Implicit Likelihood Inference in Cosmology while efficiently checking for survey systematics

National Action Dark Energy Colloquium 2023
LAPP, Annecy

Tristan Hoellinger

hoellin.github.io

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

In collaboration with:
Florent Leclercq (IAP – CNRS & Sorbonne Université)

and the Aquila Consortium

www.aquila-consortium.org

5 November 2023

Challenges

- Systematic effects

  Large surveys (Euclid, LSST) will be dominated by systematic rather than statistical uncertainty.

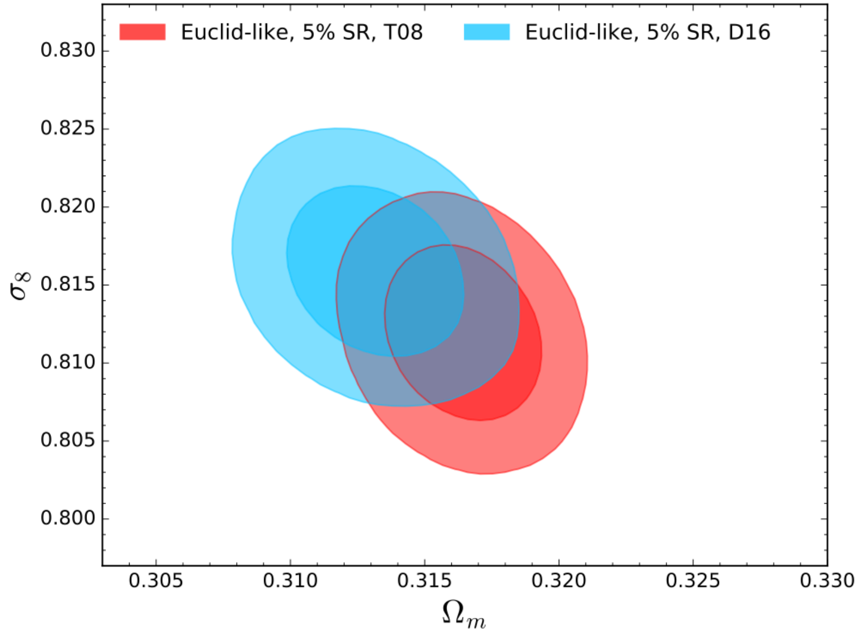
- How to integrate prior knowledge?
  - Prior on cosmology $\boldsymbol{\omega} = (h,\, \Omega_\mathrm{b},\, \Omega_\mathrm{m},\, n_\mathrm{S},\, \sigma_8)$
  - Prior on non measurable physical quantities
    - The initial matter power spectrum θ should be close to $\theta_{\mathrm{Planck,2018}}$: can we benefit from this?
    - Can we benefit from theoretical insights on θ?

Jasche & Lavaux 2017, 1706.08971

Laureijs et al., 2011, 1110.3193

- Model misspecification in Bayesian & simulation-based inference (SBI)
  - When model differ from data-generating process
  - Biased or overly concentrated posteriors

Müller 2013, 10.3982/ECTA9097



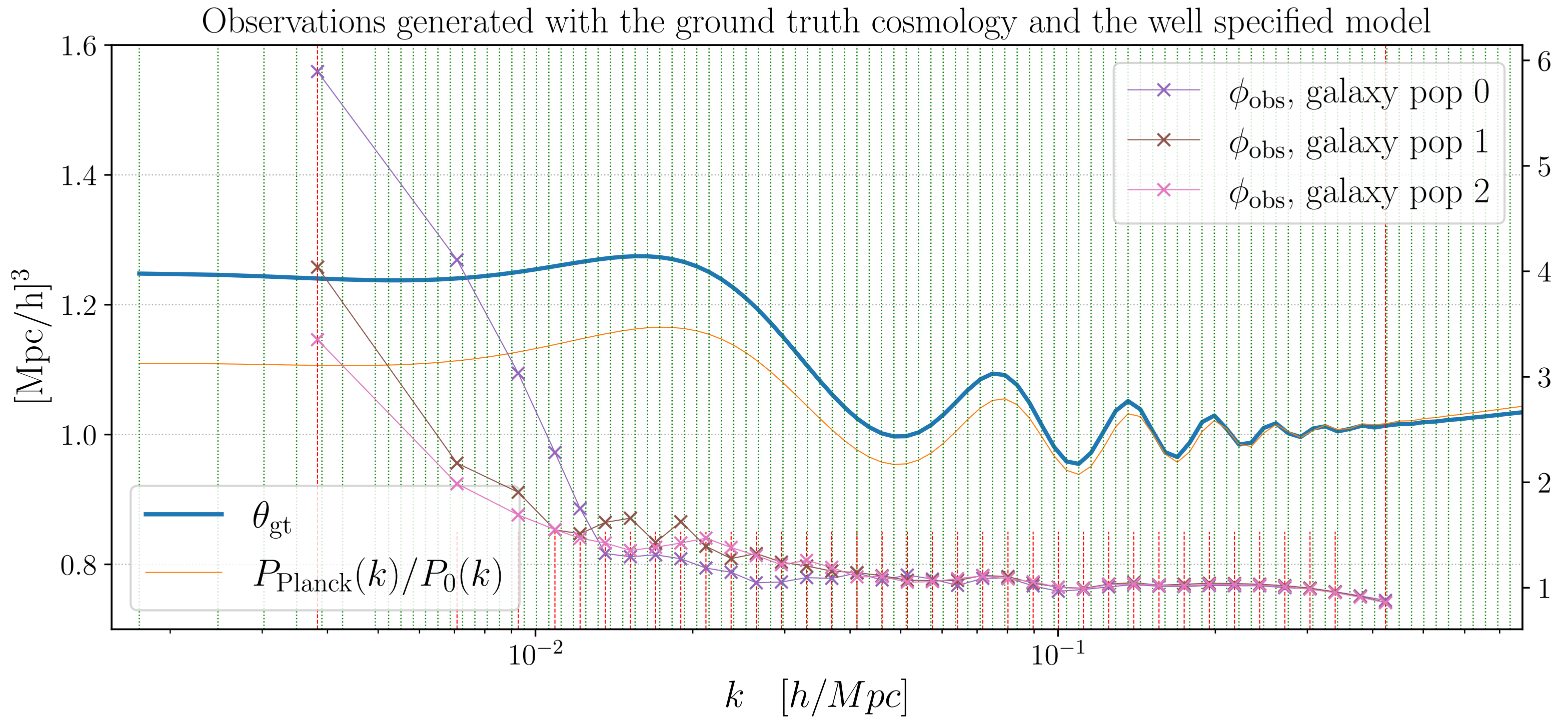
68% and 95% marginalized posterior via MCMC on simulated galaxy catalogues w/ two mass functions T08 and D16.

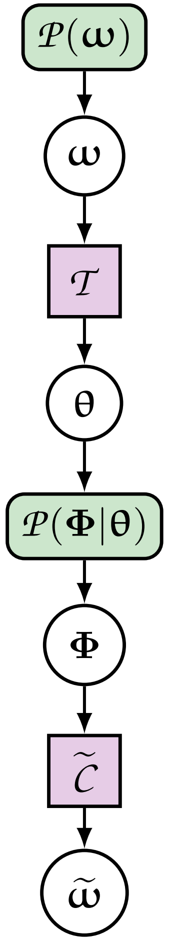
Salvati, Douspis & Aghanim 2020, 2005.10204

A general class of Bayesian Hierarchical Model of interest in cosmology

Prior on target parameters
e.g. Planck 2018

Cheap deterministic simulator
(Boltzmann solver: CLASS, CAMB)





Target parameters
$\boldsymbol{\omega} = (h,\, \Omega_{\mathrm{b}},\, \Omega_{\mathrm{m}},\, n_{\mathrm{S}},\, \sigma_8)$

Latent function
(initial matter
power spectrum)

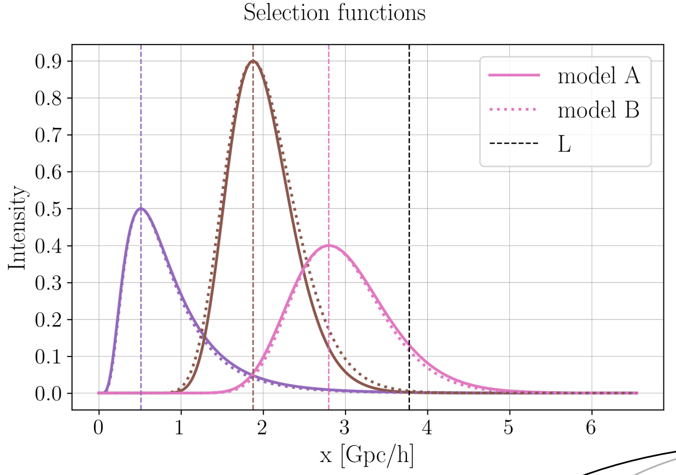
Summary statistic
from galaxy
overdensity field δ^{f}

