Implicit Likelihood Inference while efficiently checking for survey systematics

Euclid Galaxy Clustering meeting, Marseille 2024

Florent Leclercq

www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

In collaboration with:
Tristan Hoellinger (IAP)

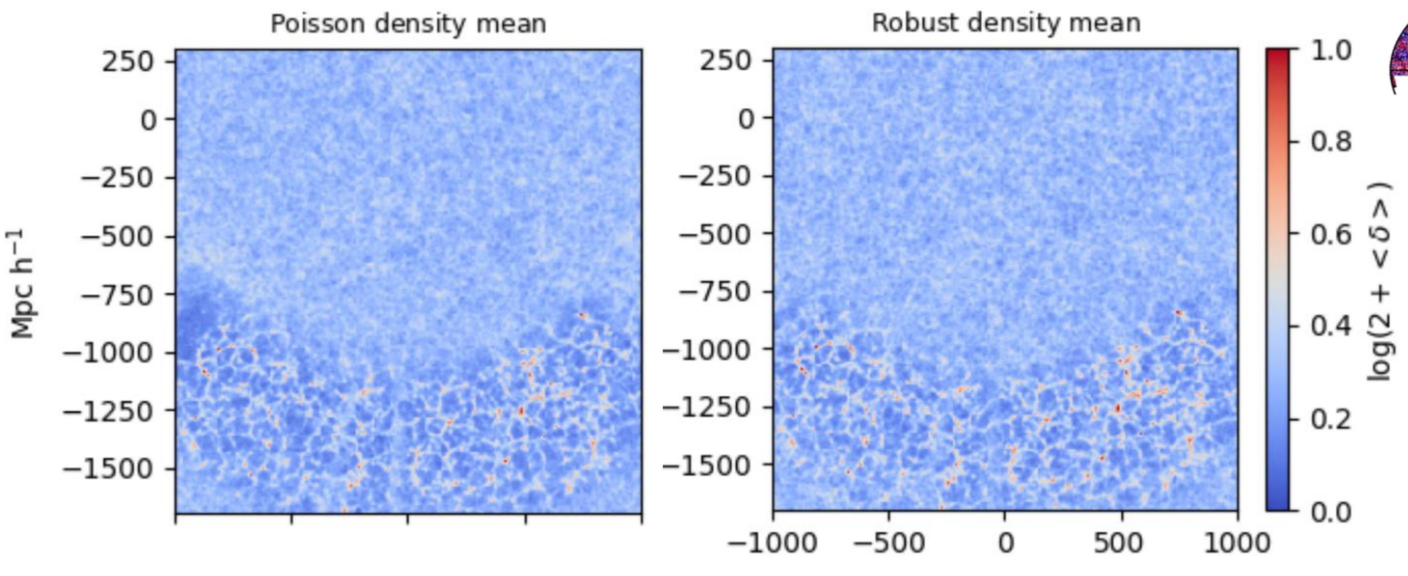
and the Aquila Consortium

www.aquila-consortium.org

1 February 2024

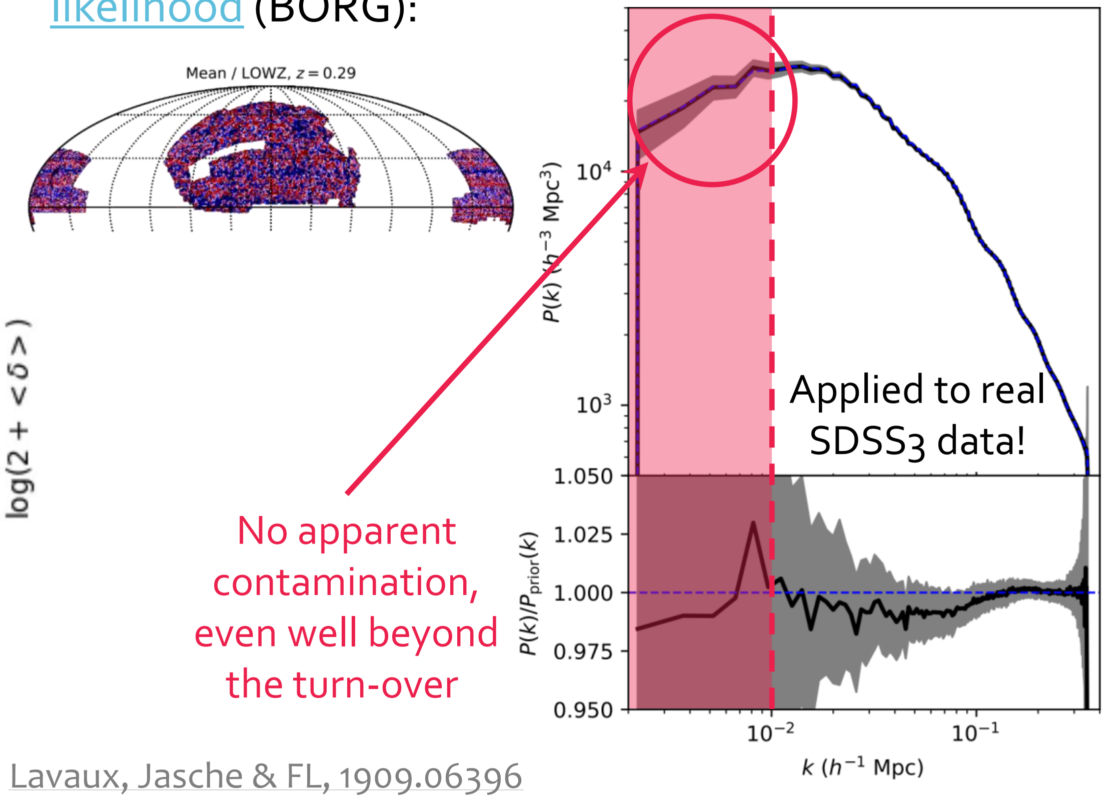
Model misspecification and unknown systematics with an explicit field-level likelihood

- Model misspecification is a long-standing problem for Bayesian inference: when the model differs from the actual data-generating process, posteriors tend to be biased and/or overly concentrated.
- This issue is particularly critical for cosmological data analysis in the presence of systematic effects.



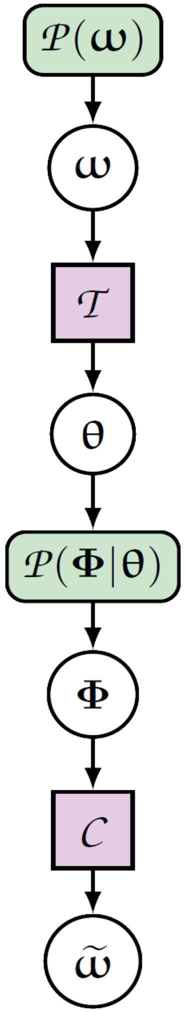
Porqueres, Ramanah, Jasche & Lavaux, 1812.05113

- In cosmology, we are sometimes unable to formulate ***any*** model that fits the data in some regimes.
- Machine-aided report of unknown systematic effects is possible with an explicit field-level likelihood (BORG):



Lavaux, Jasche & FL, 1909.06396

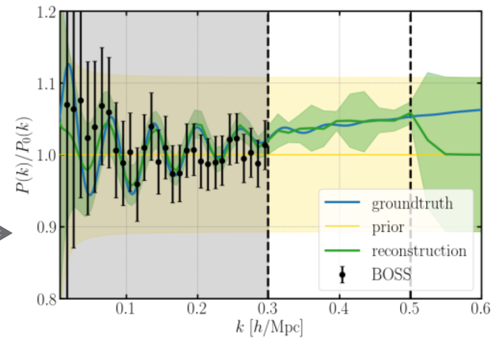
A general class of Bayesian hierarchical models (BHMs): Complex observations of a latent function controlled by top-level parameters



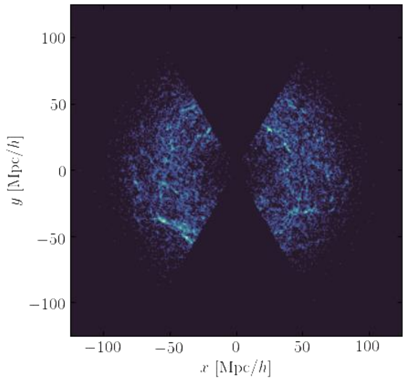
- Target parameters → ω
- Cheap (deterministic) simulator → $\mathcal{T}$
- Latent function → θ
- Complex probabilistic observational process → $\mathcal{P}(\Phi|\theta)$
- Raw data → Φ
- Compressor → $\mathcal{C}$
- Compressed summary statistics → $\tilde{\omega}$

Example from cosmology:

