# Simulation-based inference pipelines for Euclid data

Florent Leclercq

www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

In collaboration with
the Aquila Consortium

www.aquila-consortium.org

**21 June 2023**

# Markov Chain Monte Carlo is 70 years old!


THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

## Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
*Los Alamos Scientific Laboratory, Los Alamos, New Mexico*

AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

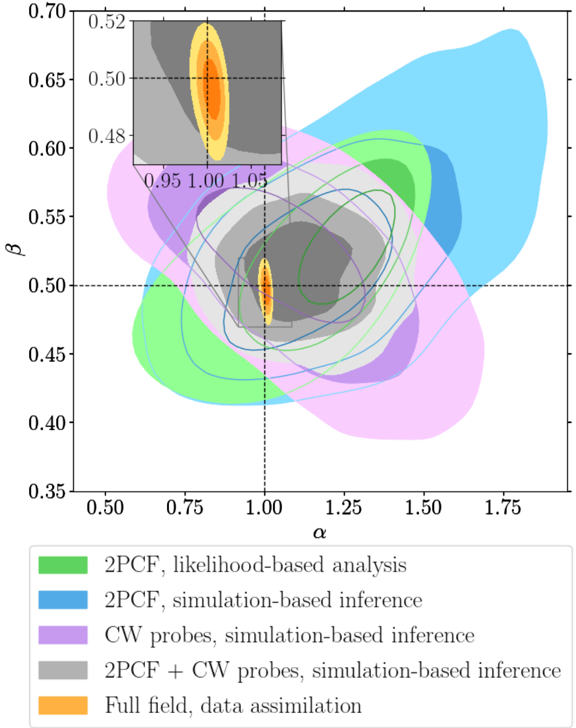
(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

# Why I decided to go "likelihood-free" for galaxy clustering additional probes

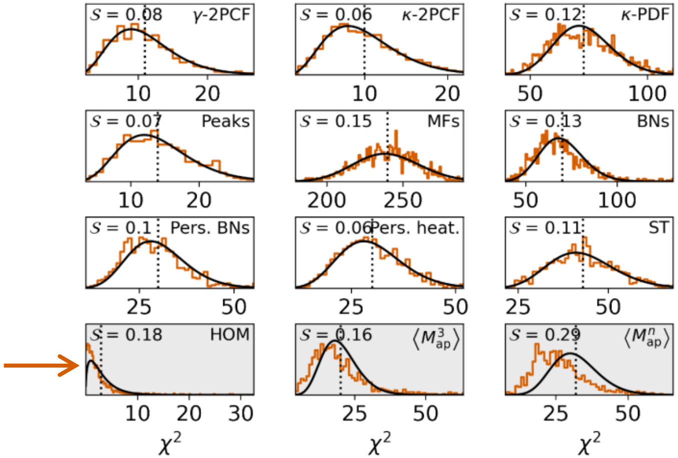
Note: likelihood-free inference ≈ simulation-based inference ≈ implicit likelihood inference

- A question of accuracy: first, avoid biases.




FL & Heavens, 2103.04158

- Some WL additional probes also have a non-Gaussian distribution.