Compressed data

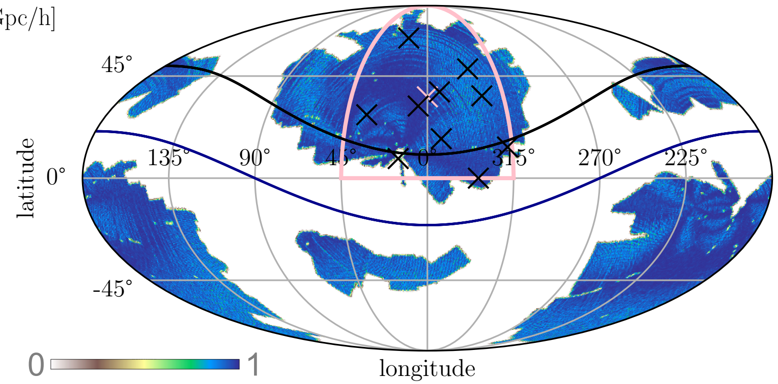
A general class of Bayesian Hierarchical Model of interest in cosmology

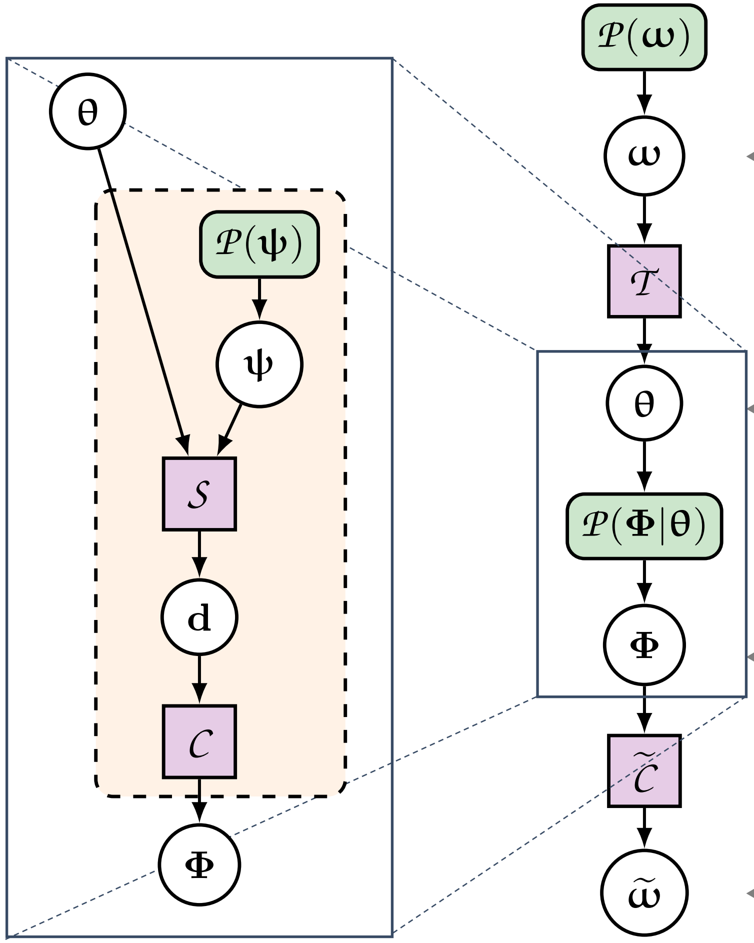
Complex probabilistic observational process

2-LPT, N-body, galaxy formation
Nuisance parameters $\boldsymbol{\psi}$ (galaxy biases, ...)



Instrumental effects :
radial selection, mask...





Target parameters
$\boldsymbol{\omega} = (h, \Omega_\mathrm{b}, \Omega_\mathrm{m}, n_\mathrm{S}, \sigma_8)$

Latent function
(initial matter
power spectrum)

Summary statistic
from galaxy
overdensity field δ^f

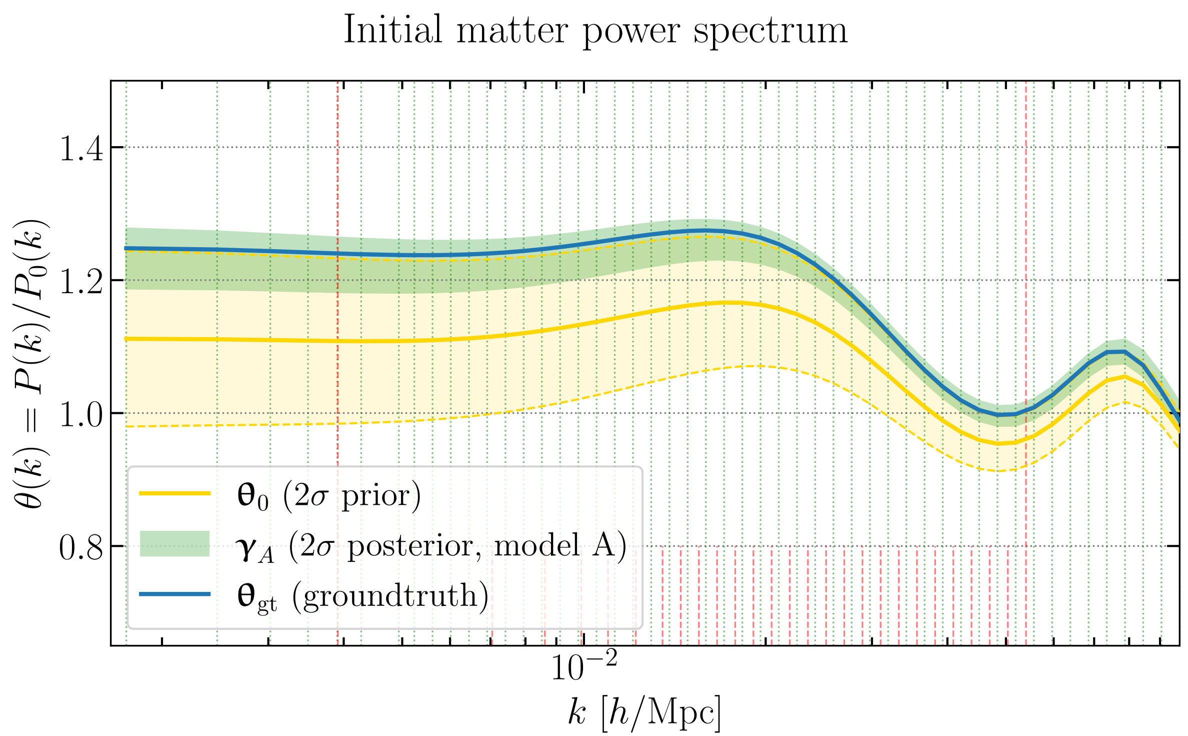
Compressed data

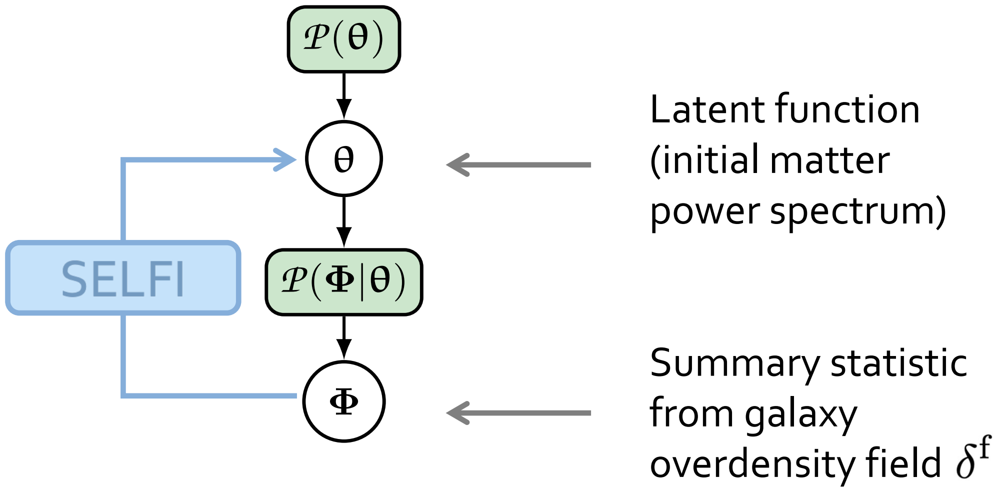
SELFI (Simulator Expansion for Likelihood-Free Inference)