- Cosmological parameters → $\Omega_\mathrm{m}, \Omega_\mathrm{b}, H_0, n_\mathrm{S}, \sigma_8, \tau$
- Solver of the Boltzmann equation (CAMB, CLASS)
- Initial power spectrum →

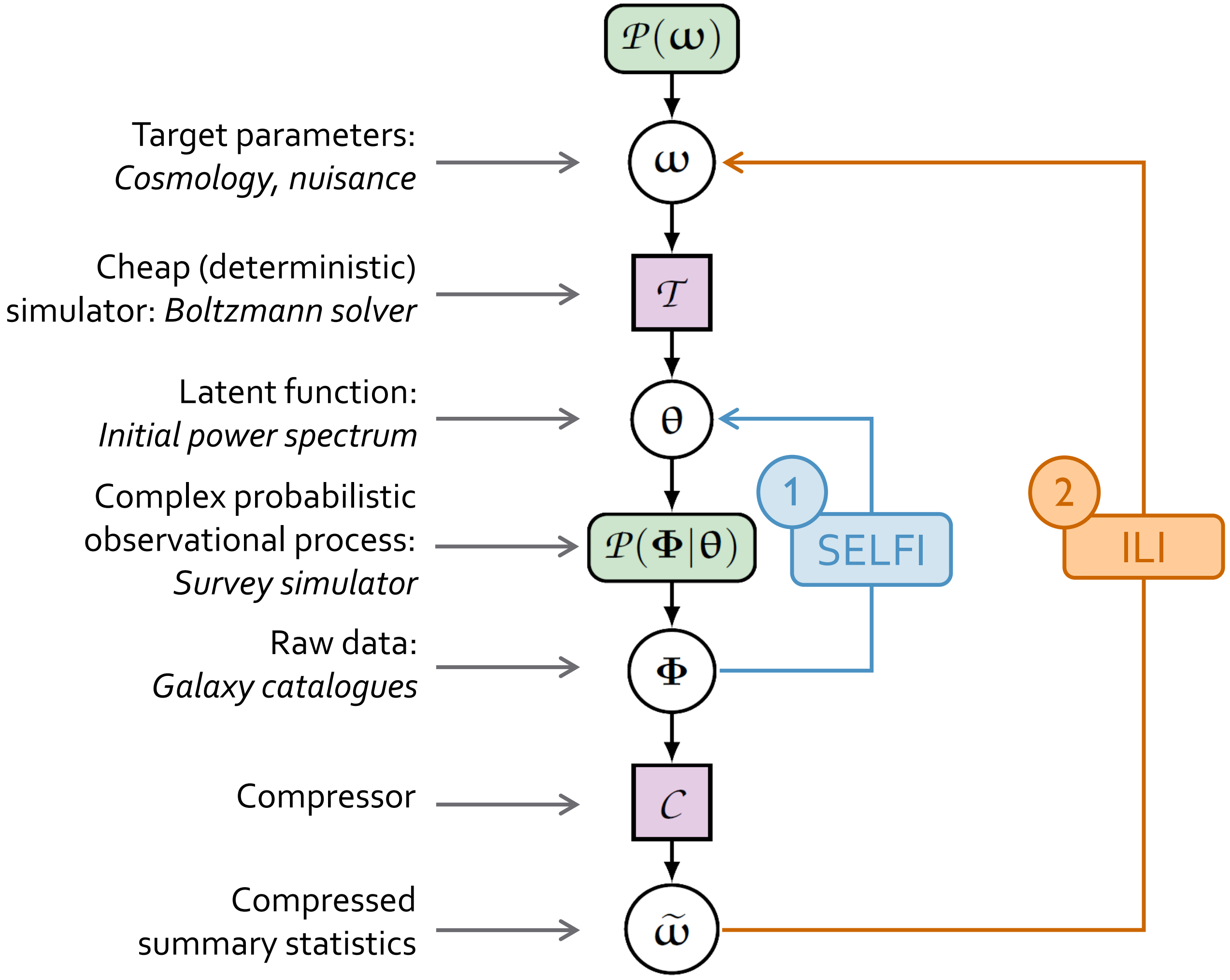


- *N*-body simulation, galaxy formation, survey modelling…
- Galaxy survey data →



- Score compression
- Quasi maximum-likelihood estimators → $\widetilde{\Omega_\mathrm{m}}, \widetilde{\Omega_\mathrm{b}}, \widetilde{H_0}, \widetilde{n_\mathrm{S}}, \widetilde{\sigma_8}, \widetilde{\tau}$

Key idea: a two-step implicit likelihood inference (ILI) process that recycles simulations



(1) Inference of the latent function θ, to check for model misspecification:

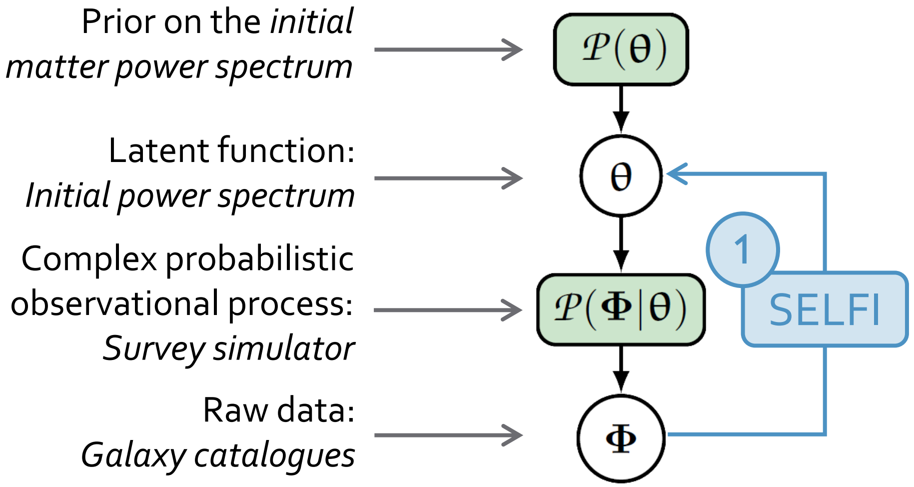
- SELFI algorithm

(2) Implicit likelihood inference of ω:

- Approximate Bayesian Computation (ABC), Likelihood-Free Rejection Sampling
- Density/ratio estimation (DELFI / NRE)
- Bayesian optimisation (BOLFI)
- others…

Important: the simulations necessary for step (1) are recycled for data compression, which is required for step (2)

Initial power spectrum inference: the SELFI approach (*Simulator Expansion for Likelihood-Free Inference*)



- Linearisation of the black-box:

$$\hat{\Phi}_\theta \approx \mathbf{f}_0 + \nabla\mathbf{f}_0 \cdot (\theta - \theta_0)$$

- Further assume:
  - Gaussian prior: $\mathcal{P}(\theta) = \mathcal{G}(\theta_0, \mathbf{S})$
  - Gaussian effective likelihood: $\mathcal{P}(\Phi|\theta) = \mathcal{G}\left[\mathbf{f}(\theta), \mathbf{C}_0\right]$

- The posterior is Gaussian and analogous to a Wiener filter:

expansion point — observed summaries

mean: $\gamma \equiv \theta_0 + \Gamma\,(\nabla\mathbf{f}_0)^\intercal\,\mathbf{C}_0^{-1}(\Phi_\mathrm{O} - \mathbf{f}_0)$

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

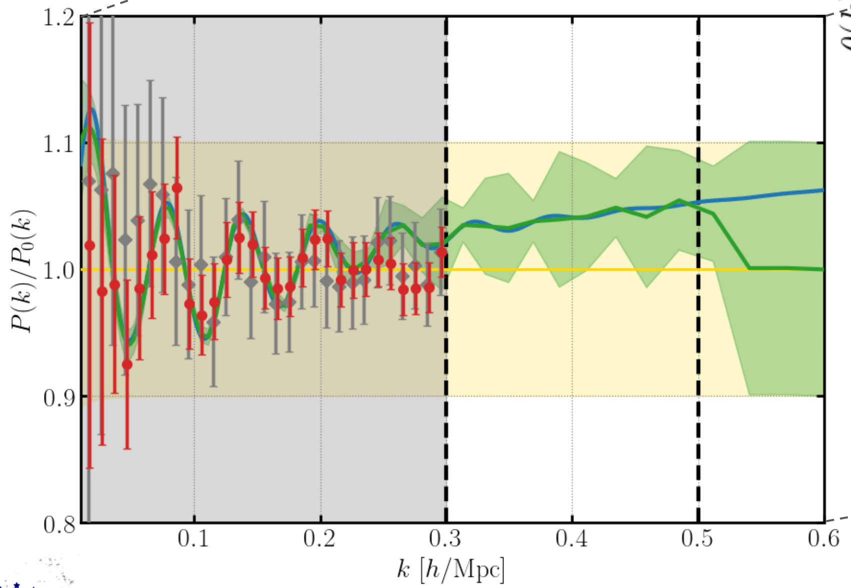
covariance of summaries — gradient of the black-box — prior covariance