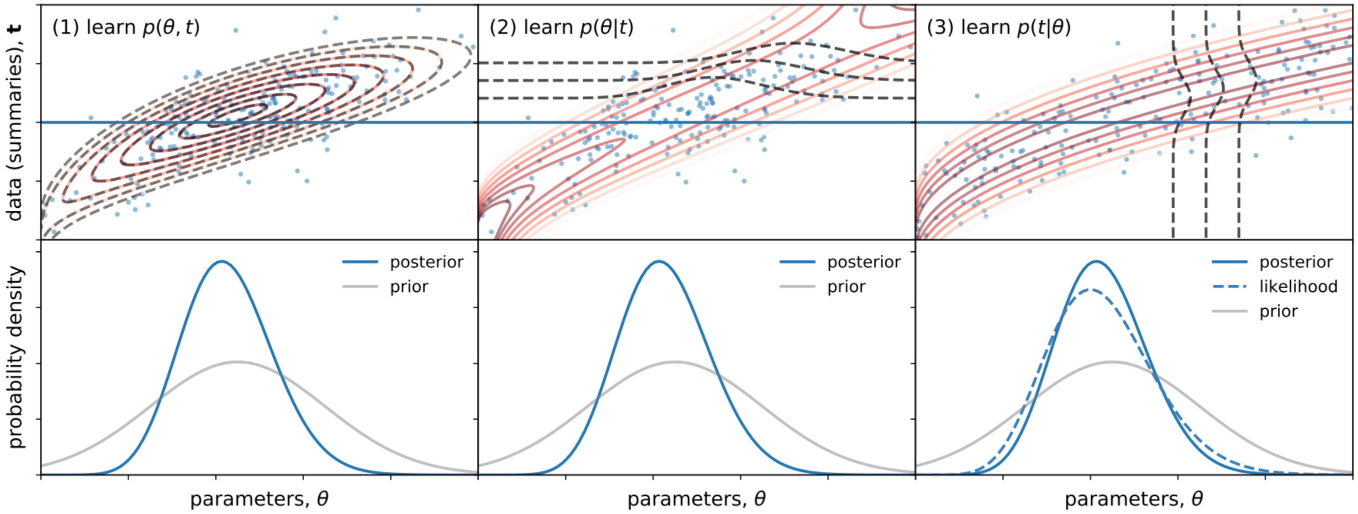



- A question of precision: can numerical forward models be used to push further than $k \gtrsim 0.15\ h/\mathrm{Mpc}$? The full field contains much more information.

Euclid HOWLS-KP paper 1, Ajani *et al.*, 2301.12890

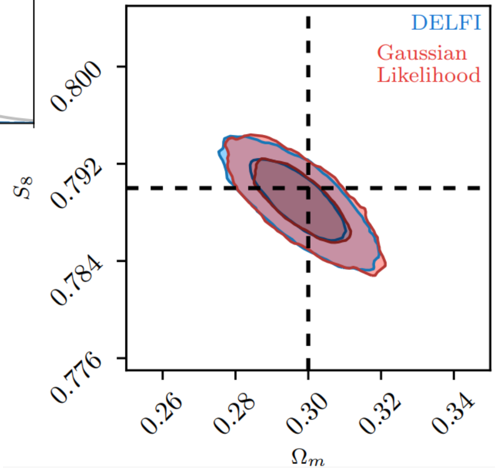
# DELFI (*Density Estimation Likelihood-Free Inference*): Inference from cosmic shear log-normal forward models

(WL: Forward Modelling)




Alsing, Wandelt & Feeney 2018, 1801.01497
Alsing, Charnock, Feeney & Wandelt 2019, 1903.00007