Leclercq et al. 2019, 1902.10149

1. Infer a latent function, in this work the initial matter power spectrum θ

θ: initial matter powerspectrum normalized by BBKS spectrum

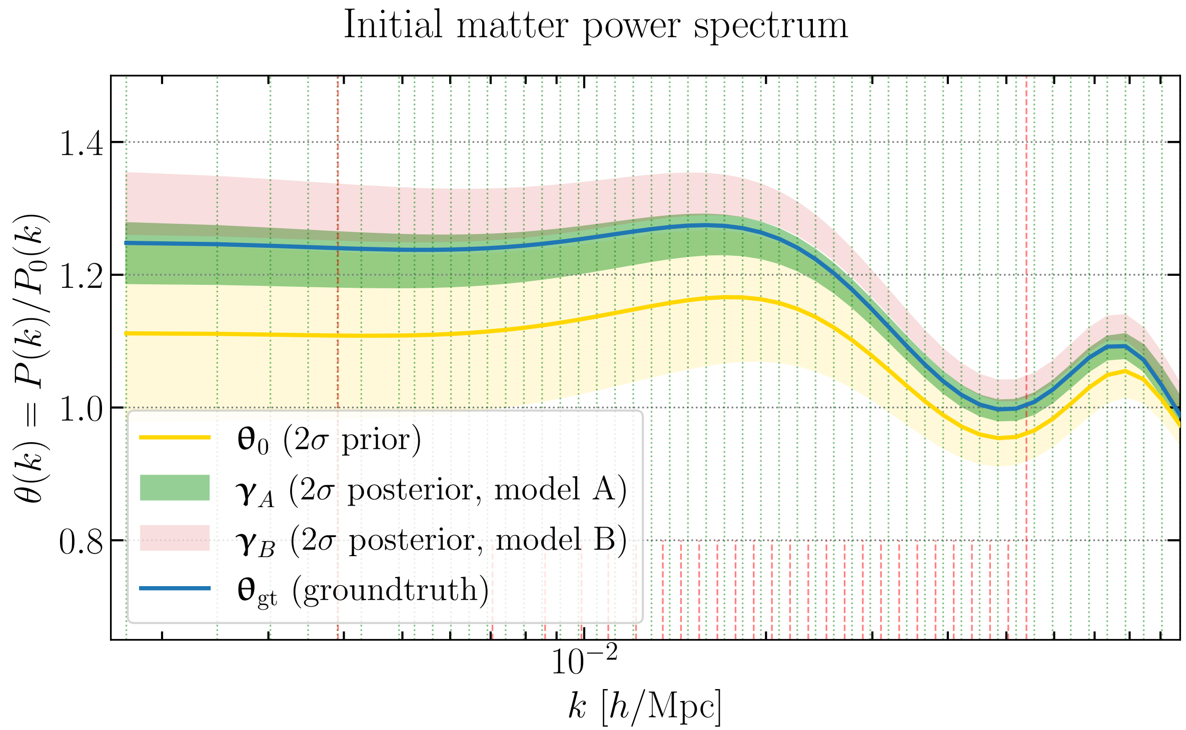




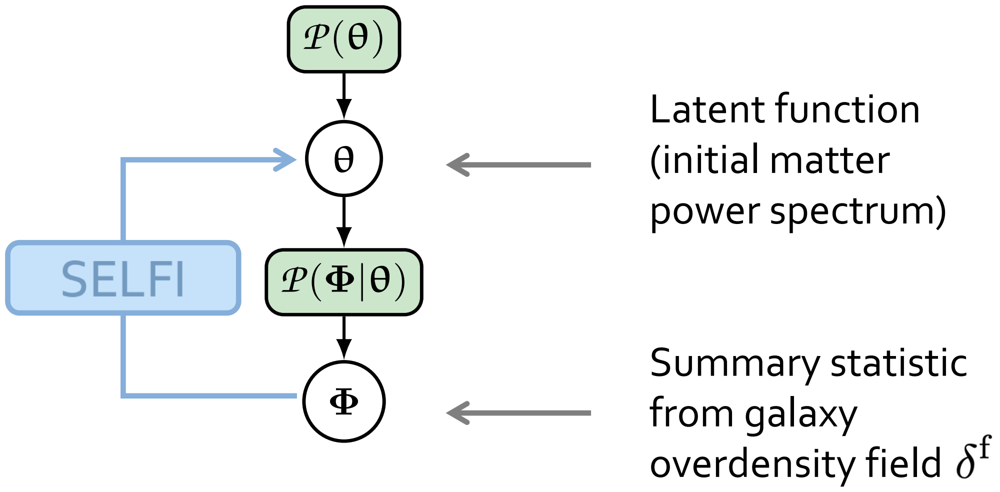
SELFI (Simulator Expansion for Likelihood-Free Inference)

Leclercq et al. 2019, 1902.10149

1. Infer a latent function, in this work the initial matter power spectrum θ
2. Utilize the posterior on θ to check for model misspecification



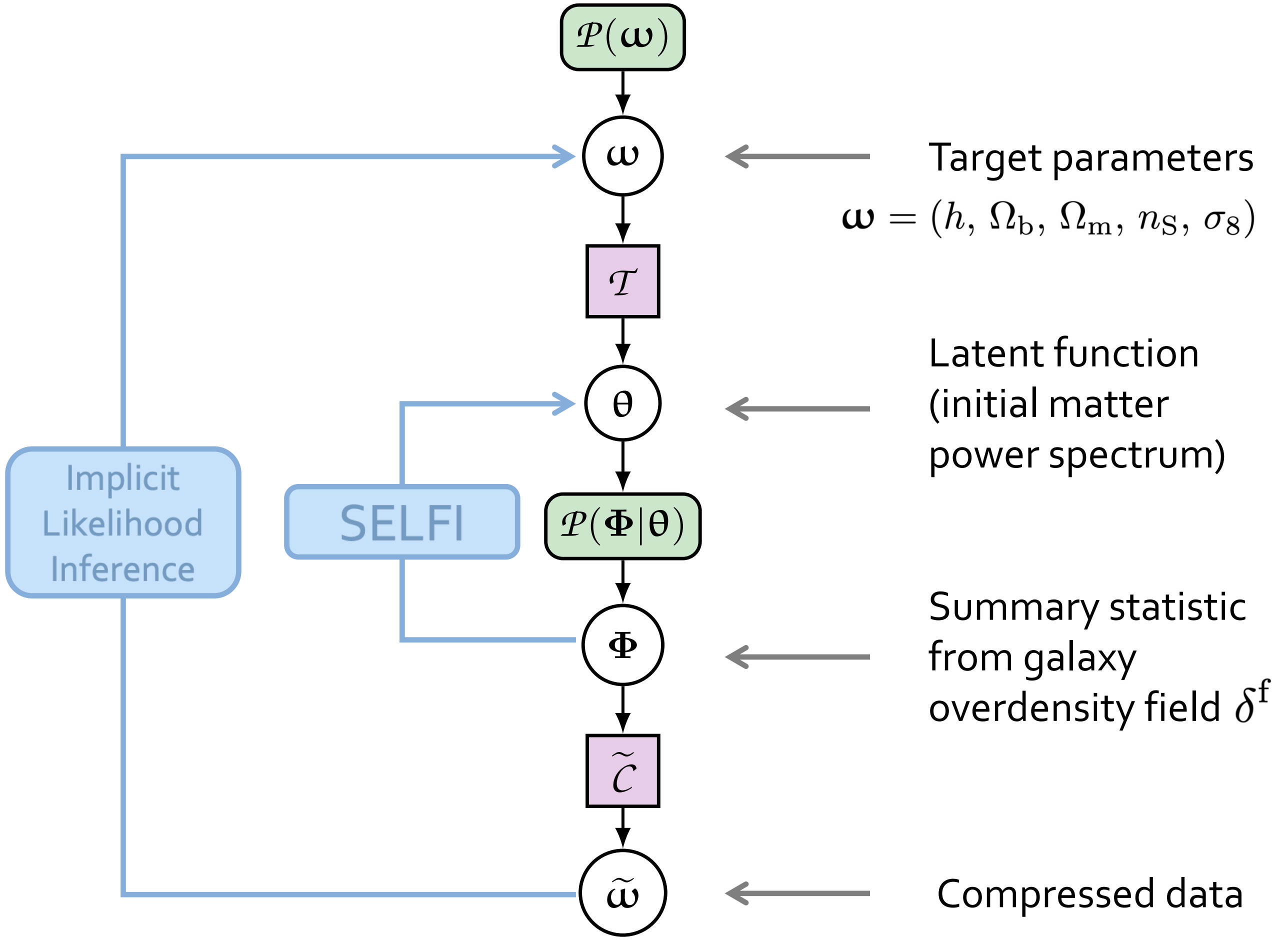
θ: initial matter powerspectrum normalized by BBKS spectrum