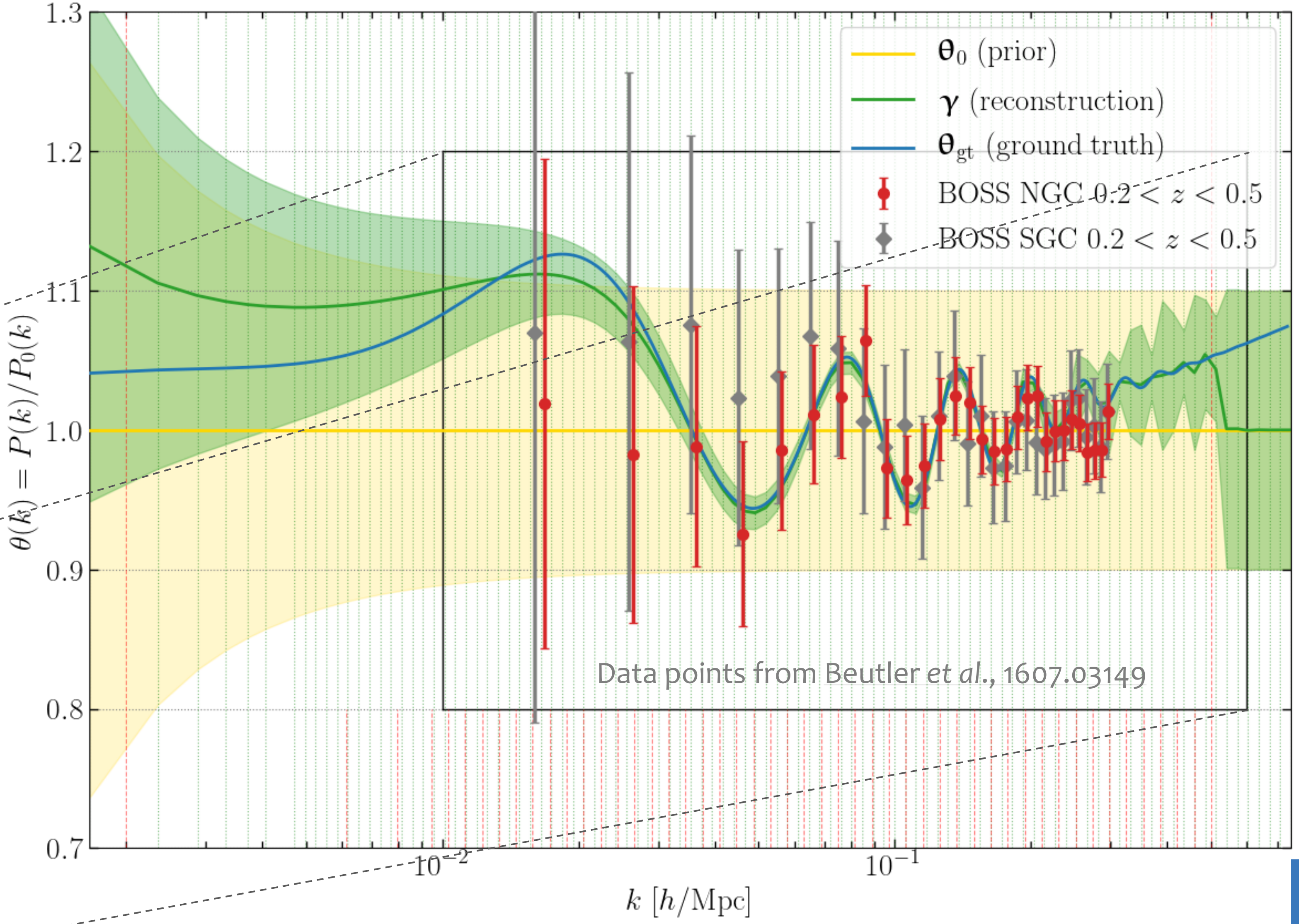
- $\mathbf{f}_0$, $\mathbf{C}_0$ and $\nabla\mathbf{f}_0$ can be evaluated through simulations only.
- The number of required simulations is fixed *a priori* (contrary to MCMC).
- The workload is perfectly parallel.

FL, Enzi, Jasche & Heavens, 1902.10149

SELFI (*Simulator Expansion for Likelihood-Free Inference*): ILI of the initial power spectrum
Euclid forecast vs BOSS data

- Numerical data models allow using the galaxy power spectrum as summary statistics up to at least $k \gtrsim 0.5\ h/\mathrm{Mpc}$ safely
- $N_\mathrm{modes} \propto k^3$: 5 times more modes are used in the analysis.





FL, Enzi, Jasche & Heavens, 1902.1014; FL, 2209.11057; Hoellinger & Leclercq, in prep.

Check for model misspecification

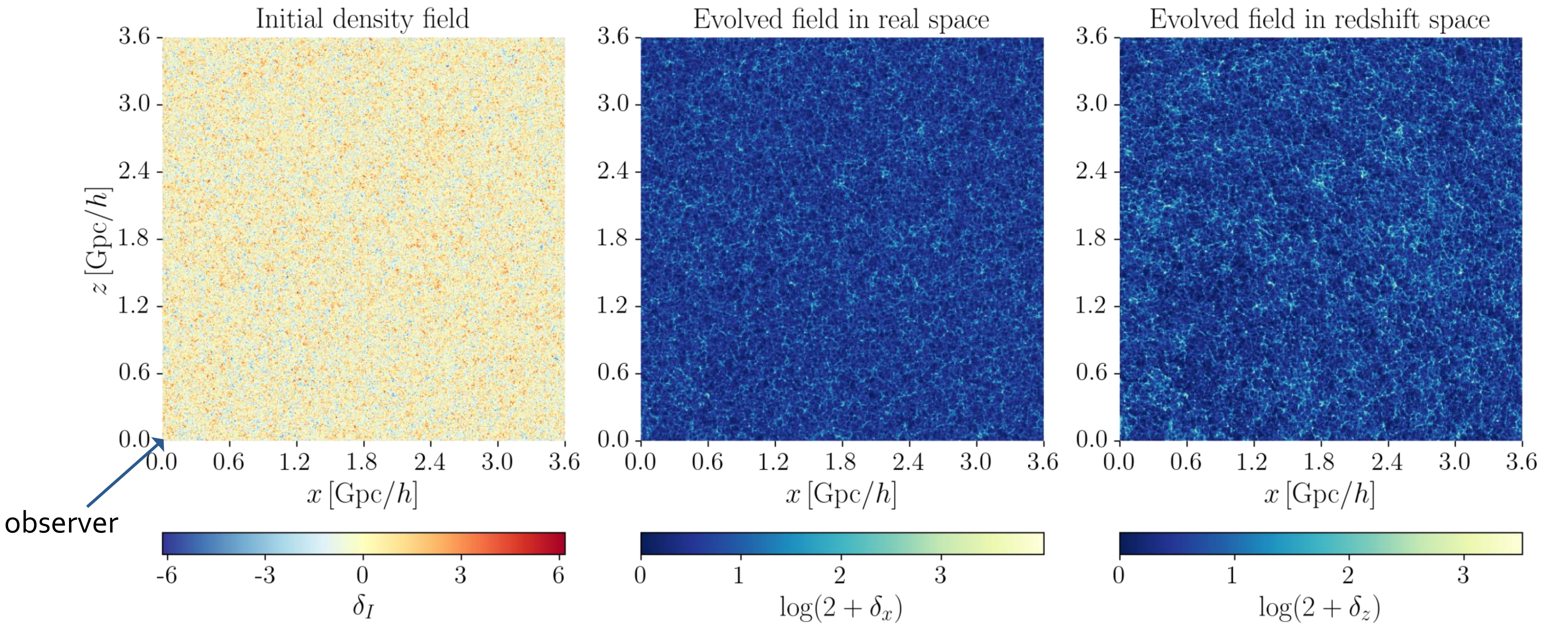
- Qualitatively: the shape of the reconstructed initial power spectrum $\boldsymbol{\theta}$ is useful as a check for unknown systematics / model misspecification (using our independent theoretical understanding).
- Quantitatively: we can use the Mahalanobis distance between the reconstruction $\boldsymbol{\gamma}$ and the prior distribution $\mathcal{P}(\boldsymbol{\theta})$:

$$d_\mathrm{M}(\boldsymbol{\gamma}, \boldsymbol{\theta}_0|\mathbf{S}) \equiv \sqrt{(\boldsymbol{\gamma} - \boldsymbol{\theta}_0)^\intercal \, \mathbf{S}^{-1}(\boldsymbol{\gamma} - \boldsymbol{\theta}_0)}$$