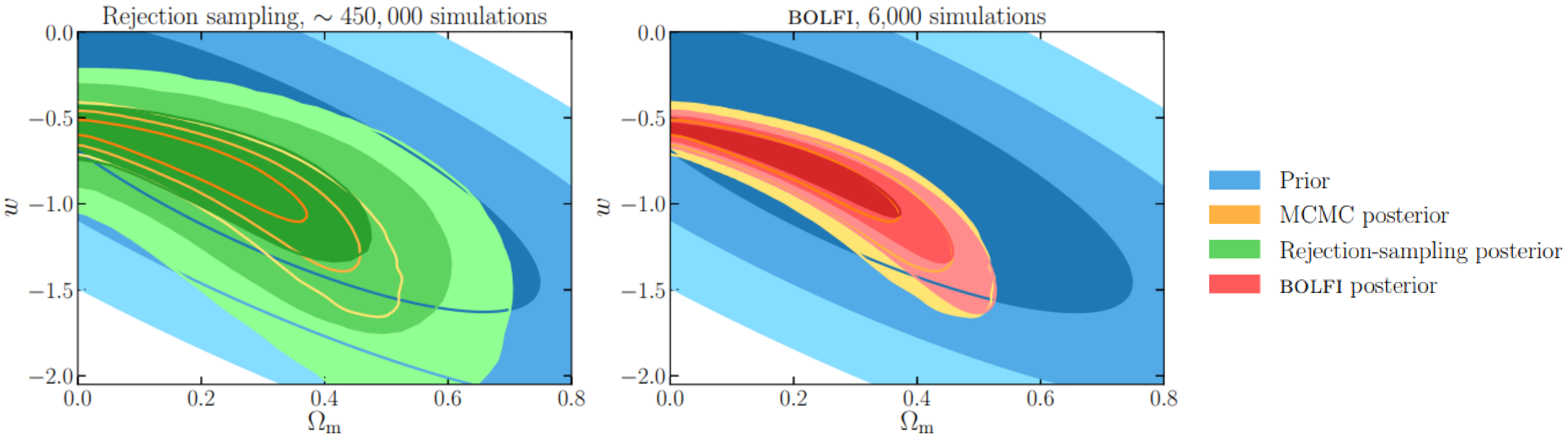
Taylor *et al.* 2019, 1904.05364



# BOLFI (*Bayesian Optimisation for Likelihood-Free Inference*): Re-analysis of the JLA supernova sample

(GC: Additional Probes)

Betoule et al. 2014, 1401.4064

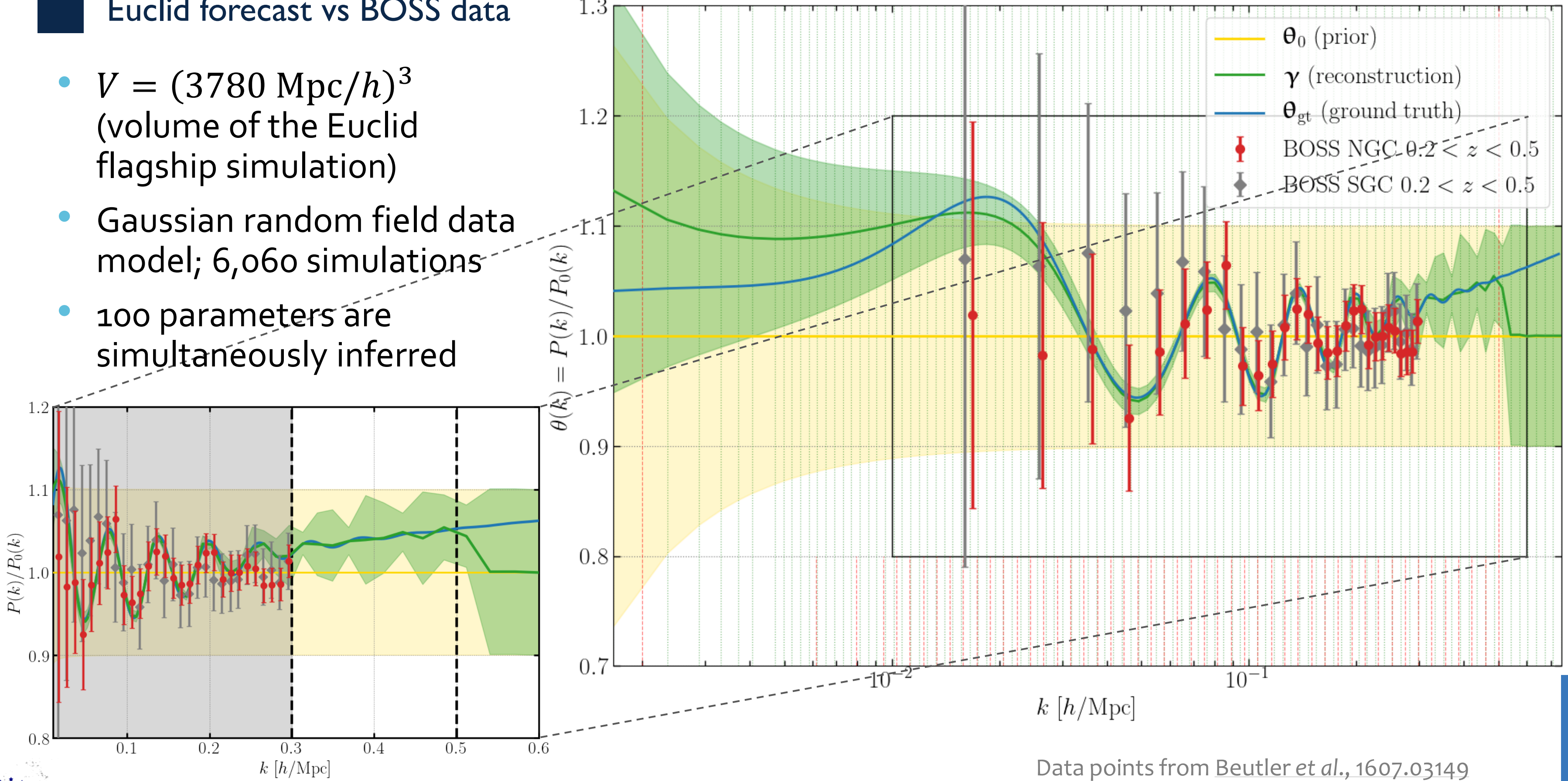



- The number of required simulations is reduced by 2 to 3 orders of magnitude with respect to likelihood-free rejection sampling or MCMC
- Bayesian optimisation can also be applied to the “true” likelihood (if known) or to iteratively build an emulator of the data model