SELFI (Simulator Expansion for Likelihood-Free Inference)

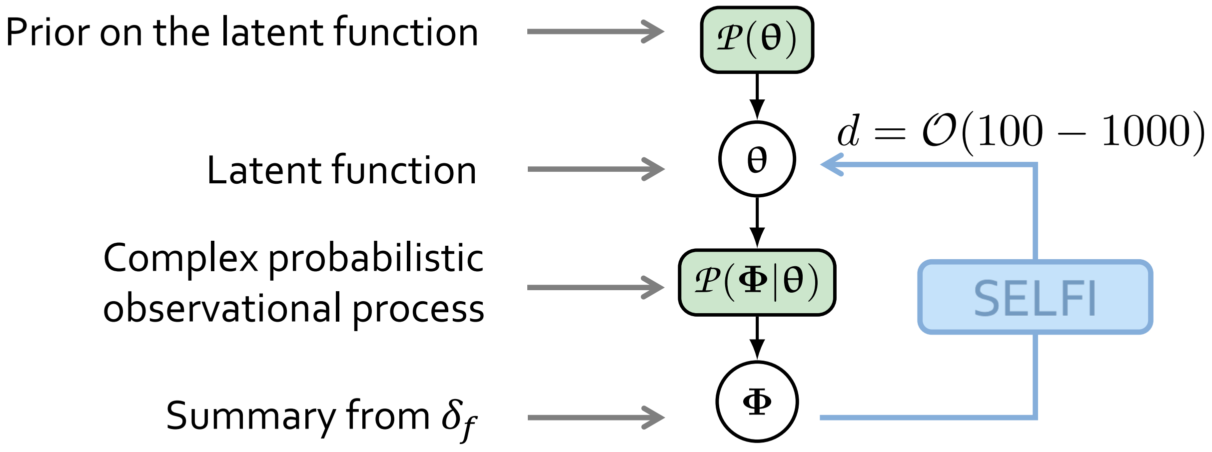
Leclercq et al. 2019, 1902.10149

1. Infer a latent function, in this work the initial matter power spectrum $\boldsymbol{\theta}$
2. Utilize the posterior on $\boldsymbol{\theta}$ to check for model misspecification
3. Infer the top-level cosmology $\boldsymbol{\omega}$
   - Recycle simulations from step 1. for optimal data compression
   - Use implicit likelihood inference (ABC, Bayesian optimization: BOLFI, other methods)



SELFI (Simulator Expansion for Likelihood-Free Inference)

Leclercq et al. 2019, 1902.10149



Assumptions

- Linearization of the black-box data model around an expansion point θ_0

$$\hat{\Phi}_\theta \approx f_0 + \nabla f_0 \cdot (\theta - \theta_0) \equiv f(\theta)$$

- For step 1. & data compression only, assume:
  - Gaussian prior
  - Gaussian effective likelihood

➢ Effective posterior:

expansion point

observed summaries

Mean: $\boldsymbol{\gamma} \equiv \boldsymbol{\theta}_0 + \boldsymbol{\Gamma} \, (\nabla \mathbf{f}_0)^\intercal \, \mathbf{C}_0^{-1} (\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}_0)$

Covariance: $\boldsymbol{\Gamma} \equiv \left[(\nabla \mathbf{f}_0)^\intercal \, \mathbf{C}_0^{-1} \nabla \mathbf{f}_0 + \mathbf{S}^{-1} \right]^{-1}$

covariance of summaries

prior covariance

gradient of the black-box

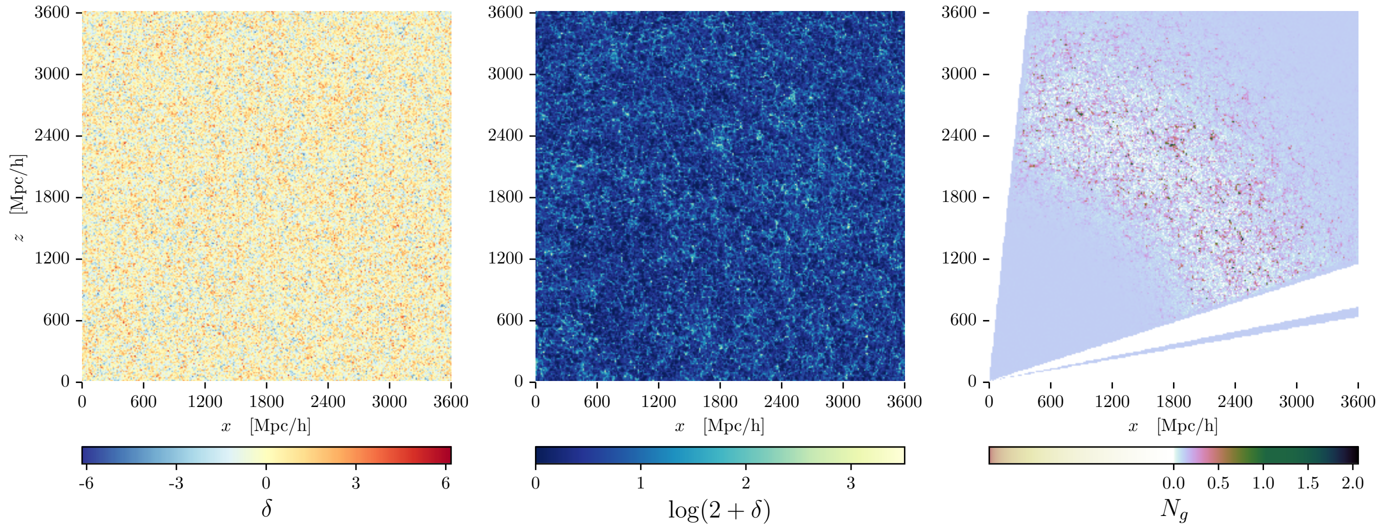
- $\mathbf{f}_0, \mathbf{C}_0$ and $\nabla \mathbf{f}_0$ evaluated through simulations
- The number of simulations is fixed *a priori* (contrary to MCMC)

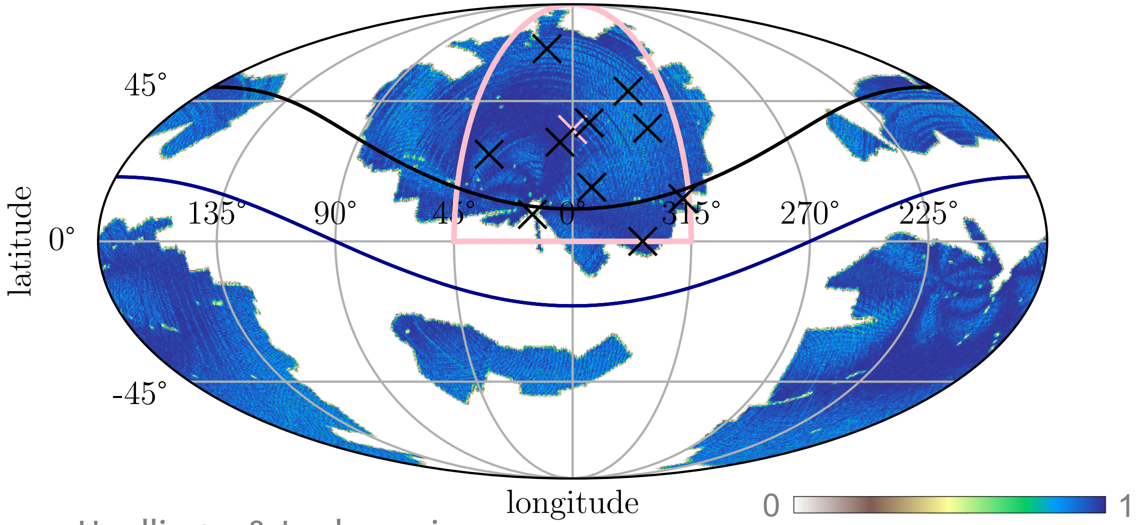
Investigating the impact of systematics with SELFI

- Grid specifications:
  - θ defined on S = 100 support wavenumbers
  - 512^3 grid, comoving L = 3.6 Gpc/h
- Gravitational evolution with Simbelmynë:

  Leclercq, Jasche & Wandelt 2015, 1502.02690
  - Flat Λ-CDM, initial δ^{i} with CLASS
  - 512^3 DM particles, 2LPT up to z = 19
  - PM grid of 1024^3 voxelx, COLA to z = 0

From initial matter overdensity field to observed galaxy counts



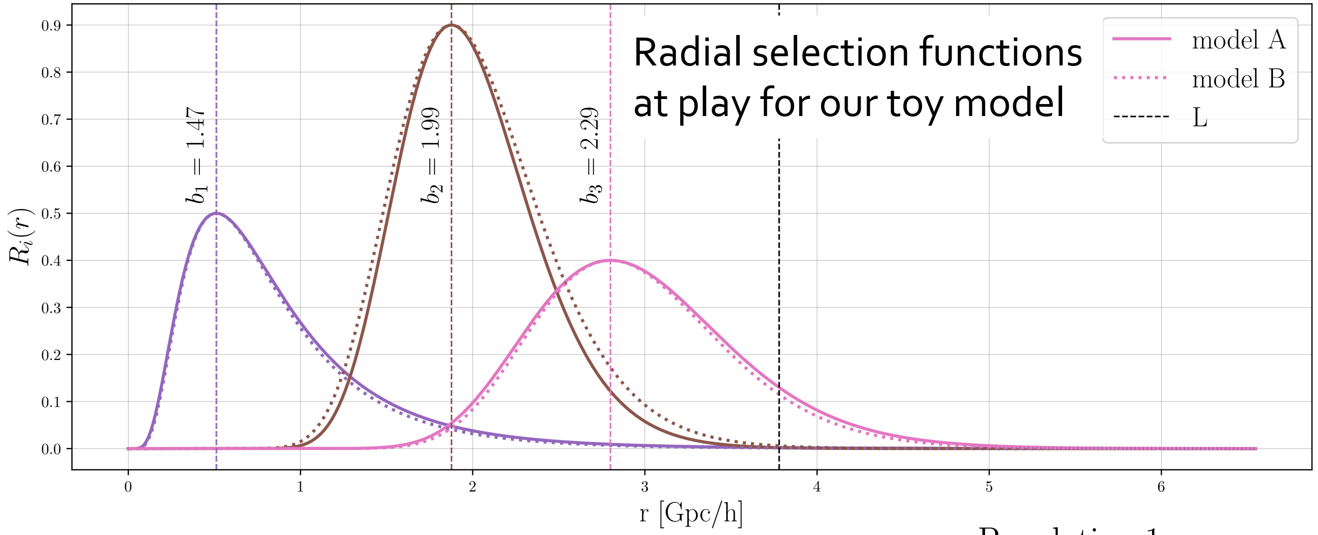


Hoellinger & Leclercq, in prep.

- The observer is at the corner of a cubic box covering 1 octant of the sky, with Euclid-like mask

Model A	10 additional masked areas, extinction near galactic plane
Model B	no such effects, lower resolution

Investigating the impact of systematics with SELFI



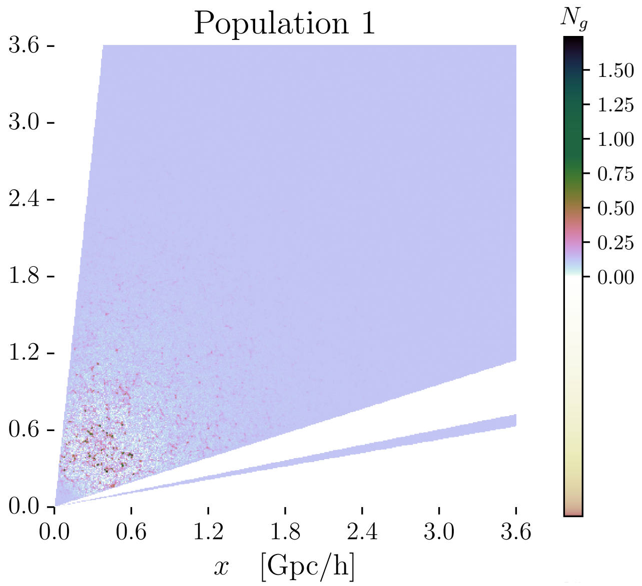
Model A

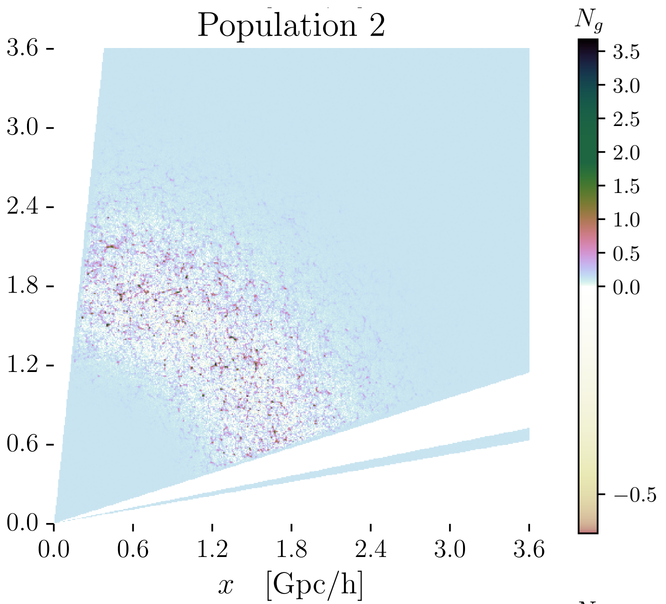
- Lognormal selection functions
- Luminosity-dependent galaxy biases