FL, 2209.11057

- θ defined on S = 100 support wavenumbers
- Flat ΛCDM assumed

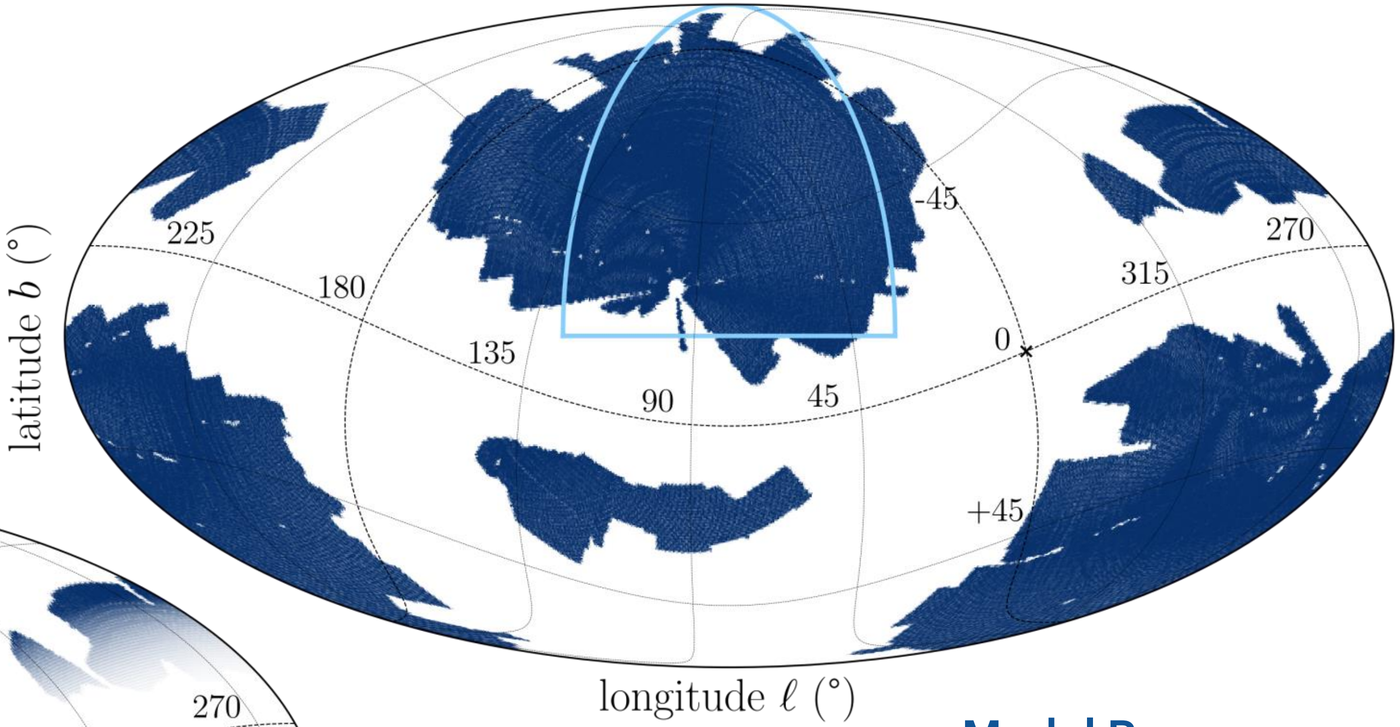
- Gravitational evolution (*N*-body) using Simbelmynë
  Leclercq, Jasche & Wandelt, 1502.02690; http://simbelmyne.florent-leclercq.eu
  - 512^3 dark matter particles, 2LPT up to $z = 19$
  - Particle-mesh grid of 1024^3 voxels, COLA to $z = 0$



Hoellinger & Leclercq, in prep.

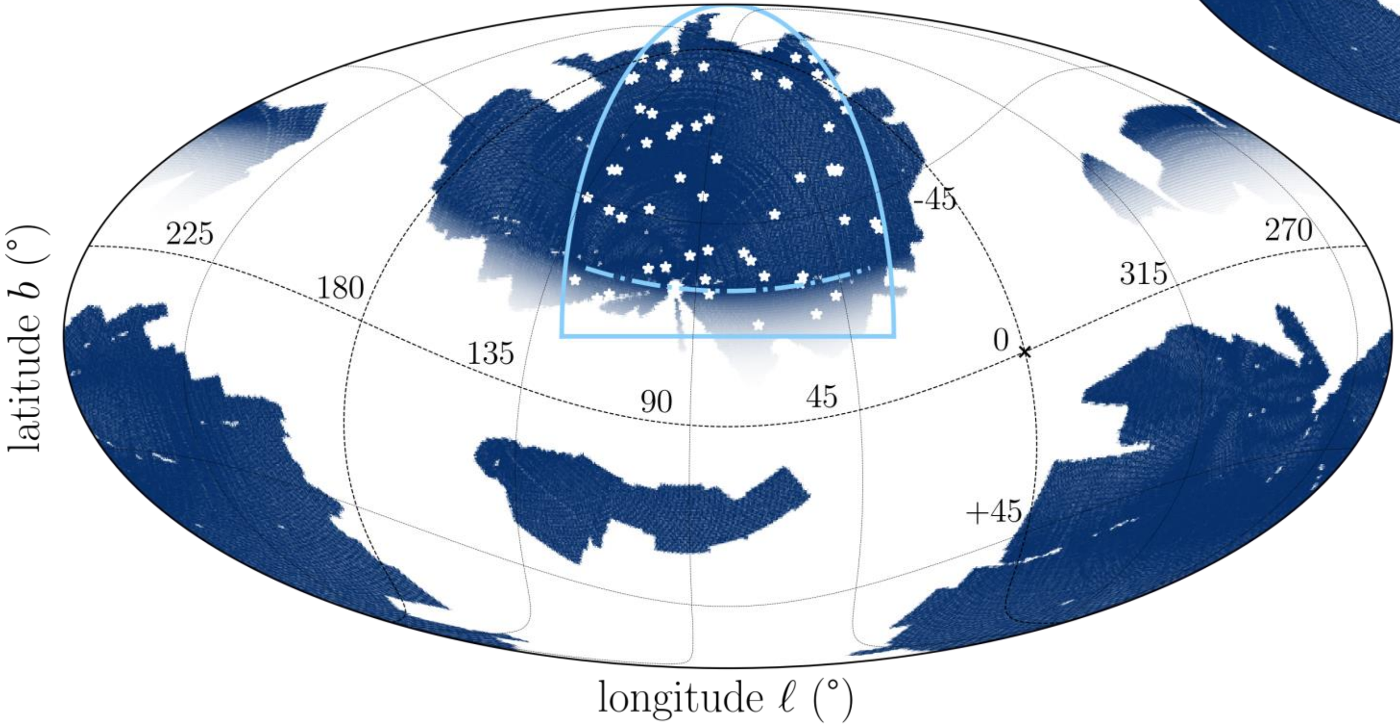
Systematic effect n°1: survey mask

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



Model B

no such effects



Model A

80 additional holes

extinction from -30° to 0° latitude (galactic)

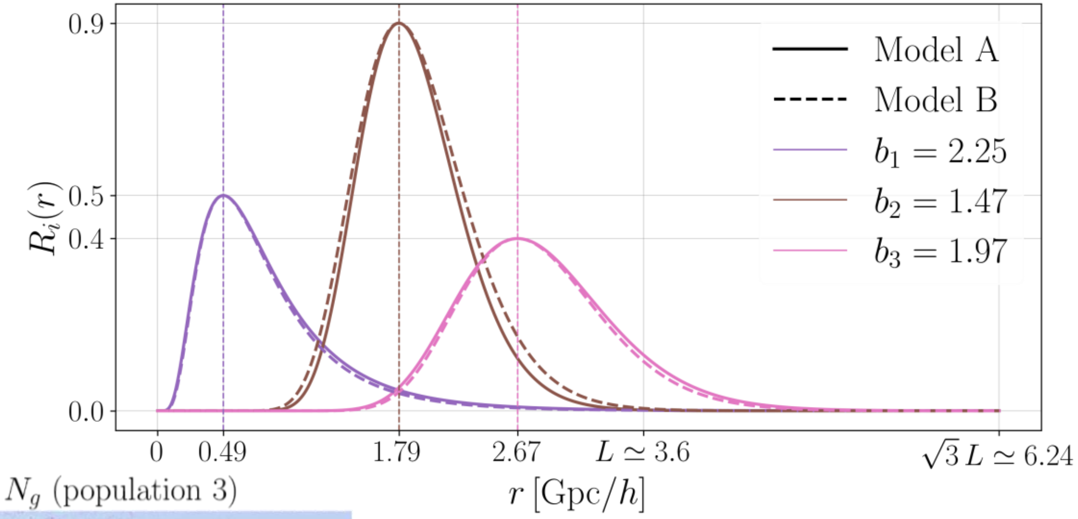
Hoellinger & Leclercq, in prep.