FL 2018, 1805.07152

# SELFI (*Simulator Expansion for Likelihood-Free Inference*): Euclid forecast vs BOSS data

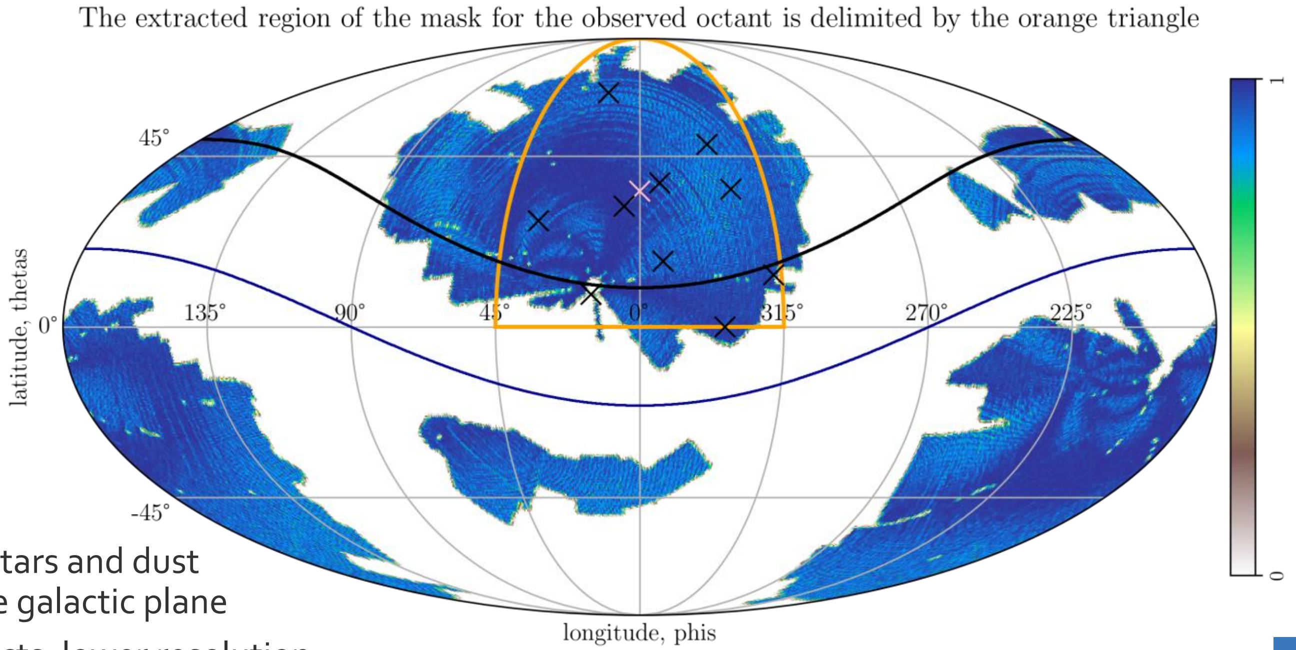
- $V = (3780\ \mathrm{Mpc}/h)^3$ (volume of the Euclid flagship simulation)
- Gaussian random field data model; 6,060 simulations
- 100 parameters are simultaneously inferred




Data points from Beutler *et al.*, 1607.03149

(GC: Additional Probes)