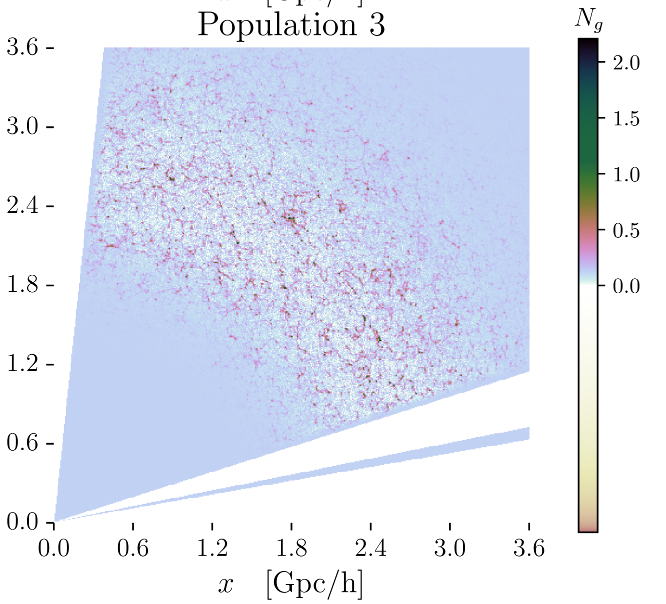
Model B

- Misspecified selections and biases
- Effects of order $\mathcal{O}(1\%)$

$N_e + S \times N_s$ = 10,200 simulations
with the full forward data model

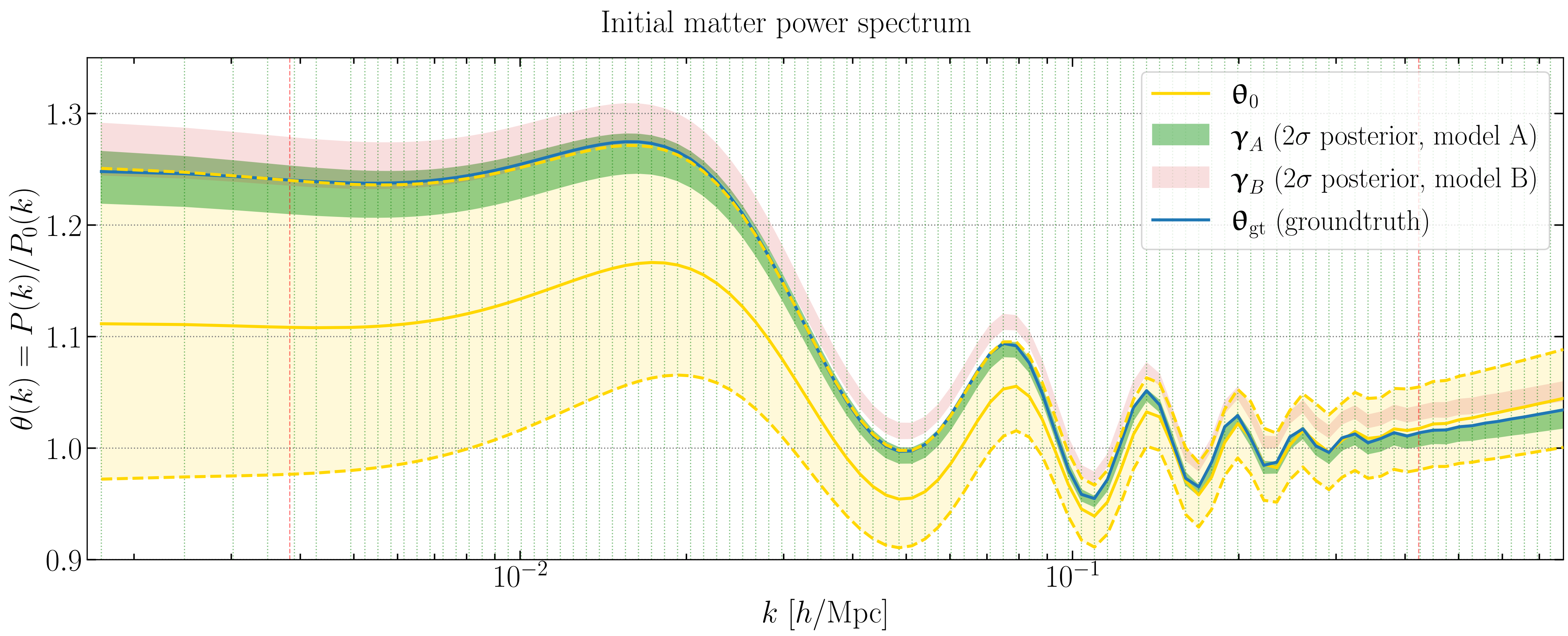






Hoellinger & Leclercq, in prep.

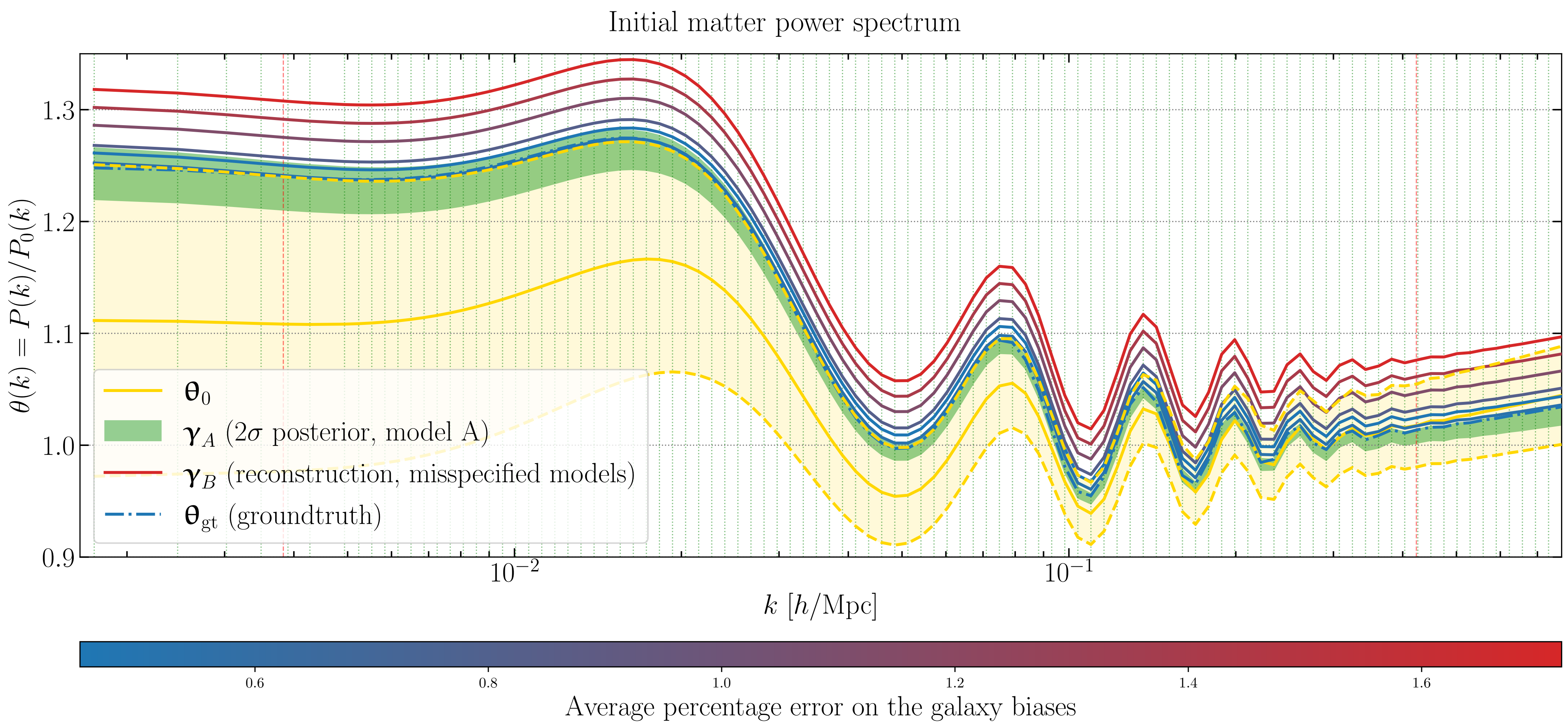
Investigating the impact of systematics with SELFI



Mahalanobis distances to prior: **Model A** $\longrightarrow$ 1.96 **Model B** $\longrightarrow$ 2.91

Hoellinger & Leclercq, in prep.

Investigating the impact of systematics with SELFI



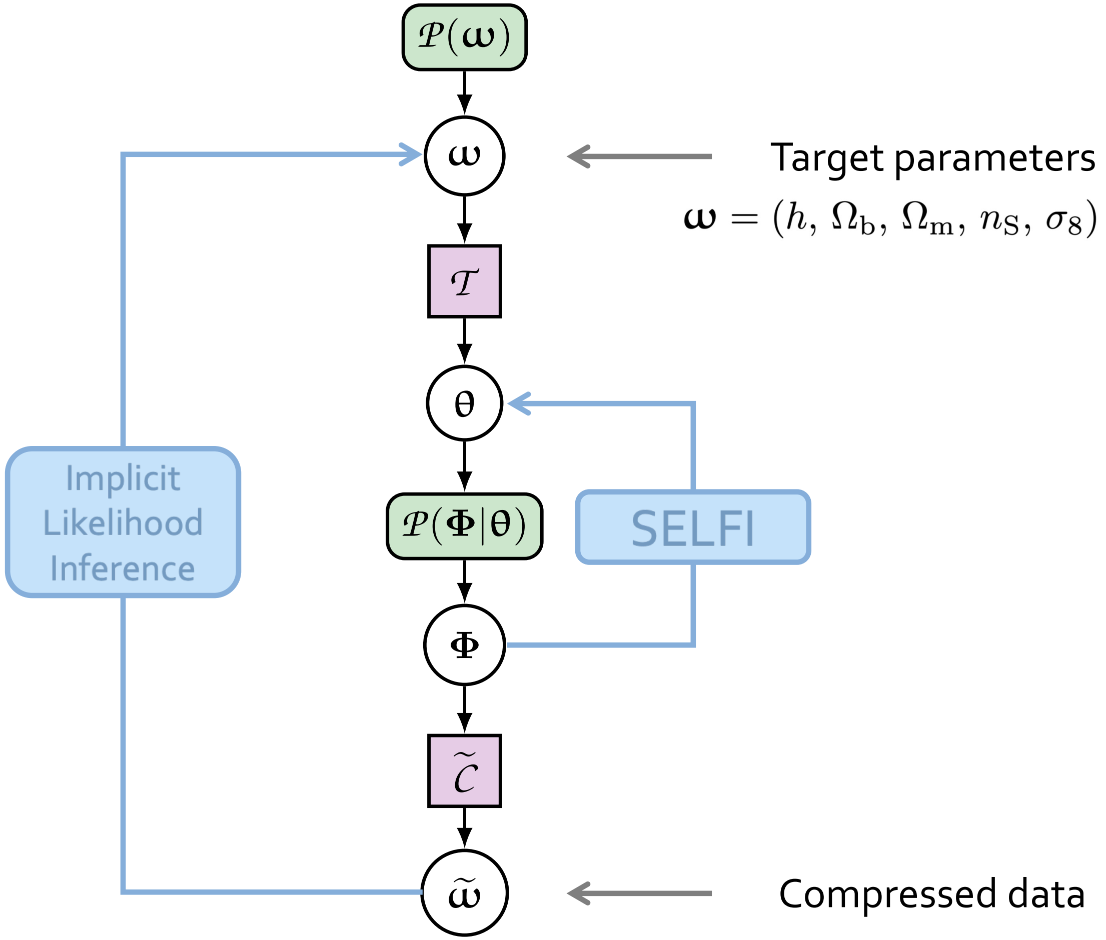
Hoellinger & Leclercq, in prep.