Systematic effect n°2: linear galaxy biases and selection functions

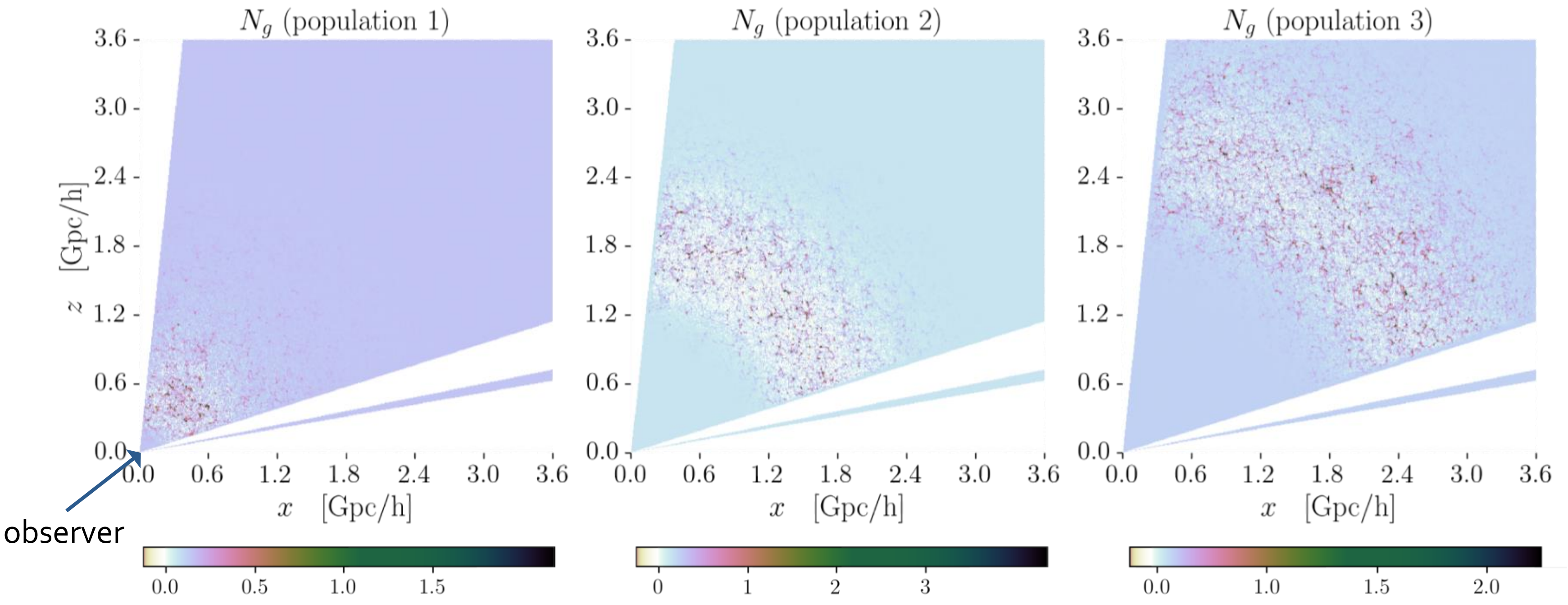
Model A

3 simulated populations of galaxies (1 nearby + 2 LRGs) with

- Log-normal selection functions
- Luminosity-dependent galaxy biases



Biases based on:
Howlett *et al.*, 1409.3238
Gil-Marín *et al.*, 1407.5668

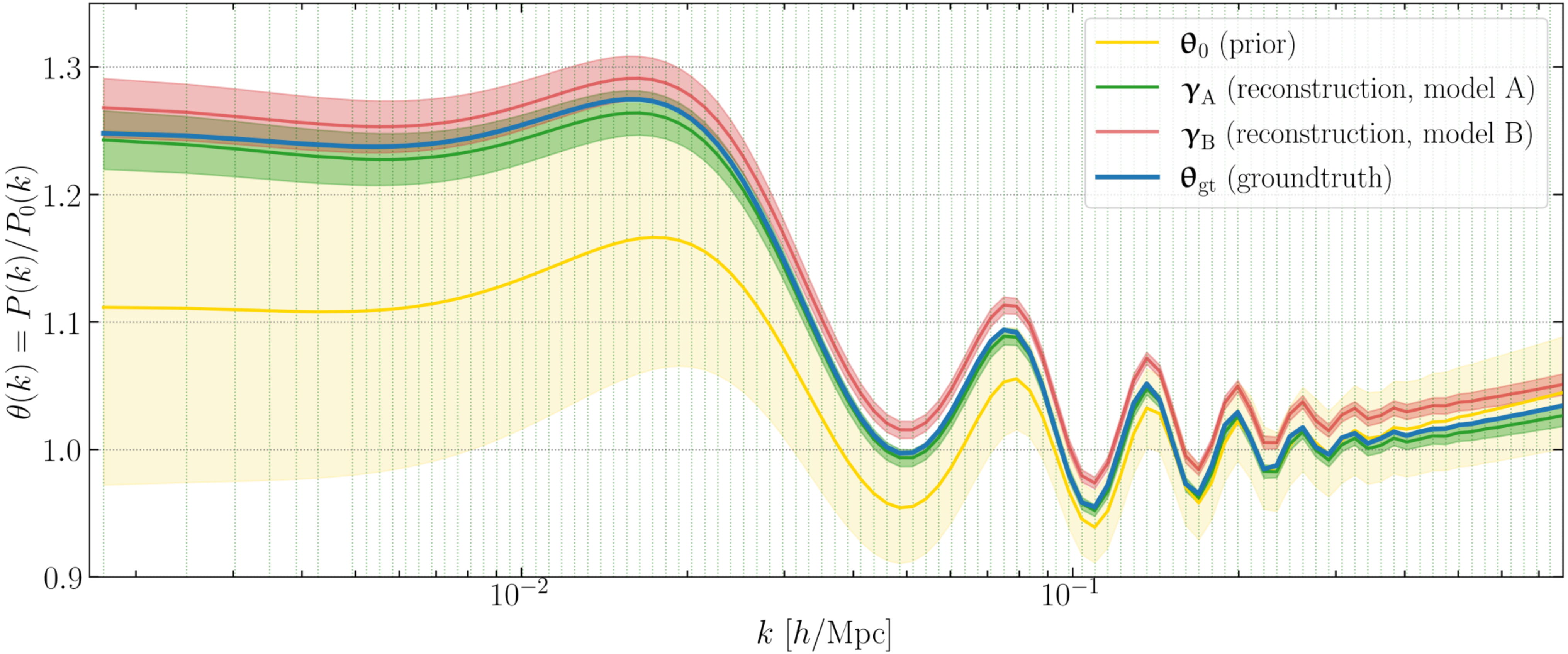


Model B

- Misspecified selections functions
- Misspecified biases
- Effect sizes $\mathcal{O}(1\%)$

Hoellinger & Leclercq, in prep.

