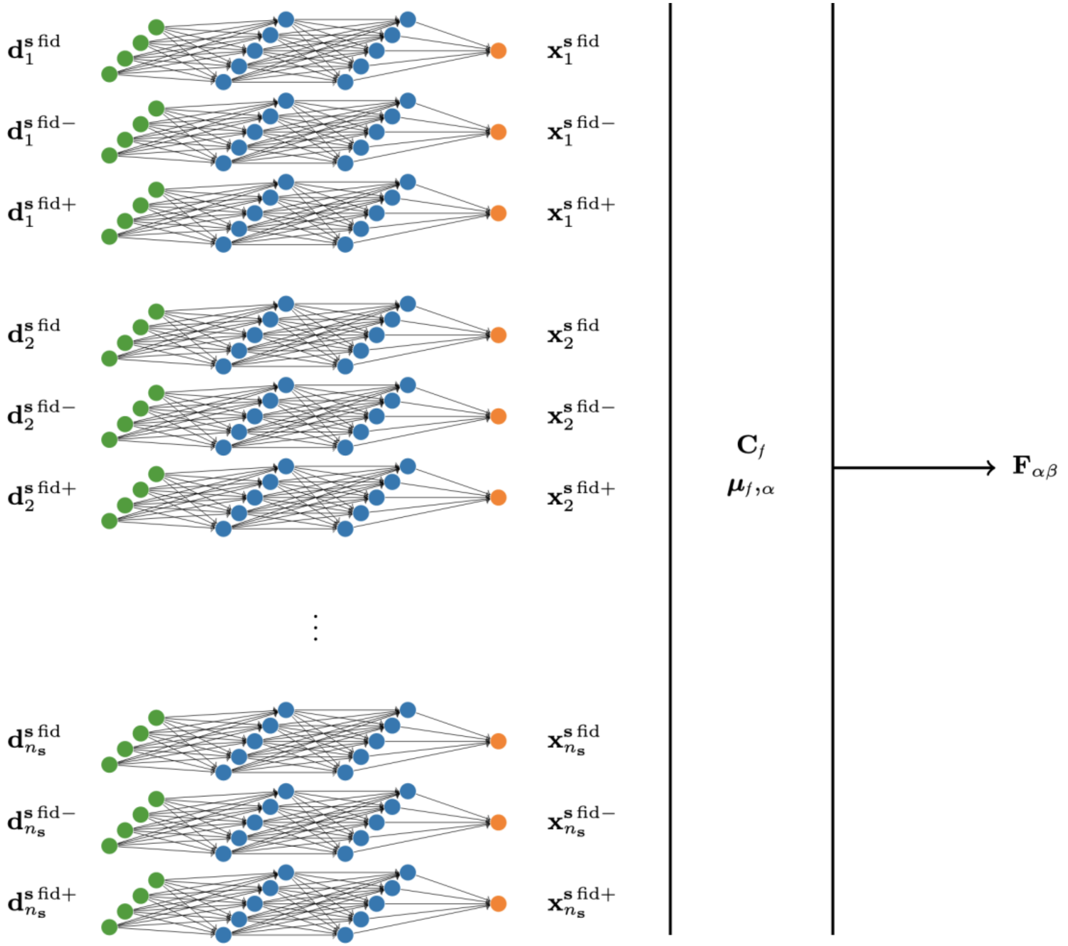
4.1 Information Maximizing NN

- Unsupervised algorithm
- Simulate the data at a fiducial parameter set: $\mathbf{d}(\boldsymbol{\theta}_{\text{fid}})$
- Simulate around the fiducial parameters: $\mathbf{d}(\boldsymbol{\theta}_{\text{fid}}^{+}), \mathbf{d}(\boldsymbol{\theta}_{\text{fid}}^{-})$
- Calculate compressed summary:

$$\boldsymbol{s} = NN(\boldsymbol{d})$$

- Maximize Fisher information:

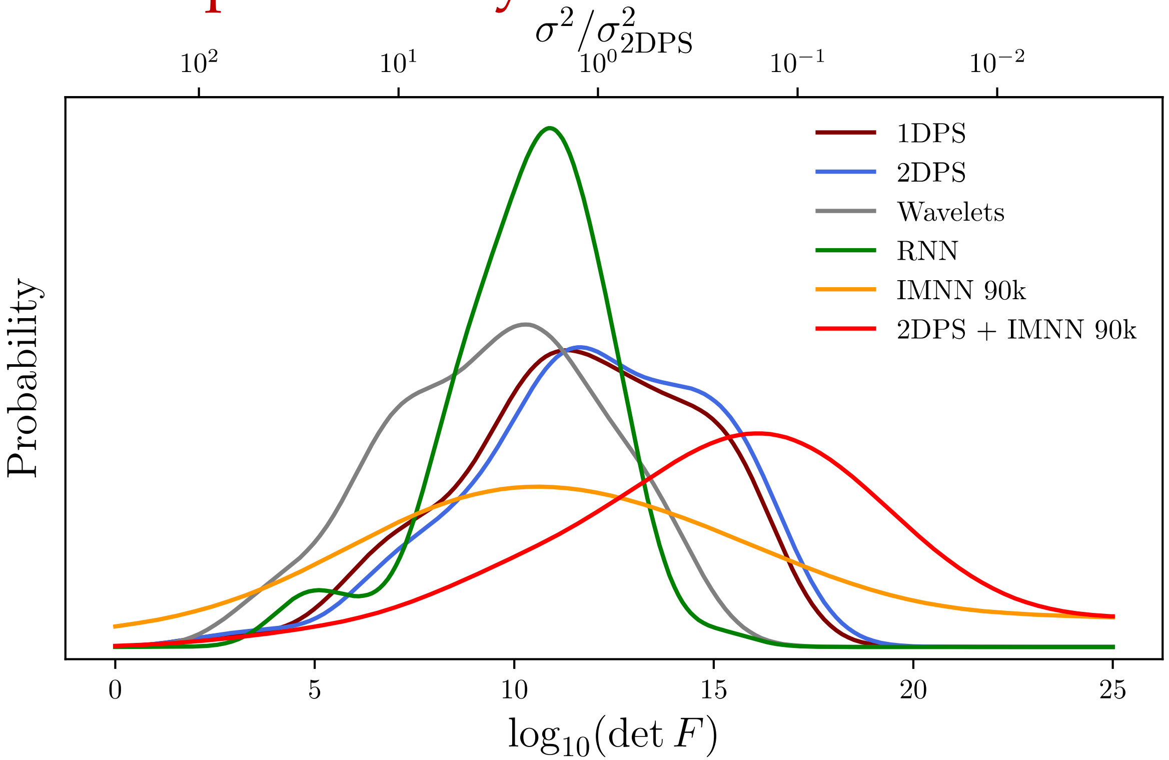
$$\mathcal{L} = -\ln(\det \boldsymbol{F})$$


$\mathbf{d}_1^{\mathbf{s}\,\text{fid}}$
$\mathbf{x}_1^{\mathbf{s}\,\text{fid}}$
$\mathbf{d}_1^{\mathbf{s}\,\text{fid}-}$
$\mathbf{x}_1^{\mathbf{s}\,\text{fid}-}$
$\mathbf{d}_1^{\mathbf{s}\,\text{fid}+}$
$\mathbf{x}_1^{\mathbf{s}\,\text{fid}+}$
$\mathbf{d}_2^{\mathbf{s}\,\text{fid}}$
$\mathbf{x}_2^{\mathbf{s}\,\text{fid}}$
$\mathbf{d}_2^{\mathbf{s}\,\text{fid}-}$
$\mathbf{x}_2^{\mathbf{s}\,\text{fid}-}$
$\mathbf{d}_2^{\mathbf{s}\,\text{fid}+}$
$\mathbf{x}_2^{\mathbf{s}\,\text{fid}+}$
$\vdots$
$\mathbf{d}_{n_\mathbf{s}}^{\mathbf{s}\,\text{fid}}$
$\mathbf{x}_{n_\mathbf{s}}^{\mathbf{s}\,\text{fid}}$
$\mathbf{d}_{n_\mathbf{s}}^{\mathbf{s}\,\text{fid}-}$
$\mathbf{x}_{n_\mathbf{s}}^{\mathbf{s}\,\text{fid}-}$
$\mathbf{d}_{n_\mathbf{s}}^{\mathbf{s}\,\text{fid}+}$
$\mathbf{x}_{n_\mathbf{s}}^{\mathbf{s}\,\text{fid}+}$
$\mathbf{C}_f$
$\boldsymbol{\mu}_{f,\alpha}$
$\mathbf{F}_{\alpha\beta}$

Charnock+2018
Makinen+2023
Prelogović+2024
Maitra+2024

5. Results

- 1DPS and 2DPS clear winners
- Combining 2DPS + IMNN
  - IMNN extracts complementary information to the PS



Prelogović&Mesinger 2024

Conclusions

- SBI – current and future frontier in the 21-cm inference
  - Cheaper and more precise, by recovering a data-driven likelihood
  - Convergence / performance tests crucial!

- Fisher distribution – information-based metric for a summary quality
  - Hard to beat the PS
  - Combination of classical + neural summaries as a powerful way forward

Thank you!