Optimal data compression

3. Infer top-level cosmology $\boldsymbol{\omega}$

We rely on score compression to compress the summaries from $\dim(\boldsymbol{\Phi}) = 111$ to $\dim(\widetilde{\boldsymbol{\omega}}) = \dim(\boldsymbol{\omega}) = 5$

- Score function $\nabla_{\boldsymbol{\omega}} \hat{\ell}_{\boldsymbol{\omega}_0} \longrightarrow$ steepness of $\hat{\ell}$
- It is a sufficient statistic for $\boldsymbol{\omega}$ for the the linearized log-likelihood, hence it is a natural way to compress the data



Alsing & Wandelt 2018, 1712.00012

Optimal data compression

$$\mathcal{C}(\boldsymbol{\Phi}) = \tilde{\boldsymbol{\omega}} \equiv \boldsymbol{\omega}_0 + \mathbf{F}_0^{-1}\left[(\nabla_{\boldsymbol{\omega}}\mathbf{f}_0)^\intercal \boxed{\mathbf{C}_0^{-1}}(\boldsymbol{\Phi} - \boxed{\mathbf{f}_0})\right]$$

Fisher matrix: $\mathbf{F}_0 = (\nabla_{\boldsymbol{\omega}}\mathbf{f}_0)^\intercal \boxed{\mathbf{C}_0^{-1}} \nabla_{\boldsymbol{\omega}}\mathbf{f}_0$

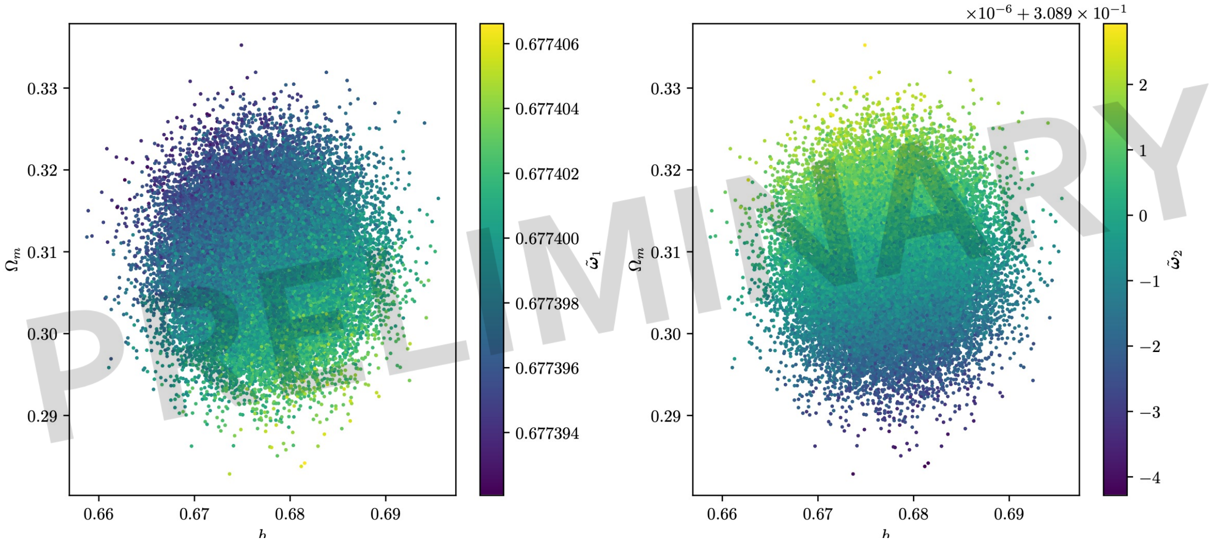
$$\nabla_{\boldsymbol{\omega}}\mathbf{f}_0 = \boxed{\nabla\mathbf{f}_0} \cdot \boxed{\nabla_{\boldsymbol{\omega}}\mathcal{T}_0}$$

Already computed for SELFI

Cheap via finite differences

- The compression is optimal in the sense that it preserves the Fisher content of the data. Hypothesis:
  - the covariance matrix does not vary close to the expansion point $\nabla_\omega \mathbf{C} = 0$
  - The likelihood is gaussian or the following holds: $\nabla \mathbb{E}_\theta \left[\nabla^T \mathcal{L}\right] = \nabla \mathbb{E}_\theta \left[\nabla \nabla^T \mathcal{L}\right]$

- Example for Ω_m and h (simplified data model):