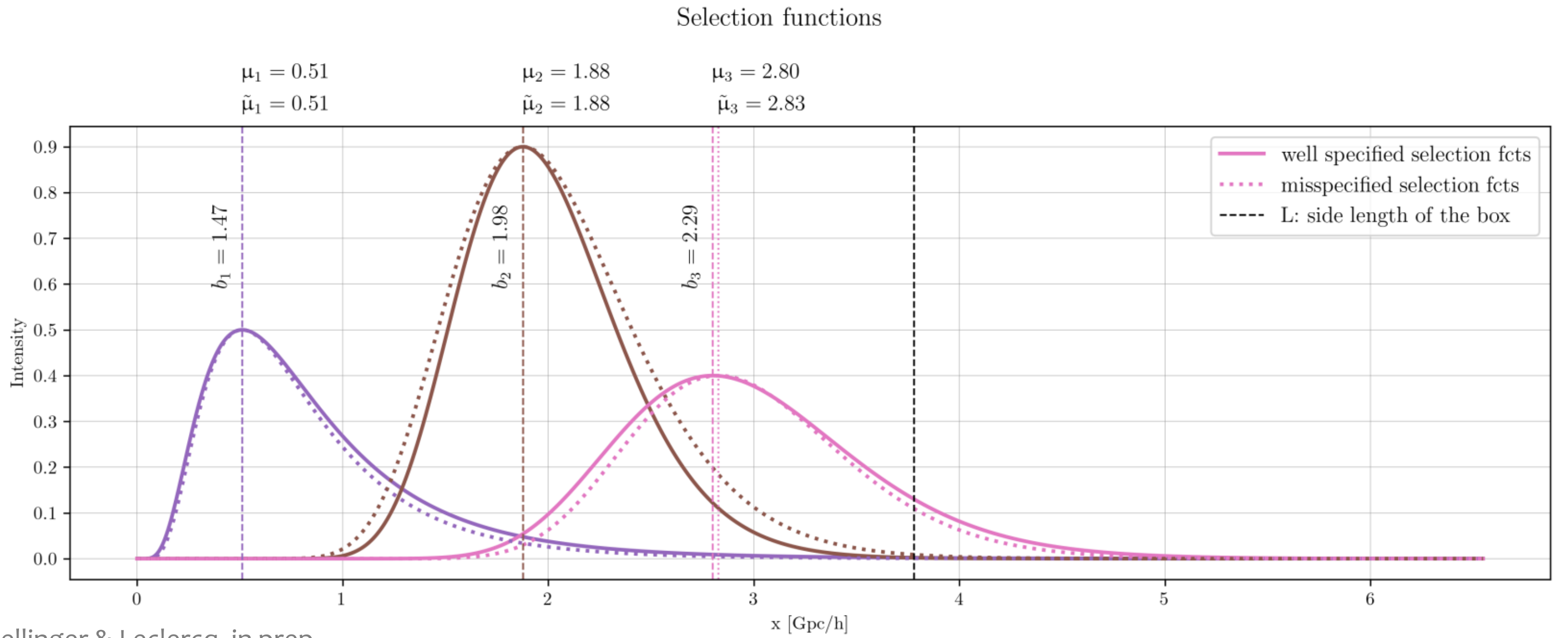
- $V = (3780\ \mathrm{Mpc}/h)^3$ cubic box, covering one octant of the sky




- Two models:
  - Model A: 10 masked stars and dust extinction close to the galactic plane
  - Model B: no such effects, lower resolution

Hoellinger & Leclercq, in prep.

- Two models:
  - Model A: lognormal selection functions, luminosity-dependent galaxy bias
  - Model B: misspecified selection functions and galaxy biases