Check for model misspecification using the SELFI posterior



Mahalanobis distances to prior:

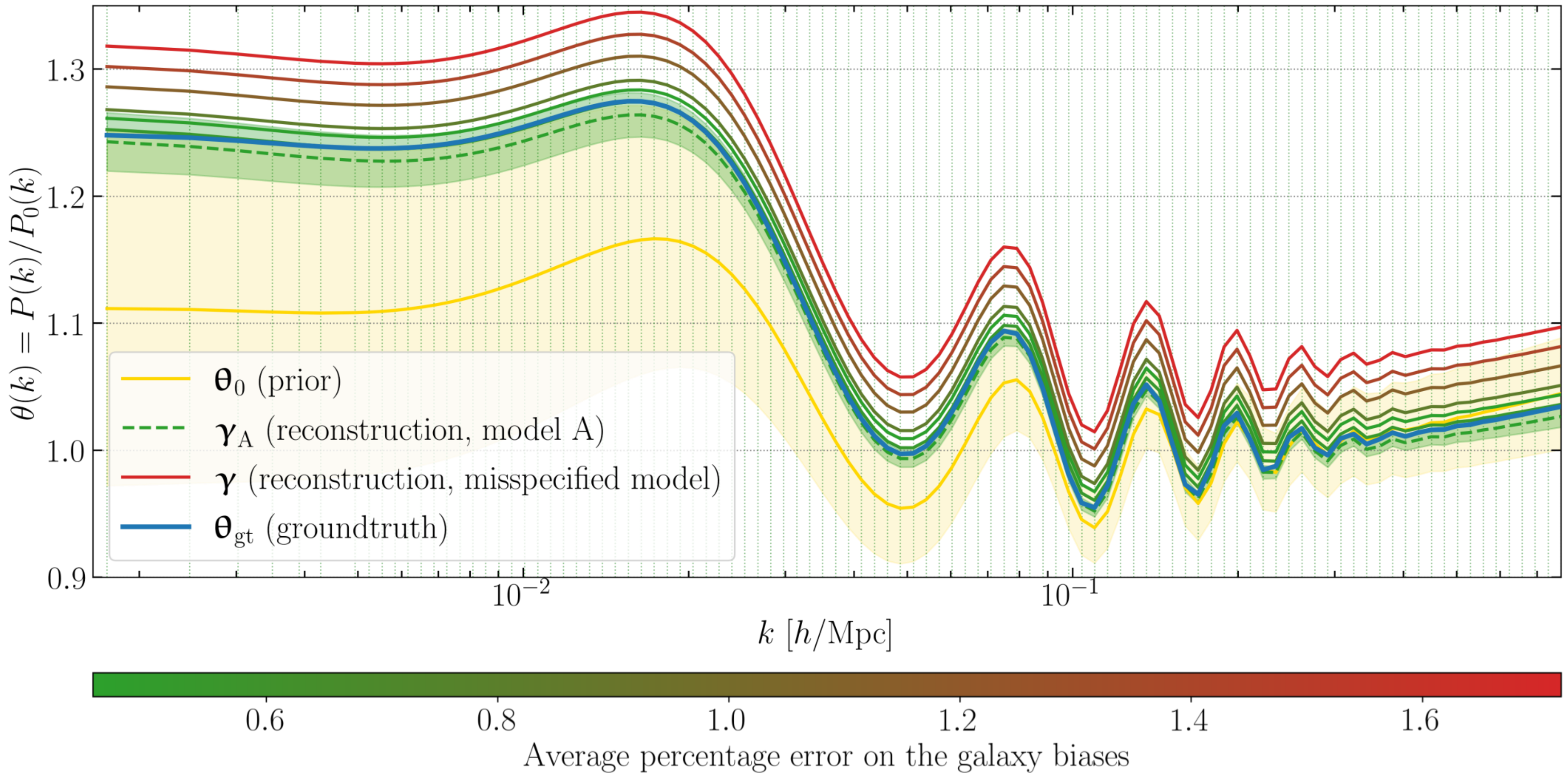
$$d_\mathrm{M}(\boldsymbol{\gamma}, \boldsymbol{\theta}_0|\mathbf{S}) \equiv \sqrt{(\boldsymbol{\gamma} - \boldsymbol{\theta}_0)^\intercal \, \mathbf{S}^{-1} (\boldsymbol{\gamma} - \boldsymbol{\theta}_0)}$$

Model A: 1.96

Model B: 2.91

Hoellinger & Leclercq, in prep.

Impact of galaxy biases on the posterior initial power spectrum

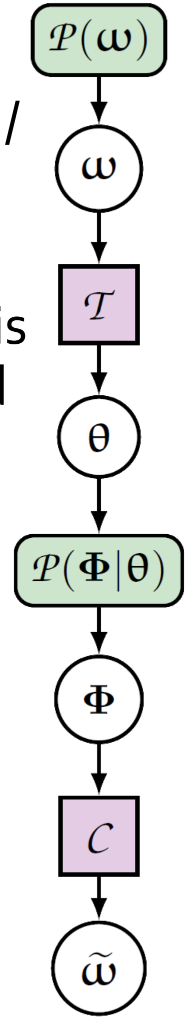


Hoellinger & Leclercq, in prep.

Check for model misspecification and data compression for ILI

- Qualitatively: the shape of the reconstructed initial power spectrum $\boldsymbol{\theta}$ is useful as a check for unknown systematics / model misspecification (using our independent theoretical understanding).
- Quantitatively: we can use the Mahalanobis distance between the reconstruction $\boldsymbol{\gamma}$ and the prior distribution $\mathcal{P}(\boldsymbol{\theta})$:

$$d_\mathrm{M}(\boldsymbol{\gamma}, \boldsymbol{\theta}_0|\mathbf{S}) \equiv \sqrt{(\boldsymbol{\gamma} - \boldsymbol{\theta}_0)^\intercal \, \mathbf{S}^{-1} (\boldsymbol{\gamma} - \boldsymbol{\theta}_0)}$$



- The score function $\nabla_{\boldsymbol{\omega}} \hat{\ell}_{\boldsymbol{\omega}0}$ is the gradient of the log-likelihood at fiducial point $\boldsymbol{\omega}_0$ in parameter space.
- A quasi maximum-likelihood estimator for the parameters is

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

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

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

Already computed for SELFI

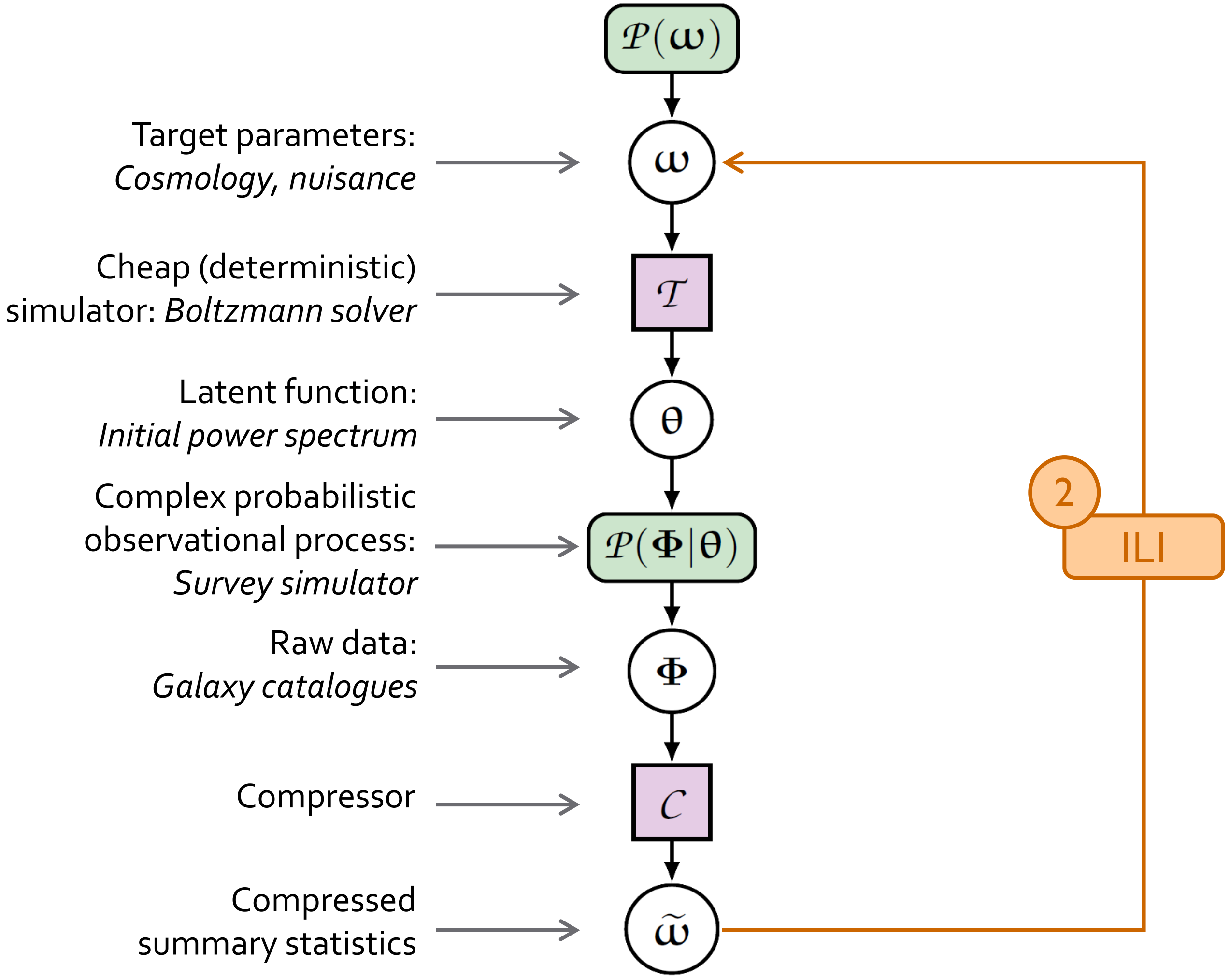
Cheap via finite differences

- Score compression is optimal in the sense that it preserves the Fisher information content of the data.