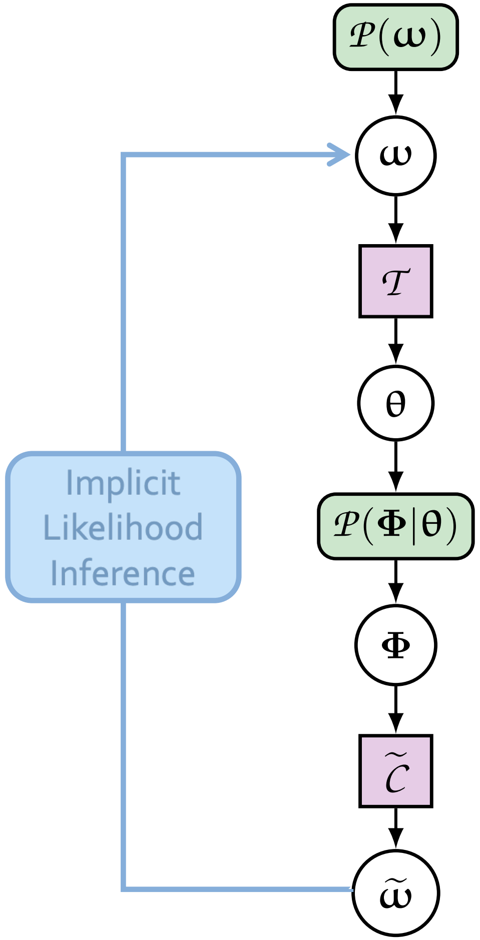
Leclercq 2022, 2209.11057

Alsing & Wandelt 2018, 1712.00012

Hoellinger & Leclercq, in prep

Optimal data compression

3. Infer top-level cosmology ω

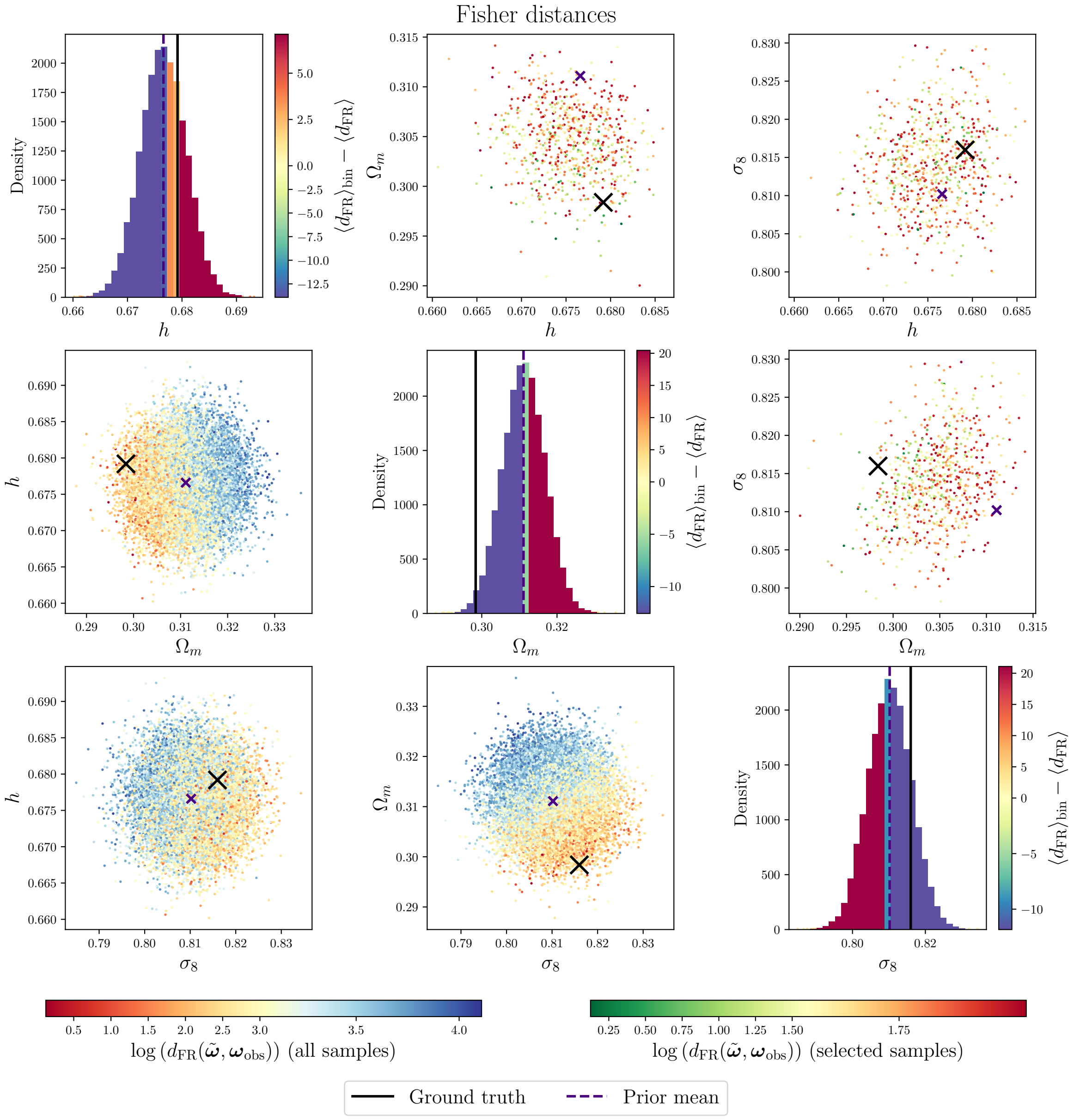


- Assumptions made for steps 1. - 2. do not impact final inference:
  - the Gaussian effective likelihood is not required to infer the cosmology except through data compression
  - lossy data compression usually leaves posteriors unbiased

- Any algorithm can be used to obtain the posterior $\mathcal{P}(\omega|\widetilde{\omega}_\mathrm{O})$
  - Non-parametric approaches can use the Fisher-Rao distance

$$d_\mathrm{FR}(\widetilde{\omega}, \widetilde{\omega}_\mathrm{O}) \equiv \sqrt{(\widetilde{\omega} - \widetilde{\omega}_\mathrm{O})^\intercal\, \mathbf{F}_0 (\widetilde{\omega} - \widetilde{\omega}_\mathrm{O})}$$

  - Example with ABC (same data model but smaller dim. for ω and 64^3 grid)



Leclercq 2022, 2209.11057

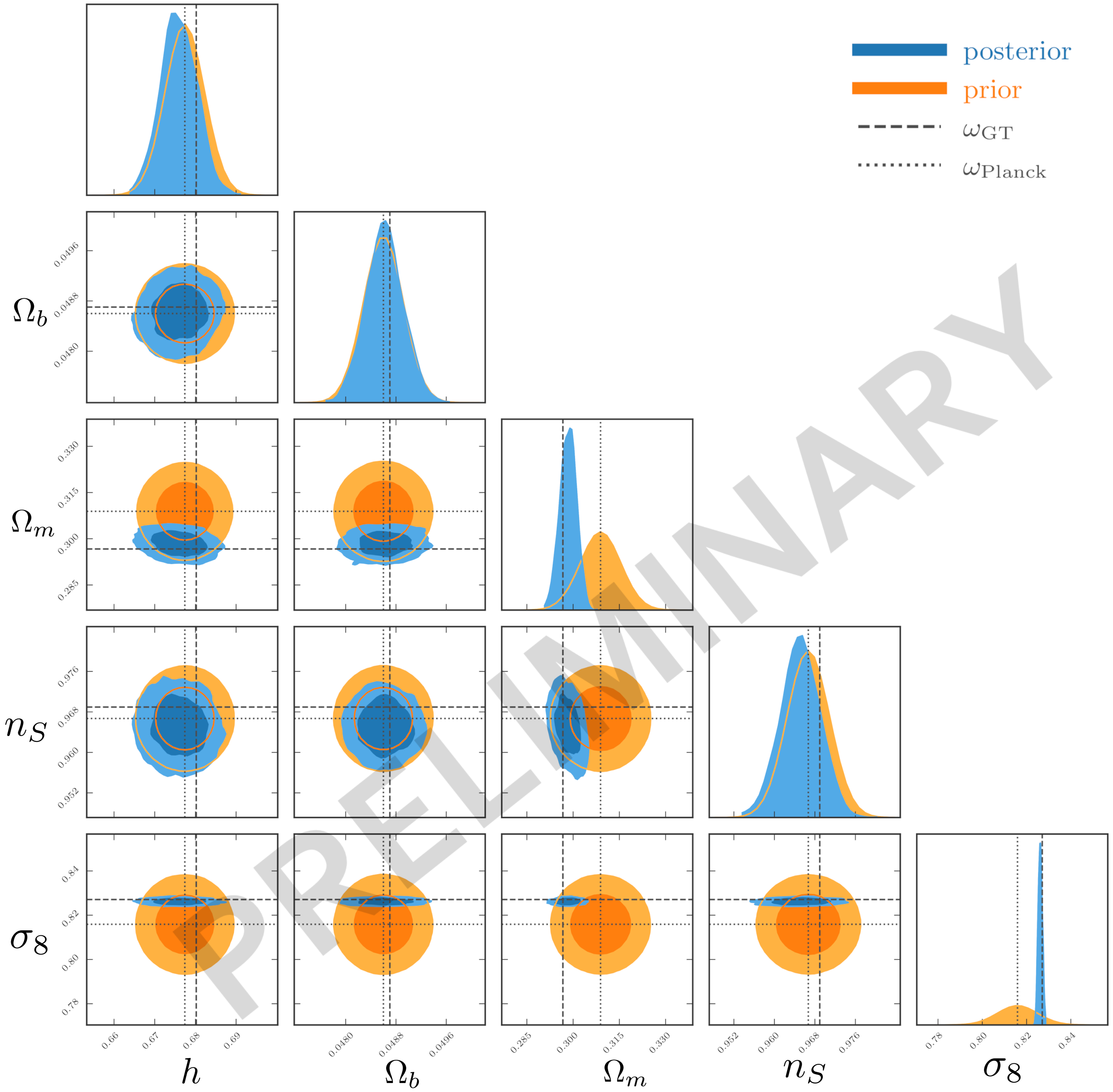
Hoellinger & Leclercq, in prep

Final posterior

- We ran the full SELFI pipeline with a simplified forward data model
  - same instrumental response as before,
  - no gravitational evolution
  - baseline ABC for step 3.

- We got unbiased posteriors for the top-level cosmological parameters

$$\boldsymbol{\omega} = (h,\, \Omega_\mathrm{b},\, \Omega_\mathrm{m},\, n_\mathrm{S},\, \sigma_8)$$



2313 selected samples out of 700K

Leclercq et al. 2019, 1902.10149

Conclusion

➢ A novel two-step simulation-based Bayesian approach, combining SELFI and SBI, to tackle the issue of model misspecification for a large class of BHMs.

- Advantages related to the first step (SELFI):
  - No need to incorporate any knowledge of the data-generating process in the analysis, even with high dimensional complex black-box simulators.
  - Number of simulations fixed a priori.
  - The computational workload is perfectly parallel.
- Advantages related to the second step (SBI):
  - The score compressor comes for free: we recycle simulations from step **1.**
  - General advantages of SBI with respect to likelihood-based methods are preserved.
  - No simplification nor inner knowledge of the forward data model required

Perspectives

- Incorporate parametrization of deviations from Λ-CDM in the top-level parameters:
  - equation of state of dark energy $w(a)$
  - total neutrino masses m_υ
  - deviations from gaussianity f_{NL}

- Jointly infer all top-level parameters and nuisance parameters such as galaxy biases
  - ➢ This means implicit likelihood inference in dimension $\mathcal{O}(10-20)$
  - ➢ We need to investigate and extend advanced Bayesian optimisation strategies to explore the parameter space to avoid curse of dimensionality

- Thanks for listening!
- Main references:

  Alsing & Wandelt 2018, 1712.00012

  Leclercq et al. 2019, 1902.10149

  Leclercq 2022, 2209.11057

  Leclercq, Jasche & Wandelt 2015, 1502.02690

- Code and data availability: wait for SELFI2 release in 2024!

hoellin.github.io

www.aquila-consortium.org