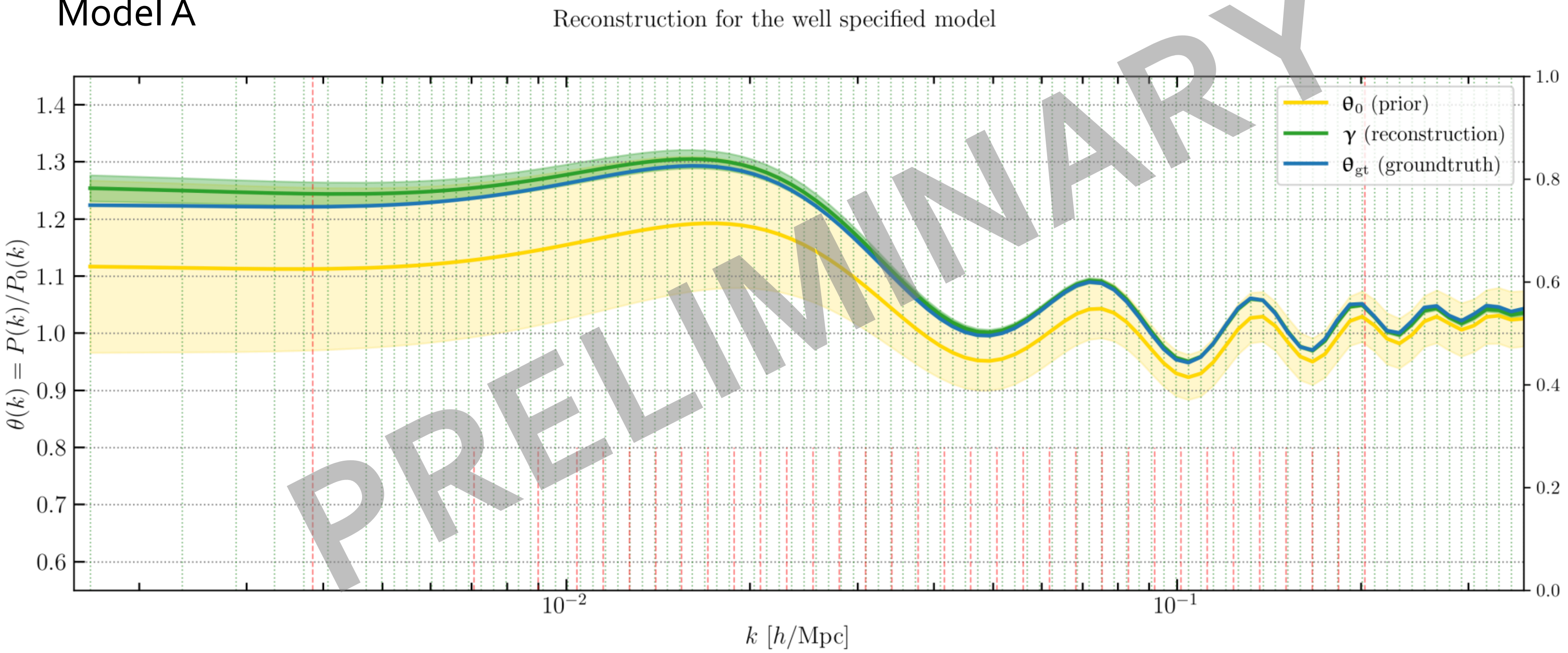



Hoellinger & Leclercq, in prep.

# SELFI posterior: reconstructed initial matter power spectrum

(GC: Additional Probes)