Alsing & Wandelt, 1712.00012

FL, 2209.11057

Implicit likelihood inference of top-level cosmological parameters

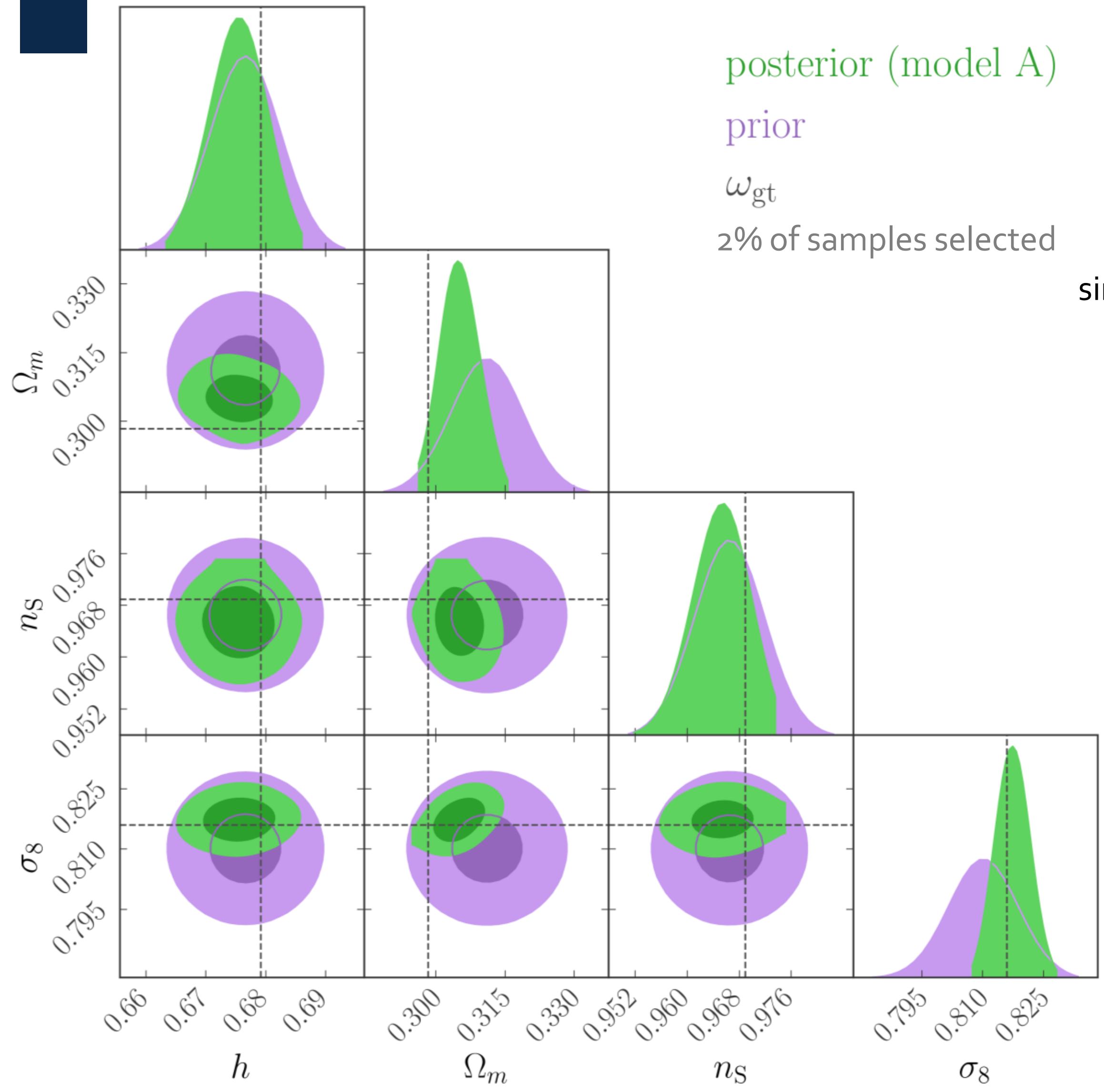


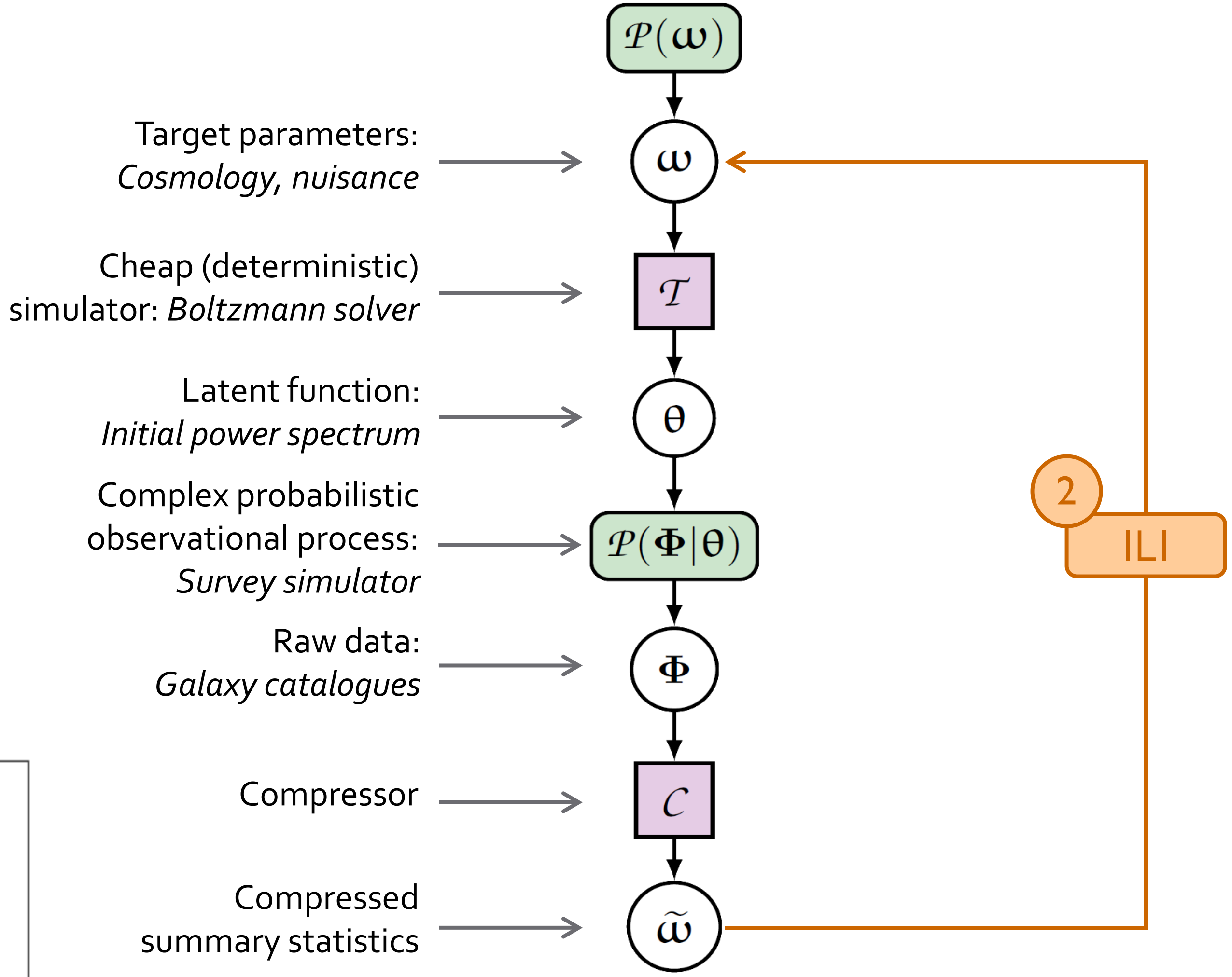
- Any ILI algorithm can be used to obtain the posterior $\mathcal{P}(\omega|\tilde{\omega}_\mathrm{O})$.
- Final inference:
  - does not depend on the assumptions made to check for model misspecification,
  - is unbiased (only more conservative) in case data compression is lossy.
- Non-parametric approaches can use the Fisher-Rao distance between simulated summaries $\tilde{\omega}$ and observed summaries $\tilde{\omega}_\mathrm{O}$:

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

FL, 2209.11057

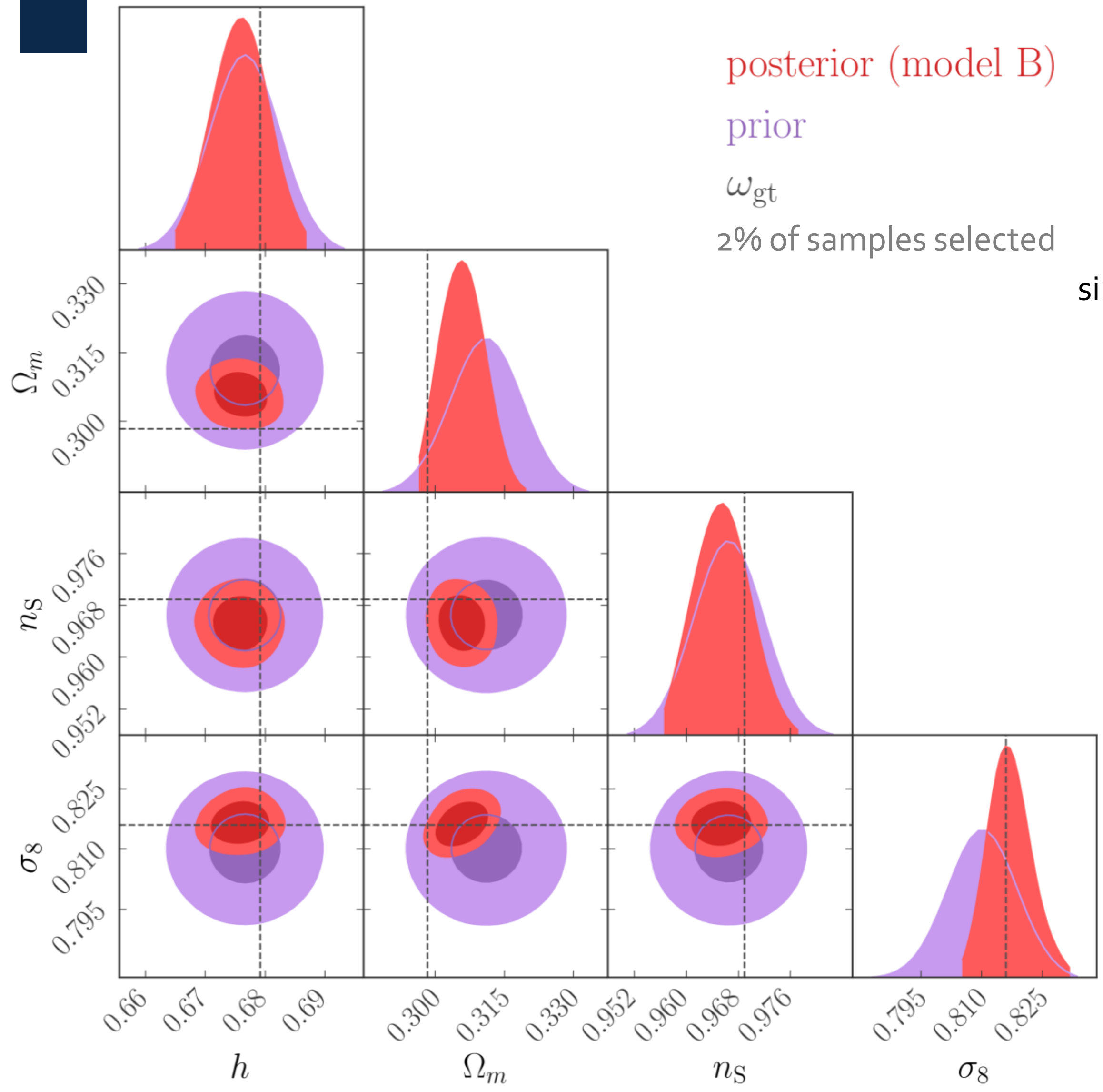
Posterior on cosmological parameters



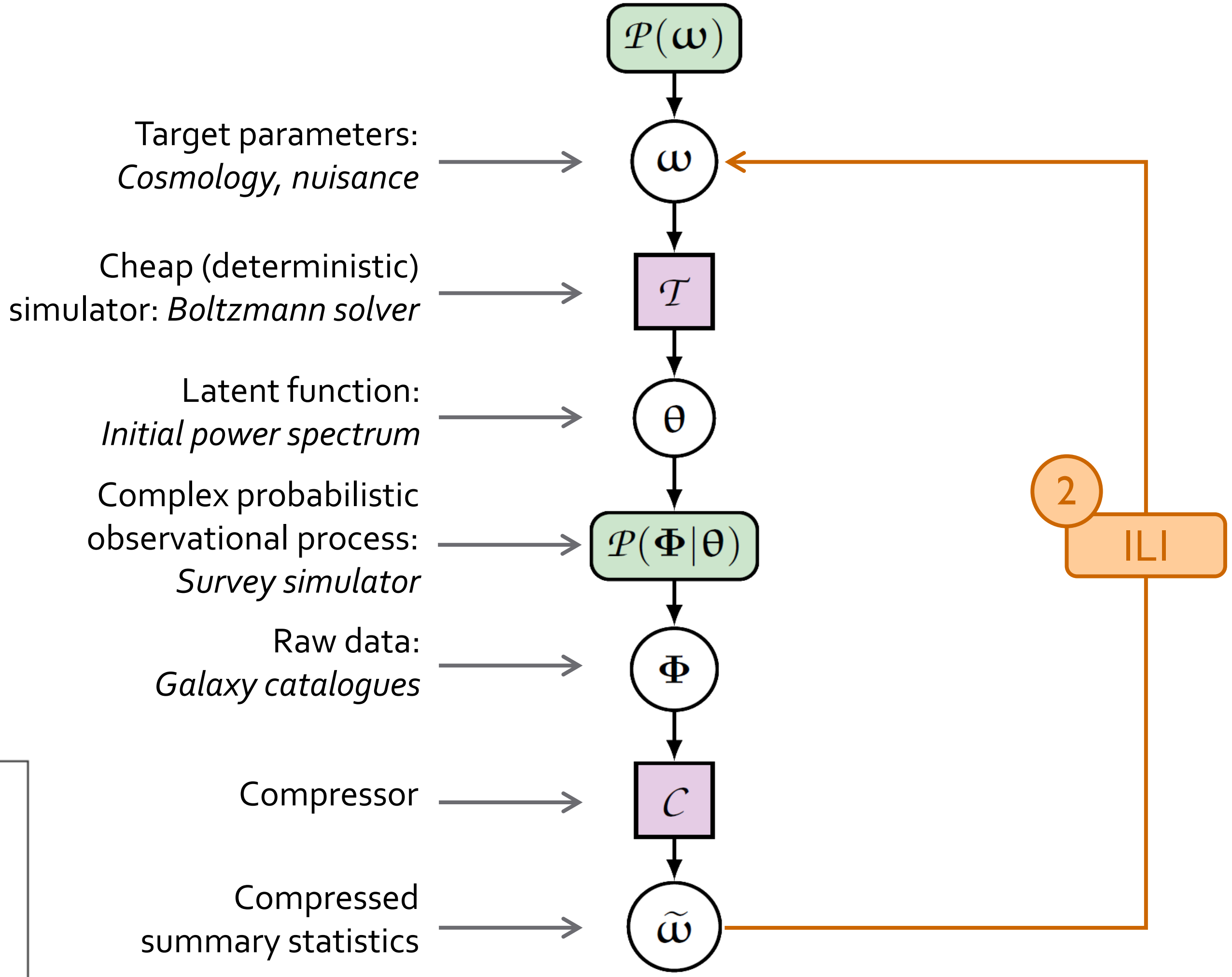


Hoellinger & Leclercq, in prep.

Posterior on cosmological parameters



Hoellinger & Leclercq, in prep.



Conclusion: the statistical framework is in place for the GC:AP pipeline

- A novel two-step simulation based Bayesian approach, combining SELFI and ILI, to tackle the issue of model misspecification for a large class of BHMs.
- Advantages of the first step (SELFI):
  - Even if the inference is in high dimension, the simulator remains a black-box.
  - The number of simulations is fixed *a priori* by the user.
  - The computational workload is perfectly parallel.
  - The linearised data model is trained once and for all independently of the data vector (amortisation).
- Advantages of the second step (ILI):
  - SELFI quantities provide a score compressor for free.
  - General advantages of ILI with respect to likelihood-based methods are preserved.
  - Inference does not depend on the assumptions made to check for model misspecification.

➢ A computationally efficient and easily applicable framework to perform ILI of BHMs while checking for model misspecification.

pySELFI is publicly available at https://pyselfi.florent-leclercq.eu.