## Model A

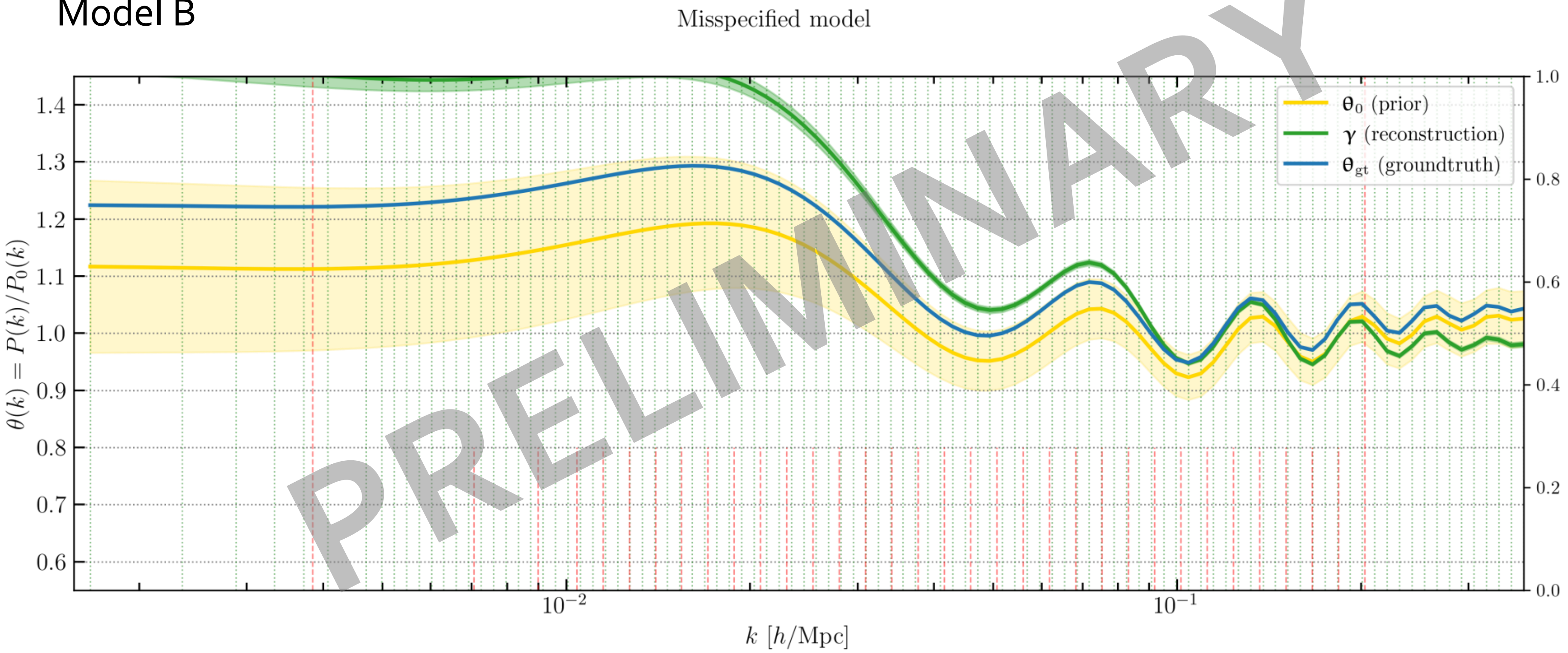



Hoellinger & Leclercq, in prep.

# SELFI posterior: reconstructed initial matter power spectrum

(GC: Additional Probes)

Model B




Hoellinger & Leclercq, in prep.

# Data compression for SBI

- Good old techniques are still quite useful…

- MOPED/score compression: compresses *n* data points to *p* summaries (where p is the number of parameters).

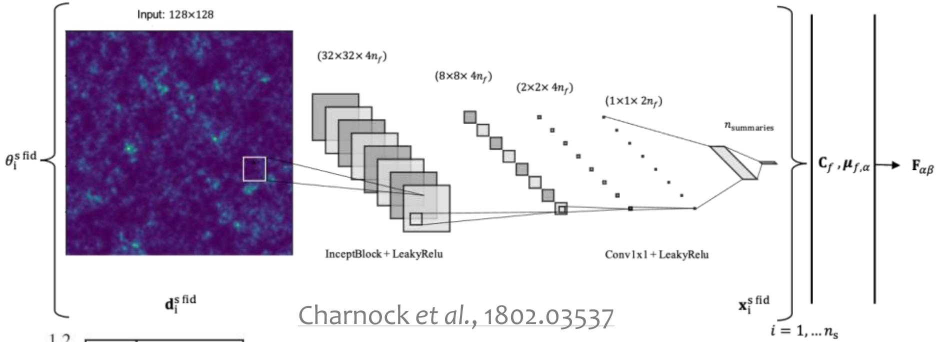
$$\mathbf{F} \equiv -\mathrm{E}_{\boldsymbol{\theta}}\left[\nabla\nabla^{\mathrm{T}}\mathcal{L}\right] = \mathrm{E}_{\boldsymbol{\theta}}\left[\nabla\mathcal{L}\nabla^{\mathrm{T}}\mathcal{L}\right]$$

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_* + \mathbf{F}_*^{-1}\nabla\mathcal{L}_*$$

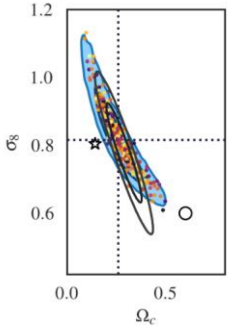
- Score compression is lossless linear data compression (optimal at first-order in the log-likelihood)

Heavens, Jimenez & Lahav 2000, astro-ph/9911102
Alsing & Wandelt 2018, 1712.00012

- Non-linear data compression: Can we saturate the information content thanks to machine learning?




Charnock *et al.*, 1802.03537




- IMNN compression gives summaries that are nearly maximally informative compared to theoretical estimates

Makinen *et al.*, 2107.07405